4th International Workshop on Perceptual Quality of Systems (PQS 2013)
2-4 September 2013, Vienna, Austria

Quality Estimation Based on Regular Perception

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

[1]Digital Signal Processing and System Theory, Christian-Albrechts-Universität zu Kiel, Germany
[2]Quality and Usability Lab, TU Berlin, Germany

cno@tf.uni-kiel.de, florian.hinterleitner@tu-berlin.de
uh@tf.uni-kiel.de, sebastian.moeller@telekom.de

Abstract

We present a novel approach for speech-quality prediction based on the observation that perception-relevant properties may cover a wide range of different values without significantly changing perceptual quality. To account for this nonlinear phenomenon, a semi-supervised discretization concept is proposed which is applied to temporal as well as aggregated properties as a preprocessing step prior to conventional modelling. The idea of *perceptual regularization* allows for integrating an assumed perceptual reference into the modelling process without neglecting the empirical cognition effects coded in the subjective test data. For the example of synthetic-speech quality we will demonstrate how the two main goals of quality estimation, namely interpretability and robustness, are addressed through the presented approach.

Index Terms: Speech-quality prediction, perceptual quality modelling, text-to-speech (TTS), synthetic-speech quality, perceptual regularization

1. Introduction

Quality is the result of a complex perceptual process which is difficult to investigate. In particular, when no explicit reference can be named or identified to which test subjects make their ratings, it is challenging to find a unifying model which (i) accurately approximates the outcome of the perception and assessment process and (ii) mimics the essential perceptual mechanisms in a meaningful way. It is, however, by no means useless to try to model a process which is apriori not understood in depth. In fact, the key for understanding quality lies in an insistent investigation of the relationship between the average quality impression of an object (or object class) and its underlying physical properties ("features"). At the same time, structuring the average quality impression in a hierarchical or multidimensional sense can help to disentangle the "perceptual mixing" and weighting.

In the area of *speech-quality estimation*, these principles have found wide acceptance in recent years, see, e.g. [1] for an overview. From the range of different model types, full-reference (intrusive) instrumental models have been addressed most often. These models are based on a signal comparison, where one of the comparing signals serves as an "explicit" quality reference. For this model type, high levels of performance are reported in the literature for many different algorithms and speech signal classes, see, e.g., [1, 2], so a sound degree of maturity can be assumed with regard to the underlying concepts and approaches. In contrast, so-called reference-free, or non-intrusive models, do not yet appear to be as elaborate. So far, the most common approach is to identify relevant measurands ("features") and to map them onto a subjective rating scale according to a trained model. This approach has gained significant attention in recent years, e.g., as a valid tool to address all sorts of speech processing topics, e.g., speech-intelligibility estimation for healthy and disordered voices, voice-personality and emotion classification, and speech-quality prediction. As a common tendency, it is interesting to observe that the larger the amount of considered measurands grows, e.g., > 500, the less seems to be known about the perceptual mechanisms. Indeed, complex machine-learning tools, such as support vector machines, can compensate lacking detail knowledge when large feature pools are used in which the major information about the subjective perception process is coded. However, it appears to be unclear to what extent this approach is suited to derive *robust* models for comparably complex signal classes, such as speech signals. The prominent risk of overfitting becomes evident as feature choices and combinations are often only comprehensible from a formal statistical viewpoint. Of course, increasing the sample size of the subjective tests remains to be the gold standard to alleviate overfitting, however, this is not a feasible option in general research projects.

Still, the (numerous) measurands necessarily form the foundation for quality prediction, and the question what should be measured in what way can practically only be answered through subjective test data which exhibits sample sizes well below the number of potential predictors. In this context, the aim of this paper is to foster thinking about how to model perceptual mechanisms in a more direct manner. We introduce the theory of *regular perception* as a basic principle within the process of subjective quality perception. It is applied to estimate synthetic-speech quality in order to show the benefits and drawbacks in a practical use-case.

