# Low Complexity, Non-Intrusive Speech Quality Assessment

Volodya Grancharov, *Student Member, IEEE,* David Y. Zhao, *Student Member, IEEE,* Jonas Lindblom, *Member, IEEE,* 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 a low complexity algorithm for monitoring the speech quality over a network. 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 modeling. The results from the performed experiments indicate that the proposed non-intrusive 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 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.

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 |

Subjective tests are believed to give the ”true” speech quality. However, the involvement of human listeners makes them expensive and time consuming. Such tests 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 intrusive 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.

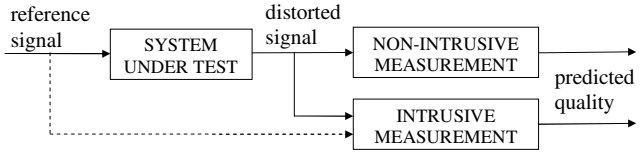


Fig. 1. Intrusive and Non-intrusive type of quality assessment. Non-intrusive algorithms do not have access to the reference signal.

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 hu-

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.

man perception [7]. The distinguishing characteristic of both waveform comparison and frequency domain techniques is that they are equipped with a simple schemes for combining the estimated per-frame distortions. In addition, they do not contain mappings that are trained by databases.

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 one or more databases. These measures 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 databases.

An early attempt towards non-intrusive speech quality measure based on a 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 estimate the perceived quality by means of physiologically based rules.

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 as a parametric representation of the speech signal. 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 and the 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] and [21].

The International Telecommunication Union (ITU) standard for non-intrusive quality assessment, ITU-T P.563, can be found in [22]. 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 a so-called 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. Such algorithms will be referred to as *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, e.g., [23]. However, the latter 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 (referred to as per-frame features), without an 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.

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. 2. The key principles of perceptual transform are signal masking, critical band spectral resolution, equal-loudness curves, and intensity loudness 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 perceptually-consistent dimension reduction on the speech signal. Ideally, a perceptual transformation retains all perceptually-relevant information, and discards all perceptually-irrelevant information. In practice, approximations and simplifications must be made and this goal may not be met. In some cases, perceptual

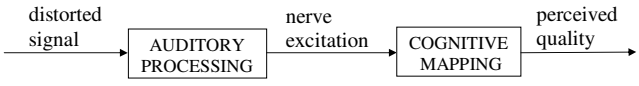


Fig. 2. Human perception of speech quality involves both hearing and judgment.

transformations may have high computational cost. To avoid these potential limitations, the proposed LCQA algorithm does not perform a perceptual transform, but instead the dimensionality is reduced simultaneously with the optimization of the mapping function coefficients. The goal is to minimize the 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 [22], 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. While this direct approach is very efficient, the internal workings of the LCQA algorithm do not lend themselves to interpretation as readily as an algorithm that uses explicit perceptual transformations and distortion modeling.

In some methods the speaker-dependent information is removed [18], [16]. However, it is known that telephony systems provide higher quality scores for some voices than for other voices [26]. Therefore, if the algorithm is to be used for continuous network monitoring, and balanced speech material for averaging cannot be guaranteed, the speaker-dependent information is relevant. 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. In this paper with utterance we denote a pair of short sentences separated by a pause of 0.5 seconds. The total length of an utterance is approximately 8 s.

The LCQA algorithm predicts speech quality 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 core of the LCQA algorithm is the 11-dimensional per-frame feature vector $\mathbf{\Phi}$, with the components defined in section III-A. The speech quality is not predicted directly from the per-frame vector, but from its global statistical properties, described as mean, variance, skew, and kurtosis of the per-frame features. The statistical properties of the per-frame features (referred to as global features $\Psi$) form the input for GMM mapping, which estimates the speech quality level on a MOS scale.

An interesting property of the LCQA algorithm is that the per-frame vector can be derived from the variance of the excitation of the AR model, the pitch period, and the ten-dimensional vector of line-spectral frequency (LSF) coefficients, $\{\mathrm{E}^e, \mathrm{T}, \mathbf{f}\}$, calculated over 20 ms speech frames. Since $\mathrm{E}^e$, T, and $\mathbf{f}$ are readily accessible in the network in the case of Code-Excited Linear Prediction (CELP) coders [27], the LCQA algorithm has additional flexibility not to extract the per-frame vector from the signal waveform, but from the network parameters.

