Predicting the Quality of Text-To-Speech Systems from a Large-Scale Feature Set

Florian Hinterleitner[1], *Christoph R. Norrenbrock*[2], *Sebastian Möller*[1], *Ulrich Heute*[2]

[1]Quality and Usability Lab, TU Berlin, Germany
[2]Digital Signal Processing and System Theory, CAU Kiel, Germany
florian.hinterleitner@tu-berlin.de, cno@tf.uni-kiel.de,
sebastian.moeller@telekom.de, uh@tf.uni-kiel.de

Abstract

We extract 1495 speech features from 2 subjectively evaluated text-to-speech (TTS) databases. These features are extracted from pitch, loudness, MFCCs, spectrals, formants, and intensity. The speech material is synthesized using up to 15 different TTS systems, some of them with up to 8 different voices. We develop quality predictors for TTS signals following two different approaches to handle the huge set of speech features: a three-step feature selection followed by a stepwise multiple linear regression and an approach based on support vector machines. The predictors are cross-validated via 3-fold cross validation (CV) and leave-one-test-out (LOTO) CV. Due to the high number of features we apply a strict CV method where the partitioning is realized prior to the feature scaling and feature selection steps. In comparison we also follow a semi-strict approach where the partitioning effectively takes place after these steps. In the 3-fold CV case we achieve correlations as high as .75 for strict CV and .89 for semi-strict CV. The more ambitious LOTO CV yields correlations around .80 for the male speakers whereas the results for the female voices show the need for improvement.

Index Terms: quality prediction, text-to-speech (TTS), cross-validation

1. Introduction

Synthetic speech has attained a level of quality that no longer reminds listeners of robot-like voices but of real human speakers. This makes TTS applicable in various different applications, e.g. email readers, information systems, and smart-home assistants. Especially, the boom in e-books and the implied possibility to synthesize the books's content as well as iPhone's spoken dialog system Siri brought numerous people in contact with synthetic speech in their everyday life. The emerging new use cases implicate the necessity for efficient quality assessment methods for TTS signals.

Depending on the aspect of interest a variety of different listening tests can be conducted. Most listening tests however, e.g. the ITU-T Rec. P.85 [1], capture the overall quality as well as inherent quality attributes of the synthesized speech signals. Since listening tests are very time-consuming and additionally extremely cost-intensive a frequent assessment of the quality of TTS systems during development is mostly not feasible. In cases like these instrumental methods that predict the perceived quality of the user without the need for actual human listeners would come in handy.

Previous research in this direction yielded several different helpful speech features. In [2] and [3] the benefit for instrumental quality assessment of prosodic as well as MFCC parameters has been explored. In [4] a large-scale feature set [5] was used for quality prediction. The results already revealed the relevance of this feature set. In this paper we present further research on TTS quality prediction on the basis of the mentioned feature set. We present two new approaches for quality estimation. Both predictors were developed applying either 3-fold CV or LOTO CV and tested on two comprehensive subjectively evaluated German TTS databases.

The following section presents the TTS databases that were used within this study. Section 3 outlines the feature extraction algorithm. In Section 4 we illustrate different approaches on model assessment. The two quality predictors are presented in Section 5 while the results and the discussion of these approaches are shown in Section 6. Finally, Section 7 concludes the outcome of this paper and gives a perspective to future work.

2. Databases

This section presents two subjectively evaluated TTS databases which were compiled during two extensive studies on quality dimensions of synthetic speech.

2.1. Test 1

The Test 1 database resulted from a study in which the inherent quality dimensions of state-of-the-art TTS systems were investigated [6]. 14 female and 15 male synthesizers were used to generate 2 samples per system configuration. All stimuli were downsampled to f_S=16kHz and level normalized to -26dBov prior to listener presentation. The average sample duration was 9-10s. In a first pretest 2179 quality describing attributes were collected out of which 296 unique descriptions were found. These attributes were condensed into 44 scales. In a second pretest this set of attribute scales was narrowed down to 16. In the main test 30 listeners (15 female, 15 male, mean age: 27.9 years) assessed all 60 stimuli (30 female, 30 male) on the overall impression scale and on the 16 attribute scales from the pretest. All participants were native German speakers, non of them had any known hearing disabilities. The stimuli were presented via head-phones (Sennheiser HD 485) and a high-quality sound device (Roland Edirol UA-25) in a soundproof booth.

