

Low Complexity, Non-Intrusive Speech Quality Assessment

Volodya Grancharov, *Student Member, IEEE*, David Y. Zhao, *Student Member, IEEE*, Jonas Lindblom, and W. Bastiaan Kleijn, *Fellow, IEEE*

Abstract—Monitoring of speech quality in emerging heterogeneous networks is of great interest to network operators. The most efficient way to satisfy such a need is through non-intrusive, objective speech quality assessment. In this paper we describe an algorithm for monitoring the speech quality over a network with extremely low complexity 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 is not performed. The critical advantage of the approach lies in generating quality assessment ratings without explicit distortion modelling. The results from the performed simulations indicate that the proposed output-based objective quality measure performs better than the ITU-T P.563 standard.

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

I. INTRODUCTION

Speech quality assessment is an important problem in mobile communications. 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 subjective test, speech is played to a group of listeners, who are asked to rate the quality of this speech signal [1], [2].

The most common measure for user opinion is the *mean opinion score* (MOS), obtained by averaging the absolute category ratings (ACR). In ACR, listeners compare the distorted signal with their internal model of high quality speech. In degradation MOS (DMOS) tests, the subjects listen to the original speech first, and then are asked to select the degradation category rating (DCR) corresponding to the distortion of the processed signal, see Table I. DMOS tests are more common in audio quality assessment [3], [4].

Assessment of the *listening quality* [1]–[4] is not the only form of quality of service (QoS) monitoring. In many cases *conversational* subjective tests [2] are the preferred method of subjective evaluation, where participants hold conversations over a number of different networks and vote on their perception of conversational quality. An objective model of conversational quality can be found in [5]. Yet another class

V. Grancharov, D. Y. Zhao, and W. B. Kleijn are with the Sound and Image Processing Lab, Royal Institute of Technology, Stockholm, Sweden (e-mail: volodya.grancharov@ee.kth.se; david.zhao@ee.kth.se; bastiaan.kleijn@ee.kth.se; phone: +46 87908819; fax: +46 87917654).

J. Lindblom was with the Sound and Image Processing Lab, Royal Institute of Technology and is currently with the Skype Technologies (e-mail: jonas.lindblom@skype.net).

This work was funded by Wireless@KTH.

TABLE I
GRADES IN MOS AND DMOS

Grade	ACR(MOS)	DCR(DMOS)
5	Excellent	Inaudible
4	Good	Audible, but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very annoying

of QoS monitoring consists of *intelligibility* tests. The most popular intelligibility tests are the Diagnostic Rhyme Test (DRT) and Modified Rhyme Test (MRT) [6]. In this paper we will not further discuss intelligibility and conversational quality tests, and will focus entirely on ACR listening quality, denoted for simplicity as subjective quality.

Subjective tests are believed to give the “true” speech quality. However, the involvement of human listeners makes them expensive and time consuming. They can be used only in the final stages of developing the speech communication system and are not suitable to measure QoS on a daily basis.

Objective measures use mathematical expressions to predict speech quality. Their low cost means that they can be used to continuously monitor the quality over the 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 simplest class of in-

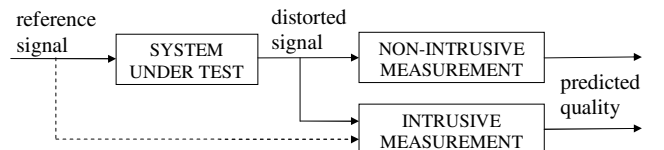


Fig. 1. Intrusive and Non-intrusive type of quality assessment. Non-intrusive algorithms do not have access to the reference signal.

trusive objective quality measures are waveform-comparison algorithms, such as signal-to-noise ratio (SNR) and segmental signal-to-noise ratio (SSNR). The waveform-comparison algorithms are simple to implement and require low computational complexity, but they do not correlate well with subjective measurements if different types of distortions are compared.

Frequency-domain techniques, such as the Itakura - Saito (IS) measure, and the spectral distortion (SD) measure are widely used. Frequency-domain techniques are not sensitive

to a time shift and are generally more consistent with human perception [7]. The distinguishing characteristic of both waveform comparison and frequency domain techniques is that they are equipped with a very simple error pooling schemes and that they do not contain mappings that are trained by databases. With error pooling we denote the final stage of all quality metrics that has to combine the estimated per-frame distortions in a single value.

A significant number of intrusive perceptual-domain measures has been developed. These measures incorporate knowledge of the human perceptual system. Mimicry of human perception is used for dimension reduction and a "cognitive" stage is used to perform the mapping to a quality scale. The cognitive stage is trained by means of a data base. These include the Bark Spectral Distortion (BSD) [8], Perceptual Speech Quality (PSQM) [9], and Measuring Normalizing Blocks (MNB) [10], [11]. Perceptual evaluation of speech quality (PESQ) [12] and perceptual evaluation of audio quality (PEAQ) [13] are standardized state-of-the-art algorithms for intrusive quality assessment of speech, and audio respectively.

