PREDICTING THE QUALITY OF SYNTHESIZED SPEECH USING REFERENCE-BASED PREDICTION MEASURES

Florian Hinterleitner[1], *Steve Zabel*[2], *Sebastian Möller*[1],
Lutz Leutelt[3] *and Christoph Norrenbrock*[4]

[1]*Quality and Usability Lab, Deutsche Telekom Laboratories, TU Berlin, Germany*
[2]*Beuth University of Applied Sciences Berlin, Germany*
[3]*Hamburg University of Applied Sciences, Germany*
[4]*Digital Signal Processing and System Theory, CAU Kiel, Germany*
florian.hinterleitner@telekom.de

Abstract: This paper presents research on the use of methods for end-to-end speech quality assessment for the perceptual evaluation of text-to-speech (TTS) systems. We analyze the ITU-T Rec. P.862.2 Wideband Perceptual Evaluation of Speech Quality (WB-PESQ) as well as its wideband optimized successor ITU-T Rec. P.863 Perceptual Objective Listening Quality Assessment (POLQA), and the Diagnostic Instrumental Assessment of Listening quality (DIAL) algorithm (Côté et al., PQS 2010). All measures were originally optimized for the evaluation of telephone networks and speech codecs; thus they are used out of their original domain. In addition to the to-be-evaluated TTS signal all measures also need a natural speech reference as input. The quality estimate is calculated by comparing the natural reference with its corresponding synthetic speech signal.
The measures are tested on data collected by the Blizzard Challenge (BC) in the past years. BC is a competition for developers of speech synthesis systems with the intent to train different systems on the same speech corpus and evaluate their performance. Thus, all synthesizers were built on the same voice.
We use the natural speech reference and its corresponding TTS signal as input for the above mentioned measures. The correlation between the calculated mean opinion score (MOS) and the perceptually evaluated quality rating serves as an indicator for the accuracy of the prediction.
The achieved results were disappointing throughout all databases. The main problem of all 3 algorithms seems to be an inaccurate time alignment between the natural speech file and its corresponding TTS sample. To fix this problem we propose a Dynamic Time Warping between both signals prior to the reference-based evaluation.

1 Introduction

The quality of text-to-speech (TTS) systems improved dramatically over the past years, especially with the trend towards database driven synthesis algorithms like unit selection and HMM synthesis. Even though TTS systems can still easily be distinguished from human speech they found their way into various different everyday applications like short message services, information systems, or smart-home assistents.
With the rise of new applications further improvements will be necessary. Thus, methods to efficiently assess different quality dimensions are an important tool. Depending on the quality aspect to be assessed, different kinds of listening tests can be carried out. Since extensive listening tests are extremely time-consuming as well as cost-intensive it is not practical to conduct

auditory experiments after every step in the development process of a TTS system. Therefore, methods of instrumentally assessing the quality of speech synthesizers could greatly support the development of high-quality TTS systems.
In 2001 Chu and Peng proposed an average concatenative cost function as an instrumental measure for naturalness of synthetic speech [1]. Even though this method reached correlations between the MOS score and the average concatenative cost of -0.872 this approach suffers from the constraint that it is developed for the prediction of the quality of unit selection synthesizers only. Another approach uses Mel Frequency Cepstral Coefficients (MFCC) of natural speech to train Hidden Markov models (HMM) [2]. A spectral distance between these HMMs and MFCCs extracted from the to-be-estimated TTS signals leads to a quality estimate. This model achieved good results on some databases but still has major difficulties especially with female voices. In an attempt to use an algorithm developed for the quality prediciton of telephone channel introduced distortions in natural speech the reference-free ITU-T Rec. P.563 [3] was applied on different TTS databases [4]. In further investigations the internal features of P.563 were analyzed [5]. Combinations of such reference-free measures as well as their underlying parameters have been investigated in [6] [7] [8]. Even though all these approaches perform well in their domain all of them have some drawbacks: either they have problems predicting the quality of female TTS, or they perform well on some databases while dropping off on others.
To get an impression of the maximum accuracy that could be reached with reference-free quality prediction measures a comparison with reference-based quality prediction measures has to be performed. Several of these models have been developed to predict distortions in natural speech introduced by transmission channels of telephone networks. These algorithms use a clean reference signal and evaluate the perceptual distance to the distorted signal. The ITU-T Rec. P.862 Perceptual Evaluation of Speech Quality (PESQ) [9] has already been tested on narrowband TTS signals [10]. The impressing performance of PESQ on single-word TTS signals lead to the approach of verifying these results and comparing the performance of PESQ with the performance of two other state-of-the-art algorithms developed for the evaluation of wideband speech signals. These are the wideband optimized successor of PESQ, ITU-T Rec. P.863 Perceptual Objective Listening Quality Assessment (POLQA) [11] and the Diagnostic Instrumental Assessment of Listening quality (DIAL) [12]. Even though these algorithms are designed for a different domain their performance on synthetic speech can give an impression of what is possible in the area of quality estimation of TTS systems.
The three aforementioned approaches are introduced in Section 2. The databases that were used for evaluation are presented in Section 3. Section 4 discusses the results that could be achieved. Finally, Section 5 gives a perspective to future work.

