Comparison of Approaches for Instrumentally Predicting the Quality of Text-to-Speech Systems: Data from Blizzard Challenges 2008 and 2009

Florian Hinterleitner[1], *Sebastian Möller*[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

florian.hinterleitner@gmail.com, sebastian.moeller@telekom.de,
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 synthesized speech signals 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 auditory evaluated synthetic speech databases from the Blizzard Challenges 2008 and 2009. The results show that auditory quality judgments can be predicted with a sufficiently high accuracy and reliability. Especially the possibility to rank different synthesizer systems by their quality comes within reach.

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

1. Introduction

Text-to-speech (TTS) systems have reached a quality level that no longer limits them to be used as an aid for visually impaired but allows to apply them to services used by an unlimited group of users like email and short message service readers, foreign language education and information systems. With the development of new applications further improvements of the TTS systems are to be expected, which will be reflected in a number of perceptual dimensions, especially with respect to the naturalness and the overall quality of the synthesized speech. As a consequence, methods for efficiently assessing these quality dimensions are of great interest.

Evaluating synthetic speech, however, is not an easy task. Depending on the quality aspects of interest, different types of tests are recommended: articulation and intelligibility tests assess whether the synthetic speech signal is able to carry information on a segmental or supra-segmental level [1]; comprehension tests investigate whether the content provided via the synthesized speech signal can be discerned [2]; and overall quality tests, for example the one described in the ITU-T Rec. P.85 [3], are used to determine global aspects of the synthesized speech signal, such as naturalness, pronunciation, intonation, speech rate, voice pleasantness, etc. Although doubts have been cast on the test protocol [4][5] this is still the common way of auditorily measuring the overall quality of synthetic speech. The major drawback to all these methods is that they are very cost-intensive as well as time-consuming which makes it hard for developers of synthetic speech to evaluate the quality of their systems after every step in the developement process. Therefore a method for instrumentally predicting the quality of synthetic speech could greatly support the development of high-quality TTS systems.

Several proposals have been made to estimate the percieved quality of synthesized speech, however, a universal method for quality prediction has not yet been established. Most measures use a natural reference signal and evaluate the spectral distance between the synthesized signal and its natural counterpart. Cernak [6] used the ITU-T P.862 PESQ measure [7], an objective method for end-to-end speech quality assessment of narrow-band networks and speech codecs, to predict the quality of concatinative speech synthesizers. Furthermore Chu and Peng [8] developed a method to predict synthesized speech quality through a concatenative cost function. Even if these approaches reached very high correaltions between their output values and the corresponding subjective mean opinion scores (MOS) of auditory evaluated test databases, they are only rarely applicable. Firstly, a natural speech reference, spoken by the same speaker as the to-be-evaluated synthetic speech samples, has to be available, which is normally only the case if corpus-based synthesizers are evaluated. Moreover both approaches are only able to capture concatination-linked distortions, while e.g. perceptual degredations originated in unnatural prosody on the sentence level are out of focus. To overcome these limitations, a reference-free approach is required.

Mariniak [9] proposed a perception-based analysis of speech samples from many natural speakers. In this manner a reference feature space could be build and compared with features from synthesized speech signals. Synthetic speech samples get classified with regard to this natural feature set, and a distance measure is computed. To our knowledge, this approach has never been implemented by Mariniak, but it has recently been taken up in [10], using Mel-frequency Cepstral Coefficients (MFCCs) as features and Hidden Markov Models (HMMs) with Gaussian Mixture densities for a temporal-spectral comparison of features. Perceptual features, extracted from the synthesized signal, are assessed against the reference models via the log-likelihood measure. This approach led to very promising results on the evaluated test databases. Correlations between the log-likelihood and the corresponding auditory MOS reached values from 0.54 to 0.81 for different quality dimensions.