The general scheme of the LCQA algorithm is shown in Fig. 3. The module "Speech Encoder Parameterization" stands for calculation of the $\mathrm{E}^e$, T, and $\mathbf{f}$. In the case of CELP coders, this operation is not performed, and the per-frame vector is directly calculated from the bitstream parameters. In the experiments described in this paper, however, the per-frame vector was always calculated from the speech waveform using the single technique described in section III-D. Note that the speech quality is predicted once per utterance, from the global features.

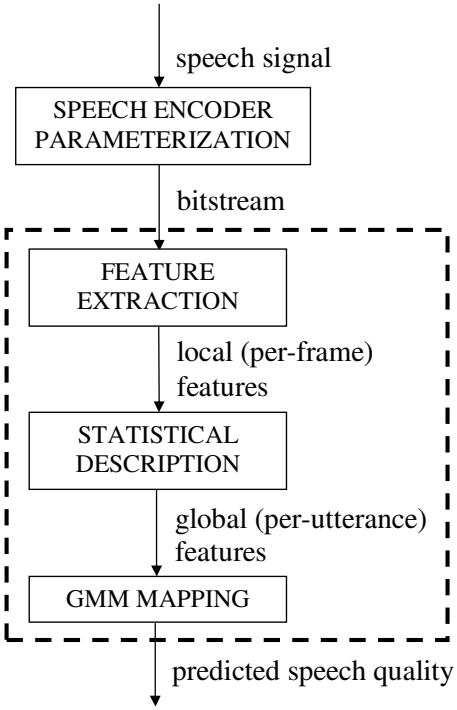


Fig. 3. The structure of the LCQA algorithm. Dashed area represents the LCQA mode optimal for the CELP coders. In any other environment the LCQA can extract the required per-frame features from the waveform.

### 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 aims 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 judgment relies on patterns extracted from information describing this structure. In this section we list the per-frame features that we have selected.

The spectral flatness measure [28] is related to the strength of the resonant structure in the power spectrum:

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

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

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

The frame index is denoted by $n$, and $p$ is the order of linear prediction analysis, typically set to ten for 8 kHz sampled signal.

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

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

The spectral dynamics feature has been studied and successfully used in speech coding [29], [30] and speech enhancement [31].

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

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

and it is also frequently used as an approximation of a measure of perceptual "brightness". The last three per-frame features are the variance of the excitation of the AR model $\mathrm{E}_n^e$, the speech signal variance $\mathrm{E}_n^s$, and the pitch period $\mathrm{T}_n$. They will be denoted as $\Phi_4(n)$, $\Phi_5(n)$, and $\Phi_6(n)$, respectively.

The per-frame features presented above, and their first time derivatives (except the derivative of the spectral dynamics) are grouped in an 11 dimensional per-frame feature vector $\mathbf{\Phi}(n)$. To clarify the notation, the elements of the per-frame feature vector are shown in Table II.

TABLE II
ELEMENTS OF PER-FRAME FEATURE VECTOR

| Description | Feature | Time derivative of feature |
|---|---|---|
| Spectral flatness | $\Phi_1$ | $\Phi_7$ |
| Spectral dynamics | $\Phi_2$ | - |
| Spectral centroid | $\Phi_3$ | $\Phi_8$ |
| Excitation variance | $\Phi_4$ | $\Phi_9$ |
| Speech variance | $\Phi_5$ | $\Phi_{10}$ |
| Pitch period | $\Phi_6$ | $\Phi_{11}$ |

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, skewness, and kurtosis. The moments are calculated independently for each per-frame feature, and this gives a set of features that globally describe one speech utterance (global features):

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

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

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

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

With $\tilde{\Omega}$ we denote the set of frames, of cardinality $|\tilde{\Omega}|$, used to calculate statistics for each of the per-frame features $\Phi_i(n)$. The global features are grouped into one global feature set $\Psi = \{\mu_{\Phi_i}, \sigma_{\Phi_i}, s_{\Phi_i}, k_{\Phi_i}\}_{i=1}^{11}$. 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 $\Omega$, 2) and transforms global feature set $\Psi$ into global feature set $\tilde{\Psi}$ of low dimensionality.

