Monitoring of In-Service Perceptual Speech Quality in Modern Cellular Radio Systems

Behrooz Rohani, Bijan Rohani, Manora Caldera, and Hans-Jürgen Zepernick

***Abstract*—A method for in-service monitoring of the end-user perceptual speech quality in modern cellular radio systems is proposed. This method incorporates the perceptual evaluation of speech quality (PESQ) algorithm to monitor the quality experienced by the end-user. Here, the monitoring is carried out at the transmitting side. In this case, the speech signal received by the end-user is approximated at the transmitter in accordance with a feedback signal. The performance of the proposed scheme has been investigated through extensive computer simulations for the Universal Mobile Telecommunication System (UMTS) using different speech coding rates and channel conditions. The results indicate that the proposed scheme can predict the end-user quality with a root-mean-squared error (RMSE) of at most 0.15 using the mean opinion score (MOS) rating scheme. Such accuracy can be beneficial in applications such as network maintenance and radio resource management for satisfying desired level of quality of service.**

I. Introduction

CUSTOMERS have become increasingly selective with the quality of service (QoS) they experience with the cellular telephony service providers. One of the key performance indicators in voice grade services is the speech quality. Consequently, in order to meet their QoS commitments to the customers, the service providers have been driven to monitor and maintain the speech quality in their networks.

The true measure of speech quality is the subjective mean opinion score (MOS) which is determined by a panel of trained listeners under controlled conditions [1], [2]. There are a variety of subjective tests and scoring schemes [1]-[5] but in practice the listening quality (LQ) test based on the 5-point absolute category rating (ACR) is widely used [1]. In this case, the listeners rate the speech material on a scale of 1 (bad quality) to 5 (excellent quality). The average of scores for all listeners represents the MOS-LQ.

Listening tests are not practicable for live network monitoring. As such, objective methods for estimation of MOS-LQ have been developed [6]-[10]. Objective methods can be categorized into parametric methods, e.g. [6]-[8], and psychoacoustic methods such as [9], [10]. The telecommunication industry developed parametric objective speech quality measurement algorithms for monitoring network speech quality. In this case, the speech quality is indirectly estimated based on a set of network parameters that can affect speech quality. One of the prominent cellular mobile network monitoring tools incorporates a parametric algorithm for calculation of the so-called speech quality index (SQI) [6]. The SQI has been formulated on the premise that almost all speech quality degradations in a cellular network are due to radio transmission imperfections. In effect, SQI estimates speech quality by computing an equivalent modulated noise reference unit (MNRU) [11] from a number of radio link parameters including the bit error rate (BER), frame erasure rate (FER), and stolen frames, in addition to codec related information. Similar measures have been investigated in [7] and [8]. Parametric methods can be misleading at times as they derive speech quality indirectly without analyzing the speech signal. In this case, a call could be indicated as technically successful even though its clarity may be poor.

The alternatives to parametric methods are psychoacoustic quality assessment techniques which have their origin in codec performance testing. Such techniques use sophisticated models of human auditory system and cognitive processes to derive a MOS-LQO (listening quality objective measure) directly from the speech signal [12]-[14]. Less than a decade ago, the perceptual speech quality measure (PSQM) was standardized by the International Telecommunication Union (ITU) as the recommended algorithm for objective speech quality assessment of codecs [15]. Further improvements to PSQM followed, culminating in the ITU-T Recommendation P.862, perceptual evaluation of speech quality (PESQ), which is applicable not only to speech codecs but also to end-to-end network measurements [16]. The performance of PESQ has been verified for a range of codecs in the presence of some common transmission impairments such as bit errors, frame erasures and delay variations.

Both PESQ and its predecessor PSQM are “intrusive” algorithms in that they require the original speech signal (the reference) beside its distorted version for calculating the MOS-LQO. This imposes limitations on the functionality of PESQ for many real-world applications where both signals cannot be present at the point of measurement. In particular, for cellular network quality monitoring, PESQ has been restricted to off-line measurements, whereby test signals are transmitted for recording during drive-test campaigns. Such

Behrooz Rohani is with the ERG Group, 247 Balcatta Road, Balcatta, WA 6021, Australia (e-mail: behrooz.rohani@erggroup.com).

Bijan Rohani is with the Western Australian Telecommunications Research Institute, 39 Fairway, Nedlands, WA 6907, Australia (bijan@watri.org.au).

Manora Caldera is with Gibson Quai-AAS Consulting, 30 Richardson Street, Perth, WA 6005, Australia (e-mail: mcaldera@gqaas.com.au).

Hans-Jürgen Zepernick is with the Blekinge Institute of Technology, SE-372 25 Ronneby, Sweden (e-mail: hans-jurgen.zepernick@bth.se).

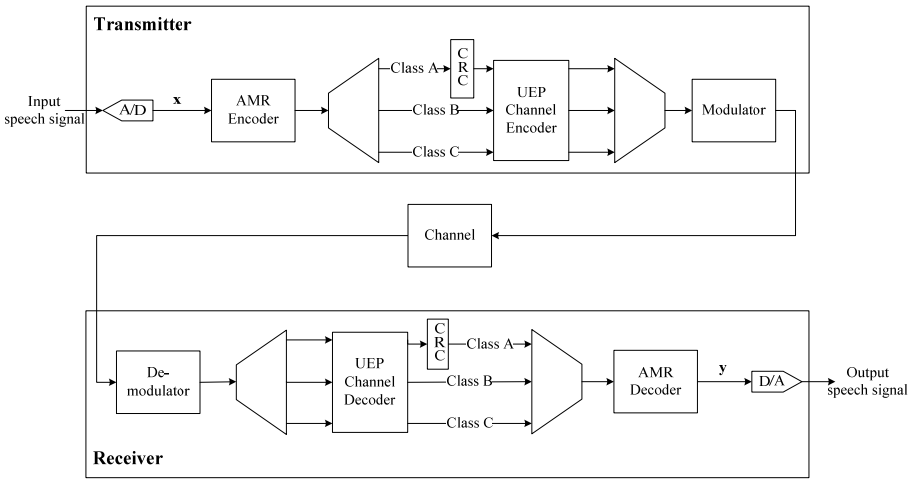


Fig. 1. Simplified block diagram of the UMTS speech link.

measurement campaigns are costly and generate additional network traffic while providing only an incomplete picture of the network quality due to geographical and time constraints of the campaign.

