Improving Instrumental Quality Prediction Performance for the Blizzard Challenge

Tiago H. Falk,[1] Sebastian Möller,[2] Vasilis Karaiskos,[3] and Simon King[4]
[1] Dept. of Electrical and Computer Engineering, Queen's University, Canada
[2] Deutsche Telekom Labs, Berlin University of Technology, Germany
[3] School of Informatics, University of Edinburgh, United Kingdom
[4] Centre for Speech Technology Research, University of Edinburgh, United Kingdom

***Abstract*— In this paper, the performance of the standard instrumental quality prediction algorithm ITU-T P.563 is reported based on the 2007 and 2008 Blizzard Challenge speech data. The algorithm, which is optimized for natural speech, is shown to obtain poor correlation with subjective quality ratings. In an attempt to improve instrumental quality prediction performance for the Blizzard Challenge, modifications to the algorithm are proposed. In particular, a novel regression tree mapping is proposed based on five key features extracted by the P.563 algorithm. Experimental results on the 2008 Challenge dataset show that the performance attained with the improved algorithm substantially outperforms the original standard algorithm implementation.**

***Index Terms*— Instrumental quality assessment, Blizzard Challenge, speech synthesis, quality diagnosis, evaluation.**

I. Introduction

The Blizzard 2007 Challenge received submissions from sixteen participants who were asked to build synthetic voices using an American English speech corpus. This year, in the 2008 Challenge, twenty groups have enrolled and are synthesizing voices based on a British English speech corpus; a subset of the participants are also building voices using a Mandarin corpus. In the Challenge, synthesized voices are subjectively evaluated using a multi-section listening test. In this paper, we are interested in the English language corpora and in the mean opinion score (MOS) section of the listening tests where subjects were asked to rate the *naturalness* of the synthesized voices. In the test, listeners used a five-point scale with a rating of 1 indicating "completely unnatural" synthetic speech and a rating of 5 indicating "completely natural."

As reported in [1], the subjective evaluation is commonly performed online over the course of several weeks and listeners usually consist of speech experts, undergraduate students, and volunteers. In the 2007 Challenge, a total of 498 listeners signed up, of which 306 completed all sections of the listening test. The Challenge entry fee was used to pay a subset of the undergraduate students to perform the test in a controlled laboratory setting. As can be seen, preparing and carrying out listening tests is costly and labour-intensive. While costs can be reduced with the use of volunteer listeners and an online evaluation system, test preparation and analysis remains a time consuming task. As a consequence, an instrumental measure, shown to correlate highly with ratings obtained from a subjective listening test, would constitute an invaluable resource for Blizzard Challenge organizers.

To date, an instrumental quality measure for *synthesized* speech has yet to emerge. Several algorithms, however, have been proposed for *natural* speech transmitted over narrowband telephone networks. Representative standard algorithms include the International Telecommunications Union ITU-T P.563 algorithm [2] and the American National Standard Institute algorithm, ANIQUE+ [3]. Recently, these two algorithms were tested on synthesized speech transmitted over different telephone channels [4]. While the measures were shown to estimate the effects of the transmission channel, poor estimation of source speech quality was attained, signaling the need for a more accurate quality measure for synthesized speech.

In this paper, the performance of ITU-T P.563 algorithm is further investigated on the 2007 and 2008 Blizzard Challenge speech data. An in-depth analysis of the algorithm's internal signal processing is carried out and the insights obtained are used to propose modifications to the algorithm. In particular, a regression tree is used to map five key features into a final objective quality rating. Experimental results show that considerable improvements in measurement performance are obtained with the proposed modifications.

