



Double-ended Prediction of the Naturalness Ratings of the Blizzard Challenge 2008-2013

Lukas Latacz¹, Werner Verhelst^{1,2}

¹Vrije Universiteit Brussel, Dep. of Electronics and Informatics, Pleinlaan 2, 1050 Brussel, Belgium

²iMinds, Multimedia Technologies Dep., Gaston Crommenlaan 8 bus 102, 9050 Gent, Belgium

{llatacz, wverhels}@etro.vub.ac.be

Abstract

In this paper we describe a double-ended (i.e. reference-based or intrusive) approach to objective quality estimation of synthetic speech that uses a linear regression model whose parameters can easily be interpreted. The model was trained and evaluated on English data from the 2008 to 2013 Blizzard Challenges (BC) [1], which is the largest publically available resource of listener-evaluated synthetic speech. To our knowledge, this is the first attempt to train and evaluate a speech quality predictor on the whole data set. Predicting the naturalness of the different participating systems in the BC is not an easy task because some of the systems are quite close in quality. Our best results correspond to a Pearson correlation coefficient of 0.60 and 0.84 for sentences and systems, respectively, using a leave-one-system-out evaluation, which by far outperformed the ITU-T standard PESQ [2] for double-ended speech quality evaluation on this data.

Index Terms: speech-quality prediction, speech synthesis, intrusive quality assessment, Blizzard Challenge

1. Introduction

The quality of synthetic speech is usually evaluated in listening experiments [3] [4]. Objective measurements of speech quality are an attractive alternative to these experiments [5] because they are cost-efficient, repeatable and they can run without human interaction. The main idea in this paper is to use listener's responses to train models that mimic human perception to map a number of acoustic features of the input speech to an output rating.

Varying successes were reported for assessing the quality of synthetic speech with standardized algorithms optimized for degraded natural speech (e.g. [6] [7] [8] [9] [10]). Synthetic speech is not simply a case of degraded natural speech due to different types of distortions and unnatural prosody. Specialized approaches (e.g. [11] [12]) therefore seem more appropriate.

Inspired by [13], we propose a simple, easily interpretable, linear model to estimate the quality of synthetic speech. The model is based on the degradation of the demiphones of the synthetic speech when compared to the corresponding demiphones of the reference signal. The 309 parameters of our model are trained using data from the BC [14] [1], a yearly-held synthesis challenge that compares the quality of speech synthesizers on common speech data. The BC ratings are the largest publically available resource of listener-evaluated synthetic speech. To our knowledge, only 3 studies have used subsets of this data ([7], [15] and [16]). Note that predicting the naturalness of the different participating systems in the BC is not an easy task because some of the systems are quite close in quality.

| Year | Speaker | # Samples | # Ratings | # Systems |
|------|------------------------|-----------|-----------|-----------|
| 2008 | Roger ^{UK, M} | 1524 | 16026 | 38 |
| 2009 | Roger ^{UK, M} | 676 | 5473 | 34 |
| 2010 | Roger ^{UK, M} | 578 | 5226 | 17 |
| 2010 | RJS ^{UK, M} | 560 | 5243 | 17 |
| 2011 | Nancy ^{US, F} | 454 | 8357 | 12 |
| 2012 | JG ^{US, M} | 210 | 5503 | 10 |
| 2013 | CB ^{US, F} | 628 | 7861 | 22 |

Table 1: Summary of BC data to train and evaluate the objective quality estimation algorithm. M: Male; F: Female.

2. Training data

We used only US and UK English synthesized speech that received a *naturalness* rating on a 5-point rating scale and for which a corresponding natural reference signal was available. All data from the sub-challenges (the "spoke"-challenges) was ignored, e.g. synthesized speech in noise (BC 2010) and synthesis that was processed by a simulated telephone line (BC 2009). Some of the remaining samples could not be used because the sample could not be automatically segmented or the required acoustic parameters could not be extracted. These problems were typically due to concatenation artifacts or weird noises in the synthesized speech. They affected only a very small percentage of all available samples and most samples could be processed automatically without any problem. A summary of the data used for training and evaluation can be found in Table 1. The data consist of 272 unique test *sentences*. We always ignored system A of each BC edition, because it represents the natural speech version that we have used as a reference in our experiments.