2.2. Test 2

The second database was gathered during a multidimensional scaling experiment [7]. To gain deeper insight into the quality

of the stimuli presented in this study a post test was conducted. The 30 female and 27 male stimuli, all synthesized by different TTS system configurations, that were evaluated in the main test were thus rated on the same scales as described in Test 1. Prior to listening, the stimuli were downsampled to f_S=16kHz and level normalized to -26dBov. 12 test participants (5 female, 7 male, mean age: 27 years), 5 expert listeners from the Quality and Usability Lab, TU Berlin and 7 naïve participants took part in the test. All of them were native German speakers. The stimuli had an average duration of 5s and were presented via head-phones (Sennheiser HD 485) and a high-quality sound device (Roland Edirol UA-25) in a quiet listening environment.

2.3. Similarities and differences between databases

10 German sentences from the EUROM.1 corpus [8] were selected as source material for Test 1 out of which each system synthesized two. One of these sentences was shortened to an approximated synthesized duration of 5s and used as source material for Test 2. Thus, all stimuli in Test 2 contain the exact same wording, while only 6 out of the 60 stimuli from Test 1 were synthesized from the same text.
In Test 1 we used most TTS systems with German voices that were available at that time, but not necessarily all available voices per system. Test 2 covers the same synthesizers with the addition of two newer systems. Moreover, Test 2 contains up to 8 different voices per synthesizer.
Finally, both tests also differ according to the invited test participants, i.e., none of the participants from Test 1 were part of the study in Test 2.

3. Feature extraction

As a basis for the quality prediction approaches in Section 5 we use the feature extraction algorithm described in [5]. The extracted features provide a broad variety of information on vocal expression patterns that are useful when classifying human emotions. As depicted in [4] the inherent information is also suitable when analyzing the quality of synthetic speech.
In the first step the following low-level audio descriptors are extracted: pitch, loudness, MFCC, spectrals, formants, and intensity. Subsequently, a statistic unit derives moments, extrema, linear regression coefficients and ranges of the respective acoustic contours. Among others this yields features like: pitch range, maximum value of the second formant, the mean of a Mel frequency cepstral coefficient, and the maximum change in spectral flux.
We applied the feature extraction on the databases from Section 2, extracted 1495 features per file and used them as an input for the quality prediction approaches in Section 5.

4. Model assessment

Different cross-validation (CV) schemes are used in order to investigate the extent to which the broad feature pool can be exploited for quality prediction. Due to the high number of available features, special care is advisable here in order to regard the overfitting problem. Random 3-fold CV [9] is used for intra-test model validation, i.e., models are trained on $^2/_3$ of the data of a given database, and tested on the remaining $^1/_3$ of that database. Second, a deterministic inter-test CV (LOTO) is performed, where the model is trained using one database, and tested on the other (and vice versa). Thus, the results of different auditory tests are compared on the basis of model generalization. Regarding feature selection, we differentiate between a strict and a semi-strict CV scenario.
Figure 1 depicts the strict CV scenario. The feature matrix is denoted by $\mathbf{X}$. The target vector $\mathbf{y}$ contains the auditory ratings. For the k-th CV partitioning, model training comprises feature normalization, supervised feature selection, and model training. The scaling information $\boldsymbol{\eta}$, the indices of the selected features $\boldsymbol{\iota}$, and the model parameter vector $\boldsymbol{\beta}$, which are evaluated using the training data, parametrize the estimation of the test data. A comparison between the estimate $\hat{\mathbf{y}}_{\text{test}}$ and the true auditory ratings $\mathbf{y}_{\text{test}}$ over all k yields the average Pearson correlation $\overline{R}$ and the average root-mean-square error $\overline{\epsilon}$.
The semi-strict CV scenario in Figure 2 differs from the strict case in that the scaling and the selection information is effectively determined before the CV partitioning, using the complete data. Hence, whereas the strict CV prohibits any prior knowledge about the data, the semi-strict CV is limited to avoiding "technical" model overfitting. Note, that the mentioned feature selection refers only to the supervised selection which involves the target ratings ($\mathbf{y}$). An unsupervised feature screening is conducted in any case, removing singular features with atypical variance.