2. Non-Intrusive Quality Assessment

2.1. Concepts

An instrumental non-intrusive assessment (NiQA) model, e.g., for speech signals, is characterized by a number of measurement parameters (measurands) which are used to estimate quality. These parameters are based on a range of *physical properties*, e.g., spectral coefficients, or fundamental frequency, which are measured from the signal under test. This is illustrated in Fig. 1. As a common statistical, i.e., data-driven approach, the measurands of the model g_0 represent the properties via simple functionals, e.g., by taking mean values or standard deviations. Alternatively, elaborate measurands can be used as they have been proposed for special NiQA scenarios, e.g., the articulation-to-nonarticulation ratio in ANIQUE [3], the speech-to-reverberation modulation energy measure [4], or the

10.21437/PQS.2013-7

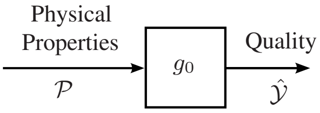

Figure 1: Overall NiQA model.

vocal-tract parameters in the P.563 model [5]. Subsequently, the measurands are mapped according to a trained estimation model, for which different modelling paradigms can be identified:

- *Plain regression approach*: Maximizing the likelihood of the data by using plain "regression" models, or more complex machine-learning methods [6]. Numerous such methods have been used for NiQA of speech signals, e.g., Bayesian methods [7], support vector regression (SVR) [8], or regression trees [9].
- *Feature-likelihood approach*: The quality estimation is based on the statistical likelihood of the measurands w.r.t. a statistically modelled reference signal class, e.g., via GMMs and HMMs. This has been investigated, e.g., for the estimation of synthetic- speech quality in [10].
- *Hierarchical approach*: An analysis of the data heterogeneity leads to subclasses of signals which can be associated with a dominant perceptual characteristic. Following a divide-and-conquer strategy, complex mappings are set up using different submodels which are then combined. Examples for this approach are P.563 [5], ANIQUE [3], and [11].
- *Multidimensional (factorial) approach*: A dimensional (factor) analysis allows to describe integral quality by perceptual (quality) dimensions which represent "holistic" quality features. For each dimension a separate NiQA model is trained. The output of the dimension estimators are used to (i) deliver diagnostic information about the quality and/or to (ii) estimate overall quality, e.g., [2], [12]. In contrast to the hierarchical approach, which is motivated by specific perceptual hierarchies in the "psycho-physical" domain, the idea of the multidimensional approach is to structure the perceptual domain in an holistic manner.

As a side remark, it should be noted that none of the mentioned approaches is per-se stronger or weaker with regard to numerical and methodical validity. However, there are clear differences w.r.t. the development effort.

2.2. Perceptual Reference

As there is no such thing as absolute or reference-free quality, one might ask how to model the perceptual reference adequately. This is a non-trivial problem since the perceptual (or internal) reference to which the listeners give their ratings is generally unknown. It is reasonable to assume that the perceptual reference should not be associated with an "optimum" object configuration, but with a broad range of equally adequate realizations, e.g., of the same speech signal. In all of the concepts mentioned in Section 2.1, the internal reference is, inevitably, implicitly modelled, either by the estimation model itself, or by using an assumed perceptual reference signal class (feature-likelihood approach). For both cases, however, a more explicit approximation of the perceptual reference from the data would be desirable in order to strive for a more comprehensible quality prediction in terms of physical properties. In the following, a novel approach is introduced which aims at estimating the perceptual reference by regular perception ranges.

3. Regular Perception Approach

3.1. Theory of Regular Perception

Two perceptual ranges constrain and define our ability to perceive physical properties of signals, e.g., acoustical signals. At the first level, bio-chemical mechanisms, e.g., in the auditory system, define a *physiological perception range* (PPR) within which the property is generally perceivable. For example, the bandwidth of the human auditory system defines the PPR of the property *frequency* to ∼20-20.000 Hz. At the psycho-physical level, the *regular perception range* (RPR) defines the range within which the property confines to some internal perceptual reference. Clearly, the RPR lies always within the PPR, see Fig. 2.

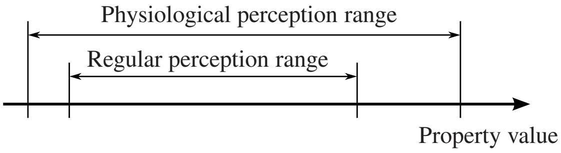

Figure 2: Illustration of perceptual ranges.