2.1. Ratings

The ratings were given on a 5-point scale between 1 (unnatural) and 5 (natural). All speech samples taken into account received multiple naturalness ratings. In almost all cases, at least one subject disagreed with the other subjects and often the variance amongst subjects was quite large. Our approach requires a single rating for training and testing. We assume that individual factors that influence the rating, such as motivation, expectations, etc. are leveled out through aggregation of the ratings of multiple subjects. In our experiments we have used both the mean and the median as the representative rating of a synthetic speech sample. The median is able to cope well with outliers and results in a discrete aggregated value. The mean on the other hand results in a continuous value.

2.2. Segmentation

All speech samples were segmented and labeled into phonemes. We used SPRAAK [17] in forced alignment mode to generate this segmentation. No external speech data or acoustic models were used. The training of the acoustic models was bootstrapped using flat-start labeling (i.e. the initial labels were evenly distributed over the whole utterance). The target text that was used as input to generate the test sentences was supplied by the BC organizers. This text was converted into a phonemic transcription using the RP and General American variants of the Unisyn lexicon [18]. Missing entries in the lexicon were manually added. Optional silences could be inserted between each word, and at the beginning and end of an utterance. Separate context-independent models were constructed, using all the samples that share a same speaker and BC edition. Enough data was available to obtain models with sufficiently good accuracy for forced alignment, except for the 2010 Roger samples. We therefore segmented these using the 2009 Roger acoustic models.

2.3. Acoustic features

The majority of the systems entered in the competitions submitted 16 kHz synthetic speech. Since 2011, speech recordings with a higher sampling frequency (48 kHz) were distributed to the participating teams and they were allowed to submit samples with these higher sample rates. 48 kHz samples were therefore downsampled to 16 kHz before extracting the acoustic parameters.

Acoustic features were extracted every 5ms. We combined the output of three pitch determination algorithms (PDAs) [19]: RAPT [20], MulticueF0 [21] and SWIPE [22]. The F0 value at a particular position was taken to be the median of the output of the three PDAs at that position. Default settings of each PDA were used, but the F0 range (minimum and maximum allowable F0) was adapted according to the speaker. A suitable F0 range was found automatically by the procedure described in [23]. STRAIGHT [24] (legacy Matlab version STRAIGHTV40pcode) was used in order to obtain a smooth, high-quality spectral representation of the speech signal. This spectrum was converted into 50 mel-generalized coefficients [25] with $\alpha = 0.42$ and $\gamma = 0$ using SPTK 3.7 [26]. These parameters were fixed for all speakers.

3. Linear regression model

3.1. Demiphone degradation

We use a linear regression model to estimate the local degradation $D_{local}(s^k, r^l)$ of the k^{th} demiphone s^k of a synthetic utterance and the corresponding l^{th} demiphone r^l of the natural reference. This degradation is estimated by looking at differences in spectrum, log F0 and duration and is calculated as:

$$D_{local}(s^k, r^l) = D_{dur}(s^k, r^l) + D_{LF0}(s^k, r^l) + D_{spec}(s^k, r^l) + W_0 \quad (1)$$

with the functions D_{dur} , D_{LF0} and D_{spec} measuring the local degradation in terms of duration, log F0 and spectrum, respectively. Each of these functions consists of a weighted sum, while the constant W_0 serves as the intercept. Notice that we deliberately avoided the term (sub)distance here. The functions that we describe here do not comply with the formal definition of a distance, such as being symmetrical. One reason why they are not symmetrical is that we can use different weights for

“positive” and “negative” degradation, with “positive” degradations occurring if a reference parameter value is larger than that of the synthesis and “negative” degradations occurring if a reference parameter value is smaller than that of the synthesis.

3.1.1. Duration

Let $dur(s^k)$ and $dur(r^l)$ be the duration of the demiphones s^k and r^l , respectively, measured in seconds. The duration degradation is calculated as:

$$D_{dur}(s^k, r^l) = w_{dur}^+ \cdot D_{dur}^{rel}(s^k, r^l) \cdot H(dur(r^l) - dur(s^k)) + w_{dur}^- \cdot D_{dur}^{rel}(s^k, r^l) \cdot H(dur(s^k) - dur(r^l)) \quad (2)$$

$$D_{dur}^{rel}(s^k, r^l) = \frac{|dur(s^k) - dur(r^l)|}{dur(r^l)} \quad (3)$$

with $H()$ the Heaviside function and w_{dur}^+ and w_{dur}^- trainable weights.

