Non-Intrusive Speech Quality Assessment with Low Computational Complexity

Volodya Grancharov, David Y. Zhao, Jonas Lindblom, and W. Bastiaan Kleijn

School of Electrical Engineering
KTH (Royal Institute of Technology)
10044 Stockholm, Sweden
{volodya.grancharov, david.zhao, bastiaan.kleijn}@ee.kth.se, jonas.lindblom@skype.net

Abstract

We describe an algorithm for monitoring subjective speech quality without access to the original signal that has very low computational and memory requirements. The features used in the proposed algorithm can be computed from commonly used speech-coding parameters. Reconstruction and perceptual transformation of the signal are not performed. The algorithm generates quality assessment ratings without explicit distortion modeling. The simulation results indicate that the proposed non-intrusive objective quality measure performs better than the ITU-T P.563 standard despite its very low computational complexity.

Index Terms: non-intrusive quality assessment, quality of service.

1. Introduction

Speech quality assessment is an important problem in communication systems. The quality of a speech signal is a *subjective* measure. It can be expressed in terms of how natural the signal sounds or how much effort is required to understand the message. In a formal subjective test, speech is played to a group of listeners, who are then asked to rate the quality of the signal. A commonly used subjective quality scale is the Mean Opinion Score (MOS) [1], [2].

Objective measures use mathematical expressions to predict speech quality. Their low cost is attractive for continuous monitoring of the quality of services (QoS) of a network. Two different test situations can be distinguished: 1) intrusive (both the original and distorted signals are available), and 2) non-intrusive (only the distorted signal is available). The methods are illustrated in Fig. 1. The original signal is typically not available in QoS monitoring, which means that *non-intrusive* quality assessment must be used.

Algorithms for non-intrusive speech quality assessment have seen rapid development over the last fifteen years. They have been based on various principles. Au and Lamb [3] partition the spectrogram and compute its variance and dynamic range on a block-by-block basis. The average variance and dynamic range is used to predict speech quality. Gray et al. [4] attempt to predict the likelihood that an audio signal is generated by the human vocal production system. The parameterized data are used to generate physiologically based rules for error assessment. Liang and Kubichek [5] compare the speech to be assessed to an artificial reference signal that is appropriately selected from a optimally clustered codebook. Recent algorithms based on Gaussian-mixture probability models (GMM) of features derived from perceptually motivated spectral-envelope representations can be found in [6] and [7]. A novel, perceptually motivated speech quality assessment algorithm based on temporal envelope representation of speech is presented in [8].

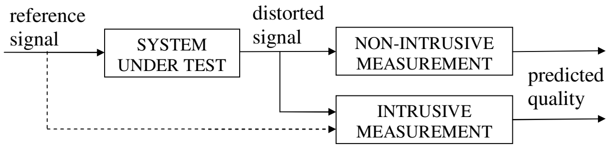

Figure 1: Intrusive and non-intrusive quality assessment. Non-intrusive algorithms do not have access to the reference signal.

The results of some of these studies have been incorporated in the International Telecommunication Union (ITU) standard for non-intrusive quality assessment, ITU-T P.563 [9]. This standard provides state-of-the-art non-intrusive speech quality assessment. A total of 51 speech features are extracted from the signal. *Key features* are used to determine a dominant distortion class, and in each distortion class a linear combination of features is used to predict the intermediate speech quality. The final speech quality is estimated from the intermediate quality and 11 additional features.

The above listed non-intrusive measures are designed to predict the effects of a large range of distortions, and they typically have high computational complexity. Non-intrusive quality prediction is possible at much lower complexity if it is assumed that the type of distortion is known [10].

We conclude that existing algorithms either have a high computational complexity and a broad range of application or a low complexity and a narrow range of application. This has motivated us to develop a low complexity quality assessment (LCQA) algorithm. The algorithm predicts speech quality from generic features commonly used in speech coding, without any assumption on the type of distortion. In extensive testing, the proposed algorithm was found to be significantly better than ITU-T P.563 (cf. section 3).