So, where the PPR describes *perceptibility*, the RPR defines perceptual *regularity* w.r.t. a subjective sensation attribute. The RPR introduces a subjective perception layer beyond classical psycho-acoustic assessment scenarios, such as just-noticeable differences (JND). Thus, a specific subjective weighting is taken into account, i.e., the question *whether* sth. is perceived is constricted towards *how* sth. is perceived. Yet, a binary, i.e., simplified perceptual nature is still enforced through the discretization of an attribute sensation which is assumed to manifest on a continuous scale, as illustrated in Fig. 3. The shape of this property-to-attribute relation is Gaussian here for illustrative purposes. As an example from psychoacoustics, consider the sensation of roughness, produced by amplitude modulation of a 1-kHz tone [13]. By varying the modulation frequency, i.e., the property of the attribute, the impression of roughness can be varied according to a Gaussian-shaped curve. Following the idea of regular perception, the RPR would be given by the range within which a subject considers the tone to be rough, i.e., the sensation is expressed as a categorical attribute, e.g., regular (1) or not (0).

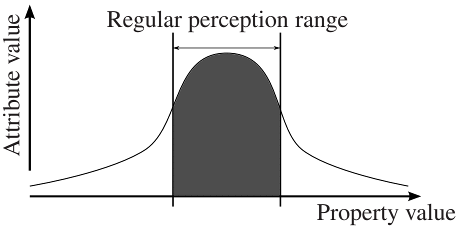

Figure 3: Attribute value as a function of the property value.

3.2. Perceptual Regularization for NiQA

The sketched theory of regular perception is used to evaluate a novel class of measurement parameters, denoted as *quality* *elements*. These parameters are derived from individual properties so as to yield a description of their *quality-indicative* effect. This will be referred to as *perceptual regularization*. The overall model becomes a two-stage model as shown in Fig. 4.

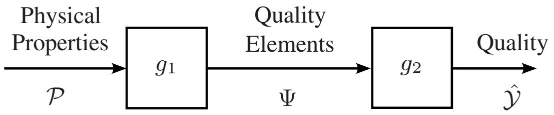

Figure 4: Two-stage model for quality estimation.

3.2.1. RPR-based Quality Elements

A quality element (Ψ) of a property ($\mathcal{P}$) is defined as a numerical description of the extent to which the property falls into its corresponding RPR according to the following classification:

$$\delta_{\boldsymbol{\xi}}(\mathcal{P}) = \begin{cases} 1, & \text{for } \xi_1 < \mathcal{P} < \xi_2 \\ 0, & \text{else.} \end{cases} \qquad (1)$$

The values of ξ_1 and ξ_2 ($\in \mathbb{R}$) are the lower and upper RP thresholds, respectively, and $\xi_1 < \xi_2$. For convenience, the RPR will be written in vector notation, $\boldsymbol{\xi} = [\xi_1, \xi_2]^{\mathrm{T}}$. For each property, the RPR defines the perceptual "non-excess" region. For time-variant properties the classifier (1) is applied to suitable analysis entities of the time-variant signal, and the classification result is averaged, yielding quality elements $\Psi \in [0, 1]$. Time-invariant, e.g., aggregated properties are represented by their plain classification result, thus $\Psi \in \{0, 1\}$.

3.2.2. Evaluation of the RPR

Given N training signals, let the vector of subjective ratings be $\mathbf{y} = [y_1, y_2, ..., y_N]^T$ and the corresponding vector of quality elements be $\boldsymbol{\psi} = [\psi_1, \psi_2, ..., \psi_N]^T$. The RPR is estimated as

$$\hat{\boldsymbol{\xi}}(\mathcal{P}) = \underset{\boldsymbol{\xi} \in \Delta\boldsymbol{\xi}_{\text{ref}}}{\arg\max} \left\{ \text{Corr}\left(\boldsymbol{\psi}(\mathcal{P}, \boldsymbol{\xi}), \mathbf{y}\right) \right\}, \qquad (2)$$

where 'Corr' denotes the correlation coefficient after Pearson. The search range for $\boldsymbol{\xi}$ is $\Delta\boldsymbol{\xi}_{\text{ref}}$. It serves as a plausibility constraint in such a way that unreasonable settings w.r.t. a reference signal class are avoided. As properties (measured equivalently for test and reference signals) are assumed to be neutral description elements without a specific perceptual valence, it is the positive correlation in (2) which is maximized. In general, the reference search range $\Delta\boldsymbol{\xi}_{\text{ref}}$ consists of two non-overlapping search ranges for the lower and upper RPR bounds, i.e., $\xi_1 \in \Delta\xi_{1,\text{ref}}$ and $\xi_2 \in \Delta\xi_{2,\text{ref}}$, which are sampled in a grid-search manner.

3.3. Motivation

Before proceeding to the practical example of the proposed approach, let us review the underlying motivation more closely.