II. Description of the ITU-T P.563 Algorithm

The ITU-T P.563 algorithm was standardized in 2004 as the first single-ended measurement algorithm for narrowband speech transmitted over telephone channels [2]. The algorithm combines three principles, as depicted in Fig. 1, to detect and quantify signal distortions [5]. First, vocal tract and linear prediction (LP) analysis is performed to detect unnaturalness in the speech signal. The vocal tract is modeled as a series of tubes of different lengths and time-varying cross-sectional areas. From the speech signal, cross-sectional areas are evaluated for unnatural behavior. Similarly, higher-order statistics (skewness and kurtosis), computed for LP coefficients and cepstral coefficients, are investigated to see if they lie within the restricted range expected for natural speech.

Second, a pseudo-reference signal is reconstructed by modifying the computed LP coefficients to fit the vocal tract model of a typical human speaker. The pseudo-reference signal serves as input, along with the degraded speech signal, to a double-ended algorithm (similar to ITU-T P.862 [6]) to generate a "basic voice quality" measure. Lastly, specific distortions such

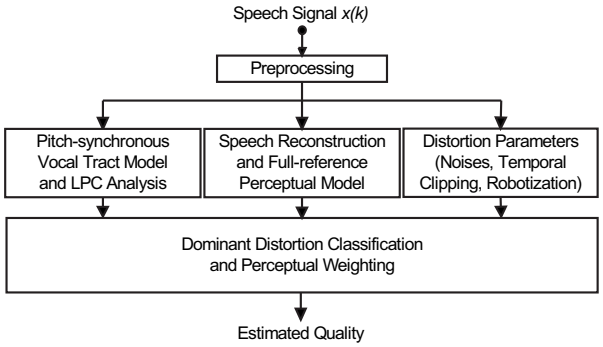


Fig. 1. Schematic representation of ITU-T P.563, taken from [4], [5].

as noise, temporal clippings, and robotization effects (voice with metallic sounds) are detected. A total of 51 characteristic signal parameters are calculated. Based on a restricted set of eight key parameters, one of six major distortion classes is detected. The distortion classes are, in decreasing order of "annoyance": high level of background noise, signal interruptions, signal-correlated noise, speech robotization, and unnatural male and female speech [5]. For each distortion class, a subset of the extracted parameters is used to compute an intermediate quality rating. Once a major distortion class is detected, the intermediate score is linearly combined with eleven other parameters to derive a final quality estimate.

It is emphasized here that synthesized speech does *not* fall within the recommended scope of the algorithm, thus the evaluation tests described herein constitute an "out-of-domain" experiment. Moreover, P.563 quality estimates are computed based on the five-point absolute category rating (ACR) scale, where a rating of 1 indicates "bad" and a rating of 5 indicates "excellent" speech quality [7]. While the ACR scale differs from the five-point naturalness scale used for the Blizzard Challenge, a recent study encompassing synthesized speech generated by six text-to-speech systems has suggested that the two quality dimensions are highly correlated [8] (correlation greater than 0.98). As such, in the sections to follow, direct comparisons between P.563 *quality* estimates and subjective *naturalness* ratings are carried out.

III. Gaining Insights: P.563 Performance on the 2007 Blizzard Challenge Dataset

As mentioned previously, the P.563 algorithm was developed for natural *transmitted* speech. As such, the algorithm first detects a major distortion class and then uses a class-specific subset of the extracted features to estimate signal quality. In the subsections to follow, an in-depth analysis of P.563's internal signal processing is carried out in order to gain insight into the strengths/weaknesses of the algorithm for quality prediction of synthesized speech. Estimation error, given by the estimated P.563 quality score minus the subjective quality score, is used as a performance metric. Performance is investigated on a "per-distortion-class" basis, on a "per-synthesizer" basis, and on an "overall" basis. Correlations between extracted features and subjective quality ratings are also analyzed in order to obtain insight into which features are most useful for the task at hand.