2 Reference-based quality prediction algorithms

This section presents the instrumental quality measures used to evaluate synthetic speech.

2.1 ITU-T Rec. P.862.2 (WB-PESQ)

The ITU-T Rec. P.862 for speech quality assessment of narrow-band telephone networks was released in 2001. With P.862.2 (WB-PESQ) [13] PESQ was extended to wideband speech signals. Figure 1 shows an overview of the PESQ model (the wideband version basically follows the same stages).
Firstly, several pre-processing steps take place: level alignment, IRS filtering, voice activity detection, followed by a time alignment between the original speech signal and the degraded signal. Secondly, both signals are compared via a perceptual model. Therefore the signals are

transformed into internal representations which emulate the ones of the human auditory system. This step comprises spectrum computation, band integration, frequency and gain compensation, as well as loudness compression. In the final step a distance between the 2 perceptually transformed signals is computed. This includes disturbance, masking and asymmetry computation, frequency and time averaging, and results in a final PESQ quality score. Subsequently, a mapping function transfers the results onto a MOS scale.
For WB-PESQ the IRS receive filter was replaced with a bandpass filter in the range 200Hz - 8000Hz and the mapping function was adjusted.

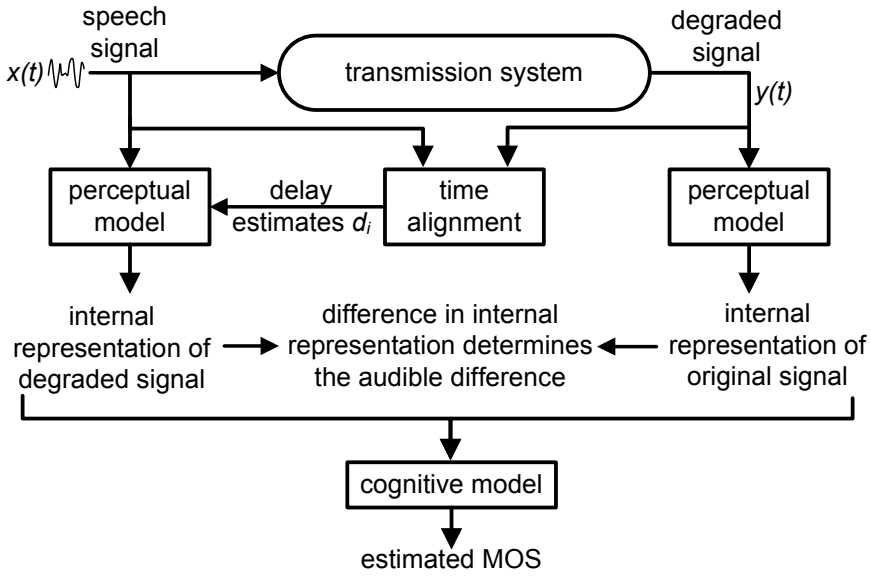

Figure 1 - Overview of the PESQ model [9]

2.2 Diagnostic Instrumental Assessment of Listening quality (DIAL)

Due to the ongoing advancements in telephone networks the ITU-T started a standardization project to come up with a new reference-based measure. As one of the contestants the DIAL model [12] was proposed. Its framework is presented in Figure 2.
The pre-processing consists of active speech level normalization, voice activity detection, and the time alignment from the PESQ model. The main model combines 3 building blocks: the **core model** estimates non-linear degradations introduced by speech processing systems. The **dimension estimators** introduced by Wältermann [14] cover the perceptual dimensions directness/frequency content (DFC), noisiness (N), loudness (L) and continuity (C) of speech signals. Finally, the **cognitive model** uses the so far computed scores to simulate the cognitive process of a human listener. Besides narrowband and wideband signals, the DIAL model can also be used in a superwideband context.