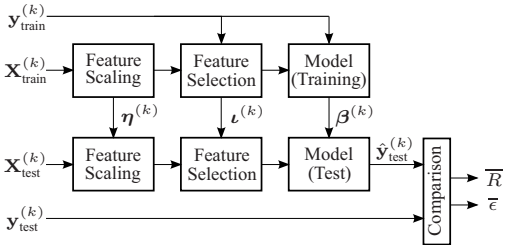

Figure 1: Strict cross-validation with feature selection.

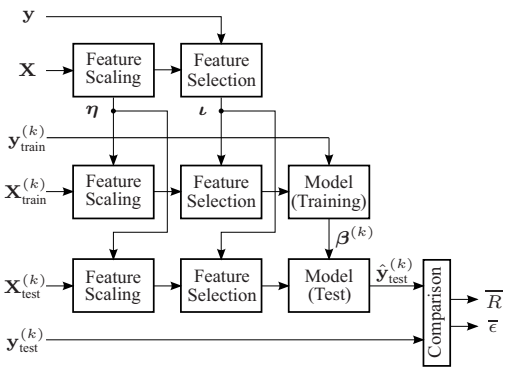

Figure 2: Semi-strict cross-validation with feature selection.

5. Predictors

In this section we present two different approaches which are able to cope with extensive feature sets as described in Section 3. Due to the differences between female and male speech and the experience that the importance of features for quality prediction heavily depends on the speaker gender of the stim-

uli [10] we develop separate quality prediction models for each gender.

5.1. Three-step feature selection and stepwise regression

The main idea of this approach is to reduce a huge feature set (in this case 1495 features) via a three-step feature selection (FS) to a small subset (fewer than 10) while keeping enough relevant information to be able to build an effective quality prediction model via a stepwise multiple linear regression (SR). The FS-SR approach is based on the algorithm described in [11] and can be seen in Figure 3.
In a first step we use the Relief algorithm [12] to omit irrelevant features by a relevance ranking of all i features in the feature set. Relief finds the nearest "hit" and the nearest "miss" for each feature x_i in the set, i.e. another sample of the same class and another sample of a different class, respectively. Subsequently, it adjusts the relevance value r of each feature x_i according to (1).

$$r = (x_i - \text{near-hit}_i)^2 + (x_i - \text{near-miss}_i)^2 \qquad (1)$$

The so-called relief-F algorithm is an adaptation of relief for the handling of noisy, incomplete as well as multi-class data sets [13]. In this paper we use the Matlab implementation of relief-F to retain the 12.5% most relevant features, e.g. 187 features, for further processing.
In a second step we remove redundancy from this feature subset by applying the k-means algorithm [14] to cluster features into groups of similar features. We build 10 feature clusters and select the feature with the highest relief-F relevance value as a representative of this cluster. The k-means cluster solution is strongly dependent on the randomly chosen starting points of the cluster centers. Thus, each execution of k-means yields slightly different clusters and representatives. Therefore, we count the occurrence of representatives throughout 2000 executions of k-means and select all features with an occurrence rate above 30%. This reduces the number of features below 15 throughout all databases.
In a third and final step we use a stepwise multiple linear regression to select relevant features from the previous subset and to build a prediction model.
Thus, the steps relief-F, k-means and the stepwise multiple linear regression from Figure 3 are all part of the *feature selection* blocks in Figures 1 and 2. Moreover, the stepwise multiple linear regression also develops the final prediction model as shown in the blocks *model (training)* and *model (test)*.