2. Low-complexity quality assessment

The objective of the proposed LCQA algorithm is to provide a estimate of the MOS for each utterance, based on a set of features that is readily available from speech coders used in a communication network. The algorithm has low computational complexity, to make it useful for practical applications.

2.1. Speech Features

Automatic quality analysis systems are based on the extraction of a feature vector. The set of *per-block features* used in LCQA aims to capture low-level aspects of the speech-signal structure that are likely relevant to human quality judgment. In this section we discuss the per-block features that we have selected.

The spectral flatness measure is related to the intensity of the resonant structure in the power spectrum:

$$\Phi_1(n) = \frac{\exp\left(\frac{1}{2\pi}\int_{-\pi}^{\pi}\log\left(P_n(\omega)\right)d\omega\right)}{\frac{1}{2\pi}\int_{-\pi}^{\pi}P_n(\omega)d\omega}, \tag{1}$$

where $P(\omega)$ is the autoregressive (AR) model spectral envelope, which is available in virtually all speech coders.

As a second per-block feature we use spectral dynamics, defined as

$$\Phi_2(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\log P_n(\omega) - \log P_{n-1}(\omega)\right)^2 d\omega, \tag{2}$$

where n is a block index that is consistent with that of speech coders (the typical separation of blocks is 20 ms). Spectral dynamics have been studied and successfully used in speech coding [11], and speech enhancement [12].

The spectral centroid [13] determines the frequency area around which most of the signal energy concentrates

$$\Phi_3(n) = \frac{\int_{-\pi}^{\pi}\omega\log\left(P_n(\omega)\right)d\omega}{\int_{-\pi}^{\pi}\log\left(P_n(\omega)\right)d\omega}, \tag{3}$$

and it is considered to be a measure of perceptual "brightness".

The final three per-block features are the variance of the excitation of the AR model E_n^e, the speech signal variance E_n^s, and the pitch period T_n. They are denoted as $\Phi_4(n)$, $\Phi_5(n)$, and $\Phi_6(n)$.

The per-block features and their first time derivatives (we do not use the derivative of the spectral dynamics) form an 11-dimensional per-block feature vector signal $\mathbf{\Phi}(n)$ with a sampling rate that depends on the speech coder.

We hypothesize that the speech quality of an utterance can be estimated from statistical properties of the per-block feature vector signal $\mathbf{\Phi}(n)$ over a subset of *relevant* (cf. section 2.2) signal blocks of the utterance. We characterize the probability distribution of $\mathbf{\Phi}(n)$ with the mean, variance, kurtosis, and skewness of its components. Thus, the moments are calculated independently for each feature. These per-utterance statistical features are grouped into a single *global feature set* Ψ.

2.2. Dimensionality Reduction

It is common in the quality assessment literature to remove signal blocks based on a voice activity detector or an energy threshold [14]. We propose a generalization of this concept by including block-activity thresholds in all feature dimensions. The scheme, presented in Table 1 excludes speech active signal blocks if they do not contribute to the accuracy of speech quality prediction.

From Table 1 we see that a vector of thresholds to select the active blocks is $\mathbf{\Theta} = \{\Theta_i^L, \Theta_i^U\}_{i=1}^{11}$. Let $\tilde{\Omega}$ denote the relevant blocks. We search for the threshold vector $\mathbf{\Theta}$ that minimizes the criterion ε:

$$\mathbf{\Theta} = \arg\min_{\mathbf{\Theta}^*} \varepsilon(\tilde{\Omega}(\mathbf{\Theta}^*)), \tag{4}$$

where ε is the (experimental) root mean square error in the MOS over the utterances in the training database (cf. section 3). Thus, the threshold vector is selected to maximize the performance of the LCQA algorithm over the training database.