Another approach is to extract parameters that are directly related to the degradations in the synthetic speech signal. It is motivated by the quality prediction model for transmitted natural speech, given in ITU-T Rec. P.563 [11]. This model com-

bines three principles for evaluating distortions: it performs an LPC analysis on a model of the human vocal tract; the second principle reconstructs a clean reference signal from the degraded input signal; and the third principle is to identify and to classify distortions typically encountered in voice transmission channels. During all three steps a large number of internal features is generated, weighted and combined, to finally result in a predicted objective MOS. Applying this model to synthesized speech [12][13], the results were not as promising as those obtained with the HMM-based approach but gave a detailed view on which of the generated P.563 internal features could be useful for TTS quality prediction. Furthermore, it could be observed [14] that correlations between the internal features and auditory MOS differed highly for files of different speaker gender.

In an attempt to compare and combine the feature-comparison and the parametric approaches in order to increase the prediction performance and robustness, HMM-based features, P.563 internal parameters as well as general speech parameters were extracted from three German TTS databases. The parameters which correlated well with auditory test results were combined to three quality estimators [15]. The correlations between predicted MOS and auditory test results showed that the prediction accuracy differs between the three approaches, between the TTS databases as well as between speaker gender of the TTS databases. Furthermore, a combination of the different approaches showed slight improvements both for male and female data. For male files correlations above 0.82 and for female data above 0.70 could be achieved [16]. However, those estimators were highly optimized on the given data, thus a generalization for other TTS databases could not be stated.

Our aim is to use these three approaches to predict the quality for the TTS files from the Blizzard Challenges. For this purpose, HMM-based feature comparison and parametric approaches have been used as they are described in Section 2. In Section 3 an analysis of the used databases is given. Applying the approaches to these databases, 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 Figure1 and the individual parts are described in the following subsections.

2.1. HMM-based approach

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 a level normalization to -26dB below the overload point of the digital system, using the active speech-level meter defined in ITU-T Rec. P.56 [17] followed by downsampling to 8kHz samplingrate. A standard telephone bandpass filtering as it was used in previous analysis [10] did not take place. Moreover, since we are only interested in the quality of the TTS system, only active speech segments were analyzed, using a simple energy thresholding Voice Activity Detection (VAD) algorithm to remove silence intervals longer than 75ms; this duration was empirically chosen as to avoid artificial discontinuities introduced by possible VAD errors.

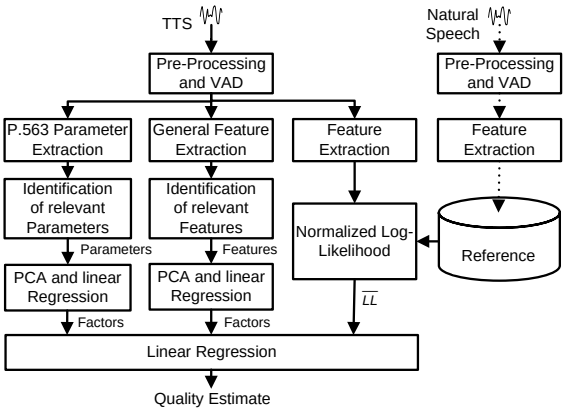

Figure 1: Modeling approach [16]. Solid lines refer to the evaluation phase, dashed lines to the training phase.

12^{th} order MFCCs are then computed both during the training and the evaluation phase using 25ms windows and 10ms 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.

Since we consider the temporal dynamics to be important for perceived quality, we use HMMs trained with natural reference features to quantify differences between naturally-produced and synthesized speech. 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 the state transition probabilities, initial state probabilities and output distribution parameters, are computed using the expectation-maximization algorithm [18]. The perceptual similarity is then expressed as a Log-Likelihood (LL) value computed using the so-called forward-backward procedure described in [18]. Normalization is performed based on the number of active-speech frames in the signal under test; the normalized log-likelihood is refered to as $\overline{LL}$ in Figure1.

2.2. P.563 internal features

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.