5.2. Support Vector Regression

Support-vector regression (SVR) adopts the support-vector-machine (SVM) principle for function estimation [9]. Aiming at a simple, i.e. flat, regression function, which approximates the training data with limited precision ϵ, a model can be found by minimizing the empirical risk (error) [15]. The tradeoff between model complexity and training error is adjusted via a constant C which is kept fixed throughout this study. We use a special case of SVR, called ν-SVR [15], [16], where ϵ is adaptively determined through a fixed constant ν. The radial basis function is chosen as the kernel type. Features and auditory ratings are scaled to $[0, 1]$ for model training and testing. Supervised feature selection is performed prior to model training. Only features with a minimum correlation magnitude, i.e., $|R| \geq 0.4$, are used for the evaluation of the SVR model.

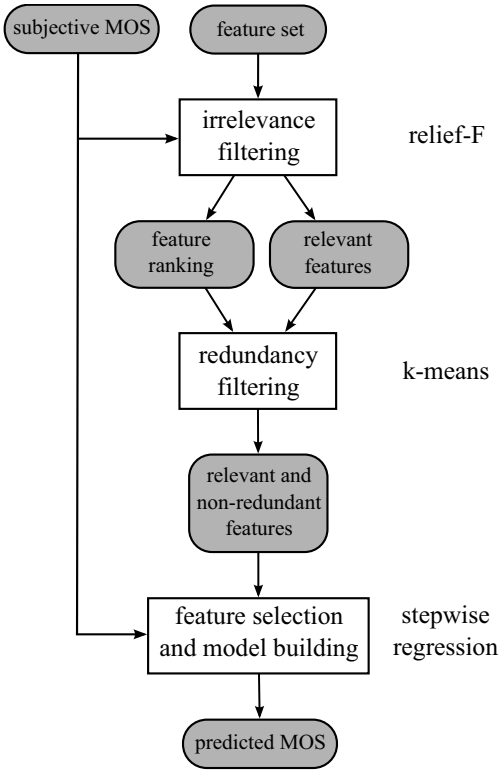

Figure 3: Three-step feature selection and model building

6. Results and discussion

We built models considering the model assessment techniques presented in Section 4. The results can be seen in Tables 1 and 2. As can be seen from Table 1, the average correlation between auditory and predicted mean opinion score (MOS) for the 3-fold CV varies between .35 and .89. The error range is 0.39-0.94, where the MOS ratings have been scaled to the common absolute-category-rating (ACR) scale $[1, 5]$ beforehand. Note that good predictive performance of the model is indicated by high correlations and low error values. In all cases, the semi-strict CV case yields better figures, which reveals the influence of feature selection outside the CV loop.
Moreover, this table shows that the SVR approach outperforms the FS-SR in 3-fold CV, especially for the case of the strict CV. When comparing the results across speaker gender both models achieve a higher prediction accuracy for the male data except in the case of a semi-strict CV on Test 1 data.
Turning to the results of the LOTO CV in Table 2, the semi-strict CV likewise leads to evidently better correlations than the strict CV independent of the prediction model and the speaker gender with the exception of the FS-SR approach on female data. The correlations vary between .43 and .89 while the error range is 0.45-0.88. This shows, compared to the results of the 3-fold CV, that the quality prediction task with LOTO CV is more ambitious because the models are trained on one database and tested on the other. Strikingly, the correlations for both models with LOTO CV on the male data exceed the averaged correlations that are achieved via a 3-fold CV. This effect applies especially for the FS-SR approach with an average 3-fold correlation of .51 in the strict CV case and .70 in the semi-strict CV case compared to correlations of .74 and .81 respectively in the

LOTO CV. A possible cause lies in the nature of the databases. As mentioned in Section 2.3 both databases contain a very similar set of TTS systems and are thus not completely independent. Moreover, the models built during the 3-fold CV are always trained on only $^2/_3$ of one database. This implicates that models in some of the 3-fold CV loops are built on a training set that contains no stimuli of synthesizer A while the test set contains only stimuli of synthesizer A (in extreme cases this can happen with entire synthesizer types e.g. no unit selection, diphone or HMM-synthesizer). This can lead to very low correlations in some 3-fold CV loops which affect the average correlation of the model to a degree that occasionally makes the 3-fold CV inferior to the LOTO CV. Moreover, since the models in the LOTO CV are trained on the whole database they are theoretically more stable then the models from the 3-fold CV and as a result they can lead to better correlations on other databases.
Furthermore, it is noticeable that compared to the 3-fold CV the performance gap between FS-SR and SVR for the male data is considerably smaller. Additionally, the prediction accuracy of both models is substantially better for the male voices than for the female.