2.3 ITU-T Rec. P.683 (POLQA)

The POLQA model was released in 2011 and defines an algorithm for speech quality assessment of state-of-the-art telephony systems. Being the successor of PESQ its framework is very similar to the one in Figure 1. Moreover it allows quality estimation of superwideband speech signals and the assessment of networks and codecs that introduce time warping.
Firstly, a time alignment takes place. Therefore both signals are split up into small chunks. The delay between each chunk of the reference signal and the distorted signal is calculated and used to adjust the sampling rate of the degraded signal. Subsequently both signals are transformed

into an internal representation of the human auditory system similar to the PESQ model. Additionally low levels of noise in the reference signal which might lead to the impression of a signal of minor quality are eliminated. This represents the idealization process that subjects usually go through during their quality judgement. Then the cognitive model uses 6 indicators (frequency response, noise, room reverberation and 3 indicators that describe the internal difference in the time-pitch-loudness domain) to compute a final estimated MOS score.

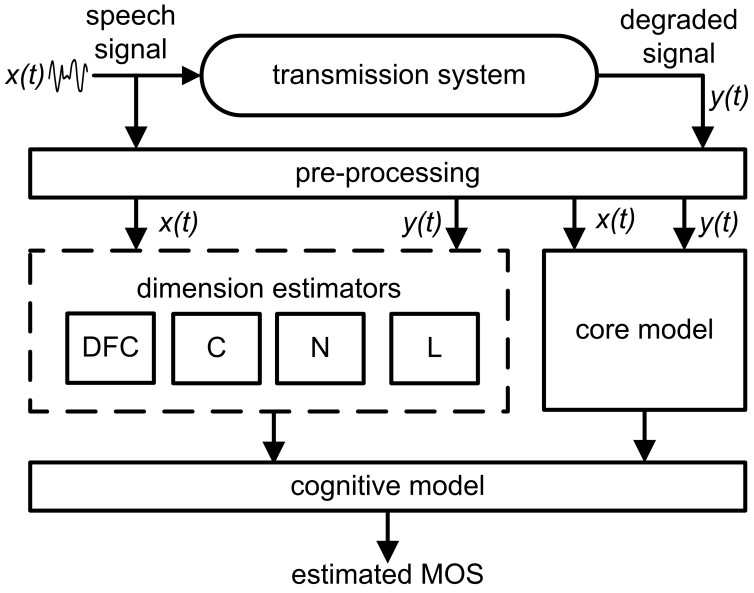

Figure 2 - Overview of the DIAL model [12]

3 Blizzard Challenge databases

The Blizzard Challenge is an annual contest for developers of UK English and Mandarin Chinese speech synthesis systems. Participants in the English part of the challenge are provided with the 'Roger' corpus of the University of Edinburgh (*full corpus*) as well as two different subcorpora of the 'Roger' corpus (*ARCTIC corpus* and *small corpus*). These corpora are used to build the TTS voices. A set of test sentences is released to the contestants, who are asked to submit synthesized versions within a limited time interval. An online listening test is conducted to evaluate naturalness, intelligibility and the degree of similarity to the original speaker. In the following the challenges of the years 2008, 2009, and 2010 are described.

3.1 Blizzard Challenge 2008 (BC 2008)

The Blizzard Challenge 2008 database [15] consists of 18 speech synthesis systems, 1 natural speaker and 2 systems from participants from previous challenges (a Festival-based system from CSTR and the HTS system from the Blizzard Challenge 2005). In an attempt to calibrate the results from year to year, the latter systems were used as benchmarking systems. For every synthesizer, 42 files were evaluated during the listening tests.