3.3.1. Internal Reference

One main motivation is to identify a perceptual reference on a physical level. A property which is subject to varying (individual) perception, is coded as a measurand with a defined perceptual valence through a form of "majority vote". Clearly, this can only be an approximation since the perceptual relations and interactions between the properties are modelled in a separate step (see Fig. 4). Thus, quality elements are to be seen as basic quality indicators rather than fully estimated quality features.

3.3.2. Perceptual nonlinearities

It can often be observed that a given property can cover a wide range of values which appear to be equally *adequate* for a certain quality impression. This can be, e.g., due to dominant perceptual characteristics which can mask other relevant properties. In fact, these nonlinearities represent a form of *perceptual invariance* which can, e.g., be tackled by the hierarchical approach discussed in Section 2.1. The idea of RPR-based quality elements is similar, as "perceptual outliers" which often impose numerical problems during the modelling process are handled in a meaningful way.

3.3.3. Interpretability

As mentioned in the introduction, using many measurands often leads to complex models with limited diagnostic insight. To enhance interpretability, a property which is coded as a RPR-based quality element can be advantageous because, practically, one is mainly interested whether a specific measurand has a negative or positive impact on the quality. This appears to be a practical step towards property-based diagnosis for NiQA: localizing the problem first, then looking at the actual property characteristics. The aim is to further improve the link between the physical and the perceptual world beyond the dimension-based diagnostic.

4. Example: Quality Estimation for Synthetic Speech

4.1. Perception of Synthetic Speech

As pointed out in several studies, e.g., in [14, 15], quality assessment of synthetic speech is a comparably complex task. This is because synthetic speech is an extensively designed compound which calls for an holistic assessment scheme. Following the multidimensional approach, the quality space of TTS can be described by 5 quality dimensions [16] which are (1) naturalness of voice, (2) prosodic quality, (3) fluency and intelligibility, (4) absence of disturbances, and (5) calmness.
In order to demonstrate the principal modelling aspects, it is shown in the following how to build an instrumental model for the dimension 'naturalness of voice'. It has been shown in previous work that this dimension is closely related to prosodic features, such as pitch perturbation features, e.g., smoothed cepstral-peak prominence (CPPS) [17]. Beyond fundamental frequency, MFCCs have been identified as useful for estimating this dimension, however, to a reduced degree. In total, 149 speech-signal properties are considered which are described in greater detail in [17, 18] and the references therein.

4.2. Perceptual Regularization for Synthetic Speech

The perceptual reference for synthetic speech can be assumed to be related to natural speech, however, previous work has shown that the physical feature likelihood of natural speech is apparently not the only factor which is taken into account when rating synthetic speech signals. It seems that also complex redundancy issues, e.g., the extent to which the feature characteristics generally fit together, play a role, which suggests a more abstract shape of the perceptual reference. According to the regular perception approach, the RPRs for the mentioned properties are derived from the observed property ranges of the synthetic speech signals (used in the subjective tests) and the natural speech signals which have been recorded for reference purposes [17].

4.3. Test Databases

4.3.1. Test I

Test I has been carried out within the scope of a Semantic-Differential scaling task [19]. Ten short paragraphs were compiled, each consisting of 2-3 simple sentences. The average number of words per paragraph is $\sim$27 and the average spoken length amounts to $\sim$10 s. For synthesis, 15 TTS systems per speaker gender (male and female voices) were used. Each system configuration is represented by two stimuli containing different paragraphs which were selected at random from the available ten. In total, 60 stimuli were evaluated on 16 different attribute scales which were derived from 2 separate pretests, see [19]. All speech files were processed towards a common sampling frequency of 16 kHz and level-normalized to -26 dBov. Thirty naïve listeners (native German speakers) took part in the test which was carried out at the Quality and Usability Lab and the Telekom Innovation Laboratories, Berlin, Germany.

4.3.2. Test II

In Test II, the same 16 attribute scales as in Test I have been used. 20 TTS systems, largely the same as in Test I but partly with different voice-system combinations, have been employed for synthesis of one fixed German sentence (14 words, $\sim$5 s spoken duration). The listening test conditions were essentially the same as in Test I. The test was conducted in two sections (one with male and one with female stimuli), where the stimuli (16 kHz sampling rate, -26 dBov normalization) were presented in randomized order. 12 native German speakers participated in the test (5 expert listeners, 7 naïve listeners, 7 male and 5 female, mean age was 27 years). The test took place at the Quality and Usability Lab and the Telekom Innovation Laboratories, Berlin, Germany.

