Towards Perceptual Quality Modeling of Synthesized Audiobooks – Blizzard Challenge 2012

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

[1]Digital Signal Processing and System Theory, Christian-Albrechts-University of Kiel, Germany
[2]Quality and Usability Lab, Telekom Innovation Laboratories, TU Berlin, Germany
{cno, uh}@tf.uni-kiel.de, {florian.hinterleitner, sebastian.moeller}@telekom.de

Abstract

This paper reports on recent advances in the field of instrumental quality evaluation of text-to-speech (TTS) synthesis. In particular, a wide range of acoustic quality markers are analyzed concerning their quality-describing power using the audiobook data from the Blizzard Challenge 2012. Several approaches for perceptual modeling are investigated and compared with each other. The results reveal substantial correlations as high as 0.87 between subjective ratings of overall impression and their estimates.

Index Terms: Speech quality, instrumental quality assessment, text-to-speech (TTS), audiobook.

1. Introduction

In general, the basic functionality of TTS systems, the conversion of orthographic text into spoken speech, is useful for a vast number of applications, which are rapidly emerging through the dissemination and performance of today's communication systems. Although this is certainly true for specific applications, e.g., reading systems for the blind, the mainstream adoption of TTS is mainly governed by its perceptual quality [1]. In this context, synthesizing audiobooks is probably one of the most challenging applications one can imagine, because the focus is shifted from pure functionality, e.g., reading of a weather forecast, towards enjoyment of literature.

This year's Blizzard Challenge (2012) included an evaluation section exclusively dedicated to synthesized audiobook paragraphs. In this paper an attempt is made to objectify the corresponding test results from an acoustic viewpoint. We investigate to what extent systematic patterns of quality variation could be uncovered by means of purely acoustical features from two perspectives, first on a basic level by analyzing individual features, and second by evaluating regressive modeling techniques provided by the machine-learning community. We emphasize our specific interest in seeking the potential for automatic, i.e., instrumental evaluation methods for TTS on a large non-laboratory scale, being aware that a purely acoustical approach bears challenges but, in our view, also the best perspective for the practical application. In this spirit we show work in progress of a research field which has yet to emerge. The paper is organized as follows: In Section 2 we provide relevant listening test details. Speech features are introduced and analyzed in Section 3. Sections 4 and 5 report on the regression methods and their results, respectively. In Section 6 we give insight into ongoing research of prosodic quality estimation using the Fujisaki model. A short conclusion marks the end of the paper.

2. Audiobook Listening Test

The audiobook listening test was part of the evaluation process of the Blizzard Challenge 2012. It consisted of two sections, one with natural reference signals and one without. In both sections listeners had to rate stimuli generated by 10 synthesizers named B-K, based on the same male voice corpus. A total of 231 files were evaluated. The listening test was designed between groups, i.e., all participants within one group listened to the same stimuli. 117 paid participants were spread across 11 groups. The test took place in the test lab at the University of Edinburgh. All stimuli were downsampled to a sampling rate of 16 kHz; the mean duration was 44 s. A subset of the attribute scales presented in [2] was chosen for the evaluation in this test. These scales are: overall impression, voice pleasantness, speech pauses, accentuation, intonation, emotion, and listening effort. The scores were given on a continuous scale. The participants were giving the score as the distance of the handle from the left end of the slider, but could not see the actual number. Due to the fact that this study is based on the preliminary test results from the test lab that did not include the results from the online listening test, we will restrict ourselves to the overall impression ratings of the TTS systems only ($N = 220$). In particular, we use the median values, i.e., “median opinion scores“ (MOS), $\mathbf{y} = [y_1, ..., y_N]^T \in \mathbb{R}^{N\times 1}$, with all ratings scaled to $y_n \in [1, 5] \,\widehat{=}$ [bad, excellent].

3. Speech Features

Two main feature classes are considered in this paper. The first group, named “prosodic”, consists of a range of features derived from F_0 (fundamental frequency) and temporal structure; the second group is based on mel-frequency cepstral coefficients (MFCCs). In general, a set of features is denoted as a matrix,

$$\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_i, ..., \mathbf{x}_I] \in \mathbb{R}^{N\times I}, \tag{1}$$

comprising N observations (stimuli) of I features. A full description of the used features is beyond the scope of this paper; hence we will refer to other publications and rather focus on specific features which we find important for the quality modeling, especially in view of diagnostics.

3.1. Prosodic Features

Two subgroups of prosodic features are referred to in the following: (1) Intonational (macro-prosodic) features, and (2) perturbation (micro-prosodic) features, which add up to $I = 26$ in total.