Existing intrusive objective speech quality measures may automatically assess the performance of the communication system without the need for human listeners. However, intrusive measures require the presence of the original signal, which is typically not available in QoS monitoring. For such applications *non-intrusive* quality assessment must be used. These methods often include both mimicry of human perception and/or a mapping to the quality measure that is trained using a data base.

An early attempt towards non-intrusive speech quality measure based on spectrogram of the perceived signal is presented in [14]. The spectrogram is partitioned, and variance and dynamic range calculated on a block-by-block basis. The average level of variance and dynamic range is used to predict speech quality.

The non-intrusive speech quality assessment reported in [15] attempts to predict the likelihood that the passing audio stream is generated by the human vocal production system. The speech stream under assessment is reduced to a set of features. The parameterized data is used to generate physiologically based rules for error assessment.

The measure proposed in [16] is based on comparing the output speech to an artificial reference signal that is appropriately selected from a optimally clustered codebook. The Perceptual Linear Prediction (PLP) [17] coefficients are used for parametric representation of the speech. A fifth-order all-pole model is performed to suppress speaker-dependent details of the auditory spectrum. The average distance between the unknown test vector to its nearest reference centroids provides an indication of speech degradation.

Recent algorithms based on Gaussian-mixture probability models (GMM) of features derived from perceptually motivated spectral-envelope representations can be found in [18] and [19]. A novel, perceptually motivated speech quality assessment algorithm based on temporal envelope representation of speech is presented in [20].

The International Telecommunication Union (ITU) standard for non-intrusive quality assessment ITU-T P.563 can be found

in [21]. 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 measures for quality assessment are designed to predict the effects of many types of distortions, and typically have high computational complexity. These type of algorithms will be referred to as the *general* speech quality predictors. It has been shown that non-intrusive quality prediction is possible at much lower complexity if it is assumed that the type of distortion is known [22] [23]. However, this class of measures is likely to suffer from poor prediction performance if the expected working conditions are not met.

We conclude that existing algorithms either have a high complexity and a broad range of application or a low complexity and a narrow range of application. This has motivated us to develop a speech-quality assessment algorithm with low computational complexity. The algorithm predicts speech quality from generic features commonly used in speech coding, without assumption of the type of distortion. In the proposed low-complexity, non-intrusive speech quality assessment (LCQA) algorithm an explicit distortion model is not used, but instead the quality estimate is based on global statistical properties of per-frame features. In the next section we provide the motivations for the critical choices made in the development of the LCQA algorithm, followed by a detailed algorithm description in section III. The performance of the proposed algorithm is compared with ITU-T P.563 in section IV.

II. KEY ISSUES IN OBJECTIVE QUALITY ASSESSMENT

In this section we discuss some unresolved questions in speech quality assessment. We give the reasoning for the conceptual choices behind the particular LCQA implementation, and outline the distinguished features of the algorithm.

A critical issue in the design of an automatic system for QoS monitoring is the scale of the quality mapping; continuous or discrete. In practice subjective MOS scores do not have continuous character, due to the limited number of listeners' opinions used in averaging. This behavior is demonstrated in Fig. 2. One approach proposed in [19] is to see quality prediction not as a regression problem, but rather as the classification of quality ratings in intervals of predefined range. The major disadvantage of the classification definition of the problem is that the resolution has to be set in advance, which may not be appropriate if the algorithm is to be used in different applications. Therefore, the proposed LCQA algorithm predicts the speech quality on a continuous scale, and this choice is supported by the simulations presented in section IV.

The human speech quality assessment process can be divided into two parts: 1) conversion of the received speech signal into auditory nerve excitations for the brain, and 2) cognitive processing in the brain, see Fig. 3. The key principles of perceptual transform are signal masking, critical band spectral resolution, equal-loudness curve, and intensity loudness

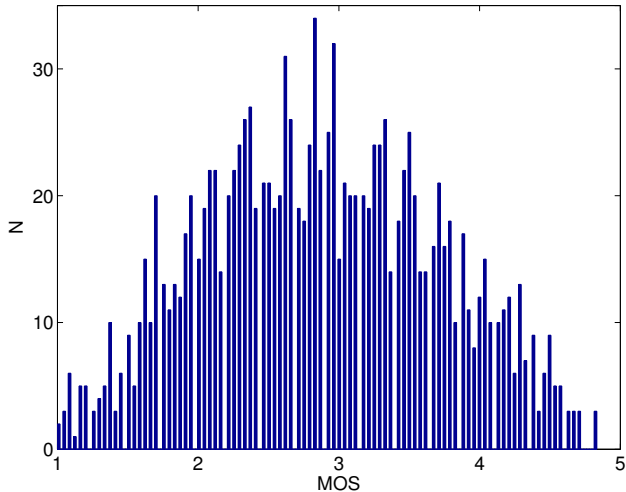


Fig. 2. Distribution of the subjective MOS scores over a database of 1000 utterances. The 200 bins histogram shows the discrete character of subjective scores.