3.1.2. log F0

We look at static and dynamic (Δ and $\Delta\Delta$) features to calculate the log F0 degradation. Undefined values (e.g. unvoiced frames, frames occurring at voiced/unvoiced borders in case of dynamic features) are set to a constant value (-1.0E+10) and are ignored in the calculation of the log F0 degradation. The i^{th} Δ feature of a parameter trajectory $f_i, i = 1 \dots N$ with length N is defined as $f_i^{\Delta} = f_{i+1} - f_i$ while the i^{th} $\Delta\Delta$ feature is defined as $f_i^{\Delta\Delta} = f_{i+2} - 2 \cdot f_i + f_{i-2}$. Let f_{LF0} , f_{LF0}^{Δ} and $f_{LF0}^{\Delta\Delta}$ be the static and dynamic log F0 contours with length N_f of the demiphone s^k . Similarly, let g_{LF0} , g_{LF0}^{Δ} and $g_{LF0}^{\Delta\Delta}$ be the static and dynamic log F0 contours with length N_g of the reference demiphone r^l . Let \hat{f}_{LF0} be the warped¹ log F0 contour so that its duration matches the duration of the reference log F0 contour g_{LF0} . Similarly, warped log F0 contours are constructed for all reference and dynamic log F0 contours. The log F0 degradation is then calculated as

$$D_{LF0}(s^k, r^l) = w_{LF0}^+ \cdot D_{LF0}^+(f_{LF0}, g_{LF0}, \hat{f}_{LF0}, \hat{g}_{LF0}) + w_{LF0}^- \cdot D_{LF0}^-(f_{LF0}, g_{LF0}, \hat{f}_{LF0}, \hat{g}_{LF0}) + w_{LF0,\Delta}^+ \cdot D_{LF0,\Delta}^+(f_{LF0}^{\Delta}, g_{LF0}^{\Delta}, \hat{f}_{LF0}^{\Delta}, \hat{g}_{LF0}^{\Delta}) + w_{LF0,\Delta}^- \cdot D_{LF0,\Delta}^-(f_{LF0}^{\Delta}, g_{LF0}^{\Delta}, \hat{f}_{LF0}^{\Delta}, \hat{g}_{LF0}^{\Delta}) + w_{LF0,\Delta\Delta}^+ \cdot D_{LF0,\Delta\Delta}^+(f_{LF0}^{\Delta\Delta}, g_{LF0}^{\Delta\Delta}, \hat{f}_{LF0}^{\Delta\Delta}, \hat{g}_{LF0}^{\Delta\Delta}) + w_{LF0,\Delta\Delta}^- \cdot D_{LF0,\Delta\Delta}^-(f_{LF0}^{\Delta\Delta}, g_{LF0}^{\Delta\Delta}, \hat{f}_{LF0}^{\Delta\Delta}, \hat{g}_{LF0}^{\Delta\Delta}) \quad (4)$$

$$D_{LF0}^+(f, g, \hat{f}, \hat{g}) = \hat{D}_{LF0}^+(f, \hat{g}) + \hat{D}_{LF0}^-(g, \hat{f}) \quad (5)$$

$$D_{LF0}^-(f, g, \hat{f}, \hat{g}) = \hat{D}_{LF0}^-(f, \hat{g}) + \hat{D}_{LF0}^+(g, \hat{f}) \quad (6)$$

$$\hat{D}_{LF0}^+(f, \hat{g}) = \frac{1}{N_f} \sum_{i=1}^{N_f} \begin{cases} |f_i - \hat{g}_i| & f_i > 0 \ \& \ \hat{g}_i > 0 \ \& \ \hat{g}_i > f_i \\ 0 & otherwise \end{cases} \quad (7)$$

$$\hat{D}_{LF0}^-(f, \hat{g}) = \frac{1}{N_f} \sum_{i=1}^{N_f} \begin{cases} |f_i - \hat{g}_i| & f_i > 0 \ \& \ \hat{g}_i > 0 \ \& \ \hat{g}_i < f_i \\ 0 & otherwise \end{cases} \quad (8)$$

with w_{LF0}^+ , w_{LF0}^- , $w_{LF0,\Delta}^+$, $w_{LF0,\Delta}^-$, $w_{LF0,\Delta\Delta}^+$ and $w_{LF0,\Delta\Delta}^-$ trainable weights.