Several non-intrusive algorithms, which calculate MOS-LQO based only on the distorted signal, have been developed over the years [17]-[21]. Subsequently, the ITU selected the single-ended assessment model (SEAM) embodied in ITU-T Recommendation P.563 as the standard non-intrusive method for objective speech quality assessment [21]. The SEAM provides a flexible method that can be used for assessing the speech quality at the output of a network without prior knowledge of the source signal and irrespective of the network type.

It is noted however that the SEAM achieves its flexibility at the expense of considerable complexity and diminished accuracy compared to PESQ. Furthermore, it still necessitates drive-tests and the restrictions associated with them. Additionally, SEAM evaluates the speech quality based only on the signal at the point of measurement. Therefore, in effect, the degradations introduced at the source, across any third party networks, and the operator's own network are lumped together. This can be of limited value to network operators who are only interested in measuring quality degradation across their radio transmission path with little interest in losses arising from a noisy source or third party networks.

In this paper, a method for in-service monitoring of speech quality in modern cellular networks is considered. This method evaluates speech quality experienced by the customer remotely, i.e., from the network side, hence eliminating the need for drive-tests. In addition, the measurements are performed on live network traffic. The proposed method calculates a reasonably close approximation of the output speech signal observed at the end-user side. This is then compared against a copy of the transmitted signal using PESQ. Even though this method has initially been studied for a limited set of conditions [22]-[24], a more comprehensive study is presented here.

In this paper, the potential application of the proposed scheme for the third generation (3G) Universal Mobile Telecommunication System (UMTS) as an example of a typical cellular system is considered. However, this scheme can be adapted to many modern communication systems. The paper is organized as follows. In Section II, the model of UMTS is described. This is followed by the explanation of the proposed quality measurement scheme in Section III. The details of the computer simulation model and the methodology used are given in Section IV and the corresponding results are presented in Section V. The paper is concluded with some remarks in Section VI.

II. SYSTYEM MODEL

A basic block diagram of a UMTS link is shown in Fig. 1. The link consists of the transmitter, the channel, and the receiver. At the transmitter, the input speech signal is divided into 20 ms frames (160 samples at 8 kHz) for encoding by the adaptive multi-rate (AMR) speech codec [12]. Input frames are encoded at one of eight selectable output bit rates $R \in \{4.75, 5.15, 5.90, 6.70, 7.40, 7.95, 10.20, 12.2\}$ kb/s. Although a different codec rate can be selected for each transmitted frame, in practice, the codec rate is changed at a much slower rate to suit network loading conditions, extend uplink coverage, or maintain speech quality through controlling the link FER [25], [26]. The contents of each frame are arranged in the order of their decreasing perceptual importance into Class-A, -B, and -C bits so that unequal error protection (UEP) can be applied to them. In this way, the perceptually more important bits can be protected more effectively against transmission errors. Errors in Class-A bits can lead to undesirable artifacts in the reproduced speech at the receiver. These errors are detected through cyclic redundancy check (CRC) bits that are attached to the Class-A bits for each frame before UEP is applied [27], [28]. Subsequently, the output of the channel encoder is processed in a chain of physical layer stages which for the sake of brevity are not discussed here. The encoded bit stream is finally modulated and transmitted over the wireless channel.

The wireless channel in modern communication systems is characterized by fast and slow fading as well as multiple access interference (MAI) from other users in the system. Essentially, these affect the integrity of the transmitted bits such that some of the demodulated bits arrive in error. The errors uncorrected by the channel decoder lead to distortion of the output speech signal and degradation of its perceptual quality. These errors are especially detrimental if they affect Class-A bits. Presence of Class-A errors is detected by testing the received CRC bits. Individual frames are flagged as "bad" or "good" depending on the outcome of the test. A binary flag, namely the bad frame indicator (BFI), is associated with each frame to indicate to the AMR decoder if the received frame is bad. The AMR decoder invokes an error concealment procedure (ECP) in the case of bad frames. During the ECP, a bad frame is "erased" and a replacement frame is calculated based on an algorithm not discussed here.