In order to extract the relevant information for the given task from this set of parameters, we employed a sequential feature selection (SFS) algorithm followed by a Principal Component

Analysis (PCA). The SFS used a correlation-based cost function where features with an average Spearman rank-order correlation between the two databases of $|\rho| \geq 0.40$ on a per synthesizer basis were kept. The determinated features are listed in Table 1. 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.

parameter	BC 2008	BC 2009	average ρ
DistortionClass	-0.77	-0.41	-0.59
SpeechInterruptions	-0.61	-0.37	-0.49
ArtAverage	0.68	0.26	0.47
UnnaturalBeepsMean	-0.29	-0.61	-0.45
MuteLength	-0.59	-0.28	-0.44
SharpDeclines	-0.48	-0.38	-0.43
UnnaturalBeeps	-0.28	-0.57	-0.42
ConsistentArtTracker	-0.36	-0.45	-0.40
UnnaturalBeepsAffectedSamples	-0.23	-0.57	-0.40

Table 1: Spearmans rank-order coefficient per synthesizer between the internal features of P.563 and the MOS score, calculated for the Blizzard Challenge 2008 and 2009 data.
(ArtAverage: averaged section of the back cavity of the vocal tract model; ConsistentArtTracker: describes how well the back and the middle cavity of the vocal tract model correlate; Unnatural beeps: voiced parts in the signal, that are too short to be of natural origin; For detailed information see [11])

2.3. General speech features

As a third basis for the quality prediction we calculated a large set of 1567 general parameters [19] which provide a broad variety of information about vocal 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 again employed a SFS algorithm followed by a PCA. The SFS used a correlation-based cost function where features with an average Spearman rank-order correlation between the two databases of $|\rho| \geq 0.40$ on a per synthesizer basis were kept. The resulting 38 features were then processed by a PCA to come up with a small set of relevant factors which are used for the quality estimation function.

2.4. Linear combination

Finally, a quality estimate is calculated from either $\overline{LL}$, the factors of the principal component analysis of the extracted features, or both. 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. The target value to be estimated was the "naturalness" score of the auditory tests. 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.

3. Databases

The aforementioned approaches where tested on data from the Blizzard Challenges 2008 and 2009. Participants of those challenges could submit synthesized speech files for different tasks as well as different languages. In the following, only files that were built on the full 15 hour recordings of a UK English male speaker with a fairly standard Received Pronunciation (RP) accent (Roger Corpus) were used for the evaluation. Since the purpose is to predict the quality of synthesized speech all natural speech files were omitted from the test databases.

3.1. Blizzard Challenge 2008 (BC 2008)

The Blizzard Challenge 2008 database [20] 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 [21] 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). 40 files generated by each system were judged during the evaluation phase.

3.3. Quality evaluation

Both listening tests were carried out online using a design developed for Blizzard 2007. Various listener types were employed in both years 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. The listener gender was anonymized, thus gender-related aspects could not be analyzed. Both 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.

4. Results and discussion

4.1. HHM-training

As preceding studies have shown [15] the selection of data used for the HMM-training has a remarkable influence on the accuracy of the quality estimations. Thus several different HMMs were trained and tested on the available data. HMMs were trained on 20min and 180min of speech files randomly chosen from the Roger Corpus. Furthermore two HMMs were trained on the natural speech files used in the Blizzard Challenges 2008 and 2009 which comprise the same phonetic content as the evaluated synthetic speech files, and which were uttered by the speaker the TTS corpus has been built from. The performance of the quality prediction based on different HMMs is assessed by Pearson's Correlation Coefficient R between the normalized Log-Likelihood x_i and the auditory MOS y_i where $\overline{x}$ und $\overline{y}$ are the corresponding average values.

$$R = \frac{\sum_{i=1}^{N}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \overline{x})^2} \cdot \sqrt{\sum_{i=1}^{N}(y_i - \overline{y})^2}}$$