### *B. Dimensionality Reduction*

In this work dimensionality reduction is attained through a combination of frame selection and global feature selection. The dimensionality reduction algorithm is important to the practical performance of quality assessment systems. The main purpose of the dimensionality reduction algorithm is to improve predictive accuracy of the quality assessment system by removing irrelevant and redundant data. The dimensionality reduction algorithms presented in this section are based on a training procedure that will be described in detail in section IV.

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]. We propose a generalization of this concept by considering activity thresholds in all per-frame feature dimensions. The scheme, presented in Table III allows speech active frames to be excluded if they do not carry information that improves the accuracy of speech quality prediction. The concept of the frame selection algorithm is to accept only frames where the per-frame feature vector $\Phi$ lies inside or on the surface of the 11-dimensional hypercube defined by threshold vector $\mathbf{\Theta}$.

TABLE III
THE OPTIMAL SET OF FRAMES AS A FUNCTION OF A THRESHOLD VECTOR $\mathbf{\Theta}$

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: $\tilde{\Omega} = \tilde{\Omega} + \{n\}$

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

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

The criterion $\varepsilon$ is calculated as the root-mean-square error (RMSE) performance of the LCQA algorithm:

$$\varepsilon = \sqrt{\frac{\sum_{i=1}^{N}(Q_i - \hat{Q}_i)^2}{N}}, \tag{10}$$

where $\hat{Q}$ is the predicted quality, and $Q$ is the subjective quality. Here $N$ is the number of MOS labeled utterances used in evaluation, see section IV. The optimization of the threshold $\mathbf{\Theta}$ is based on the entire set of global features $\Psi$.

The optimization of $\varepsilon$ in (9), with the frame selection algorithm, described in Table III, results in the following criterion for the acceptance of the $n$-th frame:

$$\Phi_5(n) > \Theta_5^L \ \& \ \Phi_1(n) < \Theta_1^U \ \& \ \Phi_2(n) < \Theta_2^U, \tag{11}$$

with the threshold values $\Theta_5^L = 3.10$, $\Theta_1^U = 0.67$, and $\Theta_2^U = 4.21$. We see that only three per-frame features have significant impact on the frame selection, namely speech variance $\Phi_5$, spectral flatness $\Phi_1$, and spectral dynamics $\Phi_2$. The first and second inequalities in (11) accept only frames with high-energy and a 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. Perhaps the latter result is due to distortion being more easily perceived in steady-state regions of the speech signal.

The criterion (11) reduces significantly the number of frames processed by the LCQA algorithm. The number of selected frames varies with speakers and sentences, and typically $\tilde{\Omega}$ contains between 20% and 50% of the total frame set $\Omega$.

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

$$\tilde{\Psi} = \underset{\tilde{\Psi}^* \subset \Psi}{\arg\min}\, \varepsilon(\tilde{\Psi}^*). \tag{12}$$

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) global 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 experiments 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 subsets are better than the previously evaluated ones, see Table IV.

After optimization of $\varepsilon$ in (12), the dimensionality of the global feature set is reduced from 44 to 14, i.e., $|\tilde{\Psi}| = 14$, and these elements are:

$$\begin{aligned} \tilde{\Psi} &= \{s_{\Phi_1}, \sigma_{\Phi_2}, \mu_{\Phi_4}, \mu_{\Phi_5}, \sigma_{\Phi_5}, s_{\Phi_5}, \\ &= \mu_{\Phi_6}, s_{\Phi_7}, \mu_{\Phi_8}, \mu_{\Phi_9}, \sigma_{\Phi_9}, s_{\Phi_9}, \mu_{\Phi_{10}}, \mu_{\Phi_{11}}\}. \end{aligned} \tag{13}$$

We observe that all per-frame features are present (through their global features statistical representation) in the set $\tilde{\Psi}$, but the speech signal variance $\Phi_5$, and the derivative of the variance of the excitation signal $\Phi_9$ are most frequent. Another interesting observation is that global speech features based only on the first three moments are present, and the global features based on kurtosis seem to be less important.