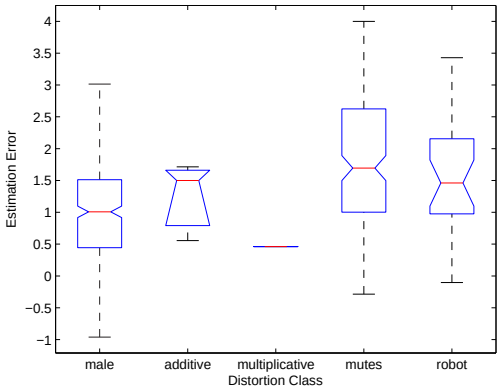
As mentioned previously, for the 2007 Challenge participants synthesized voices using subsets of the ATR English speech corpus; voice A used an eight-hour subset of the dataset, voice B used the one-hour ARTIC subset, and voice C used a participant-selected one-hour subset. In the per-synthesizer analysis to follow, the algorithms from 16 participants are available for voices A and B, whereas only 11 are available for voice C. Participants are labeled A-Q, where the label "G" is omitted as it corresponded to natural speech. We have chosen to omit natural speech from the analysis as, unlike synthesized speech, we have found the correlation between the P.563 *quality* ratings and the subjective *naturalness* ratings to be low for naturally produced speech. Such results are expected as, for example, encoding artifacts may reduce perceived speech quality but not alter the perceived naturalness of the speech signal.

A. Per-Distortion-Class Analysis

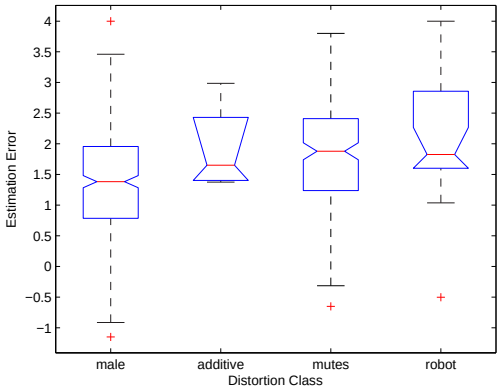
In this section, estimation errors are analyzed on a per-distortion-class basis. The goal is to investigate if major distortion classification is useful for synthesized speech. The number of speech signals used in the analysis is 544, 544, and 264 for voices A-C, respectively. The box-and-whisker plots in Figure 2 (a)-(c) depict the estimation error for voices A-C, respectively. The boxes have lines at the lower quartile, median, and upper quartile values; the whiskers extend to 1.5 times the interquartile range. Outliers (data with values beyond the ends of the whiskers) are represented by the symbol "+". The plots are computed using the Matlab function "boxplot" and the notches display the variability of the median between samples. The width of a notch is computed so that box plots whose notches do not overlap have different medians at the 5% significance level.

For voice A, it is observed that the male unnaturalness class is detected almost 64% of the time, whereas classes additive noise, multiplicative noise, mutes/interruptions, and robotization are detected approximately 0.6%, 0.2%, 31%, and 5% of the time, respectively. For voice B, distortion classes unnatural male, additive noise, mutes/interruptions, and robotization are detected approximately 63%, 0.8%, 32%, and 4% of the time, respectively. For voice C, in turn, the aforementioned classes are detected 63%, 0.8%, 34%, and 3% of the time, respectively. For voices B and C, multiplicative noise distortions are not detected.

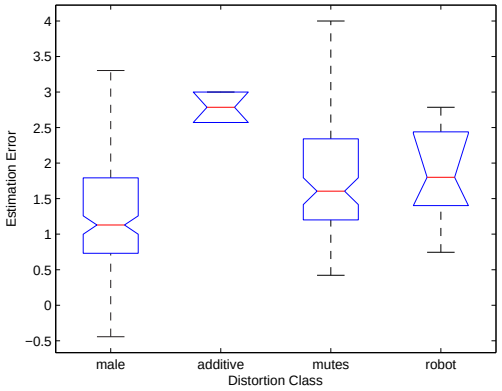
As can be observed from the plots, estimation errors are mostly positive, signaling that P.563 is overestimating synthesized speech quality. This is expected, as P.563 has been trained to disregard the source speech content and to focus on distortions introduced by the transmission channel. With synthesized speech, however, quality degradations are due to the source material per se. Moreover, if focus is placed on the unnatural male and mutes distortion classes, which occur a majority of the time, it can be observed that median errors and error spreads are similar, suggesting that poor performance is attained irrespective of the detected class. As such, it is proposed to remove the distortion classification from the P.563 algorithm, as described in Section IV.