3.1.1. *Macro-Prosodic*

In previous studies, a range of F_0-related features has been analyzed with respect to their quality-describing power, see, e.g. [3]. Along with basic F_0 features (mean, standard deviation, Δ) and known rhythm parameters, nonlinear parameters, derived from F_0 dynamics in voiced sections have been identified as useful. All features from [3] are evaluated and utilized in the present study.

In the following we illustrate the concept of auditory thresholding for prosodic quality evaluation. For notation, let $F_0(l, v)$ be the pitch contour of the l-th voiced segment ($F_0(l, v) \neq 0$), with $l = 1, 2, ..., L$ and $v = 1, 2, ..., V_l$. L is the number of voiced segments per signal and V_l denotes the number of F_0 samples evaluated every 10 ms. The *variability ratio* (VR) is defined as the relative number of segments with a minimum mean derivative:

$$\mathrm{VR} = \frac{1}{L}\sum_{l=1}^{L}\delta_\xi\left(\frac{1}{V_l-1}\sum_{v=1}^{V_l-1}|\Delta_v F_0(l,v)|\right). \qquad (2)$$

The step function δ_ξ is 1 beyond a threshold $\xi \in \mathbb{R}^+$, and 0 otherwise. The (discrete) delta operator is evaluated w.r.t. v.

Furthermore, we introduce a variant of (2) which we call the *weighted variability ratio* (WVR):

$$\mathrm{WVR} = \frac{1}{L}\sum_{l=1}^{L}\ln(V_l)\delta_\xi\left(\frac{1}{V_l-1}\sum_{v=1}^{V_l-1}|\Delta_v F_0(l,v)|\right). \qquad (3)$$

The weighting factor $\ln(V_l)$ accounts for the length of the voiced sections, thus emphasizing longer sections which can be assumed to be perceptually more relevant than very short segments. The optimum threshold ξ is evaluated by simple grid-search, using the Pearson correlation (see Section 4) as a criterion:

$$\xi_{\text{opt}} = \arg\max_{\xi} |\mathrm{R}(\mathbf{x}_{\mathrm{VR}}(\xi), \mathbf{y})| . \qquad (4)$$

The result can be traced in Figure 1 for both VR and WVR. The optimum threshold is found at $\xi \approx 0.7$, which matches quite well the value of 0.65 found for male German voices [3]. A detailed scatter plot for WVR is given in Figure 2. The per-stimulus (R_{stim}) and per-system (R_{sys}) correlations are reported along with their significance (p). With regard to the TTS systems (B-K) a clear clustering effect can be noted.

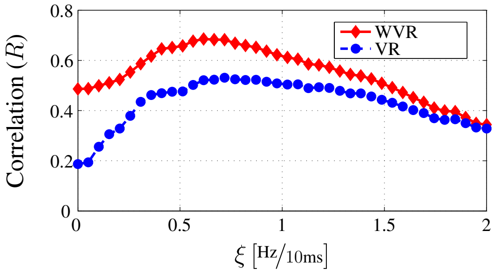

Figure 1: Evaluation of the optimum threshold ξ w.r.t. correlation with MOS.

3.1.2. *Micro-Prosodic*

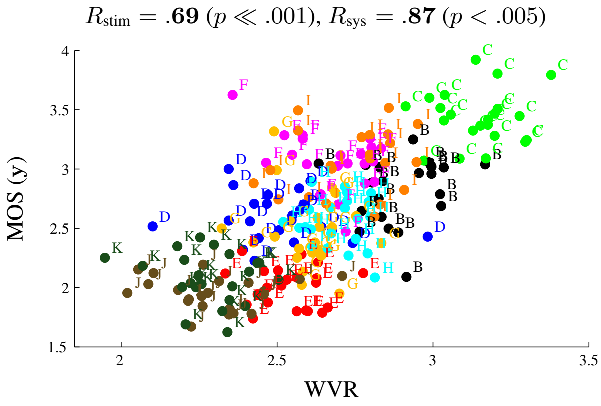

Figure 2: Weighted Variability Ratio vs. MOS.