Table 1: Selection of relevant signal blocks.

```
Initialize: $\tilde{\Omega} = \{\emptyset\}$
for $n \in \Omega$
        if $\Phi_1(n) \in [\Theta_1^L, \Theta_1^U]$ & ... & $\Phi_{11}(n) \in [\Theta_{11}^L, \Theta_{11}^U]$
            Accept the n-th frame: $\tilde{\Omega} = \tilde{\Omega} + \{n\}$
```

For our implementation (cf. section 2.4), the optimization of ε with the frame selection algorithm, described in Table 1, led to the following acceptance criterion the n-th frame:

$$\Phi_5(n) > \Theta_5^L \ \&\ \Phi_1(n) < \Theta_1^U \ \&\ \Phi_2(n) < \Theta_2^U, \tag{5}$$

with threshold values $\Theta_5^L = 3.10$, $\Theta_1^U = 0.67$, and $\Theta_2^U = 4.21$. Only the speech variance Φ_5, spectral flatness Φ_1, and spectral dynamics Φ_2), have significant impact on the block selection. The first and second inequalities in (5) accept only frames with high-energy and clear formant structure. This suggests that the LCQA algorithm extracts information about the speech quality predominately from voiced speech regions. The third inequality selects only stationary speech regions. A possible explanation is that distortion is more easily perceived in steady-state regions of speech.

Once the optimal subset of blocks $\tilde{\Omega}$ is found, we search for the optimal subset of features $\tilde{\Psi}$. (We do not optimize the sets of $\tilde{\Omega}$ and $\tilde{\Psi}$ jointly because of the associated high computational complexity.) This optimization step is defined as follows: given the original set of features Ψ of cardinality $|\Psi|$, select a subset of features $\tilde{\Psi} \subset \Psi$ of cardinality $|\tilde{\Psi}| < |\Psi|$ that is optimized for the performance of the LCQA algorithm:

$$\tilde{\Psi} = \arg\min_{\tilde{\Psi}^* \subset \Psi} \varepsilon(\tilde{\Psi}^*). \tag{6}$$

A full search is the only dimensionality reduction procedure that guarantees that a global optimum is found. Non-optimal methods with lower complexity such as the well-known Sequential Forward Selection and Sequential Backward Selection and the more advanced (L,R) algorithm [15] are commonly used. The Floating Search methods [16] are extensions of the (L,R) search methods. In our simulations we used the Sequential Floating Backward Selection procedure, which consists of applying after each backward step a number of forward steps as long as the resulting subset are better than the previously evaluated ones.

We reduced the dimensionality of the global feature set from 44 to 14, i.e., $|\tilde{\Psi}| = 14$. The final feature set is

$$\begin{aligned}\tilde{\Psi} = \{&s(\Phi_1), \sigma(\Phi_2), \mu(\Phi_4), \mu(\Phi_5), \sigma(\Phi_5), s(\Phi_5), \mu(\Phi_6),\\ &s(\Phi_7), \mu(\Phi_8), \mu(\Phi_9), \sigma(\Phi_9), s(\Phi_9), \mu(\Phi_{10}), \mu(\Phi_{11})\},\end{aligned} \tag{7}$$

where $\mu(\cdot)$ indicates the mean, $\sigma(\cdot)$ the variance, and $s(\cdot)$ the skewness of the per-block features. We observe that all per-block features are represented in the set $\tilde{\Psi}$, and that the speech signal variance Φ_5, and the time-derivative of the variance of the excitation signal Φ_9 are most frequent. Another interesting observation is that the kurtosis is found to be less important.

2.3. Mapping the Global Feature Set to Speech Quality

Let Q denote the MOS of an utterance as obtained from the MOS labeled training database. We construct an objective estimator $\hat{Q}$ of the subjective quality as a function of the global feature set, i.e., $\hat{Q} = \hat{Q}(\tilde{\Psi})$, and search for the mapping that minimizes the mean square error

$$\hat{Q}(\tilde{\Psi}) = \arg\min_{Q^*(\tilde{\Psi})} \mathrm{E}\{(Q - Q^*(\tilde{\Psi}))^2\}, \tag{8}$$

where $\mathrm{E}\{\}$ is the expectation operator. Equation (8) is minimized by the conditional expectation $\hat{Q}(\tilde{\Psi}) = \mathrm{E}\{Q|\tilde{\Psi}\}$ and the problem reduces to the estimation of the conditional probability density. To this purpose, we model the joint density of the feature variables with the subjective MOS scores as a GMM