Since R is extremly sensitive to outliers we furthermore compute Spearman's rank-order correlation ρ where $rk(x_i)$ and $rk(y_i)$ are the ranks of x_i respectively y_i, $\overline{rk_x}$ and $\overline{rk_y}$ are the average values of the ranks of x and y.

$$\rho = \frac{\sum_{i=1}^{N}(rk(x_i) - \overline{rk_x}) \cdot (rk(y_i) - \overline{rk_y})}{\sqrt{\sum_{i=1}^{N}(rk(x_i) - \overline{rk_x})^2} \cdot \sqrt{\sum_{i=1}^{N}(rk(y_i) - \overline{rk_y})^2}}$$

We furthermore compute the root-mean-square error ($RMSE$) which assesses the accuracy of the achieved quality predictions. In order to achieve meaningful results, the $\overline{LL}$ values were linearly transformed to the range between the lowest and highest auditory MOS prior the $RMSE$ computation.

$$RMSE = \sqrt{\left(\frac{\sum_{i=1}^{N}(x_i - y_i)^2}{N}\right)}$$

In addition, we compute the average $\overline{LL}$ which shows how similar the evaluted synthetic speech files were rated in comparison to the natural speech files used in the training process. The analysis is carried out on a per-stimulus and on a per-synthesizer basis. The results are shown in Table 2.

Comparing the results from Table 2 with the correlations achieved in [16] on three German test databases an inferior performance per-stimulus as well as per-synthesizer can be stated. This is due to the fact that the quality range of TTS synthesizer in the Blizzard database is closer and thus a quality prediction is more challenging. The performances of the "Roger 20min"-HMM and the "Roger 180min"-HMM show no further gain in prediction accuracy by increasing the amount of data used during the HMM-training process. Furthermore, it can be stated that the HMM trained on the natural speech files from BC 2008 leads to the best correlations per stimulus with $R = 0.30$ as well as per synthesizer with $R = 0.64$ on the data from 2008. In addition, this HMM also implicates the lowest average $\overline{LL}$. According to this the synthetic speech files get rated more similar to the training data from this HMM than to the data used in the other training processes. The results show that a training basis which consists of speech files with the same sentence structure as the files to-be-evaluated leads to better prediction results than HMMs trained on a randomly chosen subset of the Roger Corpus. Surprisingly the "BC 08"-HMM outperforms the "BC 09"-HMM on the data from 2009, therefore the "BC 08"-HMM will be used for all further computations on the 2009 dataset.

4.2. Results

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-square error $RMSE$. Since the rating scale has not really interval level we again 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. As mentioned before the analysis 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 all quality predictors are shown in Table 3 while scatter plots of the three models can be seen in Figure 2, 3 and 4. Comparing the performance of all three models on the given databases shows that no model leads to satisfying results on a per-stimulus basis. The highest correlation could be achieved by the general parameters on the data from 2009 with $R = 0.50$. However, this approach fails on the 2008 data. A combination of all three approaches leads to a more stable performance on both databases.

On a per-synthesizer basis, the performance of the estimators increase. Apparently, the differences between individually synthesized speech samples are averaged out in the per-synthesizer analysis. This shows that the models above have difficulties in predicting the quality of single speech samples but lead to very promising results when predicting the quality of synthesizers. The best results of a single model could be achieved from the P.563 approach with $R = 0.78$ for 2008 data while the general speech features score $R = 0.84$ on the 2009 data. Again a more stable prediction is achieved with a combination of all three models. This approach leads to $R = 0.70$ on the 2008 data respectively $R = 0.77$ on 2009 data.

5. Conclusions and future work

We compared and combined three approaches for instrumentally predicting TTS quality on the two auditory test databases from the Blizzard Challenges 2008 and 2009. Over all databases the combination of all three models achieved the best performance. While the correlations on a per-stimulus basis were quite disappointing compared to the results those models achieved on three German test databases [16], the per-synthesizer scores lead to very promising results. 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. This shows that it is not yet possible to predict the quality of single stimuli however a ranking of different synthesizer systems comes within reach.