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 $\tilde{\Psi}$. At the first stage we optimize the thresholds $\mathbf{\Theta}$ for frame selection procedure given the entire set of global features $\Psi$. At the second stage we reduce dimensionality of the global feature set based on the optimal subset of frames $\tilde{\Omega}(\mathbf{\Theta})$. The reason for using this sub-optimal procedure is the infeasibility of the joint optimization. However, the experiments presented in section V show that the proposed training scheme is sufficient to outperform the reference quality assessment methods.

TABLE IV

THE SEQUENTIAL FLOATING BACKWARD SELECTION PROCEDURE CONSISTS OF APPLYING AFTER EACH BACKWARD STEP A NUMBER OF FORWARD STEPS AS LONG AS THE RESULTING SUBSETS 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 global feature

$$\Psi_{i^-} = \underset{\Psi_i \in \tilde{\Psi}}{\arg\min}\, \varepsilon(\tilde{\Psi} - \{\Psi_i\})$$

Exclude the feature

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

**while** ***error decreases by more than a threshold***

Inclusion Steps:

Find the most significant global feature

$$\Psi_{i^+} = \underset{\Psi_i \notin \tilde{\Psi}}{\arg\min}\, \varepsilon(\tilde{\Psi} + \{\Psi_i\})$$

Include the feature

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

### C. Quality Estimation Given the Global Feature Set

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

criterion

$$\hat{Q}(\tilde{\Psi}) = \arg\min_{Q^*(\tilde{\Psi})} E\{(Q - Q^*(\tilde{\Psi}))^2\}, \quad (14)$$

where $E\{\}$ is the expectation operator. The above defined criterion is the probabilistic measure corresponding to (12).

It is well known, e.g., [37], that equation (14) is minimized by the conditional expectation

$$\hat{Q}(\tilde{\Psi}) = E\{Q|\tilde{\Psi}\}, \quad (15)$$

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

$$f(\varphi|\lambda) = \sum_{m=1}^{M} \omega^{(m)} \mathcal{N}(\varphi|\mu^{(m)}, \Sigma^{(m)}), \quad (16)$$

where $\varphi = [Q, \tilde{\Psi}]$, $m$ is the mixture component index, $\omega^{(m)}$ are the mixture weights, and $\mathcal{N}(\varphi|\mu^{(m)}, \Sigma^{(m)})$ are multivariate Gaussian densities, with $\mu^{(m)}, \Sigma^{(m)}$ being the means and covariance matrices of the Gaussian densities.

The GMM is completely specified by a set of $M$ mean vectors, covariance matrices and mixture weights

$$\lambda = \{\omega^{(m)}, \mu^{(m)}, \Sigma^{(m)}\}_{m=1}^{M}, \quad (17)$$

and these coefficients are estimated off-line from a large training set using the expectation maximization (EM) algorithm [38]. Details on the data used for training are presented in section IV. Our experiments showed that it is sufficient to use 12 full-covariance matrices (14 x 14), i.e., for dimensionality $K = 14$ and $M = 12$ Gaussians, this is $M * (1 + K + K * (K + 1)/2) = 1440$ training parameters.

Using the joint Gaussian mixture model, the conditional expectation can solved to be a weighted sum of component-wise conditional expectations, well-known for the Gaussian case [39]. Hence, the optimal quality estimator (15) is expressed as

$$E\{Q|\tilde{\Psi}\} = \sum_{m=1}^{M} u^{(m)}(\tilde{\Psi}) \mu_{Q|\tilde{\Psi}}^{(m)} \quad (18)$$

where

$$u^{(m)}(\tilde{\Psi}) = \frac{\omega^{(m)} \mathcal{N}(\tilde{\Psi}|\mu_{\tilde{\Psi}}^{(m)}, \Sigma_{\tilde{\Psi}\tilde{\Psi}}^{(m)})}{\sum_{k=1}^{M} \omega^{(k)} \mathcal{N}(\tilde{\Psi}|\mu_{\tilde{\Psi}}^{(k)}, \Sigma_{\tilde{\Psi}\tilde{\Psi}}^{(k)})}, \quad (19)$$

and

$$\mu_{Q|\tilde{\Psi}}^{(m)} = \mu_{Q}^{(m)} + \Sigma_{\tilde{\Psi}Q}^{(m)} (\Sigma_{\tilde{\Psi}\tilde{\Psi}}^{(m)})^{-1} (\tilde{\Psi} - \mu_{\tilde{\Psi}}^{(m)}), \quad (20)$$

with $\mu_{\tilde{\Psi}}^{(m)}, \mu_{Q}^{(m)}, \Sigma_{\tilde{\Psi}\tilde{\Psi}}^{(m)}, \Sigma_{\tilde{\Psi}Q}^{(m)}$ being the means, covariance and cross-covariance matrices of $\tilde{\Psi}$ and $Q$ of the $m$-th mixture component.