4.4. Model Types

In order to analyze the regular-perception approach, two different model types are used in two regularization modes. The regularization modes are (i) RP0: In this case, time-variant properties are represented by their mean-values, and time-invariant properties are taken without change, and (ii) RP1: all properties are represented by quality elements. The model types which are used are (i) ν-Support Vector Regression (SVR) with radial basis function (RBF) kernel, as described in [20] and implemented in [21], and (ii) the Regular Perception Model (RPM). The RPM is an experimental model which is considerably simpler than the SVR model. The model estimate of the RPM is defined as the adjusted average of positively correlating quality elements. The result is mapped according to a logistic mapping:

$$\varphi(\tilde{y}) = \frac{1}{1+\exp(a\tilde{y}+b)}, \tag{3}$$

where $\tilde{y}$ denotes the unmapped average. The free parameters a and b are optimized using a nonlinear least-squares method. Measurand scaling ("feature normalization") is applied so that all variables are within the interval $[0, 1]$. In all cases, a correlation-based selection of measurands ("feature selection") is carried out prior to model training in order to identify their quality-relevance w.r.t. a minimum correlation $R_{\min} \in (0, 1)$. In the RP0 case, the magnitude correlation is chosen as the criterion; in the RP1 case, only positively correlating measurands are selected.

4.5. Model Assessment

The model assessment is carried out within the scope of a *leave-one-test-out* cross-validation (CV). The setup is shown in Fig. 5. The variable $k = \{1, 2\}$ indicates the k-th CV partitioning of the available data into training and test set, applicable for the quality ratings $\mathbf{y}$ and the measurand matrix $\mathbf{X}$. The feature-normalization information $\boldsymbol{\eta}^{(k)}$, the indices of the selected features $\boldsymbol{\iota}^{(k)}$, and the *model parameter* vector $\hat{\boldsymbol{\beta}}^{(k)}$ are evaluated on the training set, and then used for predicting the quality of the test-set signals. The parameters in (3) are assigned to $\hat{\boldsymbol{\beta}}^{(k)}$. All hyperparameters and the selection threshold are kept fixed throughout this study. The figures of merit are the averaged Pearson correlation $\overline{R}_{\text{CV}}$ and the averaged root-mean-square error $\overline{\epsilon}_{\text{CV}}$. For the evaluation, the quality ratings are re-scaled to the interval $[1, 5]$.

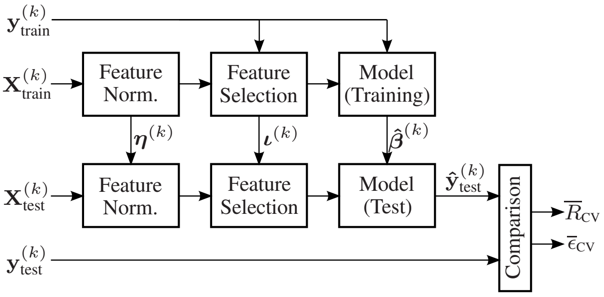
$\mathbf{y}^{(k)}_{\text{train}}$
$\mathbf{X}^{(k)}_{\text{train}}$
Feature Norm.
Feature Selection
Model (Training)
$\boldsymbol{\eta}^{(k)}$
$\boldsymbol{\iota}^{(k)}$
$\hat{\boldsymbol{\beta}}^{(k)}$
$\mathbf{X}^{(k)}_{\text{test}}$
Feature Norm.
Feature Selection
Model (Test)
$\hat{\mathbf{y}}^{(k)}_{\text{test}}$
Comparison
$\overline{R}_{\text{CV}}$
$\overline{\epsilon}_{\text{CV}}$
$\mathbf{y}^{(k)}_{\text{test}}$

Figure 5: Cross-validation set up.

4.6. Results

The results are summarized in Tab. 1. In general, better results are obtained for male voices which can be explained with the comparably less precise measurement and/or more complex perception characteristics of higher-frequency voices. A considerable improvement of correlation and error is achieved in the RP1 case, demonstrating the usefulness of the quality elements here. The correlation difference between SVR and RPM is insignificant. However, this result shows that the potential interactions between quality elements, as they can be modelled through the RBF kernel in the SVR case, do not seem to lead to a better performance here.

Table 1: Figures of merit (FoM) of the leave-one-test-out CV.