(a)

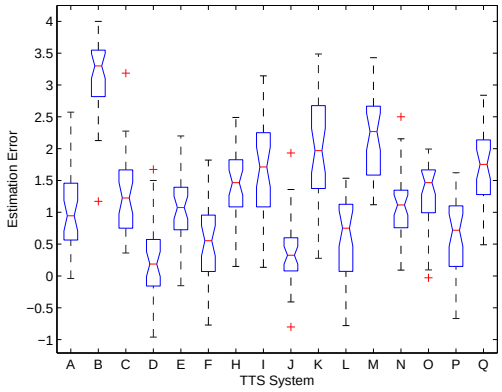


(b)

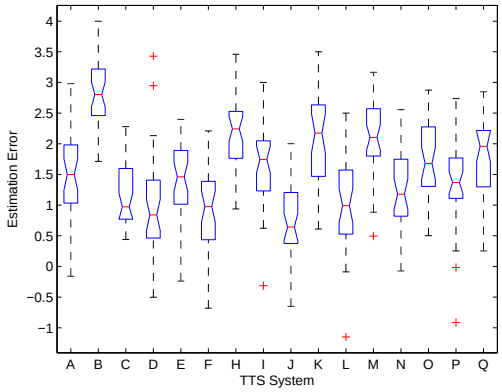


(c)

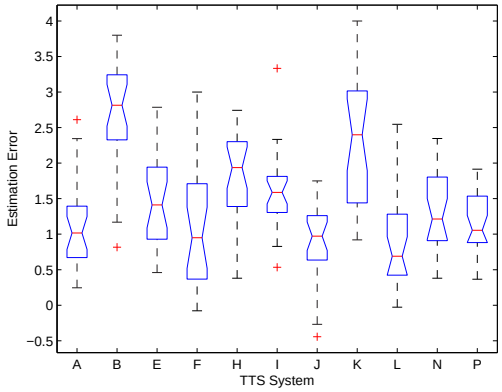
Fig. 2. Estimation errors on a per-distortion-class basis for voices (a) A, (b) B, and (c) C.



(a)



(b)



(c)

Fig. 3. Per synthesizer estimation error for voices (a) A, (b) B, and (c) C.

B. Per-Synthesizer Analysis

In this section, estimation errors are analyzed on a per-synthesizer basis. The goal is to investigate if larger errors occur for specific text-to-speech system configurations. A total of 34 sentences are analyzed per speech synthesizer. The box-and-whisker plots depicted in Fig. 3 illustrate estimation errors on a per-speech-synthesizer basis. As observed, median estimation errors are higher for system B, which is a unit selection system based on phone and syllable units. This is followed by systems K (unit selection, half-phone) and M (hidden Markov model based, phone), suggesting that errors are not dependent on synthesizer type or sub-word units; similar trends are observed for all three voices.

C. Overall Analysis

For overall analysis, Pearson correlation (ρ), root-mean-square error (ϵ), and Spearman rank-order correlation (ρ_S) are used as performance metrics. The measures are computed between the P.563 estimated quality scores and the subjective naturalness ratings. Spearman correlation is computed in a

TABLE I
P.563 OVERALL PERFORMANCE FOR VOICES A-C OF THE 2007 BLIZZARD CHALLENGE SPEECH DATA.