### *D. Implementation Details*

In this section we describe how per-frame features, for the $n$-th frame, are calculated, based entirely on $\{\mathrm{E}_n^e, \mathrm{T}_n, \mathbf{f}_n\}$ and $\{\mathrm{E}_{n-1}^e, \mathrm{T}_{n-1}, \mathbf{f}_{n-1}\}$. Then we show how the global statistical properties are calculated recursively, without storing the per-frame features in a buffer. We calculate the pitch period $\mathrm{T}_n$ according to [40], 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 bitstream, and signal reconstruction is avoided.

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

$$\Phi_1(n) = \frac{\mathrm{E}_n^e}{\mathrm{E}_n^s}. \quad (21)$$

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

$$e_k = s_k - \sum_{i=1}^{10} a_i s_{k-i}, \quad (22)$$

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:

$$\Phi_2(n) = (\mathbf{f}_n - \mathbf{f}_{n-1})^T \mathbf{W}_n (\mathbf{f}_n - \mathbf{f}_{n-1}), \quad (23)$$

where the inverse harmonic mean weight [41] is defined by the components of the LSF vector:

$$W_n^{(ii)} = (f_n^{(i)} - f_n^{(i-1)})^{-1} + (f_n^{(i+1)} - f_n^{(i)})^{-1} \quad (24)$$

$$W_n^{(ij)} = 0 \quad (25)$$

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

$$\Phi_3(n) = \frac{\sum_{i=1}^{10} i W_n^{(ii)}}{\sum_{i=1}^{10} W_n^{(ii)}}. \quad (26)$$

The spectral flatness and spectral dynamics are well approximated by (21) and (23). The correlation of the true with the approximated per-frame values, over the training set of TIMIT database [42], are $0.98$ and $0.95$ respectively. The approximation of spectral centroid with (26) is less accurate. The averaged correlation over the TIMIT database is $0.72$. The difference of per-frame values of (4) and (26) are shown in Fig. 4.

The approximation of (1), (3), and (4) with (21), (23), and (26) is advantageous in the case of evaluation of CELP coders. It is possible that in another environment the best option is to use different approximation of (1), (3), and (4).

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

$$\mu_\Phi(n) = \frac{n-1}{n} \mu_\Phi(n-1) + \frac{1}{n} \Phi(n) \quad (27)$$

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

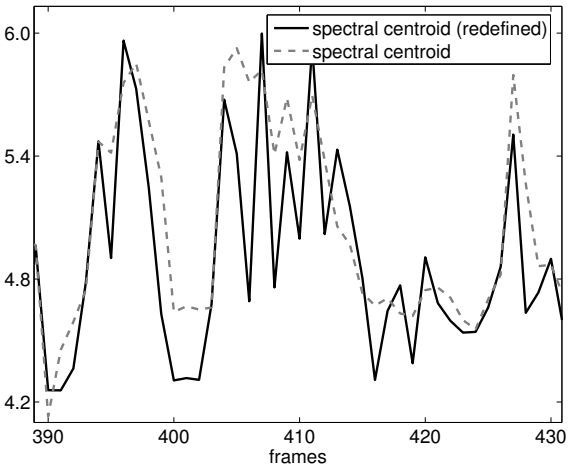


Fig. 4. Spectral centroid (4) and redefined spectral centroid (26) values over speech frames.