¹We use simple linear interpolation to modify the log f0 contour. We believe that the benefits of using a more accurate warping path are probably quite small due to the small size of the demiphone units.

3.1.3. Spectrum

The spectral degradation is calculated similarly as the calculation of the F0 degradation, except that unvoiced frames are taken into account too. Let $f_{spec,j}$ and $g_{spec,j}$ be the j^{th} spectral trajectory of the demiphone s^k and reference demiphone r^l , respectively. We assumed that the spectral representation has diagonal covariance, and calculated the spectral degradation as the sum of the degradations for each individual spectral trajectory:

$$D_{spec}(s^k, r^l) = \sum_{j=1}^{m_{spec}} D_{spec,j}(s^k, r^l) \quad (9)$$

with m_{spec} the order of the spectrum. Both static and dynamic (Δ and $\Delta\Delta$) features are taken into account:

$$\begin{aligned} D_{spec,j}(s^k, r^l) = & \\ & w_{spec,j}^+ \cdot D_{spec,j}^+(f_{spec,j}, g_{spec,j}, \hat{f}_{spec,j}, \hat{g}_{spec,j}) \\ & + w_{spec,j}^- \cdot D_{spec,j}^-(f_{spec,j}, g_{spec,j}, \hat{f}_{spec,j}, \hat{g}_{spec,j}) \\ & + w_{spec,j,\Delta}^+ \cdot D_{spec,j,\Delta}^+(f_{spec,j}^\Delta, g_{spec,j}^\Delta, \hat{f}_{spec,j}^\Delta, \hat{g}_{spec,j}^\Delta) \\ & + w_{spec,j,\Delta}^- \cdot D_{spec,j,\Delta}^-(f_{spec,j}^\Delta, g_{spec,j}^\Delta, \hat{f}_{spec,j}^\Delta, \hat{g}_{spec,j}^\Delta) \\ & + w_{spec,j,\Delta\Delta}^+ \cdot D_{spec,j,\Delta\Delta}^+(f_{spec,j}^{\Delta\Delta}, g_{spec,j}^{\Delta\Delta}, \hat{f}_{spec,j}^{\Delta\Delta}, \hat{g}_{spec,j}^{\Delta\Delta}) \\ & + w_{spec,j,\Delta\Delta}^- \cdot D_{spec,j,\Delta\Delta}^-(f_{spec,j}^{\Delta\Delta}, g_{spec,j}^{\Delta\Delta}, \hat{f}_{spec,j}^{\Delta\Delta}, \hat{g}_{spec,j}^{\Delta\Delta}) \end{aligned} \quad (10)$$

$$D_{spec}^+(f, g, \hat{f}, \hat{g}) = \hat{D}_{spec}^+(f, \hat{g}) + \hat{D}_{spec}^-(g, \hat{f}) \quad (11)$$

$$D_{spec}^-(f, g, \hat{f}, \hat{g}) = \hat{D}_{spec}^-(f, \hat{g}) + \hat{D}_{spec}^+(g, \hat{f}) \quad (12)$$

$$\hat{D}_{spec}^+(f, \hat{g}) = \frac{1}{N_f} \sum_{i=1}^{N_f} |f_i - \hat{g}_i| \cdot H(\hat{g}_i - f_i) \quad (13)$$

$$\hat{D}_{spec}^-(f, \hat{g}) = \frac{1}{N_f} \sum_{i=1}^{N_f} |f_i - \hat{g}_i| \cdot H(f_i - \hat{g}_i) \quad (14)$$

with $w_{spec,j}^+$, $w_{spec,j}^-$, $w_{spec,j,\Delta}^+$, $w_{spec,j,\Delta}^-$, $w_{spec,j,\Delta\Delta}^+$, $w_{spec,j,\Delta\Delta}^-$, $j = 1 \dots m_{spec}$ trainable weights. The ‘‘hat’’-trajectories are linearly interpolated to match the duration of either the synthesis or the reference demiphone.

3.2. Quality prediction

There does not always exist a one-to-one phonemic correspondence between the reference phonemes and these of the synthesis, due to optional silences and pronunciation variation. In order to find the correspondence between these two sets of phonemes we first remove all silences, and then apply dynamic time warping [27] to find the best corresponding path between the two sets.