3.2 Blizzard Challenge 2009 (BC 2009)

The 2009 database [16] consists of 14 speech synthesis systems, 1 natural speaker and 3 benchmarks systems (the 2 systems used during the Blizzard Challenge 2008 and the HTS system from the Blizzard Challenge 2007). 38 files generated by each system were judged during the

evaluation phase.

The voices built on the *ARCTIC* and *small corpus* in the years 2008 and 2009 were unavailable. Thus our research in these years concentrates on the TTS data on the basis of the *full corpus*.

3.3 Blizzard Challenge 2010 (BC 2010)

In 2010 18 TTS developers took part in the challenge [17]. This database consists of TTS voices built on the *ARCTIC corpus* (14 participants, 1 natural reference, and 3 benchmark systems from previous challenges) and of synthetic speech samples that were built on the *rjs* speaker provided by Phonetic Arts (15 participants, 1 natural speaker, and 2 benchmark systems). The *ARCTIC* database consists of 36 files per synthesizer and the *rjs* database holds 36 files for every contestant.

3.4 Quality evaluation

The listening tests were carried out online using a design developed for Blizzard 2007. Various listener types were employed spanning from volunteers recruited via the Challenge participants, mailing lists, blogs to speech experts, and paid UK undergraduates. Since the results of all listeners were used during the evaluation, there will be no further differentiation. In Blizzard 2008 438 listeners finished the whole test procedure whereas 365 completed the test in 2009 and 363 in 2010. The listener gender was anonymized, thus gender-related aspects could not be analyzed. The tests consisted of different sections where listeners had to rate differences in similarity, naturalness and intelligibility. Only the mean opinion scores (MOS) for the naturalness rating will be analyzed here. The evaluated files from these sections consisted of sentences from the genres news and novel and were sampled at 16kHz. Since the purpose is to predict the quality of synthesized speech all natural speech files will only be used as reference signals in the estimation process.

4 Results and discussion

To evaluate the accuracy of the predicted results we compute Pearson's Correlation Coefficient R and Spearman's Rank-Order Correlation ρ between the predicted MOS and the auditory MOS scores average per system. The results can be seen in Table 1. Correlations above 0.40 are marked in bold.

		WB-PESQ		DIAL		POLQA	
	database	R	ρ	R	ρ	R	ρ
per system	BC 2008	0.17	0.12	**0.49**	0.40	**0.46**	**0.55**
	BC 2009	0.19	0.23	0.25	0.09	0.33	0.40
	BC 2010 rjs	0.02	0.06	-0.15	-0.36	0.35	0.12
	BC 2010 arctic	0.21	0.33	-0.14	-0.17	-0.08	-0.09

Table 1 - Correlations between predicted MOS and auditory MOS scores (per synthesizer)

None of the algorithms achieves satisfying results. Only for database BC 2008 DIAL and POLQA reach correlations above 0.40. Taking a look at Table 2 shows that the mean predicted MOS value is usually between 1 and 2 with a variance just above 0. Obviously all three algorithms detect major distortions in all tested TTS systems. This leads to constant low MOS values with no distinction between good and bad sounding TTS.

	WB-PESQ		DIAL		POLQA	
database	$\overline{MOS}$	variance	$\overline{MOS}$	variance	$\overline{MOS}$	variance
BC 2008	1.38	0.22	1.97	0.63	1.22	0.06
BC 2009	1.32	0.17	2.09	0.03	1.08	0.02
BC 2010 rjs	1.11	0.00	1.91	0.03	1.34	0.09
BC 2010 arctic	1.13	0.03	2.00	0.03	1.20	0.05

Table 2 - Mean values and variances of the predicted MOS

Usually WB-PESQ, DIAL, and POLQA are used to evaluate audiomaterial of a length of at least 8-9s. Because most of the databases consist of TTS files with a length of 2-3s we supposed that this would be one source for the very low MOS values. Hence we concatenated groups of 3 TTS files of the same system from the database BC 2008 and used them as input for WB-PESQ and DIAL. The resulting scores were averaged per synthesizer, and R as well as ρ were computed. The results (Figure 3) showed little improvement for DIAL but none for the results from WB-PESQ. Also $\overline{MOS}$ and variance remain on a very low level.