In [4] a range of so-called perturbation measures (e.g., jitter, shimmer) have been analyzed. These measures were designed to capture the excitation-related aperiodicity of the vocal source, thus describing, at least partly, perceived voice quality. It has been found that high-rated TTS signals often exhibit much higher perturbation than the lower rated ones [4], hence more than 60% of the perceived quality variance could be explained. One useful parameter to evaluate the (inverse) aperiodicity without pitch marking is by means of the *cepstral peak prominence* (CPP) which is defined as the average relative hight of the first rahmonic above a normalizing regression line through the cepstrum. Smoothing across time and quefrency yields the smoothed CPP (CPPS) which we found to be preferable. For the present data, Figure 3 illustrates a marked negative correlation, confirming our previous findings [4]. A closer look reveals that systems E, H, and K, for which we find the voice quality most “buzzy” *and* “muffled“, exhibit high CPPS values. It can be hypothesized that when a certain level of periodic dominance is exceeded, the quality drops markedly and pulls down overall quality. However, a clear borderline cannot be inferred here which could be due to several reasons, e.g., influence of macroprosody (see WVR). More generally, finding reliable voice-quality descriptions for connected speech in TTS is challenging: On the one hand, the ambiguity between macro- and micro-F_0 movement can hardly be totally resolved, on the other hand, the vocal-tract filtering of the source signal influences its perception, which in turn might affect voice naturalness, since, e.g., nonlinear couplings between glottis and vocal-tract are usually not explicitly modeled. Apart from this, other distortions which contaminate the measurements ought to

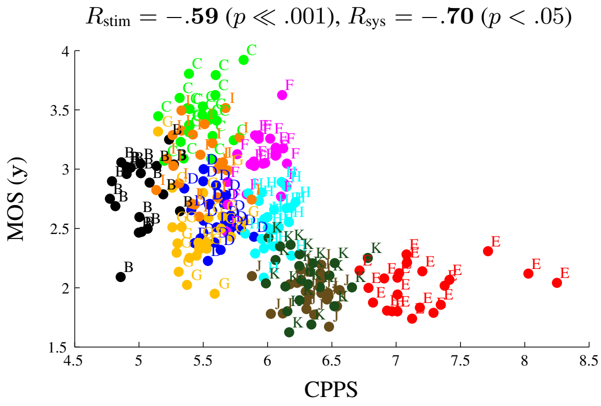

Figure 3: Smoothed CPP vs. MOS.

be considered also.

3.2. MFCCs

The wide-spread MFCC description (e.g., [1]) is effectively complementing our feature set in terms of amplitude modulation; contrasting to the prosodic features since the harmonic (F_0) resolution is essentially dropped through mel-band filtering (not through the “cepstral“ DCT (discrete cosine transformation)). The MFCC representation describes the shape of the mel-spectrum through its DCT spectrum, which has low resolution (e.g., 20 bins at 8 kHz sampling frequency). As such, the MFCCs are the resulting weights of basis (cosine) functions. We use 12th-order MFCCs, $\mathbf{c}_m = \{c_0, ..., c_{12}\}_m$, calculated for the m-th speech frame; the used features are the mean (μ) and standard deviation (σ) of each coefficient, its delta (Δ) and delta-delta ($\Delta^{(2)}$) values, calculated over all active speech frames per signal. Thus, we have:

$$\mathbf{x}_{\text{MFCC}} = [\mu(\mathbf{c}, \boldsymbol{\Delta}, \boldsymbol{\Delta}^{(2)}), \sigma(\mathbf{c}, \boldsymbol{\Delta}, \boldsymbol{\Delta}^{(2)})] \in \mathbb{R}^{1\times 78}. \quad (5)$$

Considering individual feature correlations, the σ values show some consistent relationship with quality, with varying correlation as shown in Figure 4. Note that a correlation of ±0.3 is already highly significant ($p < 10^{-4}$). In Figure 5, the σ values of the fifth delta MFCC is shown. From the negative correlation it can be inferred that the variation of the (mel) spectral differences between successive speech frames is inversely related to the quality. In our view, this corresponds to the impression of articulatory smoothness and continuity. A comparison to the research line of acoustic-distance join costs in TTS [1] is interesting, since the issue of measuring perceptual continuity with spectral distance measures remains only partly solved. The correlations (with MOS) that we notice for our data are comparable to those for measuring the ”goodness of join“ ($R \approx .6$) [1]. From Figure 4 one can also loosely hypothesize that ”envelope continuity“ is most relevant in the DCT-frequency range 3-9.