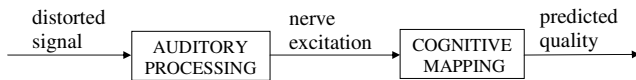


Fig. 3. Human perception of speech quality involves both hearing and judgement.

law, e.g., [24]. These principles are well studied and in most existing quality assessment algorithms a perceptual transform is a pre-processing step. The main implicit purpose of the perceptual transform is to perform dimension reduction on the speech signal. An advantage of this mimicry-motivated approach is that it reduces the need for a sophisticated feature selection mechanism based on a database. However, this comes at the cost of a high computational expense of the perceptual transform. More-over, mimicry may result in the removal of relevant information. Therefore, the proposed LCQA algorithm does not perform a perceptual transform, but instead the dimensionality is reduced jointly with optimizing the mapping function coefficients. This guarantees minimum loss of relevant information. Our approach is consistent with the recent emergence of algorithms performing quality assessment without a perceptual transform in image quality assessment [25].

Many of the existing quality assessment algorithms are based on specific models of distortion, i.e., level of background noise, multiplicative noise, presence of ringing tones [21], or simulate a known distortion like handset receiver characteristics [12]. The LCQA algorithm does not incorporate an explicit model of the distortion. The speech quality estimate is based entirely on the statistics of a processed speech signal, and the distortion is implicitly assessed by its impact on these statistics. As a result, the LCQA algorithm is easily adapted to the next generation communication systems that will likely produce new types of distortions.

In some methods the speaker-dependent information is re-

moved [18], [16]. However, it is known that telephony systems provide higher quality score for some voices over the others [26]. The algorithm presented in this paper incorporates the speaker-dependent information in the form of the pitch period and the coefficients of a tenth-order autoregressive (AR) model estimated by means of linear prediction.

III. LOW-COMPLEXITY QUALITY ASSESSMENT

The objective of the proposed LCQA algorithm is to provide an estimate for the MOS score of each utterance, using a simple set of features that is readily available from speech codecs in the network. Thus, the speech quality is predicted at low computational complexity, which makes the method useful for practical applications.

The dotted area in Fig. 4 shows extraction of the per-frame feature vector from the compact speech representation, used in Code-Excited Linear Prediction (CELP) coders [27]. Each 20

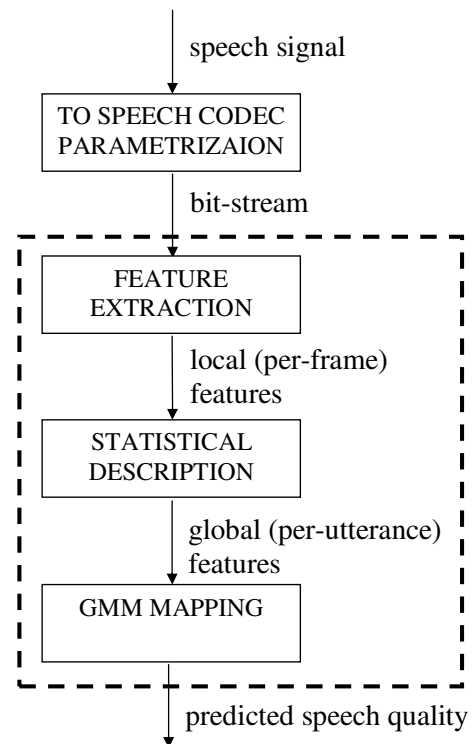


Fig. 4. The structure of the LCQA algorithm. Doted area represents the LCQA mode optimal for the CELP coders. In any other environment the LCQA can extract the required features from the waveform.

ms speech frame is represented by the variance of the excitation of the AR model, the pitch period, and a ten-dimensional vector of line-spectral frequencies (LSF) coefficients [28], $\{E_n^e, T_n, \mathbf{f}_n\}$, where n is the frame index. We hypothesize that such a compact speech representation, successfully used in speech coding, is likely to give us meaningful features. The LCQA algorithm predicts the speech quality from the global speech statistics. The statistical properties of per-frame features, captured by per-utterance features, form the input for GMM mapping, which estimates the speech quality level on a MOS scale.

A. Speech Features

The basis of any type of automatic quality analysis system is the extraction of a feature vector. The set of features used in LCQA aim to capture the structural information from a speech signal. This is motivated by the fact that the natural speech signal is highly structured, and it is likely that human quality judgement relies on patterns extracted from information describing this structure. In this section we list the features that we have selected.

The spectral flatness measure [29] is related to the amount of resonant structure in the power spectrum:

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

where the AR envelope $P(\omega)$ is defined as the frequency response of the AR model with coefficients a_k

$$P(\omega) = \frac{1}{|1 + \sum_{k=1}^p a_k e^{-j\omega k}|^2}. \quad (2)$$

As a second feature we use spectral dynamics, defined as

$$\Phi_2(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} (\log P_n(\omega) - \log P_{n-1}(\omega))^2 d\omega. \quad (3)$$

The spectral dynamics have been studied and successfully used in speech coding [30], and speech enhancement [31].