The degradation $D_{sent}(s, r)$ of a particular sentence s compared to a natural reference signal r is then calculated as the *average* of the degradations between the corresponding demiphones of the two signals. Note that averaging might not be the best option here because large degradations might influence the perception more than small degradations. In the present study we adopted the average because parameter optimization can be performed relatively straightforwardly. The degradation $D_{sent}(s, r)$ is calculated as:

$$D_{sent}(s, r) = \frac{1}{n} \sum D_{local}(s^k, r^l) \quad (15)$$

with n the length of the warping path in terms of phonemes, s^k the k^{th} demiphone of s and r^l the l^{th} demiphone of r . ‘‘Silence’’ phones are ignored as we assume that their contribution to the perceived quality is quite small. Demiphone degradations that could not be calculated (e.g. due to phones with zero duration) are also taken into account equating their degradations with the largest degradations of the whole sentence. Insertions and deletions are handled in the same manner. The quality $Q_{sent}(s, r)$ is the inverse of the sentence degradation and is calculated as

$$Q_{sent}(s, r) = 5 - D_{sent}(s, r) \quad (16)$$

$$Q_{sent}(s, r) = 5 - \mathbf{d} \mathbf{w} \quad (17)$$

with \mathbf{w} an N -dimensional column vector containing all weights, \mathbf{d} an N -dimensional row vector containing the *degradation* features calculated using the sentence s and reference r and N the total number of weights including the intercept. Optimal weights are found by minimizing the errors between the subjective and objective ratings. The number of training sentences M is higher than the number of weights, leading to an overdetermined linear system of equations that can be solved by least-squares regression. Some of the degradation features are highly-correlated (e.g. between delta and delta-delta features, and ‘‘positive’’ and ‘‘negative’’ degradations) and this may result in numerical instability [28]. We used ridge regression to improve the conditioning of the system. This introduces a penalty parameter λ (the ridge parameter), which regularizes the size of the weights:

$$\mathbf{w}_{ridge} = \arg \min_{\mathbf{w}} \{ \|\mathbf{a} - (\mathbf{5} - \mathbf{D} \mathbf{w})\| + \lambda \|\mathbf{w}\| \} \quad (18)$$

with \mathbf{a} an M -dimensional column vector containing all aggregated ratings, $\mathbf{5}$ an M -dimensional column vector containing only fives and \mathbf{D} a matrix with dimension $M \times N$ containing all degradation features. The optimal value of the ridge parameter can be found using cross-validation [29]. Flooring or ceiling is applied in order to ensure that the estimated quality value is always within the [1, 5] range. The quality Q_{sys} of a particular system (i.e. a unique combination of synthesizer setup and a particular speech database) is calculated by averaging the quality of its evaluation sentences. A monotonic mapping function can be trained in order to obtain an estimate closer to the average subjective measurements: $Q'_{sys} = w_{0,sys} + w_{0.5,sys} \sqrt{Q_{sys}} + w_{1,sys} Q_{sys} + w_{2,sys} Q_{sys}^2 + w_{3,sys} Q_{sys}^3$ with $w_{i,sys} \geq 0, i = 0.5 \dots 3$.

4. Evaluation

We evaluated our objective quality estimation algorithm using a leave-one-system-out cross-validation approach. At each iteration, all data from one system of the BC training data was held-out during training and used only for evaluation purposes. The ITU-T standard PESQ [2] was used as a baseline. It is a standardized algorithm for objective measurement of speech quality of degraded natural speech. PESQ predicts the overall impression rather than the naturalness of the input speech. These two quality subdimensions are typically highly correlated so we expected that the predictions for overall impression should be close to these of naturalness. Each sample was first downsampled to 16 kHz before applying the wideband (16 kHz) version of PESQ. No further preprocessing was applied.