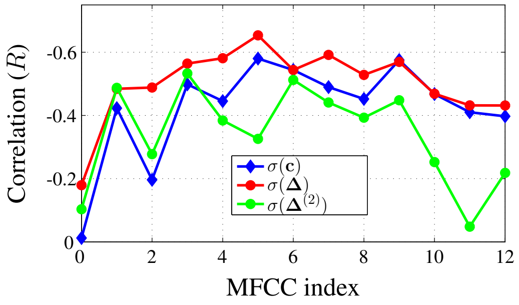

Figure 4: Correlation of standard deviations of MFCCs (Note the flipped y axis for better readability in terms of magnitude).

4. Instrumental Quality Modeling

The ultimate goal of instrumental quality prediction is to replace auditory tests as much as possible, hence a high scientific reliability needs to be guaranteed. In Section 3 it has been shown that research on individual features is necessary in order to gain true insight into the nature of quality perception. Yet, due to the multidimensional nature of perception and the inter-rater noise, appropriate models have to account for the non-obvious (but systematic) pattern in the data, thereby closing the variance gap as much as possible. For the purpose of feature-based quality modeling, a vast number of machine-learning techniques have been proposed [5], from which we selected a few for the present study in order to investigate any major differences between them.

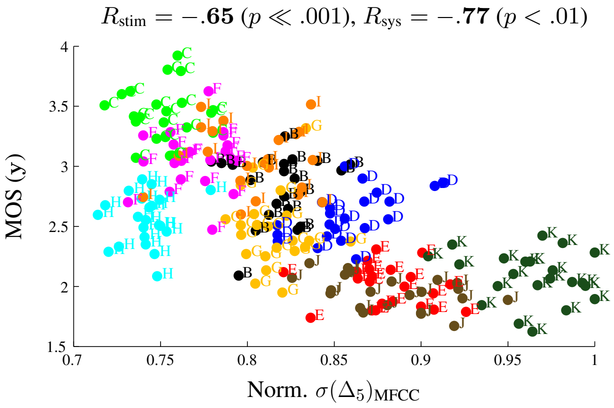

Figure 5: Standard deviation of the fifth delta MFCC vs. MOS.

In general, we aim at solving the standard regression problem by means of supervised learning, that is to find a model $f(\mathbf{X})$ which minimizes the error between subjective auditory ratings $\mathbf{y}$ and their estimates $\hat{\mathbf{y}}$:

$$\|\mathbf{y} - f(\mathbf{X})\|_2^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|_2^2 \overset{!}{=} \text{min}. \quad (6)$$

For realistic datasets (samples) we can usually only achieve an estimate $\hat{f}$ of the true model f. Assuming $\Delta\mathbf{y} \neq 0$, the model fit can be expressed as normalized root-mean-square-error (RMSE):

$$\epsilon(\mathbf{y}, \hat{\mathbf{y}}) = \frac{\|\mathbf{y} - \hat{f}(\mathbf{X})\|_2}{\sqrt{N}|\max(\mathbf{y}) - \min(\mathbf{y})|}, \quad (7)$$

where $\max(\cdot)$ and $\min(\cdot)$ give the maximum and minimum entry of the argument vector, respectively. Another criterion, which specifically focuses on linearity, is the Pearson correlation which we like to maximize:

$$|R(\mathbf{y}, \hat{\mathbf{y}})| = \left| \frac{(\mathbf{y} - \mathbf{y}_\mu)^T \cdot (\hat{\mathbf{y}} - \hat{\mathbf{y}}_\mu)}{\|\mathbf{y} - \mathbf{y}_\mu\|_2 \cdot \|\hat{\mathbf{y}} - \hat{\mathbf{y}}_\mu\|_2} \right| \overset{!}{=} \text{max}, \quad (8)$$

where the index μ indicates a vector of replicating mean. In the following we describe the regression methods used in the experiments. Since not all methods are equivariant under scaling, any feature matrix $\mathbf{X}$ and target vector $\mathbf{y}$ is z-score-normalized (standardized) before processing, thus the regressive intercept does not need to be explicitly considered.

4.1. Cross-Validated Feature Selection (CV-FS)

Correlation-based feature selection is applied according to the wrapper approach [6]. Two cross-validation (CV) loops are adopted which both involve linear multiple regression modeling; in the inner loop, repeated feature selection is conducted [3] where the average feature importance (AFI) serves as selection criterion for the features used in the outer loop, denoted as $\tilde{\mathbf{X}} \in \mathbb{R}^{N\times P}$, with $P \leq I$. Using (6), a model $\hat{f}(\tilde{\mathbf{X}}) = \tilde{\mathbf{X}}\hat{\beta}$ can be found by the least-squares solution:

$$\hat{\beta}^{\text{LS}} = (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\mathbf{y}, \quad (9)$$

where $\hat{\beta}$ denotes the vector of regression coefficients. By using only the selected features $\tilde{\mathbf{X}}$, we practically found the matrix rank issues to be uncritical, which would not necessarily be the case when using the full matrix $\mathbf{X}$, due to highly correlated columns ("multicollinearity").