III. The Proposed Speech Quality Monitoring Scheme

Remote monitoring of the network quality is when measurement of the receiver speech quality is performed at the transmitting end. An ideal scenario for remote monitoring of the perceptual speech quality is shown in Fig. 2(a). In this case, PESQ resides on the transmitting side and computes the LQO of the distorted output $\mathbf{y}=[y_1\ y_2\ \ldots\ y_M]$ by comparing it against the corresponding transmitted signal $\mathbf{x}=[x_1\ x_2\ \ldots\ x_M]$, where x_m and y_m for m=1,2, …, M are the m^{th} samples of the input and output signals, respectively. Let the calculation of LQO be represented by the perceptual distance function D, thus the quality q obtained from comparing $\mathbf{y}$ against $\mathbf{x}$ for the period of measurement can be expressed as

$$q = D(\mathbf{x}, \mathbf{y}) \tag{1}$$

For calculating q in (1), both $\mathbf{x}$ and $\mathbf{y}$ are required by the PESQ algorithm at the transmitting side. Since the distorted output $\mathbf{y}$ is not available in real time at the transmitter, the situation depicted in Fig. 2(a) cannot be realized in practice unless the recordings of the signals $\mathbf{x}$ and $\mathbf{y}$ are used. Apparently, this is not a suitable solution for in-service network monitoring with live traffic. To remedy this, an approximation $\hat{\mathbf{y}}$ of the output signal $\mathbf{y}$ can be used for quality measurement as shown in Fig. 2(b). The output signal can be approximated by distorting the transmitted signal $\mathbf{x}$ according to a distortion function ψ such that

$$\hat{\mathbf{y}} = \psi(\mathbf{x}, \mathbf{y}_r) \tag{2}$$

where $\mathbf{y}_r$ is a distortion vector derived from the output $\mathbf{y}$. In this case, the quality is approximated by $\hat{q}$ as

$$\hat{q} = D(\mathbf{x}, \hat{\mathbf{y}}) \tag{3}$$

The distortion function ψ represents the overall signal distortion from the input of the AMR encoder at the transmitter to the output of the AMR decoder at the receiver as shown in Fig. 3. Here, the input signal $\mathbf{x}$ is divided into N_F frames each containing 160 speech samples for coding by the AMR encoder, i.e. $\mathbf{x}=[\mathbf{x}^{(1)}\ldots\ \mathbf{x}^{(n)}\ \ldots\ \mathbf{x}^{(NF)}]$ where $\mathbf{x}^{(n)}$ is the n^{th} frame. The output of the encoder for the n^{th} speech frame $\mathbf{x}^{(n)}$ can be expressed as

$$\mathbf{x}_Q^{(n)} = Q\left(\mathbf{x}^{(n)}, R^{(n)}\right) \tag{4}$$

where Q represents the AMR transcoding function and $R^{(n)}$ is the coding rate used for the n^{th} speech frame $\mathbf{x}^{(n)}$.

The speech coded frame $\mathbf{x}_Q^{(n)}$ consists of Class-A bits $\mathbf{x}_A^{(n)}$, their corresponding CRC bits $\mathbf{p}^{(n)}$, Class-B and -C bits $\mathbf{x}_B^{(n)}$ and $\mathbf{x}_C^{(n)}$, respectively. This is represented as

$$\mathbf{x}_Q^{(n)} = \left[\mathbf{x}_A^{(n)}\ \ \mathbf{p}^{(n)}\ \ \mathbf{x}_B^{(n)}\ \ \mathbf{x}_C^{(n)}\right] \tag{5}$$

The overall effect of the communication chain between the AMR encoder and the decoder is represented by bit errors affecting the transmitted bits in each frame $\mathbf{x}_Q^{(n)}$. In this case, the received speech frame at the input of the AMR decoder is given by

$$\hat{\mathbf{x}}_Q^{(n)} = \mathbf{x}_Q^{(n)} + \mathbf{e}^{(n)} \tag{6}$$

where $\mathbf{e}^{(n)}$ represents the transmission errors associated with the n^{th} frame. Specifically, transmission errors can be divided into Class-A errors $\mathbf{e}_A^{(n)}$, CRC bit errors $\mathbf{e}_p^{(n)}$, Class-B and -C errors $\mathbf{e}_B^{(n)}$ and $\mathbf{e}_C^{(n)}$, respectively. Thus, $\mathbf{e}^{(n)}$ can be represented as

$$\mathbf{e}^{(n)} = \left[\mathbf{e}_A^{(n)}\quad \mathbf{e}_p^{(n)}\quad \mathbf{e}_B^{(n)}\quad \mathbf{e}_C^{(n)}\right] \tag{7}$$

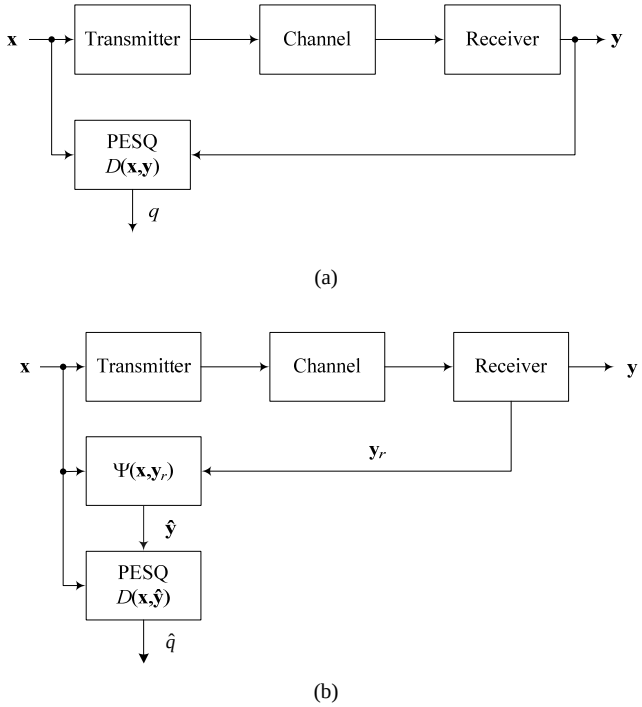


Fig. 2. Scenarios showing transmitter remotely measuring the perceptual speech quality of the end-user: (a) Ideal case, (b) Based on approximation of the output signal.