$$\sigma_\Phi = \mu_{\Phi^2} - (\mu_\Phi)^2 \quad (28)$$

$$s_\Phi = \frac{\mu_{\Phi^3} - 3\mu_\Phi\mu_{\Phi^2} + 2(\mu_\Phi)^3}{\sigma_\Phi^{3/2}} \quad (29)$$

$$k_\Phi = \frac{\mu_{\Phi^4} - 4\mu_\Phi\mu_{\Phi^3} + 6(\mu_\Phi)^2\mu_{\Phi^2} - 3(\mu_\Phi)^4}{\sigma_\Phi^2}. \quad (30)$$

Table V gives a short overview of the proposed LCQA algorithm. Note that for the experiments presented in section V, the required per-frame features are extracted from the waveform using the single case of 20 ms non-overlapping windows as described in section III-D. The per-frame features and are not obtained directly from the network. Thus, we only show that the chosen per-frame and global features allow the accurate prediction of speech quality. This provides proof-of-concept, but if actual CELP coder parameters are to be extracted from the network, then additional sources of variability may require further LCQA design and optimization work to attain accurate speech quality predictions. Some examples of these sources of additional variability are frame lengths and overlaps, window shapes, LPC orders and algorithms, pitch algorithms, post-filtering, and packet loss concealment behavior.

TABLE V
OVERVIEW OF THE GENERIC LCQA ALGORITHM

1) For the $n$-th speech frame calculate $\{\mathrm{E}_n^e, \mathrm{T}_n, \mathbf{f}_n\}$ from the waveform or extract from the bitstream.
2) Calculate per-frame feature vector $\mathbf{\Phi}(n)$, based on $\{\mathrm{E}_n^e, \mathrm{T}_n, \mathbf{f}_n\}$ and stored in a buffer $\{\mathrm{E}_{n-1}^e, \mathrm{T}_{n-1}, \mathbf{f}_{n-1}\}$.
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 $\mathbf{\Theta}$.
4) At the end of the utterance calculate global feature set $\tilde{\Psi} = \{\mu_{\Phi_i}, \sigma_{\Phi_i}, s_{\Phi_i}, k_{\Phi_i}\}$ as mean, variance, skew, and kurtosis of per-frame features.
5) Predict the speech quality as a function of the global feature set $\hat{Q} = \hat{Q}(\tilde{\Psi})$, through GMM mapping.

## IV. Speech Databases and Training

In this section we first discuss speech databases used in the training and evaluation process of the LCQA algorithm. Then we reveal the details of the training process, used for the dimensionality reduction procedure (section III-B), and the performance evaluation (section V).

### A. Speech Databases

For the training and evaluation procedure we used 11 MOS labeled databases provided by Ericsson AB and 7 similarly labeled databases from the ITU-T P.Supp 23 [43]. Data with DMOS scores were excluded from our experiments, e.g., from ITU-T P.Supp 23 we excluded Experiment 2. The speech material in these databases contains utterances in the following languages: English, French, Japanese, Italian and Swedish. The databases contain a large variety of distortions, such as: different coding, tandeming, and modulated noise reference unit (MNRU) [44] conditions, as well as packet loss, background noise, effects of noise suppression, switching effects, different input levels, etc. The total size of the union of databases is 7646 utterances with averaged length 8s.

### B. Training

We split the available databases into two parts, *test set* and *training set*. The test set is based on 7 databases from ITU-T P.Supp 23 (1328 utterances) and the training set is based on 11 Ericsson's databases (6318 utterances). The test set is not available during the training, but used only for evaluation. The training, used for the dimensionality reduction scheme and performance evaluation experiments is based entirely on the training set. To improve generalization performance we use a *training with noise* procedure [45]. We create virtual ("noisy") training patterns, by adding zero mean white Gaussian noise, at 20 dB SNR to the global feature set $\Psi$. In this manner for each global feature set we create four virtual sets, and the training is based on the union of the "original" and "noisy" features.