Since the two parametric approaches were highly optimized on the available data a generalization for other databases can not be given. To verify the results we plan to extend the analysis to the data of Blizzard Challenge 2010. We will also test the prediction algorithms on the telephone bandpass-filtered speech samples from Blizzard Challenge 2009, since those files passed the same preprocessing steps as the synthesized speech files used in [16] which lead to the best prediction accuracy. Furthermore, we want to combine the general speech parameters with the HMM-based approach using the feature extraction described in [19] as input for both the normalized Log-Likelihood measure as well as the HMM-training.

Finally, we need to test our model on independent data, e.g from future Blizzard Challenges or other evaluated synthetic speech databases in order to analyze the robustness of our approach.

6. Acknowledgement

The authors would like to thank Simon King and Vasilis Karaiskos from the Blizzard Challenge organization team for making available the data from previous challenges and supporting us throughout our research.

	training data	database							
		BC 2008				BC 2009			
		R	$RMSE$	ρ	$\overline{\overline{LL}}$	R	$RMSE$	ρ	$\overline{\overline{LL}}$
per stimulus	Roger Corpus 20min	0.13	1.03	0.13	-24.19	-0.15	1.33	-0.17	-23.67
	Roger Corpus 180min	0.12	1.04	0.11	-22.88	-0.16	1.29	-0.19	-22.70
	BC 08 natural speech files	**0.30**	0.80	0.31	-18.72	**0.09**	1.34	0.08	-18.53
	BC 09 natural speech files	0.25	0.87	0.26	-19.20	-0.06	1.22	-0.06	-18.66
per synthesizer	Roger Corpus 20min	0.14	0.65	0.06	-24.19	-0.10	1.09	-0.19	-23.67
	Roger Corpus 180min	0.22	0.68	0.15	-22.88	-0.14	1.03	-0.23	-22.70
	BC 08 natural speech files	**0.63**	0.53	0.47	-18.72	**0.17**	0.81	0.14	-18.53
	BC 09 natural speech files	0.49	0.46	0.58	-19.20	0.02	0.94	-0.04	-18.66

Table 2: Correlations and prediction error between $\overline{LL}$ and MOS score for different HMMs

	database					
	BC 2008			BC 2009		
model	R	$RMSE$	ρ	R	$RMSE$	ρ
per stimulus						
$\overline{LL}$	0.30	0.80	0.31	0.09	1.34	0.08
P.563 parameters	0.29	1.48	0.27	0.26	1.75	0.26
general parameters	-0.02	0.93	-0.02	0.50	0.76	0.48
$\overline{LL}$ + P.563 parameters	0.38	1.30	0.39	0.26	1.67	0.23
$\overline{LL}$ + general parameters	0.18	0.95	0.16	0.46	0.78	0.44
P.563 parameters + general parameters	0.21	1.12	0.19	0.49	1.25	0.51
$\overline{LL}$ + P.563 parameters + general parameters	**0.30**	1.11	0.28	**0.49**	1.09	0.48
per synthesizer						
$\overline{LL}$	0.63	0.53	0.47	0.17	0.81	0.14
P.563 parameters	0.78	0.71	0.83	0.52	0.85	0.44
general parameters	-0.03	0.98	0.05	0.84	0.48	0.93
$\overline{LL}$ + P.563 parameters	0.82	0.51	0.67	0.45	0.77	0.38
$\overline{LL}$ + general parameters	0.42	0.54	0.29	0.73	0.50	0.73
P.563 parameters + general parameters	0.58	0.65	0.51	0.81	0.69	0.90
$\overline{LL}$ + P.563 parameters + general parameters	**0.70**	0.42	0.60	**0.77**	0.62	0.79

Table 3: Correlations and prediction error for all models and every possible combination

7. References

[1] R. Van Bezooijen and V.J. van Heuven. Assessment of speech output systems. In D. Gibbon, R. Moore, and R. Winski, editors, *Handbook of Standards and Resources for Spoken Language Systems*, pages 481–563, Berlin, 1997. Mouton de Gruyter.