As described in Section II, the BFI flag $I^{(n)}$ for the n^{th} received frame can be derived from the Class-A errors and the CRC bit errors as

$$I^{(n)} = f\left(\mathbf{e}_A^{(n)}, \mathbf{e}_p^{(n)}\right) \tag{8}$$

Here, $f(\cdot,\cdot)$ is a binary function whose value is "0" when the polynomial

$$E(X) = \left[\mathbf{e}_A^{(n)} \quad \mathbf{e}_p^{(n)}\right] \begin{bmatrix} X^{N_A+N_p-1} \\ X^{N_A+N_p-2} \\ \vdots \\ X \\ 1 \end{bmatrix} \tag{9}$$

is divisible by the CRC polynomial [28], and its value is "1" otherwise. Here, N_A and N_p are the number of Class-A bits and the CRC bits, respectively. Subsequently, as depicted in Fig. 3, the output $\mathbf{y}^{(n)}$ can be obtained according to

$$\mathbf{y}^{(n)} = Q^{-1}\left(\hat{\mathbf{x}}_Q^{(n)}, R^{(n)}, I^{(n)}\right) \tag{10}$$

where Q^{-1} representing the AMR decoding.

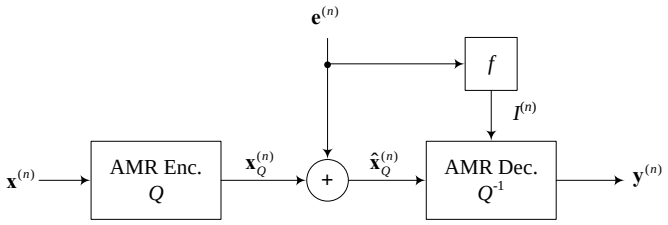


Fig. 3. Modeling of the output signal in terms of AMR codec functions and transmission errors.

It is noted that $I^{(n)}$=1 invokes the error concealment procedure in which case a substitution frame is calculated based on the most recent frame with error-free Class-A bits [29]. In this case, the resulting quality degradation could be significantly higher than when only $\mathbf{e}_B^{(n)}$ and $\mathbf{e}_C^{(n)}$ are present. Quality degradation is even more severe in the event of successive frame erasures, whereby the ECP gradually reduces the output signal power whose equivalent psychoacoustic effect is loss of signal loudness [29]. The occurrences of frame erasures and their pattern are both represented by the binary sequence $I^{(n)}$, n=1,2,…,N_F. Accordingly, $I^{(n)}$ may be considered sufficient for estimating the perceptual distortion of $\mathbf{y}^{(n)}$. Therefore using the error-free $\mathbf{x}_Q^{(n)}$ to substitute for $\hat{\mathbf{x}}_Q^{(n)}$ in (10) should lead to a reasonable perceptual approximation of the output $\mathbf{y}^{(n)}$ as

$$\hat{\mathbf{y}}^{(n)} = Q^{-1}\left(\mathbf{x}_Q^{(n)}, R^{(n)}, I^{(n)}\right) \tag{11}$$

It is therefore proposed that the distortion function is modeled as follows

$$\psi\left(\mathbf{x}^{(n)}, \mathbf{y}_r^{(n)}\right) = Q^{-1}\left(\mathbf{x}_Q^{(n)}, R^{(n)}, I^{(n)}\right) \tag{12}$$

where $\mathbf{y}_r^{(n)} = I^{(n)}$, and $\mathbf{x}_Q^{(n)}$ is obtained from (4).

The block diagram of the proposed method is shown in Fig. 4. Here, the error-free signal $\mathbf{x}_Q^{(n)}$ is used for $\hat{\mathbf{x}}_Q^{(n)}$ in (11) and (12). In this case, the distortion function ψ only takes into account the distortion due to frame erasures by virtue of $I^{(n)}$ and ignores Class-B and -C errors. The exact quality degradation due to these errors cannot be calculated in the absence of information regarding the error patterns $\mathbf{e}_B^{(n)}$ and $\mathbf{e}_C^{(n)}$, as such, the average degradation for a given codec rate $R^{(n)}$ can be calculated based on Class-B and -C bit error rates. In addition, because these BERs are directly related to the FER, a mapping function has been employed in Fig. 4 for correcting the quality estimate $\hat{q}$ by factoring in the FER. It is noted that the FER is fixed by the network for the outer loop power control [25] and it can also be estimated based on averaging $I^{(n)}$ values. In Fig. 4, the inaccuracy in the mapping function g is shown by ε.

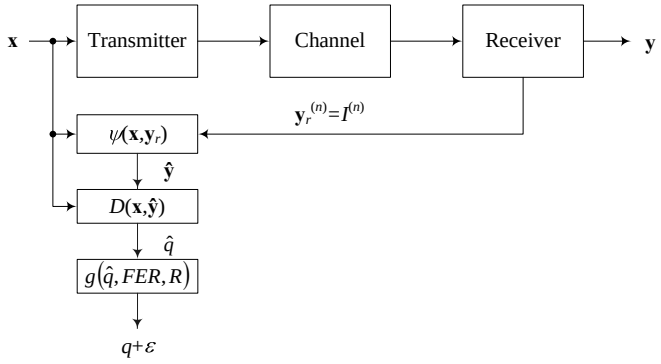


Fig. 4. Block diagram of the proposed scheme for remote monitoring of the output speech quality.