The spectral centroid [32] determines the frequency area around which most of the signal energy concentrates

$$\Phi_3(n) = \frac{\int_{-\pi}^{\pi} \omega \log(P_n(\omega)) d\omega}{\int_{-\pi}^{\pi} \log(P_n(\omega)) d\omega}, \quad (4)$$

and it is also frequently used as an approximation for a measure of perceptual "brightness". The last three 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 will be denoted as $\Phi_4(n)$, $\Phi_5(n)$, and $\Phi_6(n)$, respectively. The presented above features and their first time derivatives (except the derivative of the spectral dynamics) are grouped in a 11 dimensional per-frame feature vector $\Phi(n)$.

We hypothesize that the speech quality can be estimated from statistical properties of these per-frame features, and describe their probability distribution with the mean, variance, kurtosis, and skewness. The moments are calculated independently for each feature, and this gives a set of features that globally describe one speech utterance:

$$\mu_{\Phi_i} = \frac{1}{|\tilde{\Omega}|} \sum_{n \in \tilde{\Omega}} \Phi_i(n) \quad (5)$$

$$\sigma_{\Phi_i} = \frac{1}{|\tilde{\Omega}|} \sum_{n \in \tilde{\Omega}} (\Phi_i(n) - \mu_{\Phi_i})^2 \quad (6)$$

$$s_{\Phi_i} = \frac{1}{|\tilde{\Omega}|} \frac{\sum_{n \in \tilde{\Omega}} (\Phi_i(n) - \mu_{\Phi_i})^3}{\sigma_{\Phi_i}^{3/2}} \quad (7)$$

$$k_{\Phi_i} = \frac{1}{|\tilde{\Omega}|} \frac{\sum_{n \in \tilde{\Omega}} (\Phi_i(n) - \mu_{\Phi_i})^4}{\sigma_{\Phi_i}^2}. \quad (8)$$

With $\tilde{\Omega}$ we denote the frames set, of cardinality $|\tilde{\Omega}|$, used to calculate statistics for each of the per-frame features $\Phi_i(n)$.

The global features are grouped in one feature vector $\Psi = \{\mu_{\Phi_i}, \sigma_{\Phi_i}, s_{\Phi_i}, k_{\Phi_i}\}_{i=1}^{11}$. In the next subsection we describe a two-step dimensionality reduction procedure that 1) extracts the "best" subset of frames $\tilde{\Omega}$ out of all available frames Ω , 2) and transforms feature vector Ψ into feature vector $\tilde{\Psi}$ of low dimensionality.

B. Dimensionality reduction

The feature selection algorithm is important to the practical performance of quality assessment systems. The main purpose of the feature selection algorithm is to improve predictive accuracy of the quality assessment system by removing irrelevant and redundant data. A commonly used approach, in the quality assessment literature, is to remove non-speech regions based on a voice activity detector or an energy threshold [33]. It is interesting to note that the removal of low energy regions can be seen as a feature selection problem. We propose a generalization of this concept by considering activity thresholds in all feature dimensions. The scheme, presented in Table II allows speech active frames to be excluded if they do not carry information that improves the accuracy of speech quality prediction.

TABLE II
THE OPTIMAL SET OF FRAMES AS A FUNCTION OF A THRESHOLD VECTOR Θ

<pre> 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 $\tilde{\Omega} = \tilde{\Omega} + \{n\}$ </pre>

From Table II we can see that the optimal set of frames is determined by the threshold $\Theta = \{\Theta_i^L, \Theta_i^U\}_{i=1}^{11}$, i.e., $\tilde{\Omega} = \tilde{\Omega}(\Theta)$. We search for the threshold Θ that minimizes the criterion ε :

$$\Theta = \arg \min_{\Theta^*} \varepsilon(\tilde{\Omega}(\Theta^*)). \quad (9)$$

The criterion ε is related to the root-mean-square error (RMSE) performance of the LCQA algorithm, and is properly defined in section IV. The choice of optimization criterion is motivated by the fact that no other objective measures than the performance of the regression function can determine the set of optimal features.

Once the optimal subset of frames $\tilde{\Omega}$ is found, we search for the optimal subset of features $\tilde{\Psi}$. This optimization step is defined as follows: given the original set of features Ψ of cardinality $|\Psi|$, and the optimal set of frames, $\tilde{\Omega}$, select a subset of features $\tilde{\Psi} \subset \Psi$ of cardinality $|\tilde{\Psi}| < |\Psi|$ that is optimized for the performance of the LCQA algorithm:

$$\tilde{\Psi} = \arg \min_{\tilde{\Psi}^* \subset \Psi} \varepsilon(\tilde{\Psi}^*). \quad (10)$$

A full search is the only dimensionality reduction procedure that guaranties that a global optimum is found. It is rarely applied due to its high computational requirements.

The well-known Sequential Forward Selection and Sequential Backward Selection, e.g., [34] are step-optimal only, since the best (worst) feature is added (discarded), but the decision cannot be corrected at a later stage. The more advanced (L,R) algorithm [35] consists of applying Sequential Forward Selection L times, followed by R steps of Sequential Backward Selection. The Floating Search methods [36] are extensions of the (L,R) search methods, where the number of forward and backward steps is not pre-defined, but dynamically obtained. In our simulations we use 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, see Table III.