[2] C. Delogu, S. Conte, and C. Sementina. *Speech Communication*, chapter Cognitive Factors in the Evaluation of Synthetic Speech, pages 153–168. 1998.

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

[4] D. Sityaev, K. Knill, and T. Burrows. Comparison of the ITU-T P.85 Standard to Other Methods for the Evaluation of Text-to-Speech Systems. *Proc. 9th International Conference on Spoken Language Processing (Interspeech 2006 - ICSLP)*, pages 1077–1080, 2006.

[5] M. Viswanathan and M. Viswanathan. 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] M. Cernak and M. Rusko. An Evaluation of Synthetic Speech Using the PESQ Measure. *Proc. European Congress of Acoustics*, pages 2725–2728, 2005.

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

[8] M. Chu and H. Peng. An objective measure for estimating mos of synthesized speech. *Proc. 7th Int. Conf. on Speech Communication and Technology (EUROSPEECH 2001)*, 3:2087–2090, 2001.

[9] A. Mariniak. A Global Framework for the Assessment of Synthetic Speech Without Subjects. *Porc. 3rd European Conference on Speech Processing and Technology (Eurospeech 1993)*, pages 1683–1686, 1993.

[10] 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.

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

[12] T. H. Falk, S. Möller, V. Karaiskos, and S King. Improving Instrumental Quality Prediction Performance for the Blizzard Challenge. *Proc. Blizzard Challenge Workshop*, 2008.

[13] S. Möller and J. Heimansberg. Estimation of tts quality in telephon environments using a reference-free quality prediction model. *Proc. 2nd ISCA/DEGA Tutorial and Research Workshop on Perceptual Quality of Systems*, pages 56–60, 2006.

[14] S. Möller and T. H. Falk. Single-Ended Quality Estimation of Synthesized Speech: Analysis of the Rec. P.563 Internal Signal Processing. *ITU-T SG12 Meeting*, 2008.

[15] F. Hinterleitner. *Vorhersage der Qualität synthetischer Sprache mittels eines signalbasierten Maßes*. Magisterarbeit, Institut für Sprache und Kommunikation, Fachgebiet Audiokommunikation, Technische Universität Berlin, 2010.

[16] S. Möller, F. Hinterleitner, T.H. Falk, and T. Polzehl. Comparison of Approaches for Instrumentally Prediction the Quality of Text-to-Speech Systems. *Proc. International Conference on Spoken Language Processing (Interspeech 2010 - ICSLP)*, 2010.

[17] ITU-T Rec. P.56. *Objective Measurement of Active Speech Level*. International Telecommunication Union, Geneva, 1993.

[18] L. Rabiner. A Tutotial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proc. of the IEEE*, 77:257–286, 1989.

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

[20] V. Karaiskos, S. King, R. A. J. Clark, and C. Mayo. The Blizzard Challenge 2008. 2008.

[21] S. King and V. Karaiskos. The Blizzard Challenge 2009. 2009.