4.2. Ridge Regression (RR)

This shrinkage method, designed to explicitly reduce the problem of multicollinearity in $\mathbf{X}$, can be seen as more continuous than feature selection since individual features are usually not dropped. The basic idea is to introduce a penalty parameter λ that regularizes the size of the regression coefficients which are obtained through

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \left\{ \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_2^2 \right\}. \tag{10}$$

The closed-form solution is

$$\hat{\beta}^{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}, \tag{11}$$

where $\mathbf{I}$ denotes the identity matrix. The effective shrinkage can be shown to depend on the eigenvalues of $\mathbf{X}$, hence those directions that have large variance are granted larger β_is than those with smaller variance [5].

4.3. The Lasso

The lasso is a modified version of Ridge Regression, whereby the L_2 ridge penalty is replaced by the L_1 lasso penalty $\|\beta\|_1 = \sum|\beta_i|$:

$$\hat{\beta}^{\text{lasso}} = \arg\min_{\beta} \left\{ \frac{1}{2}\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_1 \right\}. \tag{12}$$

The solution is nonlinear in $\mathbf{y}$ and a closed form solution is lost. Practically, the main difference that can be observed is that the lasso can shrink some β_is exactly to zero, thus performing a type of continuous subset selection.

4.4. Principal Component Regression (PCR)

Redundancy in $\mathbf{X}$ can be effectively reduced by principal component analysis (PCA), thus (9) is simply applied for a transformed feature matrix $\tilde{\mathbf{X}} = \mathbf{Z}^{\text{PCA}}_{(P)}$ using the first $P < I$ principal components of $\mathbf{X}$. Since the transformed input columns $\mathbf{z}_p = \mathbf{X}\mathbf{v}_p$ are orthogonal, with the p-th eigenvector $\mathbf{v}_p$ defining the linear combination of the $\mathbf{x}_i$, the regression is equal to the sum of univariate regressions of $\mathbf{y}$ on $\mathbf{z}_p$ [5].

4.5. Partial Least Squares (PLS)

In contrast to PCR which keys only on high variance in the feature matrix (not necessarily yielding the best model to explain $\mathbf{y}$), PLS takes into account the correlation with $\mathbf{y}$. Similar to PCR, the regression is performed using derived inputs (partial least squares directions), where a weighting of each input w.r.t. their univariate effect on $\mathbf{y}$ is incorporated. More details can be found in [5].

4.6. ν-Support Vector Regression (SVR)

In contrast to the previous methods, the modeling criterion is not evaluated for the complete data, rather the focus is shifted towards specific observations by application of the ϵ-insensitive loss function [7],

$$|y - \hat{f}(\mathbf{x})|_\epsilon = \max\{0, |y - \hat{f}(\mathbf{x})| - \epsilon\}, \tag{13}$$

which considers only major errors beyond some $\epsilon > 0$. For the sake of convenient notation, we now refer to the single observation case, i.e., our data is of the form

$$\{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n), ..., (\mathbf{x}_N, y_N) | \mathbf{x}_n \in \mathbb{R}^{1\times I}, y_n \in \mathbb{R}\}. \tag{14}$$

We now aim to find a linear inhomogeneous function $f(\mathbf{x}) = \mathbf{x}\beta + \beta_0$, $\beta \in \mathbb{R}^{I\times 1}$, $\beta_0 \in \mathbb{R}$, by minimizing the following expression:

$$\text{minimize} \quad \frac{1}{2}\|\beta\|_2^2 + C \cdot \left(\nu\epsilon + \frac{1}{N}\sum_{n=1}^{N}(\xi_n + \xi_n^*) \right) \tag{15}$$

$$\text{subject to} \quad (\mathbf{x}_n\beta + \beta_0) - y_n \le \epsilon + \xi_n \tag{16}$$

$$y_n - (\mathbf{x}_n\beta + \beta_0) \le \epsilon + \xi_n^* \tag{17}$$

$$\xi_n^{(*)} \ge 0, \ \epsilon \ge 0. \tag{18}$$

The regularizing parameter C determines the trade-off between model complexity and training error (observations lying outside the ϵ-region by $\xi_n^{(*)}$). Via the constant ν the size of ϵ is conveniently determined, since $\nu \in [0, 1]$ is the sensible setting range. Incorporating an arbitrary kernel function, the estimate of the regression function can be shown to be [7]:

$$\hat{f}(\mathbf{x}) = \sum_{n=1}^{N}(\hat{\alpha}_n^* - \hat{\alpha}_n)K(\mathbf{x}_n, \mathbf{x}) + \hat{\beta}_0, \tag{19}$$

where $\hat{\alpha}_n^{(*)}$ are constrained Lagrange multipliers. $(\hat{\alpha}_n^* - \hat{\alpha}_n)$ is typically only nonzero for part of the data, these cases form the support vectors which define the regression. The radial basis function is chosen as a kernel function, $K(\mathbf{x}_n, \mathbf{x}) = \exp(-\|\mathbf{x}_n - \mathbf{x}\|^2/2\sigma^2)$, thereby accounting for nonlinear interactions between features. We use the LIBSVR library [8] with default configuration parameter settings.

5. Results and Discussion

Figures of merit are summarized in Table 1. We give averaged correlations $\overline{R}^{(\text{CV})}$ and errors $\overline{\epsilon}^{(\text{CV})}$ from 100 random 5-fold CV trials, i.e., we report the model performance using the test sets only. The feature categories "Prosodic" and "MFCC" are used separately and in combination. The penalty parameters for RR and the lasso have been evaluated by 4-fold CV on the training sets. The number of components for PCR and PLS have been set empirically to 8 and 4 respectively. We need more components in PCR than in PLS for PCR to work reasonably well, but still PLS performs better and is thus preferable. SVR can be identified as the best modeling technique for our data. Regarding the feature groups, the MFCCs work somewhat better than the prosodic ones; in any case the combination yields the best result. These results show that the information gain from individual features to feature combination is considerable. Also, CV correlations around .85 show that modeling of the present audiobook data works much better than for the data of previous Blizzard Challenges [9]. We see one main reason for this in the greater stimulus length of the audiobook data. In Figure 6 we present the model performance for CV-FS from one complete CV trial, i.e., with averaged regression coefficients.

6. Ongoing Research on Prosodic Quality

Currently, we are investigating to what extent explicit prosody models could be beneficial for deriving features for an improved

Model	Prosodic		MFCC		Combined	
	$\overline{R}^{(\text{CV})}$	$\overline{\epsilon}^{(\text{CV})}$	$\overline{R}^{(\text{CV})}$	$\overline{\epsilon}^{(\text{CV})}$	$\overline{R}^{(\text{CV})}$	$\overline{\epsilon}^{(\text{CV})}$
CV-FS	**0.78**	0.14	**0.83**	0.13	**0.84**	0.13
Ridge	**0.79**	0.18	**0.85**	0.17	**0.87**	0.17
PCR (P=8)	**0.76**	0.15	**0.82**	0.13	**0.84**	0.13
Lasso	**0.79**	0.14	**0.83**	0.13	**0.86**	0.12
PLS (P=4)	**0.80**	0.14	**0.85**	0.12	**0.86**	0.12
ν-SVR	**0.83**	0.13	**0.86**	0.12	**0.87**	0.11

Table 1: Figures of merit of quality prediction models using different feature sets (5-fold cross-validation).

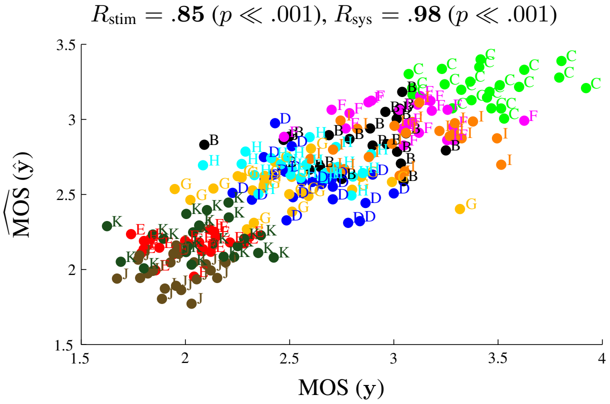

Figure 6: Subjective MOS vs. estimated MOS for one CV trial of "CV-FS" using the average regression coefficients $\hat{\beta}$.

description of prosodic quality. In [10] we introduced a TTS-quality predictor based on the Fujisaki model [11]. This section gives a brief overview of this approach and the results that could be achieved on a subset of the present audio data using an independent as well as an optimized model. The results are preliminary in that the methodology differs from the previous part of the paper for organizational reasons, however, a fusion of the approaches is scheduled.