IV. Simulation Methodology

The performance of the proposed scheme has been investigated through computer simulations. In these simulations speech files from ITU Supp. 23 [30] database have been used. Two sets of clean speech sample files, where each set contained 32 files, have been used to represent the input signals (reference signals). The files in the first set were used for obtaining the mapping functions g and hence are referred to as the training data. The purpose of the second set was to verify the accuracy of the proposed scheme based on the mapping functions derived from the training data. Half of the files in each set contained speech samples from a male talker and the other half was from a female talker. Each file contained short sentences separated with pauses. The duration of each file was 8 seconds with a voice activity factor of approximately 0.5. Hereafter, the signals $\mathbf{x}^{(n)}$, $\mathbf{y}^{(n)}$, and $\hat{\mathbf{y}}^{(n)}$ corresponding to the k^{th} speech sample file for k=1,2, …,32, are denoted as $\mathbf{x}_k^{(n)}$, $\mathbf{y}_k^{(n)}$, and $\hat{\mathbf{y}}_k^{(n)}$, respectively.

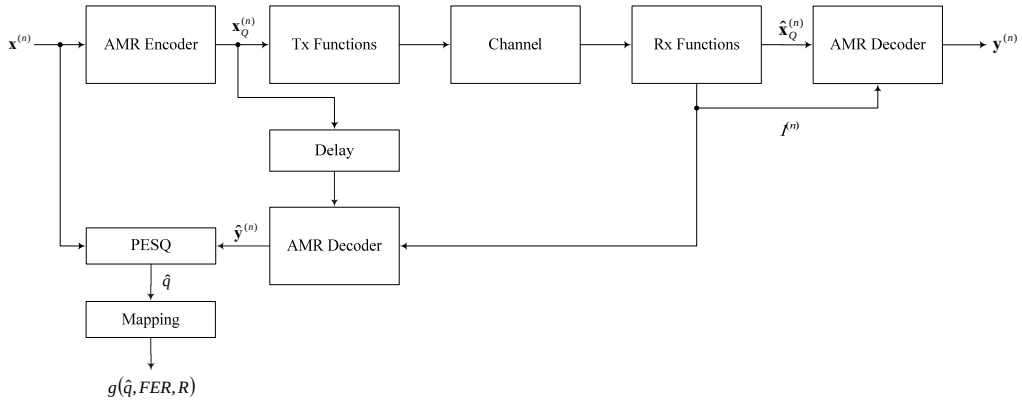


Fig. 5. Block diagram of the simulation model.

The simulations were performed for three error conditions and three codec rates for each file. The error conditions were represented by FER∈{1%, 3%, 5%} which characterizes a typical range of error conditions that may arise in a UMTS link. The FER=1% is typically recommended for maintaining an adequate level of speech quality. However, an operator may sacrifice some quality for a gain in network capacity by allowing the FER to rise up to 3%. Further increases in FER could lead to such a poor quality that the call is handed over or dropped. The simulation error condition with FER=5% was intended to represent such a scenario. It is noted that the simulations can be extended to include a wider range of FER with increased resolution as well as all codec rates. However, this has not been considered here because the results for the other conditions can be approximated based on the available results.

In addition, three vehicular speeds v=3 km/h, 50 km/h, and 120 km/h, were used in simulations for each error condition [31]. These are representative vehicular speeds for typical channel time variations.

As mentioned earlier, the AMR codec provides eight rates ranging from 4.75 kb/s to 12.2 kb/s. In this study, the minimum, maximum, and the median rates of 4.75 kb/s, 12.2 kb/s, and 7.40 kb/s were considered. In addition, as the input speech files are sufficiently short, a fixed codec rate for the entire file, i.e., $R^{(n)}=R$ for n=1, 2, … , N_F have been assumed.

The simulations were performed on a link level simulator implemented in detail according to the physical layer specifications of the third generation partnership project (3GPP) [32]. The block diagram of the simulation models is shown in Fig. 5 with some of the main simulation parameters provided in Table I. In Fig. 5, all the UMTS physical layer functions from the AMR encoder output to the input of the channel are collectively shown as "Tx Functions". The corresponding receiver functionalities are shown as "Rx Functions". The channel model included log-normal shadowing, Rayleigh fading, and MAI. Statistical independence of transmission errors across all simulations was ensured.

Table I

Summary of the main simulation parameters.

Codec Rate	4.75 kb/s, 7.4 kb/s,12.2 kb/s
Channel Coding	
Class A	Rate-1/3 CC[1] + 12 bits CRC
Class B	Rate-1/2 CC
Class C	Rate-1/2 CC
DTX[2]	ON
Trans. Time Interval (TTI)	20 ms
Power Control (PC)	
Outer Loop PC	ON – FER=1%, 3%, 5%
Inner Loop PC	ON
Channel Bit Rate	60 kb/s (Full Rate)
Channel Type	
MAI	ON
Log-normal Fading[3]	ON
Fast Fading	6-tap Vehicular A model
Vehicular Speed	3 km/h, 50 km/h, 120 km/h

Notes:
1 Convolutional Coding
2 Discontinuous Transmission
3 Standard deviation of 8 dB and decorrelation distance of 100 m used.