Metric	Voice A	Voice B	Voice C
ρ	0.33	0.42	0.35
ϵ	1.60	1.78	1.71
ρ_S	0.22	0.28	0.30
$\bar{\rho}$	0.43	0.64	0.64
$\bar{\epsilon}$	1.48	1.65	1.58
$\bar{\rho}_S$	0.11	0.27	0.49
$\bar{\rho}_{reg}$	0.56	0.67	0.72
$\bar{\epsilon}_{reg}$	0.66	0.52	0.48

manner similar to Pearson correlation, except original data values are replaced by the *ranks* of the data values. Results are reported on a per-sample basis and on a per-synthesizer basis. In the latter scenario, scores (both objective and subjective) for each system are averaged prior to calculation of the performance metrics; the overbar notation is used to represent per-synthesizer performance metrics. Additionally, as recommended in [2], a third-order monotonic regression function is used to map the estimated quality score onto the subjective scale; performance metrics with a subscript "reg" are used to distinguish metrics computed after regression. Note that monotonic regression does not alter the ranking of the scores, thus $\rho_S = \rho_{(S,reg)}$.

Table I reports performance metrics for the 2007 Blizzard Challenge data (voices A-C). As can be seen for all three voices, low correlation and high errors are attained on a per-sample basis. Somewhat improved performance is attained once analysis is performed on a per-synthesizer basis, except for the rank correlation where lower values are attained. After third order polynomial regression, superior performance is attained for voice C. The obtained performance figures, however, are considerably lower than those attained with the P.563 algorithm for natural speech, as reported in [9].

D. Correlation Between P.563 Internal Features and Subjective Quality Ratings

In order to investigate improvements to the P.563 algorithm, we compute the performance attained with each individual feature extracted by the P.563 algorithm. In previous experiments (e.g., [10]), it has been observed that the behaviour of P.563 internal features is highly dependent on the characteristics of the speech signals available in the test database. One possible cause may be that the P.563 algorithm uses different signal processing strategies for signals with, e.g., speech activity ratios greater than 80%, or signals with different low-frequency energy content (i.e., processed by filters with different lowpass cutoff frequencies). In our previous experiments, the latter has shown to cause some of the features extracted from the algorithm to obtain inconsistent trends between different databases, such as positive correlation with subjective quality on one dataset and negative correlation on another.

Table II reports the names of the features that attain $|\rho > 0.1|$ with subjective quality for the English subsets of the 2007/2008 corpora; the reader is referred to [2] for a detailed description of the features. Compared to the results described in Table I, several internal features attain higher correlations than the final P.563 quality score. Such insight suggests that improvements to the final mapping function are needed. It is also noted that the top features are gleaned from all three major processing blocks depicted in Fig. 1.

Interestingly, for the 2008 dataset, a local background noise related feature is shown to attain the highest correlation with subjective quality. According to [2], local background noise is defined as the noise between phonemes. The basic assumption used in the algorithm is that for an interval of normal speech of 1-second duration, at least four start or stop events are expected. If the number of start or stop events in one second is less than four, the algorithm assumes that local background noise is present. It is conjectured that for the synthesized speech signals under test, less than four start/stop events are detected, possibly due to conservative measures taken by the text-to-speech synthesizer in order to avoid concatenation artifacts. Such assumptions will be investigated once the speech files become publicly available in the near future.

Moreover, it is observed that the top-selected features differ from the 2007 and 2008 datasets. Features in Table II, described in bold, represent those with consistent behaviour (i.e., correlation values with same signs) across the two datasets. On the other hand, features described in italics represent features with inconsistent behavior. In our experiments, "consistent" features are used to train a regression mapping for instrumental quality measurement; data from the 2007 Challenge is used for training. Once the speech files are made publicly available, further analysis will be carried out and additional features may be incorporated into a modified regression mapping. The section to follow describes the regression mapping used in our experiments, as well as reports performance figures for the unseen 2008 Challenge speech dataset.

IV. EXPERIMENT RESULTS

We have tested different mapping functions, namely, linear regression, support vector regression, and regression trees to improve P.563 performance; in our experiments, a regression tree attained superior accuracy. The performance figures described in this section are based on the regression tree depicted in Fig. 4, which was trained on the 2007 Challenge speech data with the ten features described in bold letters in Table II. After tree pruning, only five of the top ten features are used in the final quality mapping.