TABLE III

THE SEQUENTIAL FLOATING BACKWARD SELECTION PROCEDURE 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

initialize: $\tilde{\Psi} = \Psi$
while error does not increase by more than a threshold
 Exclusion Step:
 Find the least significant feature

$$\Psi_{i-} = \arg \min_{\Psi_i \in \tilde{\Psi}} \varepsilon(\tilde{\Psi} - \{\Psi_i\})$$

 Exclude the feature

$$\tilde{\Psi} = \tilde{\Psi} - \{\Psi_{i-}\}$$

while error decreases by more than a threshold
 Inclusion Steps:
 Find the most significant feature

$$\Psi_{i+} = \arg \min_{\Psi_i \notin \tilde{\Psi}} \varepsilon(\tilde{\Psi} + \{\Psi_i\})$$

 Include the feature

$$\tilde{\Psi} = \tilde{\Psi} + \{\Psi_{i+}\}$$

The presented two-stage dimensionality reduction procedure is sub-optimal, i.e., we do not optimize jointly for the optimal sets of $\tilde{\Omega}$ and $\tilde{\Psi}$. The main reason for that is the high computational complexity. However, the simulations presented in section IV show that the proposed training scheme is sufficient to outperform the reference quality assessment methods.

C. Quality Estimation Given the Global Feature Set

Let Q denote the subjective quality of an utterance as obtained from MOS labelled training database. We construct an objective estimator \hat{Q} of the subjective quality as a function of a feature vector, i.e., $\hat{Q} = \hat{Q}(\tilde{\Psi})$, and search for the function closest to the subjective quality with respect to the criterion

$$\hat{Q}(\tilde{\Psi}) = \arg \min_{Q^*(\tilde{\Psi})} E\{(Q - Q^*(\tilde{\Psi}))^2\}, \quad (11)$$

where $E\{\}$ is the expectation operator. The above defined criterion is the probabilistic measure corresponding to (10).

It is well known, e.g. [37] that equation (11) is minimized by the conditional expectation

$$\hat{Q}(\tilde{\Psi}) = E\{Q|\tilde{\Psi}\}, \quad (12)$$

and the problem reduces to the estimation of the conditional probability. To facilitate this estimation, we model the joint density of the feature variables with the subjective MOS scores as a Gaussian mixture

$$f(\varphi|\lambda) = \sum_{m=1}^M \omega^{(m)} \mathcal{N}(\varphi|\mu^{(m)}, \Sigma^{(m)}), \quad (13)$$

where $\varphi = [Q, \tilde{\Psi}]$, $\omega^{(m)}$ are the mixture weights, and $\mathcal{N}(\varphi|\mu^{(m)}, \Sigma^{(m)})$ are multivariate Gaussian densities.

The Gaussian mixture is completely specified by the mean vector, covariance matrix, and mixture weight

$$\lambda = \{\omega^{(m)}, \mu^{(m)}, \Sigma^{(m)}\}_{m=1}^M, \quad (14)$$

and these coefficients are estimated off-line from a large training set using the EM algorithm [38].

Finally, we express the optimal quality estimator (12) in a form of a weighted sum of known quantities:

$$E\{Q|\tilde{\Psi}\} = \sum_{m=1}^M u^{(m)}(\tilde{\Psi}) \mu_{Q|\tilde{\Psi}}^{(m)} \quad (15)$$

where

$$u^{(m)}(\tilde{\Psi}) = \frac{\omega^{(m)} \mathcal{N}(\tilde{\Psi}|\mu_{\tilde{\Psi}}^{(m)}, \Sigma_{\tilde{\Psi}}^{(m)})}{\sum_{k=1}^M \omega^{(k)} \mathcal{N}(\tilde{\Psi}|\mu_{\tilde{\Psi}}^{(k)}, \Sigma_{\tilde{\Psi}}^{(k)})}, \quad (16)$$

and

$$\mu_{Q|\tilde{\Psi}}^{(m)} = \mu_Q^{(m)} + \Sigma_{\tilde{\Psi}Q}^{(m)} (\Sigma_{\tilde{\Psi}}^{(m)})^{-1} (\tilde{\Psi} - \mu_{\tilde{\Psi}}^{(m)}). \quad (17)$$

D. Implementation Details

In this section we describe how the n-th frame features are calculated, based entirely on $\{E_n^e, T_n, \mathbf{f}_n\}$ and $\{E_{n-1}^e, T_{n-1}, \mathbf{f}_{n-1}\}$. Then we show how the global statistical properties are calculated recursively, without storing the local features in a buffer. We calculate the pitch period T_n according to [39], and the AR coefficients are extracted from the speech signal every 20 ms without overlap.

To keep the complexity of the LCQA algorithm low, we redefine the per-frame features: spectral flatness, spectral dynamics, and spectral centroid. The new definitions are based entirely on the speech codec bit-stream, and signal reconstruction is avoided.