In the simulations, each transmission scenario was distinguished based on the transmitted speech file number k, error condition FER, and the AMR codec rate R. In each case, the perceptual speech qualities for both $\mathbf{y}_k^{(n)}$ and its approximation $\hat{\mathbf{y}}_k^{(n)}$ were obtained by applying the input pairs $\{\mathbf{x}_k^{(n)}, \mathbf{y}_k^{(n)}\}$ and $\{\mathbf{x}_k^{(n)}, \hat{\mathbf{y}}_k^{(n)}\}$ to the PESQ algorithm to give the respective objective LQs

$$q_k = D\left(\mathbf{x}_k^{(n)}, \mathbf{y}_k^{(n)}\right) \quad for \quad n = 1,2, \dots, N_F \tag{13}$$

and

$$\hat{q}_k = D\left(\mathbf{x}_k^{(n)}, \hat{\mathbf{y}}_k^{(n)}\right) \quad for \quad n = 1,2, \dots, N_F \tag{14}$$

Following the mapping from $\hat{q}_k$ to q_k for each error condition FER and AMR codec rate R, the actual objective

listening quality q_k and its corresponding estimation $g(\hat{q}_k, FER, R)$ would be related according to

$$q_k = g(\hat{q}_k, FER, R) + \varepsilon_{k,FER,R} \quad (15)$$

where $\varepsilon_{k,FER,R}$ signifies the estimation error. It is noted that the vehicular speed v was not incorporated in the mapping function g. Although v is a factor in the wireless signal quality and the resulting speech quality, in practice, it is not readily measurable at the transmitter.

The mapping function was derived based on the training data according to a linear regression between $\hat{q}_k$ and q_k by minimizing the mean-squared-error in (15). Subsequently, the accuracy of the mapping function was validated based on the results from verification data. This was done by considering the mean estimation error μ, the RMSE σ, and the correlation coefficient ρ, between q_k and $g(\hat{q}_k, FER, R)$; k=1,2, … ,32.

V. Simulation Results

The scatter diagrams of the actual quality q_k against the quality estimate $\hat{q}_k$ for the training data for codec rates R=4.75 kb/s, 7.4 kb/s, and 12.2 kb/s are shown in Figs. 6-8, respectively. The regression line and the ideal mapping between $\hat{q}_k$ and q_k have also been shown for each case. Note that although the scatter diagrams have been plotted on a full MOS scale of 1-5, very low or very high scores were not obtainable because of FER restriction between 1% and 5%. It is observed from Figs. 6-8 that the proposed method would lead to an overestimation of the output quality demonstrated by the bias of the scatter diagrams with respect to the ideal case. The overestimation is due to exclusion of Class-B and -C errors from the approximation of the output signal. The bias can be removed by applying an appropriate regression line (based on *FER* and *R*) for mapping $\hat{q}_k$ to q_k. The corresponding mapping functions are given by

$$g(\hat{q}_k, FER, R) = \alpha_{FER,R} \cdot \hat{q}_k + \beta_{FER,R} \quad (16)$$

where the regression line parameters $\alpha_{FER,R}$ and $\beta_{FER,R}$ for the simulated conditions are summarized in Table II.

The verification data were used for validating the accuracy of the mapping functions in (16). In this case, for each combination of *FER* and *R*, the output quality for the k^{th} verification file was computed and the resulting value $\hat{q}_k$ was corrected based on the appropriate mapping function derived previously from the training data. The estimation error between q_k and the corresponding output of the mapping function was calculated according to

$$\varepsilon_{FER,R} = q_k - g(\hat{q}_k, FER, R) \quad (17)$$

The normalised histograms of the estimation errors $\varepsilon_{FER,R}$ corresponding to the simulated codec rates are shown in Figs. 9(a)-(c). It is observed that the distribution of the errors can be approximated with a normal probability density function (pdf). The mean μ and standard deviation σ (both expressed in MOS) of each pdf for the various simulation conditions are summarized in Table III. The mean of each pdf represented the average bias in estimation of the quality q_k. It is observed from the results of Table III that distribution means were practically equal to zero, confirming that the mapping functions on average removed the quality overestimation arising from overlooking Class-B and -C errors. The standard deviations of the estimation errors, corresponding to RMSE, ranged between 0.08 and 0.15 MOS points. In each pdf, the deviations of the estimation error from the mean were largely due to the mapping function's inability to take into account the effect of Class-B and -C bit errors exactly as well as the effect of the vehicular speed, which had not been accounted for in the proposed scheme. It is, nevertheless, noted that in all cases the RMSE were sufficiently small such that the associated quality differences would not be noticeable by human ears.

Table II

Summary of regression line parameters.

R	**4.75 kb/s**		
FER	**1%**	**3%**	**5%**
$\alpha_{FER,R}$	0.91	0.83	0.85
$\beta_{FER,R}$	0.06	0.21	0.08

R	**7.4 kb/s**		
FER	**1%**	**3%**	**5%**
$\alpha_{FER,R}$	0.96	0.79	0.83
$\beta_{FER,R}$	-0.06	0.44	0.26

R	**12.2 kb/s**		
FER	**1%**	**3%**	**5%**
$\alpha_{FER,R}$	0.93	0.84	0.78
$\beta_{FER,R}$	-0.11	0.09	0.21

Table III

Mean and standard deviation of each error distribution. In addition, the correlation coefficients between the actual and estimated LQO's are included.

R	**4.75 kb/s**		
FER	**1%**	**3%**	**5%**
μ	-0.01	-0.01	0.01
σ	0.11	0.12	0.12
ρ	0.87	0.86	0.84

R	**7.4 kb/s**		
FER	**1%**	**3%**	**5%**
μ	0.01	0.00	-0.01
σ	0.10	0.10	0.08
ρ	0.89	0.91	0.93

R	**12.2 kb/s**		
FER	**1%**	**3%**	**5%**
μ	-0.01	-0.01	0.00
σ	0.15	0.13	0.11
ρ	0.82	0.85	0.89

The correlation coefficient ρ for each case has also been calculated and included in Table III. The correlation coefficients ranged between 0.82 and 0.93, corresponding to RMSE values of σ=0.15 and σ=0.08, respectively. Such high correlation coefficients further confirm the accuracy of the proposed scheme for in-service monitoring the speech quality.