Table III reports the performance of the modified and the original P.563 algorithm on the unseen 2008 Challenge dataset. As witnessed, considerable improvement in performance is attained after the proposed modifications. Note that the results obtained are only somewhat lower than the expected $\rho > 0.8$ threshold set by ITU-T during the 2004 competition which saw P.563 as the winning quality measurement system for natural speech. Moreover, Fig. 5 (a) and (b) illustrates scatter plots of subjective versus estimated quality scores, on a per-synthesizer basis, for the modified and original P.563 implementations, respectively. From the plots, it can be seen that the proposed

TABLE II

TOP FEATURES EXTRACTED BY P.563 AND THEIR RESPECTIVE CORRELATION VALUES WITH SUBJECTIVE NATURALNESS RATINGS.

	Blizzard Challenge 2007 data		Blizzard Challenge 2008 data	
Rank	Feature Name	ρ	Feature Name	ρ
1	**Mute length**	-0.41	Local background noise affected samples	0.56
2	**Speech interruptions**	-0.38	Cepstral absolute deviation	-0.39
3	**Final VTP average**	0.37	Basic voice quality	0.28
4	**Sharp declines**	-0.37	ART average	0.28
5	LPC absolute skewness	0.30	Local background noise	-0.27
6	**Local background noise affected samples**	0.29	Speech section level variation	-0.26
7	LPC kurtosis	0.27	VTP peak tracker	-0.23
8	Speech level	0.26	Final VTP average	0.17
9	**ART average**	0.23	Basic voice quality asymmetric	-0.17
10	Cepstral kurtosis	-0.22	VTP VAD overlap	-0.14
11	*Basic voice quality asymmetric*	0.20	Spectral level range	0.14
12	Spectral clarity	0.18	Pitch cross power	-0.13
13	Global background noise	-0.18	Relative noise floor	0.13
14	*Basic voice quality*	-0.17	Pitch cross correlation offset	-0.12
15	**VTP maximum tube section**	0.17	VTP maximum tube section	0.11
16	Spectral level deviation	0.15	Mute length	-0.11
17	**Pitch cross correlation offset**	-0.15	Speech interruptions	-0.11
18	**Spectral level range**	0.14	Sharp declines	-0.10
19	**Speech section level variation**	-0.13	–	–
20	*Pitch cross power*	0.11	–	–
21	Estimated segmental SNR	-0.11	–	–

TABLE III

PERFORMANCE OF THE MODIFIED P.563 ALGORITHM ON THE BLIZZARD CHALLENGE 2008 DATA; RESULTS ARE BASED ON THE REGRESSION TREE TRAINED ON THE 2007 CHALLENGE DATASET (VIDE FIG. 4).

Metric	Modified	Original
ρ	0.30	-0.07
ϵ	0.76	1.86
ρ_S	0.33	-0.10
$\bar{\rho}$	0.69	-0.20
$\bar{\epsilon}$	0.38	1.68
$\bar{\rho}_S$	0.65	0.08
$\bar{\rho}_{reg}$	0.79	-0.08
$\bar{\epsilon}_{reg}$	0.27	0.43

algorithm is capable of correctly detecting the two best systems and one of the worst systems from the unseen test dataset.

V. CONCLUSION

The ITU-T standard P.563 quality measurement algorithm, optimized for natural speech, is tested on synthesized speech data obtained from the 2007 and 2008 Blizzard Challenges. It is shown that the quality estimates obtained from the algorithm are poorly correlated with quality ratings obtained from subjective listening tests. An in-depth analysis of the algorithm is carried out and the insights obtained are used to propose modifications to the algorithm in order to improve quality measurement performance for synthesized speech. In particular, a regression tree is used to map five features computed by the P.563 algorithm into a final quality rating. The performance of the modified algorithm on synthesized speech is shown to be only somewhat lower than the expected performance threshold set by ITU-T for instrumental quality measurement of natural transmitted speech.