The evaluation was performed at two levels, namely per individual sentence and per system. The figures of merit are cal-

| | PESQ [2] | | Proposed | |
|----------|-------------|-------------|-------------|-------------|
| | Sentence | System | Sentence | System |
| ρ | 0.04 / 0.04 | 0.05 / 0.05 | 0.60 / 0.59 | 0.82 / 0.84 |
| ρ_s | 0.07 / 0.07 | 0.14 / 0.13 | 0.60 / 0.58 | 0.83 / 0.84 |
| RMSE | 1.74 / 1.80 | 0.62 / 0.70 | 0.65 / 0.75 | 0.35 / 0.38 |

Table 2: Evaluation using leave-one-system-out cross-validation on the BC data. Values on the left of each entry are calculated by mean aggregation, while these on the right are calculated by median aggregation

| | ρ | ρ_s | RMSE |
|------------------------------------|--------|----------|------|
| Proposed (Ridge Regression) | 0.59 | 0.58 | 0.75 |
| ν -SVR (linear kernel) | 0.58 | 0.57 | 0.79 |
| ν -SVR (RBF kernel) | 0.54 | 0.53 | 0.82 |
| RepTree | 0.31 | 0.30 | 0.94 |
| IB1 | 0.37 | 0.36 | 1.07 |

Table 3: Evaluation using leave-one-system-out cross-validation on the BC data. The proposed approach was compared to other models. Results shown here are for median aggregation and sentence quality prediction.

culated using the vectors \mathbf{y}_{sent} and \mathbf{y}_{sys} containing the subjective ratings for sentences and systems, respectively, and the vectors $\hat{\mathbf{y}}_{sent}$ and $\hat{\mathbf{y}}_{sys}$ containing the predicted objective ratings for sentences and systems, respectively. All experiments were performed twice, using either the mean or the median aggregated ratings of the training data. None of the held-out evaluation data was used to train the coefficients of system quality mapping function. We calculated the Pearson correlation coefficient ρ , Spearman correlation coefficient ρ_s and root-mean square error (RMSE) between the subjective and objective ratings. Comparisons with previous studies based on BC data ([7], [15] and [16]) are not straightforward, due to different evaluation strategies and because these studies were based on a much smaller part of the BC data.

Results of the leave-one-system-out evaluation are shown in Table 2 and Figure 1. The baseline system, PESQ, systematically underestimated the quality of the synthetic sentences and systems: almost 90% of the predicted values per sentence were between 1 and 2. Similar results were found for (non-intrusive) ITU-T P.563 on the BC 2008 data [7]. This is another sign that algorithms optimized for degraded natural speech, are less suitable for synthetic speech. The proposed algorithm was much more accurate than PESQ. 54% and 45% of the errors per sentence were less than 0.5 in absolute value for mean and median rating aggregation, respectively. 87% and 80% of the errors per sentence were less than 1 in absolute value, for mean and median rating aggregation, respectively. The results per system are better than those per sentence. This is in agreement with previous studies ([7] [15] and [16]). Some of the errors for individual sentences are canceled out by averaging the ratings per system. 83% and 79% of the errors per system were less than 0.5 in absolute value for mean and median rating aggregation, respectively, while all errors per system were less than 1 in absolute value. The performance of our quality prediction algorithm on unit selection and HMM-based synthesis was quite similar.

The mapping between the degradation features and the subjective ratings might be non-linear. To test this hypothesis, we performed an experiment with the WEKA machine learning toolkit [30], in which the performances of three non-linear models (IB1, a K-nearest-neighbor algorithm [31], ν -SVR with

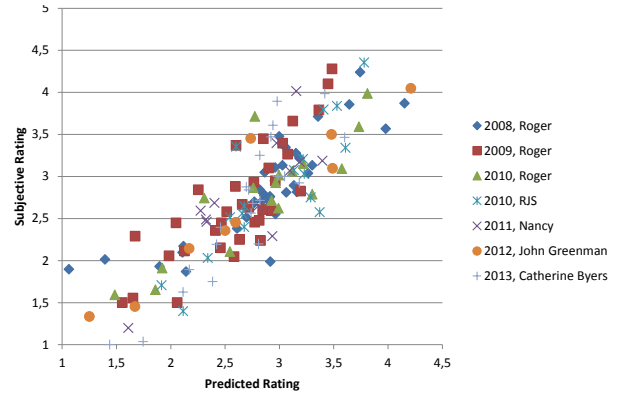


Figure 1: Scatterplot of the predicted and subjective ratings per system, grouped per BC and speaker. Results shown here are for the proposed quality prediction algorithm and median aggregation.