$$f(\varphi) = \sum_{m=1}^{M} \omega^{(m)} \mathcal{N}(\varphi|\mu^{(m)}, \Sigma^{(m)}), \tag{9}$$

where $\varphi = [Q, \tilde{\Psi}]$, $\omega^{(m)}$ are the mixture weights, and $\mathcal{N}(\varphi|\mu^{(m)}, \Sigma^{(m)})$ are multivariate Gaussian densities with mean vectors $\mu^{(m)}$ and covariance matrices $\Sigma^{(m)}$. The parameters can be trained using the well-known expectation maximization (EM) algorithm. Our experiments showed that it is sufficient to use 12 full-covariance matrices, i.e., $M = 12$.

2.4. Implementation Details

In this section we describe how the features are calculated in our experimental system, which has a block rate of 50 Hz. To obtain low complexity, we simplify the computations for the spectral flatness, spectral dynamics, and spectral centroid so that the speech coder bit-stream can be used, avoiding signal reconstruction.

We calculate the spectral flatness as the ratio of the tenth-order prediction error and the signal variance

$$\Phi_1(n) = \frac{\mathrm{E}_n^e}{\mathrm{E}_n^s}. \tag{10}$$

We calculate the signal variance without reconstructing the waveform by means of the reverse Levinson-Durbin recursion for the AR model parameters (step-down algorithm).

The AR-spectrum dynamics are implemented as a weighted Euclidean distance on the line spectral frequencies (LSF):

$$\Phi_2(n) = (\mathbf{f}_n - \mathbf{f}_{n-1})^T \mathbf{W}_n (\mathbf{f}_n - \mathbf{f}_{n-1}), \tag{11}$$

where **f** is the LSF vector and $\mathbf{W}_n$ is the diagonal matrix

$$W_n^{(ii)} \quad = \quad (f_n^{(i)} - f_n^{(i-1)})^{-1} + (f_n^{(i+1)} - f_n^{(i)})^{-1}. \tag{12}$$

These weights are also used to obtain a redefined spectral centroid:

$$\Phi_3(n) = \frac{\sum_{i=1}^{10} i W_n^{(ii)}}{\sum_{i=1}^{10} W_n^{(ii)}}. \tag{13}$$

To reduce storage, we calculate the selected global descriptor μ_Φ recursively:

$$\mu_\Phi(n) = \frac{n-1}{n} \mu_\Phi(n-1) + \frac{1}{n} \Phi(n), \tag{14}$$

where n is the index over the relevant block set $\tilde{\Omega}$. In a similar fashion, we propagate Φ^2, Φ^3, and Φ^4 to obtain the central moments. Upon completion of the utterance, these quantities are used to obtain the remaining global descriptors variance, skew, and kurtosis. Table 2 summarizes the LCQA algorithm.

3. Performance Evaluation

We compare the performance of the proposed LCQA algorithm with the ITU-T P.563 standard in terms of both per-utterance and per-condition performance metrics. The *per-utterance* quality-estimation performance is evaluated using the root mean square error

$$\varepsilon = \sqrt{\frac{\sum_{i=1}^{N} (Q_i - \hat{Q}_i)^2}{N}}, \tag{15}$$

Table 2: Overview of LCQA Algorithm.

1. For the n-th speech block calculate $\{\mathrm{E}_n^e, \mathrm{T}_n, \mathbf{f}_n\}$ from the waveform or extract from the bit-stream.
2. Calculate per-block feature vector $\mathbf{\Phi}(n)$, based on $\{\mathrm{E}_n^e, \mathrm{T}_n, \mathbf{f}_n\}$ and $\{\mathrm{E}_{n-1}^e, \mathrm{T}_{n-1}, \mathbf{f}_{n-1}\}$.
3. For the selected subset of blocks $\tilde{\Omega}$ recursively calculate the central moments. Block selection is controlled by the threshold vector $\mathbf{\Theta}$.
4. At the end of the utterance calculate global feature set $\tilde{\Psi}$ containing mean, variance, skew, and kurtosis of the local features over the relevant blocks.
5. Predict the speech quality as a function of the global feature set $\hat{Q} = \hat{Q}(\tilde{\Psi})$, through GMM mapping.