Table 1: Performance of quality prediction models on the test sets (**3-fold** cross-validation). The figures are averaged over 500 random CV partitionings.

TEST	MODEL	MALE		FEMALE	
		$\overline{R}$	$\overline{\epsilon}$	$\overline{R}$	$\overline{\epsilon}$
I	FS-SR*	**.517**	0.761	**.475**	0.784
	SVR*	**.720**	0.566	**.634**	0.648
	FS-SR**	**.644**	0.595	**.731**	0.532
	SVR**	**.855**	0.423	**.889**	0.392
II	FS-SR*	**.493**	0.767	**.354**	0.940
	SVR*	**.751**	0.681	**.571**	0.826
	FS-SR**	**.747**	0.501	**.683**	0.617
	SVR**	**.887**	0.496	**.848**	0.534

*Strict CV. **Semi-strict CV.

Table 2: Performance of quality prediction models on the test sets (**Leave-one-test-out** cross-validation).

TEST	MODEL	MALE		FEMALE	
		$\overline{R}$	$\overline{\epsilon}$	$\overline{R}$	$\overline{\epsilon}$
I,II	FS-SR*	**.739**	0.642	**.425**	0.879
	SVR*	**.796**	0.550	**.614**	0.725
	FS-SR**	**.810**	0.517	**.427**	0.839
	SVR**	**.893**	0.450	**.775**	0.594

*Strict CV. **Semi-strict CV.

Table 3 depicts the average number of features per model. While the FS-SR approach uses 4 features independent of CV method the SVR models consist of way over 100 features. Furthermore, the number of features for the SVR models depends on the type of CV: a 3-fold CV uses more features than a LOTO CV, a strict CV uses more features than a semi-strict CV.
Moreover, this table points out that one reason for the superior performance of the SVR approach lies in the number of selected features. With between 29 and 69 times the number of features of the FS-SR models the SVR models are far more comprehensive. A second advantage of the SVR approach is the ability to model non-linear relationships while the FS-SR approach relies on the linear procedure of a stepwise regression. Thus, these two facts explain most of the different prediction accuracy as seen in the Tables 1 and 2.

Table 3: Average number of features per model.

	3-fold*	3-fold**	LOTO*	LOTO**
FS-SR	4	4	4	4
SVR	277	210	219	116

*Strict CV. **Semi-strict CV.

The results reflect the range of predictor performance which may be achieved without explicit feature construction, i.e., by using a large feature pool which describes a speech signal on virtually all physical levels. From a statistical viewpoint, feature selection outside the CV loop introduces an optimistic bias since potential randomness is not necessarily leveled out. In contrast, the strict CV case can be seen as pessimistic since the simulation of "non-knowledge" may appear artificial in that the credit of empirical evidence is largely truncated. We believe that a realistic performance range can be hypothesized through the reported figures. The features that have been chosen during feature selection cover a wide range of different signal properties.

7. Conclusions and future work

We used a feature extraction algorithm to extract 1495 speech features on two different subjectively evaluated TTS databases. Subsequently, two quality prediction models were developed. The first one consists of a three-step feature selection followed by a stepwise multiple linear regression while the second one is based on support-vector machines. We experimented with different cross-validation techniques, i.e., strict and semi-strict CV. The former divides a database into training set and test set before feature scaling and feature selection steps are taken while the latter realizes the partitioning afterwards. Moreover, we built models via 3-fold CV, which leads to a high intra-test model validation, and LOTO CV, where the model is validated through a second database and thus leads to more generalizable models.
The correlations achieved by means of a 3-fold CV are as high as .75 in the strict CV case and .89 in the semi-strict case. The results for LOTO CV show correlations above .80 for semi-strict CV and the male data. The correlations on the female voices with LOTO CV show that there is still, especially for the FS-SR approach, the need for further improvements.
All in all, the present models have been validated via different CV techniques on 2 different databases, containing up to 15 different TTS systems. Thus, they represent considerable improvements in the field of quality prediction for synthetic speech.
In the future we plan to extend the feature set to various other features that have already been proven to be useful for quality prediction, e.g., additional prosodic features [17] and Fujisaki-Features [18]. Furthermore, we aim at building prediction models trained on the databases of the annual Blizzard Challenge. These models will be used to estimate the quality of the synthesizers of the current competition. In addition, we will test the generalizability of the presented predictors on independent databases.