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

$$\Phi_1(n) = \frac{E_n^e}{E_n^s}. \quad (18)$$

Given the variance of the excitation of the AR model, its definition

$$e_k = s_k - \sum_{i=1}^{10} a_i s_{k-i}, \quad (19)$$

and AR coefficients a_i , we calculate the signal variance without reconstructing the waveform s_k using the reverse Levinson-Durbin recursion (step-down algorithm).

The spectral dynamics are redefined as a weighted Euclidean distance in the LSF space:

$$\Phi_2(n) = (\mathbf{f}_n - \mathbf{f}_{n-1})^T \mathbf{W}_n (\mathbf{f}_n - \mathbf{f}_{n-1}), \quad (20)$$

where the inverse harmonic mean weight is defined by the components of the LSF vector:

$$W_n^{(ii)} = (f_n^{(i)} - f_n^{(i-1)})^{-1} + (f_n^{(i+1)} - f_n^{(i)})^{-1} \quad (21)$$

$$W_n^{(ij)} = 0 \quad (22)$$

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

$$\Phi_3(n) = \frac{\sum_{i=1}^{10} i W_n^{(ii)}}{\sum_{i=1}^{10} W_n^{(ii)}}. \quad (23)$$

We calculate the selected global descriptors recursively, i.e., the per-frame features are not stored in the buffer. Until the end of the utterance the mean is recursively updated

$$\mu_\Phi(n) = \frac{n-1}{n} \mu_\Phi(n-1) + \frac{1}{n} \Phi(n) \quad (24)$$

to obtain the desired μ_Φ . Here n is the index over the accepted frames set $\tilde{\Omega}$, as discussed earlier in this section. In a similar fashion, we propagate Φ^2 , Φ^3 , and Φ^4 to obtain the central moments μ_{Φ^2} , μ_{Φ^3} , and μ_{Φ^4} . These quantities are used to obtain the remaining global descriptors, namely variance, skew, and kurtosis:

$$\sigma_\Phi = \mu_{\Phi^2} - (\mu_\Phi)^2 \quad (25)$$

$$s_\Phi = \frac{\mu_{\Phi^3} - 3\mu_\Phi \mu_{\Phi^2} + 2(\mu_\Phi)^3}{\sigma_\Phi^{3/2}} \quad (26)$$

$$k_\Phi = \frac{\mu_{\Phi^4} - 4\mu_\Phi \mu_{\Phi^3} + 6(\mu_\Phi)^2 \mu_{\Phi^2} - 3(\mu_\Phi)^4}{\sigma_\Phi^2}. \quad (27)$$

Table IV gives a short overview of the proposed LCQA algorithm.

TABLE IV
OVERVIEW OF LCQA ALGORITHM

- 1) For the n -th speech frame calculate $\{\mathbf{E}_n^e, \mathbf{T}_n, \mathbf{f}_n\}$ from the waveform or extract from the bit-stream.
- 2) Calculate per-frame feature vector $\Phi(n)$, based on $\{\mathbf{E}_n^e, \mathbf{T}_n, \mathbf{f}_n\}$ and stored in a buffer $\{\mathbf{E}_{n-1}^e, \mathbf{T}_{n-1}, \mathbf{f}_{n-1}\}$.
- 3) From a selected subset of frames $\tilde{\Omega}$ recursively calculate the central moments $\{\mu_\Phi, \mu_{\Phi^2}, \mu_{\Phi^3}, \mu_{\Phi^4}\}$. Frames selection is controlled by the threshold Θ .
- 4) At the end of the utterance calculate global feature vector $\tilde{\Psi} = \{\mu_{\Phi_i}, \sigma_{\Phi_i}, s_{\Phi_i}, k_{\Phi_i}\}$ as mean, variance, skew, and kurtosis of local features.
- 5) Predict the speech quality as a function of the global feature vector $\hat{Q} = \hat{Q}(\tilde{\Psi})$, through GMM mapping.

IV. SIMULATIONS

In this section we discuss type of training and MOS labelled databases, used with the LCQA algorithm. We also present results from simulations, with respect to both prediction accuracy and computational complexity of the proposed algorithm.

A. Training

For the training procedure we used 11 MOS labelled databases provided by Ericsson AB and one ITU database [40]. The combined database contains utterances in the following languages: English, French, Japanese, Italian and Swedish. The database contains large variety of distortions, such as: different coding, tandeming, and modulated noise reference unit (MNRU) [41] conditions, as well as packet loss, background noise, effects of noise suppression, switching effects, different input levels, etc. The total size of the database is 7646 speech files.

In the training we use 10-fold *cross validation* with 20% of the speech material, to provide robustness in the performance evaluation [42]. To further improve generalization performance we perform *training with noise*. We create virtual training patterns, by adding zero mean white Gaussian noise to true training patterns.