FoM	Model	Male		Female	
		RP0	RP1	RP0	RP1
$\overline{R}_{\text{CV}}$	SVR	.78	.93	.66	.88
	RPM	-	.94	-	.89
$\overline{\epsilon}_{\text{CV}}$	SVR	.56	.39	.69	.47
	RPM	-	.37	-	.45

In Fig. 6, the cases $\{\text{RP0}, \text{SVR}\}$ and $\{\text{RP1}, \text{RPM}\}$ (male voices) are illustrated by scatterplots. These plots show the relation between the auditory ratings of Dimension 1 and their model-based CV estimates, where the predicted observations of Tests I and II are indicated by crosses and circles, respectively.

To further analyze the nature of perceptual regularization in terms of model structure, three time-variant properties are selected here for illustrative purposes. These are (i) the smoothed segmental cepstral-peak prominence (CPPS$_{\text{seg}}$), (ii) the regressive F_0 slope of voiced segments with declining F_0, and (iii)

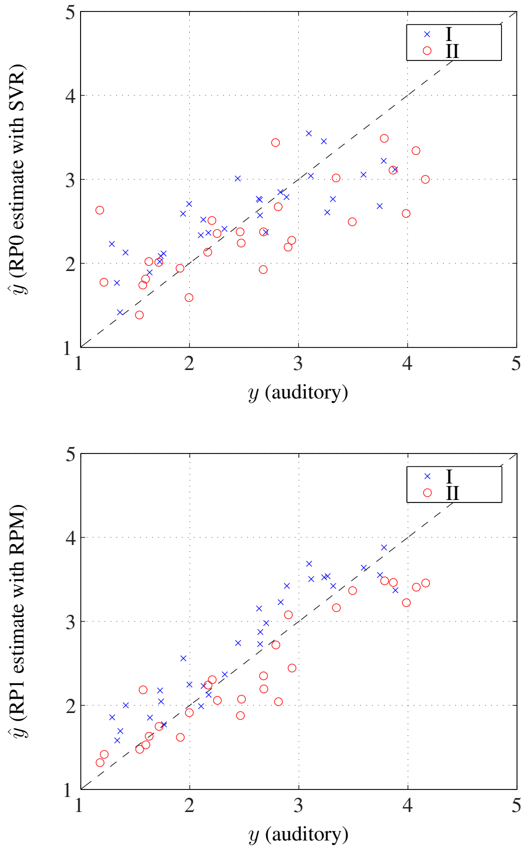

Figure 6: Scatter plots of leave-one-test out cross-validation (male voice): The upper plot shows the results for the case {RP0,SVR}. The lower plot shows the results for {RP1,RPM}. The abscissas and ordinates correspond to the auditory and estimated dimension ratings, respectively.

the fundamental frequency range ΔF_0, where the latter two properties are evaluated per voiced segment and transformed to the semitone scale [17]. The segments for CPPS$_{\text{seg}}$ are pitch-synchronous with a gender-dependent number of pitch-periods per segment. The correlation of these properties, represented as mean values $\mu(\mathcal{P})$ and quality elements $\Psi(\mathcal{P})$, is given in Tab. 2. As can be seen, the magnitude correlation improves in all cases except one. This shows that, in terms of correlation, the RPR-based representation of properties is not necessarily better for each property, yet for most properties we do see an improvement. Furthermore, it can be seen how a negatively correlating property is rendered a positively correlating quality element (for CPPS here). More insight is gained through the scatterplots in Fig. 7. Comparing the first figure row ($\mu(\mathcal{P})$) with the second ($\Psi(\mathcal{P})$) it becomes evident that especially those properties benefit from the proposed approach which already show, up to considerable outliers, a basic linear behaviour.

Table 2: Individual correlations of important prosodic measurands which are derived from time-variant properties $\mathcal{P}$. Measurands are the property means $\mu(\mathcal{P})$ in the RP0 case, and the quality elements $\Psi(\mathcal{P})$ in the RP1 case.