6.1. Fujisaki Model

Generally, F_0 contours of speech signals are characterized by a decline from the onset towards the end of an utterance. During word accents the F_0 contour is superposed by local intonation humps. The Fujisaki model follows this principle by describing an F_0 contour as a superposition of phrase (PC) and accent commands (AC) and an underlying base frequency (BF). PCs consist of several starting points, each of them with a specific amplitude. Thus, they describe a set of impulses. ACs consist of starting and ending points that describe a set of stepwise functions. The time within one pair of starting and ending points represents an accented block. The BF describes the minimum value of the logarithmized F_0 contour throughout the signal.

The PCs and ACs are the input for two critically-damped second-order linear systems to these commands (*phrase-control mechanism* and *accent-control mechanism*, respectively). The PCs and ACs are assumed to be smoothed by the low-pass characteristics of their respective control mechanisms. The output of those control mechanisms (the *phrase components* and *accent components*) and the BF are then joined to form the pitch curve of an utterance. Thus, this model reduces the complexity of a pitch contour to a minimal set of three parameters (PC, AC, and BF) that still capture the main aspects of the pitch curve.

6.2. Fujisaki Features

We used the Fujisaki model, implemented especially for the German language [12], to extract the parameters introduced in Section 6.1 for every signal under test. We then computed 47 statistical features based on these parameters. They comprise mean, minimum, maximum values as well as the variance of the extracted parameters. Moreover, we computed several features based on the quantity of increasing/decreasing (in relation to the previous command) PC/AC segments in a signal. For a detailed description of all features see [10].

6.3. General German Fujisaki Predictor (GFP)

6.3.1. *Training Database*

We used data from 4 German auditory TTS databases in the training process of a Fujisaki predictor for general German TTS samples. The databases comprised files of at least 6 different synthesizers per gender. All in all the training database consisted of 114/111 female/male TTS signals. For a detailed description see [10].

6.3.2. *Model*

We conducted one stepwise multiple linear regression analysis for each gender. The auditory mean opinion score (MOS) was used as response variable $\mathbf{y}$ while the 47 Fujisaki features ($\mathbf{X}$) described in Section 6.2 were extracted for the files from the training database and used as predictors. For both genders one significant model could be created. To test for over-fitting effects a leave-one-out CV was conducted. The R^2 values for both models could be confirmed; the root-mean-square error (RMSE) showed a minor increase. Thus, both models can be accounted to be stable.

6.4. Audiobook Fujisaki Predictor (AFP)

The same approach as Section in 6.3 was used to create a Fujisaki predictor optimized for the present (English) audiobook data.

6.4.1. *Audiobook Database*

The duration of the input files for the Fujisaki model by Mixdorff [12] is roughly limited to data on sentence level. Therefore, the audiobook files were first manually split up to meet this criterion. Due to this time-consuming process, the following preliminary analysis is based on a subset of $N = 68$ audiobook files. This lead to a set of 459 files for which the Fujisaki features (as described in Section 6.2) were computed (files with less than 3 PCs were omitted). In a next step we calculated the mean value of each Fujisaki feature for each of the 68 files.

6.4.2. *Model*

We conducted a stepwise multiple linear regression analysis. The auditory MOS ($\mathbf{y}$) was used as response variable while the 47 Fujisaki features were used as predictors ($\mathbf{X}$).

In Table 2 we list the selected features for the optimized model, its beta values (B), their standard error (SE B), and their standardized values (β). The two selected features denote the mean amplitude of PCs (*mean pc amp*) and the mean distance between starting points and ending points of the ACs in the signal (*mean dist ac sp ep*). To test for over-fitting effects, a leave-one-out CV was conducted. The R^2 and RMSE value could be confirmed. Thus, the models can be accounted to be stable.

Table 2: Results of the stepwise multiple linear regression analysis for the audiobook database. $R^2 = .31$.

FEATURE	B	SE B	β
constant	2.996	0.679	
mean pc amp	4.175	0.999	.437 ***
mean dist ac sp ep	-5.843	2.216	-.276 *

$^{*}p < .05$. $^{***}p < .001$.
Note: see text for explanation of the features.

6.5. Results

We used both predictors to compute the predicted MOS ($\hat{y}$) for the 68 files from the audiobook database. As a measure of accuracy we report on the Pearson correlation R and the normalized error per file and per synthesizer, see Table 3. The GFP achieves a medium correlation of .42 per file and a strong correlation of .61 per system. Even though the results show that this approach does not cover all relevant aspects of the overall quality, it is still encouraging to see that the GFP performs well on a completely independent database (i.e., different TTS systems, different language, different use case).