VI. Conclusions

A method for remote monitoring of the perceptual speech quality experienced by an end-user during a call has been described. This method relies on the feedback of the BFI flags from the receiver to construct an approximation of the received output signal. This approximation can be used in conjunction with the PESQ algorithm and an appropriate mapping function to provide a reasonably accurate measurement of the end-user quality. The accuracy of the proposed method was investigated through computer simulations for UMTS under a range of conditions. The results showed that a correlation coefficient of at least 0.82 was achieved between the PESQ listening quality based on the actual output signal and the listening quality obtained from the proposed method. In addition, the RMSE was at most 0.15 MOS points. Apart from its accuracy, the proposed scheme has the following desirable features. This method can be used for round-the-clock monitoring of the network speech quality without need for drive-tests. There is no additional traffic generated by this method as it passively monitors live traffic. The complexity of the scheme lies on the network side and not the end-user side. In fact all that is needed from the end-user side is the BFI flag which is generated for speech decoding in any case. The feedback of BFI is at a very low information bit rate of 50 b/s corresponding to one BFI for every 20 ms speech frame. Additionally, this method measures quality degradation due to radio link impairments only as opposed to the single-ended model ITUP.563 that evaluates the end-to-end quality inclusive of source signal degradation and loss along the signal path.

Acknowledgements

The authors would like to thank and acknowledge OPTICOM GmbH of Germany for the permission of using PESQ software for their academic research purpose. Also, we are grateful to PHYBIT, Inc. in Singapore for providing the simulation model for the UMTS link. This work was partly funded by National ICT Australia. National ICT Australia is funded through the Australian Government's Backing Australia's Ability initiative, in part through the Australian Research Council.

References

[1] "ITU-T Recommendation P.800: Methods for Subjective Determination of Transmission Quality," Aug. 1996.

[2] "ITU-T Recommendation P.830: Subjective Performance Assessment of Telephone-band and Wideband Digital Codecs," Feb.1996.

[3] "ITU-R Recommendation BS.1534: Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems," Jun. 2001.

[4] "ITU-R Recommendation BS.1284: General Methods for the Subjective Assessment of Sound Quality," Dec. 2003.

[5] W. Voiers, "Diagnostic Acceptability Measure for Speech Communication Systems," IEEE Int. Conf. Acoustics, Speech, Signal Processing, Connecticut, U.S.A., pp. 204-207, May 1977.

[6] G. H. Karlsson, T. B. Minde, M. Nordlund, and B. Timus, "Radio Link Parameter Based Speech Quality Index SQI," IEEE Speech Coding Workshop, pp. 147-149, Jun. 1999.

[7] M. Werner, K. Kamps, U. Tuisel, J. G. Beerends, and P. Vary, "Parameter-based Speech Quality Measures for GSM," IEEE Personal, Indoor and Mobile Radio Communications Conference, Beijing, China, pp. 2611-2615, Sep. 2003.

[8] M. Werner, T. Junge, and P. Vary, "Quality Control for AMR Speech Channels in GSM Networks," IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Montreal, Canada, pp.1076-1079, May 2004.

[9] A. W. Rix, M. P. Hollier, A. P. Hekstra, and J. G. Beerends, "PESQ, The New ITU Standard for Objective Measurement of Perceived Speech Quality, Part I – Time Alignment," J. Audio Eng. Soc., vol. 50, pp. 755-764, Oct. 2002.

[10] J. G. Beerends, A. P. Hekstra, A. W. Rix, and M. P. Hollier, "PESQ, The New ITU Standard for Objective measurement of perceived speech quality, Part II – Perceptual Model," J. Audio Eng. Soc., vol. 50, pp. 765-778, Oct. 2002.

[11] "ITU-T Recommendation P.810: Modulated Noise Reference Unit (MNRU)", Feb. 1996.

[12] S. Wang, A. Sekey, and A. Gersho, "An Objective Measure for Predicting Subjective Quality of Speech Coders," IEEE J. Sel. Areas in Commun., vol. 10, no. 5, pp. 819-829, Jun. 1992.

[13] J. G. Beerends and J. A. Stemerdink, "A Perceptual Speech-quality Measure based on a Psychoacoustic Sound Representation," J. Audio Eng. Soc., vol. 42, pp. 115-123, Mar. 1994.

[14] M. Hansen and B. Kollmeier, "Using a Quantitative Psychoacoustical Signal Representation for Objective Speech Quality Measurement," IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Munich, Germany, pp. 1387-1390, Apr. 1997.

[15] "ITU-T Recommendation P.861: Objective Quality Measurement of Telephone-band (300-3400 Hz) Speech Codecs, Perceptual Speech Quality Measure (PSQM)," Aug. 1996.

[16] "ITU-T Recommendation P.862: Perceptual Evaluation of Speech Quality (PESQ), An Objective Method for

End-to-end Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs," Feb. 2001.

[17] P. Gray, M. P. Hollier, and R. E. Massara, "Non-intrusive Speech Quality Assessment Using Vocal-tract Models," IEE Proc. of Vision, Image and Signal Processing, vol. 147, no. 6, pp. 493-501, Dec. 2000.

[18] J. Liang and R. Kubichek, "Output-based Objective Speech Quality," IEEE Veh. Techn. Conf., Stockholm, Sweden, pp. 1719-1723, Jun. 1994.

[19] C. Jin and R. Kubichek, "Vector Quantization Techniques for Output Based Objective Speech Quality," IEEE Int. Conf. on Acoustics, Speech, and Signal Proc., Georgia, U.S.A., pp. 7-10, May 1996.