REFERENCES

[1] M. Fraser and S. King, "The Blizzard Challenge 2007," in *Proc. Blizzard Challenge Workshop*, 2007.

[2] ITU-T P.563, "Single ended method for objective speech quality assessment in narrowband telephony applications," Intl. Telecom. Union, 2004.

[3] ATIS-PP-0100005.2006, "Auditory non-intrusive quality estimation plus (ANIQUE+): Perceptual model for non-intrusive estimation of narrowband speech quality," American National Standards Institute, 2006.

[4] 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, pp. 21–31, 2008.

[5] L. Malfait, J. Berger, and M. Kastner, "P.563 - The ITU-T standard for single-ended speech quality assessment," *IEEE Trans. on Audio, Speech and Language Proc.*, vol. 14, no. 6, pp. 1924–1934, Nov. 2006.

[6] ITU-T P.862, "Perceptual evaluation of speech quality: An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Intl. Telecom. Union, 2001.

[7] ITU-T P.800, "Methods for subjective determination of transmission quality," Intl. Telecom. Union, 1996.

[8] K. Seget, "Untersuchungen zur auditiven qualität von sprachsyntheseverfahren (Study of perceptual quality of text-to-speech systems)," July 2007, Bachelor thesis, Christian-Albrechts-University of Kiel.

[9] A. Rix, J. Beerends, D.-S. Kim, P. Kroon, and O. Ghitza, "Objective assessment of speech and audio quality - Technology and applications," *IEEE Trans. on Audio, Speech and Language Proc.*, vol. 14, no. 6, pp. 1890–1901, Nov. 2006.

[10] ITU-T Contribution COM 12-180, "Single-ended quality estimation of synthesized speech: Analysis of the Rec. P.563 internal signal processing," Intl. Telecom. Union, Geneva, (Authors: S. Möller and T. H. Falk).