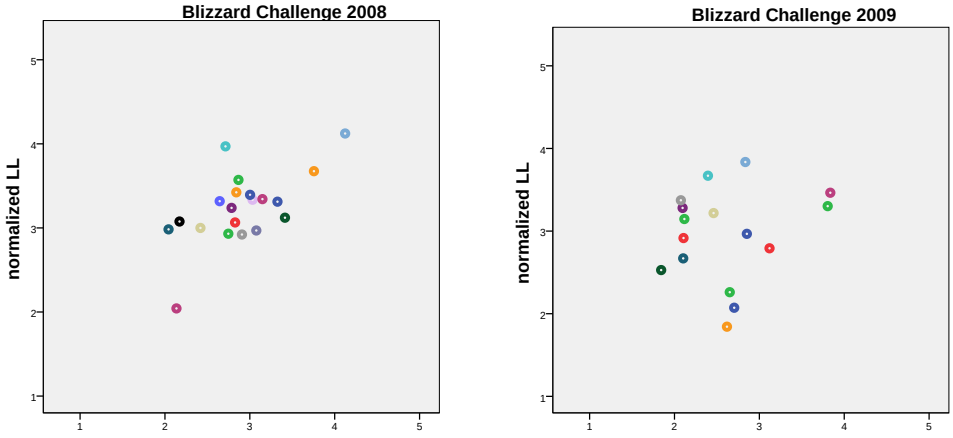

Figure 2: Scatter plots of $\overline{LL}$ and corresponding MOS scores from the BC databases per-synthesizer

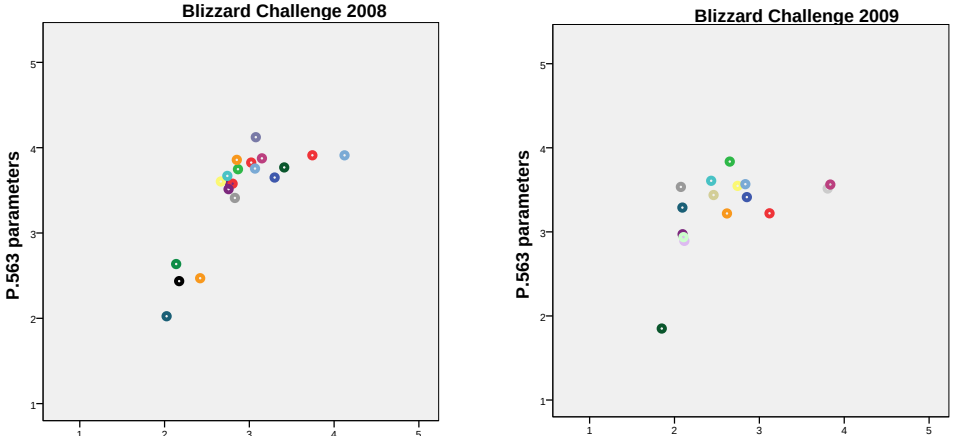

Figure 3: Scatter plots of P.563 parameters and corresponding MOS scores from the BC databases per-synthesizer

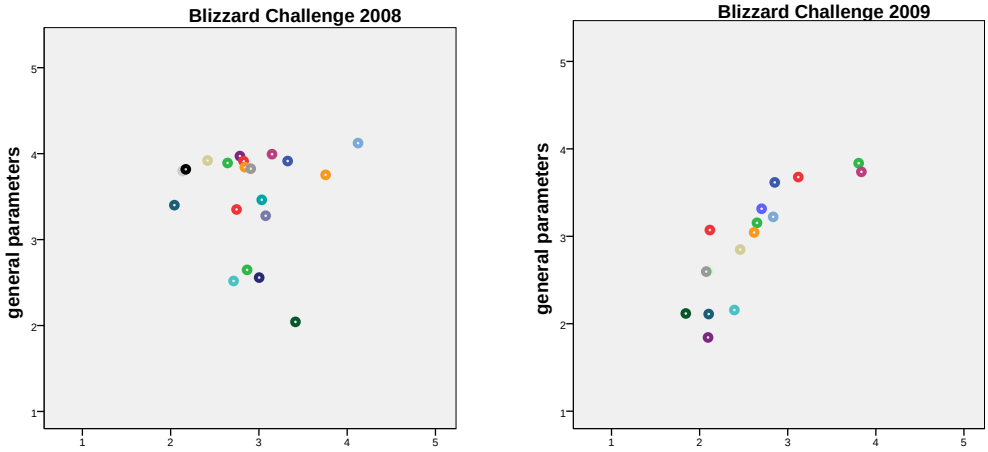

Figure 4: Scatter plots of general parameters and corresponding MOS scores from the BC databases per-synthesizer