$\mathcal{P}$	MALE		FEMALE	
	$\mu(\mathcal{P})$	$\Psi(\mathcal{P})$	$\mu(\mathcal{P})$	$\Psi(\mathcal{P})$
CPPS$_{\text{seg}}$	-.73	.84	-.65	.60
F_0 slope drop	.38	.75	.49	.63
ΔF_0	.67	.71	.32	.53

5. Discussion and Conclusion

In this paper, prediction modelling for NiQA has been addressed. We have introduced the approach of perceptual regularization and demonstrated its use for the example of quality estimation for synthetic speech signals. A novel class of measurement parameters, denoted as quality elements, have been motivated by categorical quality perception on the physical level as a means of representing physical properties in a linearized quality-indicative sense. In order to avoid unreasonable quality indications (i.e., RPR thresholds), a reference signal class (here: natural speech) is considered during the evaluation of the RPRs. The essential characteristics of the proposed approach are:

- Quality perception is modelled as a compound of regularized quality-indicative properties (quality elements).
- Quality elements describe "non-critical" ranges of individual properties through a form of majority vote. This is interpreted as a primitive mechanism within the process of quality perception which accounts for (i) the limited explanatory power of single properties, (ii) the perceptual limitations, and (iii) perceptual invariance.
- Efficient development of a NiQA model.

For future work the following aspects can be envisaged:

- Investigation of the perceptual attributes of individual properties. So far, the RPRs have been derived via the subjective quality ratings. A more detailed knowledge about how the variation of individual properties manifests perceptually, while keeping all other properties fixed, would allow for an improved validity of the RPRs. Furthermore, a shift towards specific property compounds would be necessary in order to better account for property interactions.
- Investigation of categorical quality perception. The fact that subjects often cannot even notice (subtle) quality differences (despite significant changes on the physical level), e.g., in the area of image processing, casts doubt on the adequacy of a continuous quality variable being estimated by means of 'continuous' properties or formal statistics. Modelling of the "nonlinear bendings", which arise from pure physiological or higher cognitive perception layers, appears to be necessary, and the question on which level these bendings should be addressed is overlooked too often. In any case, the art will remain to decide how much generality and diagnostic insight can be extracted from empirical evidence beyond data fitting.

6. Acknowledgement

This work was funded by the German Research Foundation (Deutsche Forschungsgemeinschaft) under grants HE 4465/4-1, HE 4465/4-2, MO 1038/11-1, and MO 1038/11-2.

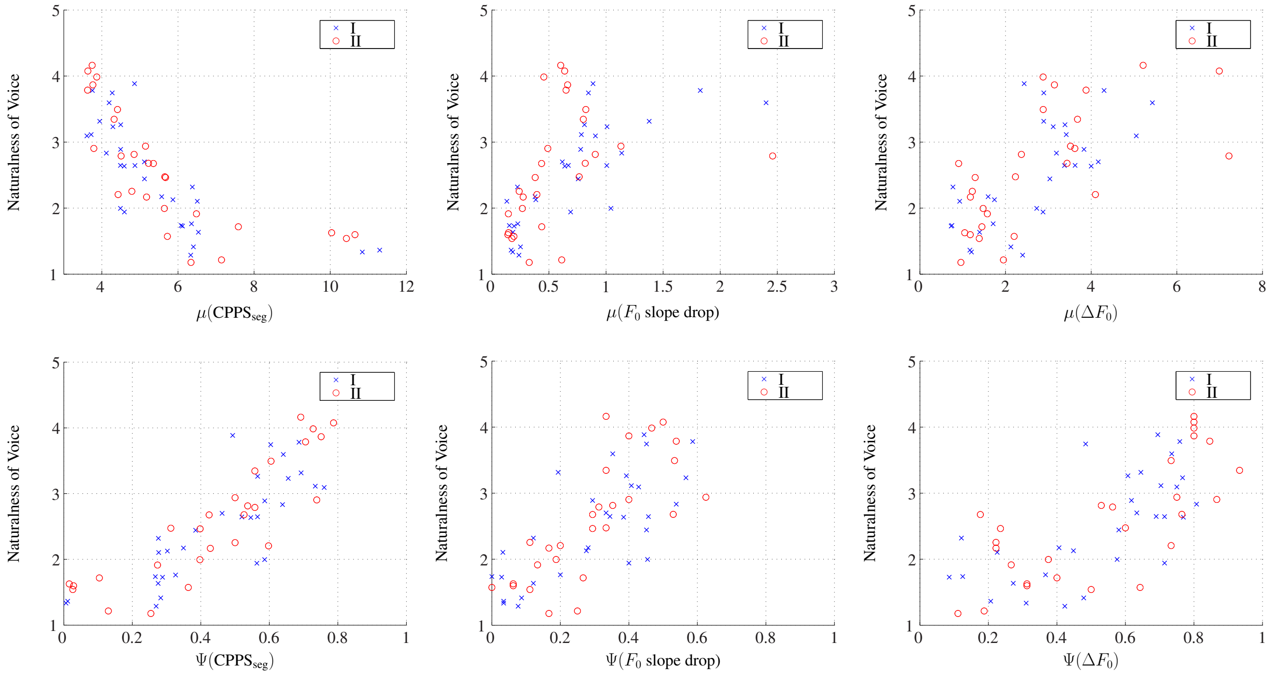

Figure 7: Scatter plots of selected property means (first row) and corresponding quality elements (second row). The abscissas and ordinates correspond to measurand values and subjective ratings, respectively.