	WB-PESQ		DIAL	
database	R	ρ	R	ρ
BC 2008	-0.18	-0.14	0.59	0.59

Table 3 - Correlations between predicted MOS and auditory MOS for concatenated TTS files (per synthesizer)

We assumed one of the reasons for the very low MOS values are failures in the time alignment between the natural speech file and the TTS signal. Compared to natural speech, TTS systems often produce signals that comprise parts that are lenghtened or shortened. This makes a time alignment more challenging than between a natural speech signal and a telephone-network-coded one. To simulate these TTS distortions we inserted 5 to 6 small pauses between 0.1s and 1s into 5 natural speech files. These files were tested with WB-PESQ and DIAL against the original natural speech data. Figure 3 shows the resulting MOS values.
The quality predicted by WB-PESQ for the natural speech files without any modifications is around 4.5. After the insertion of pauses (no matter which length) the value drops below 1.5. DIAL estimates values between 3.5 and 4 for the original speech samples. For inserted pauses longer than 0.1s the MOS score predicted by DIAL decreases to values between 1.5 and 2.5. In contrast, natural speech files with inserted pauses of 0.1s length get a slightly better rating than the original files.

The achieved correlations lag far behind the results achieved by Cernak and Rusko with the PESQ measure [10]. However, their approach differed in the point that they used only word long TTS samples which makes the time alignment between the natural speech and the TTS signal much easier. Hence, we used 5 samples from the BC 2008 database for 5 TTS systems, cut out one word per sample and computed the WB-PESQ scores. As auditory MOS we used the MOS value the whole sample had been rated with. However, the results could not be improved: $\overline{MOS}$ of 1.11, a variance of 0.03 and R=0.17.

5 Conclusions and future work

We tested 3 reference-based quality prediction algorithms for the assessment of telephone networks (WB-PESQ, DIAL, POLQA) on TTS data from the Blizzard Challenges 2008, 2009, and 2010. The correlations between the predicted MOS and the auditory MOS were disappointing throughout all databases. The best correlations were achieved by POLQA on database BC 2008

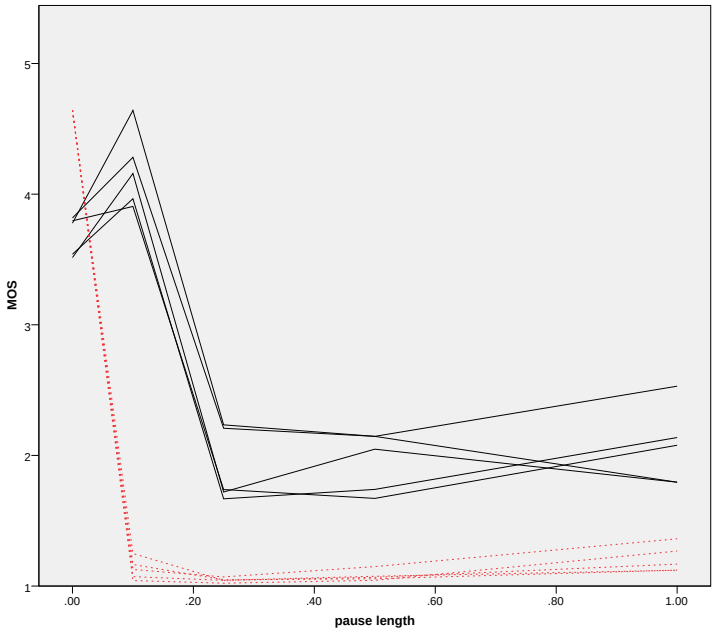

Figure 3 - Predicted MOS in relation to length of inserted pauses (DIAL: black lines, WB-PESQ: dashed red lines)

($R = 0.46$, $\rho = 0.55$). Of course it has to be noted that all of the tested predictors were used out of their original intended domain, therefore the achieved correlations do not contradict the good results attained for telephone-band speech.
One of the reasons for the overall low predicted MOS values seems to be an inaccurate time alignment between the TTS samples and the natural speech files. This is due to the non-linear distortions introduced by the TTS algorithms. For further studies a Dynamic Time Warping could be used as a pre-processing step to ensure exact time alignment between the test signals.