B. Performance Evaluation

In this section we compare the performance of the proposed LCQA algorithm with the standardized ITU-T P.563. The estimation performance is assessed using correlation coefficient R and RMSE ε , between the predicted quality \hat{Q} and subjective quality Q . The RMSE is given by

$$\varepsilon = \sqrt{\frac{\sum_{i=1}^N (Q_i - \hat{Q}_i)^2}{N}}, \quad (28)$$

and the correlation coefficient is defined as

$$R = \frac{\sum_{i=1}^N (\hat{Q}_i - \mu_{\hat{Q}})(Q_i - \mu_Q)}{\sqrt{\sum_{i=1}^N (\hat{Q}_i - \mu_{\hat{Q}})^2} \sqrt{\sum_{i=1}^N (Q_i - \mu_Q)^2}}, \quad (29)$$

where μ_Q and $\mu_{\hat{Q}}$ are the mean values of the introduced variables. Here N is the number of MOS labelled utterances used in evaluation. Table V contains the averaged results of the cross-validation tests, and Table VI contains the RMSE outliers in %. The test results clearly indicate that the proposed LCQA algorithm outperforms the standardized ITU-T P.563. In Fig. 5 we demonstrate the correlation between subjective speech quality ratings and LCQA predicted, values over a database.

TABLE V
AVERAGED PERFORMANCE IN CORRELATION AND RMSE

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

Processing time and memory requirements are important figures of merit for the quality estimation algorithms. The

TABLE VI

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

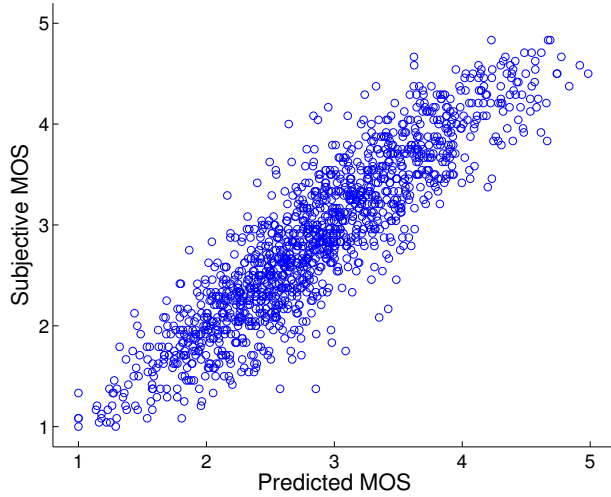


Fig. 5. Correlation between subjective and predicted MOS values for ITU-T P.23 database.

LCQA algorithm has insignificant memory requirements: a buffer of 12 scalar values, calculated from the previous and current frame is needed (future frames are not required), as well as memory for the 12 Gaussian mixtures.

In table VII we demonstrate the difference in computational complexity between the proposed LCQA and the ITU-T P.563. The comparison is between the optimized ANSI-C implementation of ITU-T P.563 and the 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 input features $\{E_n^e, T_n, \mathbf{f}_n\}$ are readily available from codecs used in the network.

TABLE VII

EXECUTION TIME (IN S) FOR UTTERANCES OF AVERAGED LENGTH 8 S

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

V. CONCLUSIONS

We demonstrated that a low-cost non-intrusive speech quality assessment algorithm can be a valuable tool for monitoring the performance of a speech communication system. The proposed quality assessment algorithm operates on a heavily restricted parametric representation of speech, without the need for a perceptual transform of the input signal. 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, at much lower complexity.

In the proposed algorithm the distortion is modeled only implicitly by its effect on the distribution of the selected speech features. Since there is no explicit distortion model, the algorithm is easily extendable towards quality assessment of future communication systems.

ACKNOWLEDGMENT

The authors would like to thank Stefan Bruhn of Ericsson AB for providing MOS databases.