8. Acknowledgements

The present study was supported by the Deutsche Forschungsgemeinschaft (DFG), grants MO 1038/11-1 and HE 4465/4-1.

9. References

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

[2] C. Norrenbrock, F. Hinterleitner, U. Heute, and S. Möller, "Quality Analysis of Macroprosodic F0 Dynamics in Text-To-Speech Signals," *Proceedings of the 13th Annual Conference of the International Speech Communication Association (Interspeech 2012)*, 2012.

[3] ——, "Towards Perceptual Quality Modeling of Synthesized Audiobooks," *Proceedings of the Blizzard Challenge Workshop*, 2012.

[4] S. Möller, F. Hinterleitner, T. Falk, and T. Polzehl, "Comparison of Approaches for Instrumentally Predicting the Quality of Text-To-Speech Systems," *Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech 2010).*, pp. 1325–1328, 2010.

[5] W. Minker, G. Lee, J. Mariani, and S. Nakamura, *Spoken Dialogue Systems Technology and Design*. Springer, 2010, ch. Salient Features for Anger Recognition in German and English IVR Portals.

[6] 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)*, pp. 2177–2180, 2011.

[7] 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)*, pp. 240–245, 2012.

[8] D. Chan, A. Fourcin, D. Gibbon, B. Grandstrom, M. Huckvale, G. Kokkonakis, K. Kvale, L. Lamel, B. Lindberg, A. Moreno, J. Mouropoulos, F. Senia, I. Trancoso, C. Veld, and J. Zeiliger, "EUROM- A Spoken Language Resource for the EU," *Proceedings of the 4th European Conference on Speech Communication and Technology (EUROSPEECH 1995)*, pp. 867–870, 1995.

[9] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, 2009.

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

[11] J. Bins and B. Draper, "Feature Selection from Huge Feature Sets," *Proceedings of the 8th IEEE International Conference on Computer Vision 2001 (ICCV 2001)*, 2001.

[12] K. Kira and L. Rendell, "The Featture Selection Problem: Traditional Methods and a New Algorithm," *Proceedings of the 10th National Conference on Machine Intelligence*, pp. 129–134, 1992.

[13] I. Kononenko, E. Simec, and M. Robnik-Sikonja, "Overcoming the Myopia of Inductive Learning Algorithms with RELIEFF," *Applied Intelligence*, vol. 7, pp. 39–55, 1997.

[14] J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," in *Proceedings of the 5th Berkley Symposium on Mathematics, Statistics and Probability*, L. LeCam and J. Neyman, Eds. University of California Press, Berkley, 1967, pp. 281–297.

[15] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett, "New Support Vector Algorithms," *Neural Computation*, pp. 1207–1245, 2000.

[16] C.-C. Chang and C.-J. Lin, "Libsvm : a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 27:1–27:27, 2011.

[17] C. Norrenbrock, F. Hinterleitner, U. Heute, and S. Möller, "Instrumental Assessment of Prosodic Quality for Text-To-Speech Signals," *IEEE Signal Processing Letters*, vol. 19, pp. 255–258, 2012.

[18] F. Hinterleitner, C. Norrenbrock, and S. Möller, "On the Use of Fujisaki Parameters for the Quality Prediction of Synthetic Speech," *Proc. of the 23th Konferenz Elektronische Sprachsignalverarbeitung (ESSV), Cottbus, (Germany)*, 2012.