## V. Performance Evaluation

In this section we present results from experiments, with respect to both prediction accuracy and computational complexity of the proposed algorithm. We compare the performance of the proposed LCQA algorithm with the standardized ITU-T P.563. The estimation performance is assessed using per-condition correlation coefficient $R$ between the predicted quality $\hat{Q}$ and the subjective quality $Q$:

$$R = \frac{\sum_i (\hat{Q}_i - \mu_{\hat{Q}})(Q_i - \mu_Q)}{\sqrt{\sum_i (\hat{Q}_i - \mu_{\hat{Q}})^2 \sum_i (Q_i - \mu_Q)^2}}, \quad (31)$$

where $\mu_Q$ and $\mu_{\hat{Q}}$ are the mean values of the introduced variables, and summation is over conditions. Table VI contains

the performance results in terms of the selected performance metric over a test set of 7 databases from ITU-T P.Supp 23. The ITU-T P.Supp 23 Exp 1 contains speech coding distortions, produced by seven standard speech codecs (predominantly using G.729 speech codec [46]) alone, or in tandem configuration. In the ITU-T P.Supp 23 Exp 3 G.729 speech codec is evaluated under various channel error conditions like frame erasure, random bit error, and background noise. The test results, presented in Table VI clearly indicate that the proposed LCQA algorithm outperforms the standardized ITU-T P.563.

TABLE VI
PERFORMANCE OF THE LCQA ALGORITHM IN TERMS OF PER-CONDITION CORRELATION COEFFICIENT OVER ITU-T P.SUPP 23 DATABASES

| Database | Language | LCQA | ITU-T P.563 |
|---|---|---|---|
| Exp 1 A | French | 0.94 | 0.88 |
| Exp 1 D | Japanese | 0.94 | 0.81 |
| Exp 1 O | English | 0.95 | 0.90 |
| Exp 3 A | French | 0.93 | 0.87 |
| Exp 3 C | Italian | 0.95 | 0.83 |
| Exp 3 D | Japanese | 0.94 | 0.92 |
| Exp 3 O | English | 0.93 | 0.91 |

Processing time and memory requirements are important figures of merit for the quality estimation algorithms. The LCQA algorithm has insignificant memory requirements: a buffer of 12+12 scalar values, calculated from the previous and current frame is needed (future frames are not required), as well as memory for the mixture of 12 Gaussians.

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 $\{\mathrm{E}_n^e, \mathrm{T}_n, \mathbf{f}_n\}$ 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 |

## VI. CONCLUSIONS

We demonstrated that a low-complexity 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 experiments over MOS labeled databases 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 per-frame 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 and also thank the anonymous reviewers whose valuable comments led to improvement of the paper.

## 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.1534-1, "Method for the subjective assessment of intermediate quality level of coding systems," 2005.
[4] ITU-R Rec. BS.1284-1, "General methods for the subjective assessment of sound quality," 2003.
[5] ITU-T Rec. G.107, "The e-model, a computational model for use in transmission planning," 2005.
[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.1387-1, "Method for Objective Measurements of Perceived Audio Quality (PEAQ)," 2001.
[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. IEE 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] D. Kim and A. Tarraf, "Enhanced perceptual model for non-intrusive speech quality assessment," in *Proc. IEEE Int. Conf. Acous., Speech, Signal Processing*, vol. 1, pp. 829–832, 2006.
[22] ITU-T P. 563, "Single ended method for objective speech quality assessment in narrow-band telephony applications," 2004.
[23] 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.

[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*, vol. 10, pp. 937–940, 1985.

[28] S. Jayant and P. Noll, *Digital Coding of Waveforms*. Englewood Cliffs NJ: Prentice-Hall, 1984.

[29] 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.

[30] F. Norden and T. Eriksson, "Time evolution in LPC spectrum coding," *IEEE Trans. Speech, Audio Processing*, vol. 12, pp. 290–301, 2004.

[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 analized 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] S. M. Kay, *Fundamentals of Statistical Signal Processing, Estimation Theory*. Prentice Hall, 1993.

[40] 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.

[41] R. Laroia, N. Phamdo, and N. Farvardin, "Robust and efficient quantization of speech LSP parameters using structured vector quantizers," in *Proc. IEEE Int. Conf. Acous., Speech, Signal Processing*, vol. 1, pp. 641–644, 1991.

[42] DARPA-TIMIT, "Acoustic-phonetic continuous speech corpus, NIST Speech Disc 1-1.1," 1990.

[43] ITU-T Rec. P. Supplement 23, "ITU-T coded-speech database," 1998.

[44] ITU-T. Rec. P.810, "Modulated Noise Reference Unit," 1996.

[45] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. Wiley-Interscience, second ed., 2001.

[46] ITU-T. Rec. G.729, "Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP)," 1996.