and the correlation coefficient

$$R = \frac{\sum_{i=1}^{N} (\hat{Q}_i - \mu_{\hat{Q}})(Q_i - \mu_Q)}{\sqrt{\sum_{i=1}^{N} (\hat{Q}_i - \mu_{\hat{Q}})^2} \sqrt{\sum_{i=1}^{N} (Q_i - \mu_Q)^2}}, \tag{16}$$

where Q_i and $\hat{Q}_i$ are the quality and the estimated quality of the i-th utterance, μ_Q and $\mu_{\hat{Q}}$ are the respective mean values and N is the number of utterances.

In the *per-condition* quality estimate, i is the condition, Q_i and $\hat{Q}_i$ are the quality and the estimated quality of the i-th condition, μ_Q and $\mu_{\hat{Q}}$ are the mean values over all conditions, and N is the number of conditions.

To improve generalization performance, we use *training with noise* procedure [17]. We created virtual ("noisy") training patterns, by adding zero mean white Gaussian noise, at 20 dB SNR (measured including silence regions), to the training patterns. In this manner we created four virtual sets for each global feature set Ψ, and the training was based on the union of the "real" and "noisy" data.

3.1. Speech Databases

For the training procedure we used 11 MOS labeled databases provided by Ericsson AB and seven such databases from ITU-T P.Supp 23 database [18]. We refer to the combined set as the *global database*. The global database contains utterances in English, French, Japanese, Italian and Swedish, with a large variety of distortions, such as various coding, tandeming, and packet loss conditions, background noise, effects of noise suppression, switching effects, and different SNR levels. The global database consisted of 7646 speech files with an average length of 8s.

3.2. Per-Utterance Performance over Global Database

We assessed the accuracy of the proposed LCQA algorithm over the global database. We used 10-fold *cross validation* with 20% of the speech material, to provide robustness in the performance evaluation. Table 5 shows the the averaged results of the cross-validation tests, and Table 4 shows the RMSE outliers in %. The LCQA algorithm significantly outperforms ITU-T P.563.

Table 3: Averaged performance in correlation and RMSE.

	R	ε
LCQA	0.89	0.39
ITU-T P.563	0.75	0.61

Table 4: Outliers in RMSE, averaged over cross-validation tests.

	Outliers (in %)		
	$\varepsilon > 0.6$	$\varepsilon > 0.8$	$\varepsilon > 1.0$
LCQA	6.1	3.9	2.6
ITU-T P.563	22.5	14.6	10.3

3.3. Per-Condition Performance over Unknown Databases

In this experiment we split the available databases into two parts, *test set* and *training set*. The test set was based on seven databases from ITU-T P.Supp 23 (1328 files) and the training set was based on 11 Ericsson databases (6318 files). The test set is not available during the training, but used only for evaluation. The training for the dimensionality reduction scheme and performance evaluation experiments was based entirely on the training set.

Table 5 shows the per-condition performance results over the seven databases from ITU-T P.Supp 23. The test results clearly indicate that the proposed LCQA algorithm outperforms the standardized ITU-T P.563.

Table 5: Performance of the LCQA algorithm in terms of per-condition correlation coefficient.

Database	LCQA	ITU-T P.563
ITU-T P.Supp 23 Exp 1 A	0.94	0.88
ITU-T P.Supp 23 Exp 1 D	0.94	0.81
ITU-T P.Supp 23 Exp 1 O	0.95	0.90
ITU-T P.Supp 23 Exp 3 A	0.93	0.87
ITU-T P.Supp 23 Exp 3 C	0.95	0.83
ITU-T P.Supp 23 Exp 3 D	0.94	0.92
ITU-T P.Supp 23 Exp 3 O	0.93	0.91

3.4. Computational Complexity and Memory Requirements

The LCQA algorithm has low computation and storage requirements. It requires a buffer of 12+12 scalar values, calculated from the previous and current block. Table 6 shows the advantage in computational complexity of LCQA over ITU-T P.563. The comparison is between the optimized ANSI-C implementation of ITU-T P.563 and a MATLAB 7 implementation of LCQA, both executed on a Pentium 4 machine at 2.8 GHz with 1 GB RAM. With LCQA-P we denote the case where the input features $\{\mathrm{E}_n^e, \mathrm{T}_n, \mathbf{f}_n\}$ are available from the speech coder. It is seen that LCQA has significantly lower computational complexity despite the usage of an interpretative language.