Table 3: Pearson Correlation and normalized error between predicted MOS and auditory MOS.

	GFP		AFP	
	R	ϵ	R	ϵ
per stimulus	$.42^{**}$	0.38	$.56^{**}$	0.30
per system	$.61^{*}$	0.35	$.71^{**}$	0.25

$^{*}p < .05$. $^{**}p < .01$.

The AFP trained on the audiobook database achieves strong correlations per file and per synthesizer. We do expect to further improve this model as soon as there is more training data available. This would lead to a model with more than two features which presumably improves the predictions. Figure 7 shows the scatter plot of subjective vs. predicted MOS using the AFP model.

7. Conclusion

In light of the results reported in this paper, the feasibility of instrumentally predicting perceptual quality of synthesized audiobooks has been demonstrated. We conclude that joint research on the feature and the model level is necessary for predictions that are robust but also interpretable. In the future we will work on the optimum fusion of all introduced features and regression methods for modeling the most relevant aspects of TTS quality, aiming for a multidimensional diagnostic prediction model that allows for an unbiased indication and understanding of perceptual quality differences in TTS.

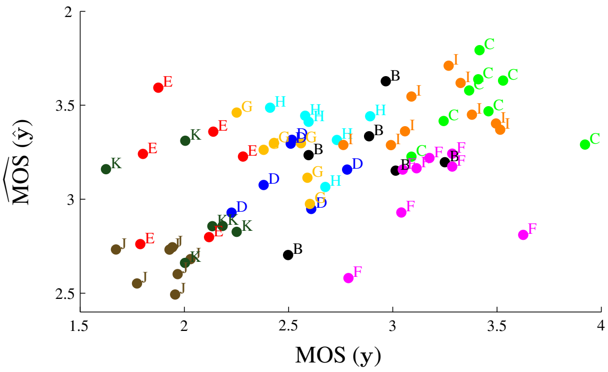

Figure 7: Subjective MOS vs. estimated MOS for the AFP.

8. Acknowledgements

The authors would like to thank the Blizzard Challenge organizers for providing the audiobook database. This work was supported by the Deutsche Forschungsgemeinschaft (DFG) under grants HE 4465/4-1 and MO 1038/11-1.

9. References

[1] P. Taylor, *Text-to-Speech Synthesis*. Cambridge University Press, 2009.

[2] F. Hinterleitner, G. Neitzel, S. Möller, and C. Norrenbrock, "An evaluation protocol for the subjective assessment of text-to-speech in audiobook reading tasks," *Proc. Blizzard Challenge Workshop. International Speech Communication Association (ISCA)*, 2011.

[3] C. R. Norrenbrock, F. Hinterleitner, U. Heute, and S. Möller, "Instrumental assessment of prosodic quality for text-to-speech signals," *IEEE Sig. Proc. Letters*, vol. 19, no. 5, pp. 255–258, 2012.

[4] C. R. Norrenbrock, U. Heute, F. Hinterleitner, and S. Möller, "Aperiodicity analysis for quality estimation of text-to-speech signals," *Proc. Interspeech,* Florence, Italy, pp. 2194–2197, 2011.

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

[6] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1, pp. 273–324, 1997.

[7] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett, "New support vector algorithms," *Neural Computation*, pp. 1207–1245, 2000.

[8] 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.

[9] F. Hinterleitner, S. Möller, T. H. Falk, and T. Polzehl, "Comparison of approaches for instrumentally predicting the quality of text-to-speech systems: Data from blizzard challenges 2008 and 2009," *Proc. Blizzard Challenge Workshop. International Speech Communication Association (ISCA)*, 2010.

[10] F. Hinterleitner, C. Norrenbrock, and S. Möller, "On the use of fujisaki parameters for the quality prediction of synthetic speech," *Proc. ESSV, Cottbus, Germany*, 2012 (to appear).

[11] H. Fujisaki, "Dynamic characteristics of voice fundamental frequency in speech and singing. acoustical analysis and physiological interpretations." *STL-QPSR*, vol. 22, no. 1, pp. 1–20, 1981.

[12] H. Mixdorff, "Manual for the fujiparaeditor - an interactive tool for extracting fujisaki model parameters," 2010. [Online]. Available: http://public.beuth-hochschule.de/~mixdorff/thesis/fujisaki.html