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] ITU-R Rec. BS.1534, "Method for the subjective assessment of intermediate quality level of coding systems," 2001.
- [4] ITU-R Rec. BS.562-3, "Subjective assessment of sound quality," 1990.
- [5] ITU-T Rec. G.107, "The e-model, a computational model for use in transmission planning," 2003.
- [6] M. Goldstein, "Classification of methods used for assessment of text-to-speech systems according to the demands placed on the listener," *Speech Communication*, vol. 16, pp. 225–244, 1995.
- [7] S. Quackenbush, T. Barnwell, and M. Clements, *Objective Measures of Speech Quality*. Prentice Hall, 1988.
- [8] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE J. Selected Areas in Commun.*, vol. 10, no. 5, pp. 819–829, 1992.
- [9] J. Beerends and J. Stemerdink, "A perceptual speech-quality measure based on a psychoacoustic sound representation," *J. Audio Eng. Soc.*, vol. 42, no. 3, pp. 115–123, 1994.
- [10] S. Voran, "Objective estimation of perceived speech quality - Part I: Development of the measuring normalizing block technique," *IEEE Trans. Speech, Audio Processing*, vol. 7, no. 4, pp. 371–382, 1999.
- [11] S. Voran, "Objective estimation of perceived speech quality - Part II: Evaluation of the measuring normalizing block technique," *IEEE Trans. Speech, Audio Processing*, vol. 7, no. 4, pp. 383–390, 1999.
- [12] ITU-T Rec. P. 862, "Perceptual evaluation of speech quality (PESQ)," 2001.
- [13] ITU-R. BS.1387, "Method for Objective Measurements of Perceived Audio Quality (PEAQ)," 1998.
- [14] O. Au and K. Lam, "A novel output-based objective speech quality measure for wireless communication," *Signal Processing Proceedings, 4th Int. Conf.*, vol. 1, pp. 666–669, 1998.
- [15] P. Gray, M. Hollier, and R. Massara, "Non-intrusive speech-quality assessment using vocal-tract models," in *Proc. IEEE Vision, Image and Signal Processing*, vol. 147, pp. 493–501, 2000.
- [16] J. Liang and R. Kubichek, "Output-based objective speech quality," *IEEE 44th Vehicular Technology Conf.*, vol. 3, no. 8–10, pp. 1719–1723, 1994.
- [17] H. Hermansky, "Perceptual linear prediction (PLP) analysis of speech," *J. Acous. Soc. Amer.*, vol. 87, pp. 1738–1752, 1990.
- [18] T. Falk, Q. Xu, and W.-Y. Chan, "Non-intrusive GMM-based speech quality measurement," in *Proc. IEEE Int. Conf. Acous., Speech, Signal Processing*, vol. 1, pp. 125–128, 2005.
- [19] G. Chen and V. Parsa, "Bayesian model based non-intrusive speech quality evaluation," in *Proc. IEEE Int. Conf. Acous., Speech, Signal Processing*, vol. 1, pp. 385–388, 2005.
- [20] D. Kim, "ANIQUE: An auditory model for single-ended speech quality estimation," *IEEE Trans. Speech, Audio Processing*, vol. 13, pp. 821–831, 2005.
- [21] ITU-T P. 563, "Single ended method for objective speech quality assessment in narrow-band telephony applications," 2004.
- [22] M. Werner, T. Junge, and P. Vary, "Quality control for AMR speech channels in GSM networks," in *Proc. IEEE Int. Conf. Acous., Speech, Signal Processing*, vol. 3, pp. 1076–1079, 2004.
- [23] A. Conway, "Output-based method of applying PESQ to measure the perceptual quality of framed speech signals," in *Proc. IEEE Wireless Communications and Networking*, vol. 4, pp. 2521–2526, 2004.
- [24] B. C. J. Moore, *An Introduction to the Psychology of Hearing*. London: Academic Press, 1989.

- [25] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, pp. 600–612, 2004.
- [26] R. Reynolds and A. Rix, "Quality VoIP - an engineering challenge," *BT Technology Journal*, vol. 19, pp. 23–32, 2001.
- [27] M. Schroeder and B. Atal, "Code-excited linear prediction (CELP): high quality speech at very low bit rates," in *Proc. IEEE Int. Conf. Acous., Speech, Signal Processing*, pp. 937–940, 1985.
- [28] F. Itakura, "Line spectrum representation of linear predictor coefficients of speech signals," *J. Acous. Soc. Amer.*, vol. 57, S35(A), 1975.
- [29] S. Jayant and P. Noll, *Digital Coding of Waveforms*. Englewood Cliffs NJ: Prentice-Hall, 1984.
- [30] H. Knagenhjelm and W. B. Kleijn, "Spectral dynamics is more important than spectral distortion," in *Proc. IEEE Int. Conf. Acous., Speech, Signal Processing*, vol. 1, pp. 732–735, 1995.
- [31] T. Quatieri and R. Dunn, "Speech enhancement based on auditory spectral change," in *Proc. IEEE Int. Conf. Acous., Speech, Signal Processing*, vol. 1, pp. 257–260, 2002.
- [32] J. Beauchamp, "Synthesis by spectral amplitude and brightness matching of analyzed musical instrument tones," *J. Audio Eng. Soc.*, vol. 30, pp. 396–406, 1982.
- [33] S. Voran, "A simplified version of the ITU algorithm for objective measurement of speech codec quality," in *Proc. IEEE Int. Conf. Acous., Speech, Signal Processing*, vol. 1, pp. 537–540, 1998.
- [34] P. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach*. London, UK: Prentice Hall, 1982.
- [35] S. Stearns, "On selecting features for pattern classifiers," in *Proc. 3th Int. Conf. on Pattern Recognition*, pp. 71–75, 1976.
- [36] 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*, pp. 279–283, 1994.
- [37] T. Soderstrom, *Discrete-time Stochastic Systems*. London: Springer-Verlag, second ed., 2002.
- [38] A. Dempster, N. Lair, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal Royal Statistical Society.*, vol. 39, pp. 1–38, 1977.
- [39] W. B. Kleijn, P. Kroon, L. Cellario, and D. Sereno, "A 5.85 kbps CELP algorithm for cellular applications," in *Proc. IEEE Int. Conf. Acous., Speech, Signal Processing*, vol. 2, pp. 596–599, 1993.
- [40] ITU-T Rec. P. Supplement 23, "ITU-T coded-speech database," 1998.
- [41] ITU-T. Rec. P.810, "Modulated Noise Reference Unit," 1996.
- [42] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. Wiley-Interscience, second ed., 2001.