a radial basis function kernel [32] and RepTree, a decision tree learner [33]) were compared to two linear models (ν -SVR with a linear kernel [32] and the proposed ridge regression model). The same evaluation setup was used (leave-one-system-out). The ratings per sentence were obtained by median aggregation. The default parameters of the algorithms were kept, but the degradation features were scaled so that 95% of the values of each feature was within the range [-1,1] in order to improve the performance of the SVR models. The results of this initial experiment indicate that linear models outperform the non-linear algorithms (see Table 3). Part of the non-linearity of the human auditory system seems to be captured by the degradation features themselves, e.g. by using log F0, MFCCs and relative durations. Performing an additional non-linear transformation does not seem to help. These results contrast with the findings of [12]: they reported that non-linear models outperformed linear models for quality estimation with single-ended models. Our results seem to imply that simpler, linear models might be sufficient if a reference signal is available. A full exploration of non-linear models is, however, beyond the scope of this paper.

5. Conclusions

We described a relatively simple linear model to predict the naturalness of synthetic speech that was trained and evaluated on English data from the BC 2008 to 2013. An initial experiment with non-linear models indicated that linear models are sufficient for objective quality prediction of synthetic speech, at least when a reference signal is available. Performance of PESQ was quite poor on the BC data. The ITU-T standard POLQA [34] might be more suitable for dealing with synthesized speech than PESQ due to its better handling of timing differences, but this still needs to be investigated. We still need further experiments to confirm the validity of our approach on different languages and voices. This could for example be done using part of the remaining BC data (e.g. English paragraphs, Mandarin Chinese, Indian languages).

6. Acknowledgements

The research reported in this paper was partly supported by the projects IWT-SBO-SPACE, iMinds-RAILS, iMinds-SEGA and EC FP7 ALIZ-E (FP7-ICT-248116).