7. References

[1] S. Möller, W.-Y. Chan, N. Côté, T. Falk, A. Raake, and M. Wältermann, “Speech quality estimation: Models and trends,” *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 18–28, 2011.

[2] K. Scholz, *Instrumentelle Qualitätsbeurteilung von Telefonbandsprache beruhend auf Qualitätsattributen*, ser. Arbeiten über Digitale Signalverarbeitung, Christian-Albrechts-Universität zu Kiel, Dissertation , U. Heute, Ed. Shaker Verlag, 2008.

[3] D.-S. Kim, “ANIQUE: An Auditory Model for Single-Ended Speech Quality Estimation,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 821–831, 2005.

[4] T. Falk and W.-Y. Chan, “A non-intrusive quality measure of dereverberated speech,” in *Proc. IWAENC*, Seattle, WA, USA, 2008.

[5] ITU-T Recommendation P.563, *Single-ended Method for Objective Speech Quality Assessment in Narrow-band Telephony Applications*, International Telecommunication Union, Geneva, Switzerland, 2004.

[6] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New York, USA: Springer, 2009.

[7] P. N. Petkov, I. S. Mossavat, and W. B. Kleijn, “A bayesian approach to non-intrusive quality assessment of speech,” in *Proc. Interspeech*, 2009, pp. 2875–2878.

[8] M. Narwaria, W. Lin, I. McLoughlin, S. Emmanuel, and L.-T. Chia, “Nonintrusive quality assessment of noise suppressed speech with mel-filtered energies and support vector regression,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 4, pp. 1217–1232, 2012.

[9] T. Falk, S. Möller, V. Karaiskos, and S. King, “Improving Instrumental Quality Prediction Performance for the Blizzard Challenge,” in *Proc. Blizzard Challenge Workshop*, Brisbane, Australia, 2008.

[10] T. 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] I. S. Mossavat, P. N. Petkov, W. B. Kleijn, and O. Amft, “A hierarchical bayesian approach to modeling heterogeneity in speech quality assessment,” *IEEE Transactions on Audio, Speech & Language Processing*, pp. 136–146, 2012.

[12] N. Côté, *Integral and Diagnostic Intrusive Prediction of Speech Quality*, ser. T-Labs Series in Telecommunication Services. Springer, 2011.

[13] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*, 2nd ed. Springer, Berlin/Heidelberg, 1999.

[14] V. J. van Heuven and R. van Bezooijen, *Quality Evaluation of Synthesized Speech*, ser. In: Kleijn, W. B., Paliwal K. K. (Eds.), Speech Coding and Synthesis. Elsevier, Amsterdam, New York, 1995, pp. 707–738.

[15] U. Jekosch, *Voice and Speech Quality Perception: Assessment and Evaluation*. Springer, Berlin/Heidelberg, 2005.

[16] F. Hinterleitner, C. Norrenbrock, and S. Möller, “Is Intelligibility Still the Main Problem? A Review of Perceptual Quality Dimensions of Synthetic Speech,” *Proc. Speech Synthesis Workshop,* Barcelona, Spain, 2013.

[17] C. Norrenbrock, F. Hinterleitner, U. Heute, and S. Möller, “Towards Perceptual Quality Modeling of Synthesized Audiobooks-Blizzard Challenge 2012,” in *Proc. Blizzard Challenge Workshop*, Portland, OR, USA, 2012.

[18] ——, “Quality Analysis of Macroprosodic F0 Dynamics in Text-to-Speech Signals,” in *Proc. Interspeech*, Portland, Oregon, USA, 2012.

[19] F. Hinterleitner, S. Möller, C. Norrenbrock, and U. Heute, “Perceptual quality dimensions of text-to-speech systems,” *Proc. Interspeech,* Florence, Italy, pp. 2177–2180, 2011.

[20] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett, “New support vector algorithms,” *Neural Computation*, vol. 12, pp. 1207–1245, 2000.

[21] 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.