Table 6: Mean execution time for utterances with 8 s mean length.

	Execution time (in s)		
	ITU-T P.563	LCQA	LCQA-P
Time	4.63	1.24	0.01

4. Conclusions

We demonstrated that a low-cost non-intrusive speech quality assessment algorithm can be a valuable tool for monitoring the performance of a communication system. By means of simulations over a large database we demonstrated that the presented algorithm predicts speech quality more accurately than the standardized ITU-T P.563 and at much lower complexity. Since the algorithm does not use an explicit distortion model, the algorithm facilitates extension to quality assessment of future communication systems.

5. Acknowledgment

We thank Stefan Bruhn of Ericsson AB for MOS databases.

6. References

[1] ITU-T Rec. P.830, “Subjective performance assessment of telephone-band and wideband digital codecs,” 1996.

[2] ITU-T Rec. P.800, “Methods for Subjective Determination of Transmission Quality,” 1996.

[3] O. Au and K. Lam, “A novel output-based objective speech quality measure for wireless communication,” *Signal Process. Proc, 4th Int. Conf.*, vol. 1, pp. 666–669, 1998.

[4] P. Gray, M. Hollier, and R. Massara, “Non-intrusive speech-quality assessment using vocal-tract models,” in *Proc. IEE Vision, Image and Signal Process.*, vol. 147, no. 6, 2000, pp. 493–501.

[5] J. Liang and R. Kubichek, “Output-based objective speech quality,” *IEEE 44th Vehicular Technology Conf.*, vol. 3, no. 8-10, pp. 1719–1723, 1994.

[6] T. Falk, Q. Xu, and W.-Y. Chan, “Non-intrusive GMM-based speech quality measurement,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, 2005, pp. 125–128.

[7] G. Chen and V. Parsa, “Bayesian model based non-intrusive speech quality evaluation,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, 2005, pp. 385–388.

[8] D. Kim, “ANIQUE: An auditory model for single-ended speech quality estimation,” *IEEE Trans. Speech, Audio Process.*, vol. 13, pp. 821–831, 2005.

[9] ITU-T P.563, “Single ended method for objective speech quality assessment in narrow-band telephony applications,” 2004.

[10] M. Werner, T. Junge, and P. Vary, “Quality control for AMR speech channels in GSM networks,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 3, 2004, pp. 1076–1079.

[11] H. Knagenhjelm and W. B. Kleijn, “Spectral dynamics is more important than spectral distortion,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, 1995, pp. 732–735.

[12] T. Quatieri and R. Dunn, “Speech enhancement based on auditory spectral change,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, 2002, pp. 257–260.

[13] J. Beauchamp, “Synthesis by spectral amplitude and brightness matching of analized musical instrument tones,” *J. Audio Eng. Soc*, vol. 30, pp. 396–406, 1982.

[14] S. Voran, “A simplified version of the ITU algorithm for objective measurement of speech codec quality,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, 1998, pp. 537–540.

[15] S. Stearns, “On selecting features for pattern classifiers,” in *Proc. 3th Int. Conf. on Pattern Recognition*, 1976, pp. 71–75.

[16] P. Pudil, F. Ferri, J. Novovicova, and J. Kittler, “Floating search methods for feature selection with nonmonotonic criterion functions,” in *Proc. IEEE Intl. Conf. Pattern Recognition*, 1994, pp. 279–283.

[17] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd ed. Wiley-Interscience, 2001.

[18] ITU-T Rec. P. Supplement 23, “ITU-T coded-speech database,” International Telecommunication Union, 1998.