6 Acknowledgements

The present study was carried out at Deutsche Telekom Laboratories, Berlin. It was supported by the Deutsche Forschungsgemeinschaft (DFG), grants MO 1038/11-1 and HE 4465/4-1.

References

[1] CHU, M. ; PENG, H.: An Objective Measure for Estimating MOS of Synthesized Speech. In: *Proc. 7th Int. Conf. on Speech Communication and Technology (EUROSPEECH 2001)* 3 (2001), S. 2087–2090

[2] FALK, T. H. ; MÖLLER, S.: Towards Signal-Based Instrumental Quality Diagnosis for Text-to-Speech Systems. In: *IEEE Signal Processing Letters* 15 (2008), S. 781–784

[3] ITU-T REC. P.563: *Single Ended Method for Objective Speech Quality Assessment in Narrow-Band Telephony*. Geneva: International Telecommunication Union, 2004

[4] MÖLLER, S. ; HEIMANSBERG, J.: Estimation of TTS Quality in Telephon Environments Using a Reference-free Quality Prediction Model. In: *Proc. 2nd ISCA/DEGA Tutorial and Research Workshop on Perceptual Quality of Systems* (2006), S. 56–60

[5] FALK, T. H. ; MÖLLER, S. ; KARAISKOS, V. ; KING, S: Improving Instrumental Quality Prediction Performance for the Blizzard Challenge. In: *Proc. Blizzard Challenge Workshop* (2008)

[6] MÖLLER, S. ; HINTERLEITNER, F. ; FALK, T.H. ; POLZEHL, T.: Comparison of Approaches for Instrumentally Predicting the Quality of Text-To-Speech Systems. In: *Proceedings of the 11th Annual Conference of the ISCA (Interspeech 2010). International Speech Communication Association (ISCA)* (2010), S. 1325–1328

[7] HINTERLEITNER, F. ; MÖLLER, S. ; FALK, T.H. ; POLZEHL, T.: Comparison of Approaches for Instrumentally Predicting the Quality of Text-to-Speech Systems: Data from Blizzard Challenges 2008 and 2009. In: *Proceedings of the Blizzard Challenge Workshop. International Speech Communication Association (ISCA)* (2010)

[8] HINTERLEITNER, F. ; MÖLLER, S. ; NORRENBROCK, C. ; HEUTE, U.: Vergleich unterschiedlicher Ansätze zur instrumentellen Vorhersage der Qualität von Text-to-Speech Systemen: Daten der Blizzard Challenge 2010. In: *Proceedings DAGA 2011* (2011)

[9] ITU-T REC. P.862: *Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs*. Geneva: International Telecommunication Union, 2001

[10] CERNAK, M. ; RUSKO, M.: An Evaluation of Synthetic Speech Using the PESQ Measure. In: *Proc. European Congress of Acoustics* (2005), S. 2725–2728

[11] ITU-T REC. P.863: *Perceptual Objective Listening Quality Assessment (POLQA)*. Geneva: International Telecommunication Union, 2011

[12] CÔTÉ, N.: *Integral and Diagnostic Intrusive Prediction of Speech Quality*. Springer Verlag Heidelberg, 2011

[13] ITU-T REC. P.862.2: *Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs*. Geneva: International Telecommunication Union, 2007

[14] WÄLTERMANN, M. ; SCHOLZ, K. ; RAAKE, A. ; HEUTE, U. ; MÖLLER, S.: Underlying Quality Dimensions of Modern Telephone Connections. In: *Proceedings of the 7th Annual Conference of the ISCA (Interspeech 2006). International Speech Communication Association (ISCA)* (2006)

[15] KARAISKOS, V. ; KING, S. ; CLARK, R. A. J. ; MAYO, C.: The Blizzard Challenge 2008. In: *Proceedings of the Blizzard Challenge Workshop. International Speech Communication Association (ISCA)* (2008)

[16] KING, S. ; KARAISKOS, V.: The Blizzard Challenge 2009. In: *Proceedings of the Blizzard Challenge Workshop. International Speech Communication Association (ISCA)* (2009)

[17] KING, S. ; KARAISKOS, V.: The Blizzard Challenge 2010. In: *Proceedings of the Blizzard Challenge Workshop. International Speech Communication Association (ISCA)* (2011)