7. References

- [1] S. King, "Measuring a decade of progress in Text-to-Speech," *Linguistics*, vol. 1, no. 1, pp. 1–12, 2014.
- [2] ITU-T Recommendation P.862, "Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech codecs," International Telecommunication Union, Geneva, Switzerland, Tech. Rep., 2001.
- [3] V. J. van Heuven and R. van Bezooijen, "Quality Evaluation of Synthesized Speech," in *Speech Coding & Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Amsterdam, The Netherlands: Elsevier, 1995, pp. 707–738.
- [4] N. Campbell, "Evaluation of speech synthesis: from reading machines to talking machines," in *Evaluation of Text and Speech Systems*, L. Dybkjær, H. Hemsén, and W. Minker, Eds. Dordrecht, The Netherlands: Springer, 2007, pp. 29–64.
- [5] S. Möller, W.-Y. Chan, N. Cote, T. Falk, A. Raake, and M. Waltermann, "Speech Quality Estimation: Models and Trends," *Signal Processing Magazine, IEEE*, vol. 28, no. 6, pp. 18–28, Nov. 2011.
- [6] M. Cernak and M. Rusko, "An evaluation of synthetic speech using the PESQ measure," in *Proc. Forum Acusticum*, Budapest, Hungary, 2005.
- [7] T. H. 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.
- [8] S. Möller, D.-S. Kim, and L. Malfait, "Estimating the Quality of Synthesized and Natural Speech Transmitted Through Telephone Networks Using Single-ended Prediction Models," *Acta Acustica united with Acustica*, vol. 94, no. 1, pp. 21–31, Jan. 2008.
- [9] D.-Y. Huang, "Prediction of perceived sound quality of synthetic speech," in *Proc. APSIPA ASC*, Xi'an, China, 2011.
- [10] P. Počta and J. Holub, "Predicting the Quality of Synthesized and Natural Speech Impaired by Packet Loss and Coding Using PESQ and P.563 Models," *Acta Acustica united with Acustica*, vol. 97, no. 5, pp. 852–868, Sep. 2011.
- [11] S. Möller, F. Hinterleitner, T. H. Falk, and T. Polzehl, "Comparison of approaches for instrumentally predicting the quality of text-to-speech systems," in *Proc. Interspeech*, Makuhari, Chiba, Japan, 2010, pp. 1325–1328.
- [12] C. R. Norrenbrock, F. Hinterleitner, U. Heute, and S. Möller, "Quality prediction of synthesized speech based on perceptual quality dimensions," *Speech Communication*, vol. 66, no. 0, pp. 17–35, Feb. 2015.
- [13] A. W. Black and P. Taylor, "Automatically Clustering Similar Units For Unit Selection In Speech Synthesis," in *Proc. EUROSPEECH*, 1997, pp. 601–604.
- [14] A. W. Black and K. Tokuda, "Large Scale Evaluation of Corpus-based Synthesizers: Results and Lessons from the Blizzard Challenge 2005," in *Proc. Interspeech*, Lisbon, Portugal, 2005, pp. 77–80.
- [15] 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," in *Proc. Blizzard Challenge Workshop*, Kansai Science City, Japan, 2010.
- [16] C. R. Norrenbrock, F. Hinterleitner, U. Heute, and S. Möller, "Towards perceptual quality modeling of synthesized audiobooks-blizzard challenge 2012," in *Proc. Blizzard Challenge Workshop*, Portland, Oregon, USA, 2012.
- [17] K. Demuyck, J. Roelens, D. Van Compernelle, and P. Wambacq, "SPRAAK: an open source "SPeech Recognition and Automatic Annotation Kit"," in *Proceedings Interspeech*, Brisbane, Australia, 2008, pp. 495–498.
- [18] S. Fitt, "Unisyn Multi-accent Lexicon, version 1.3," 2007. [Online]. Available: <http://www.cstr.ed.ac.uk/projects/unisyn>
- [19] J. Yamagishi, H. Zen, T. Toda, and K. Tokuda, "Speaker-independent HMM-based speech synthesis system - HTS-2007 system for the Blizzard Challenge 2007," in *Proc. Blizzard Challenge Workshop*, Bonn, Germany, 2007.
- [20] D. Talkin, "A Robust Algorithm for Pitch Tracking (RAPT)," in *Speech Coding & Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Amsterdam, The Netherlands: Elsevier, 1995, pp. 495–518.
- [21] H. Kawahara, A. de Cheveign, H. Banno, T. Takahash, and T. Irino, "Nearly Defect-free F0 Trajectory Extraction for Expressive Speech Modifications based on STRAIGHT," in *Proc. Interspeech*, Lisbon, Portugal, 2005, pp. 537–540.
- [22] A. Camacho, "SWIPE: A Sawtooth Waveform Inspired Pitch Estimator for Speech And Music," Ph.D. Thesis, University of Florida, 2007.
- [23] D. Hirst, "A Praat plugin for Momel and INTSINT with improved algorithms for modelling and coding intonation," in *Proc. ICPhS*, Saarbruecken, Germany, 2007, pp. 1233–1236.
- [24] H. Kawahara, "STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic de-composition of speech sounds," *Acoustical Science and Technology*, vol. 27, no. 6, pp. 349–353, 2006.
- [25] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis: a unified approach to speech spectral estimation," in *Proc. ICSLP*, Yokohama, Japan, 1994, pp. 18–22.
- [26] K. Tokuda, K. Oura, A. Tamamori, S. Sako, H. Zen, T. Nose, T. Takahashi, J. Yamagishi, and Y. Nankaku, "Speech Signal Processing Toolkit (SPTK) v.3.7," 2013. [Online]. Available: <http://sp-tk.sourceforge.net/>
- [27] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Upper Saddle River, New Jersey, USA: Prentice Hall Signal Processing Series, 1993.
- [28] N. Balakrishnan, C. Read, B. Vidakovic, and N. L. Johnson, Eds., *Encyclopedia of statistical sciences*, 2nd ed. Hoboken, New Jersey, USA: John Wiley & Sons Inc., 2006.
- [29] G. H. Golub, M. Heath, and G. Wahba, "Generalized Cross-Validation as a Method for Selecting a Good Ridge Parameter," *Technometrics*, vol. 21, no. 2, pp. 215–223, May 1979.
- [30] I. W. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed. Elsevier, 2011.
- [31] D. Aha and D. Kibler, "Instance-based learning algorithms," *Machine Learning*, vol. 6, pp. 37–66, 1991.
- [32] 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, software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [33] T. Elomaa and M. Kääriäinen, "An Analysis of Reduced Error Pruning," *Journal of Artificial Intelligence Research*, vol. 15, pp. 163–187, 2001.
- [34] ITU-T Recommendation P.863, "Perceptual objective listening quality assessment," International Telecommunication Union, Geneva, Switzerland, Tech. Rep., 2011.