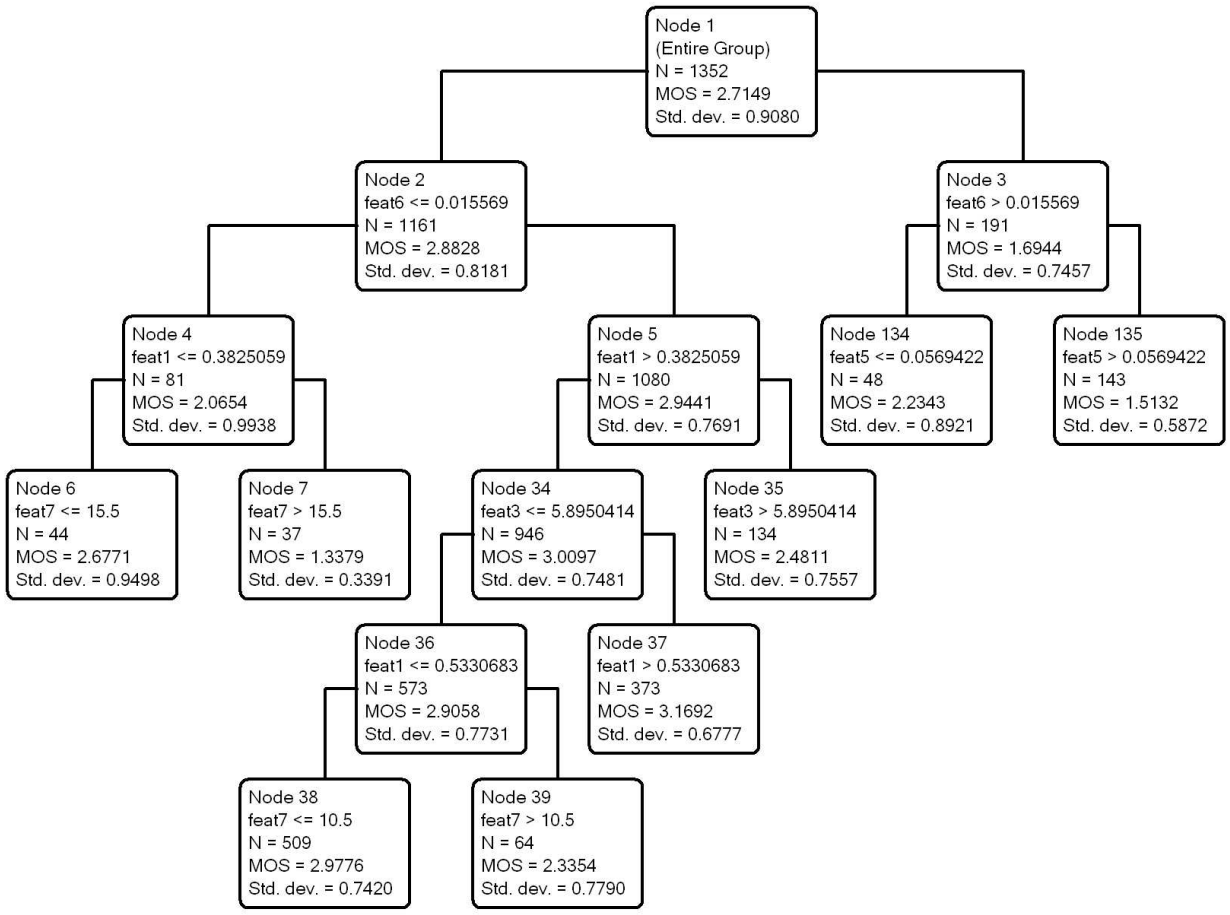
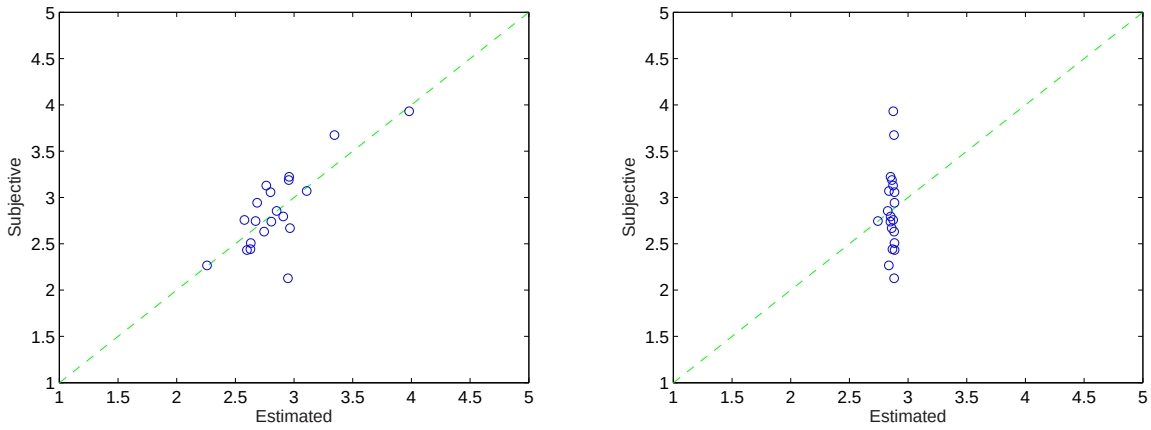


Fig. 4. Overview of the regression tree trained on the 2007 Blizzard Challenge dataset. The labels "feat$_i$" refer to features "final VTP average" ($i = 1$), "pitch cross correlation offset" ($i = 3$), "speech interruptions" ($i = 5$), "sharp decline" ($i = 6$), and "mute length" ($i = 7$).



(a) (b)

Fig. 5. Scatter plot of subjective versus estimated quality scores, on a per-synthesizer basis, after third-order polynomial regression for the (a) modified, and (b) original P.563 implementations. Plots are for the 2008 Blizzard Challenge dataset which was unseen to the modified P.563 algorithm.