[20] Doh-SukKim, "ANIQUE: An Auditory Model for Single-ended Speech Quality Estimation," IEEE Trans. on Speech and Audio Processing, vol. 13, no. 5, part 2, pp. 821-831, Sep. 2005.

[21] "ITU-T Recommendation P.563: Single-ended Method for Objective Speech Quality Assessment in Narrow-band Telephony Applications," May 2004.

[22] B. Rohani, B. Rohani, and H.-J. Zepernick, "Feedback Method for Real-time Perceptual Quality Estimation," IEE Electronics Letters, vol. 40, no. 14, pp. 913-915, Jul. 2004.

[23] B. Rohani, B. Rohani, and H.-J. Zepernick, "Frame Erasure Pattern Feedback for Real-time Perceptual Quality Estimation," Joint Int. Conf. on Information, Commun. and Signal Processing, and the Pacific Rim Conf. on Multimedia, vol. 1, pp. 110-113, Dec. 2003.

[24] B. Rohani, B. Rohani, and H.-J. Zepernick, "Application of a Perceptual Speech Quality Metric for Link Adaptation in Wireless Systems," IEEE Int. Symp. on Wireless Commun. Systems, pp. 260-264, Sep. 2004.

[25] H. Holma and A. Toskala, "WCDMA for UMTS-Radio Access for Third Generation Mobile Communications," John Wiley & Sons, Chichester, 2002.

[26] "3GPP Technical Report TR 25.922: Radio Resource Management Strategies," Mar. 2006.

[27] "3GPP Technical Report TR 25.944: Channel Coding and Multiplexing Examples," Jun. 2006.

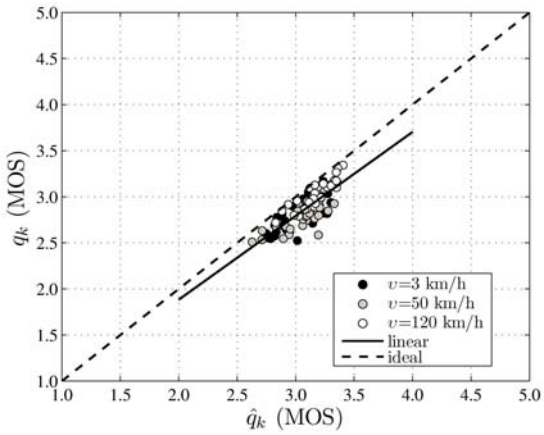
[28] "3GPP Technical Specification TS 25.212: Multiplexing and Channel Coding (FDD)," Jun. 2006.

[29] "3GPP Technical Specification TS 26.091: Adaptive Multirate (AMR) Speech Codec; Error Concealment of Lost Frames," Dec. 2004.

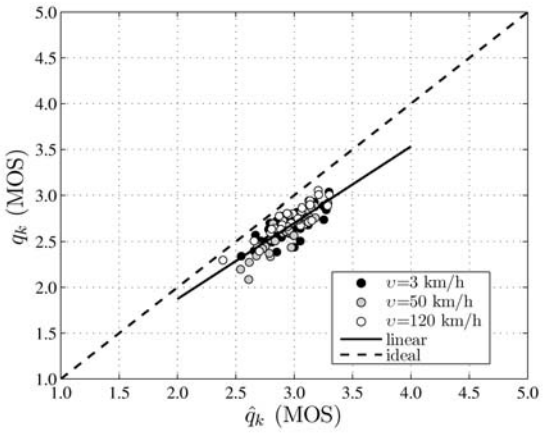
[30] "ITU-T Recommendation P.Supp 23: Coded Speech Database," Feb.1995.

[31] "3GPP Technical Specification TS 25.101: User Equipment (UE) Radio Transmission and Reception (FDD), Rel.7," Oct. 2006.

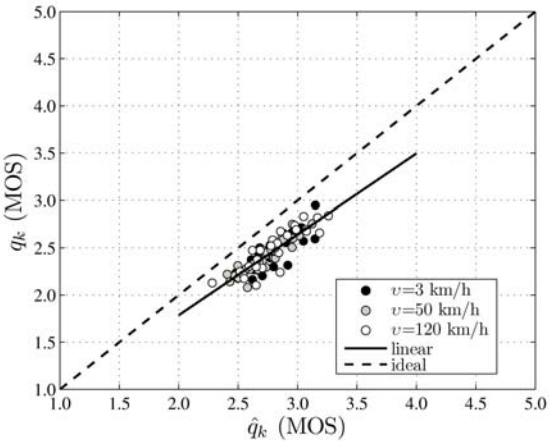
[32] "3GPP Technical Specifications, Series 25," [online] available: http://www.3gpp.org.



(a)

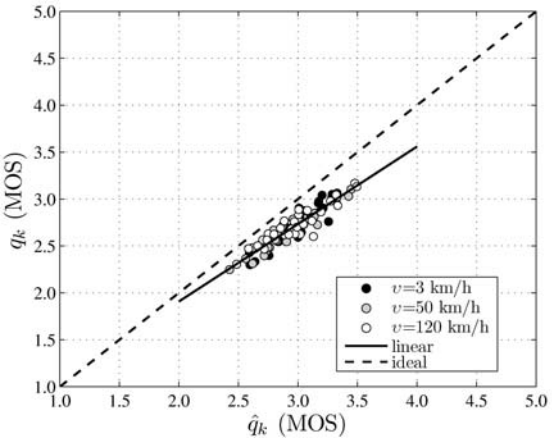


(b)

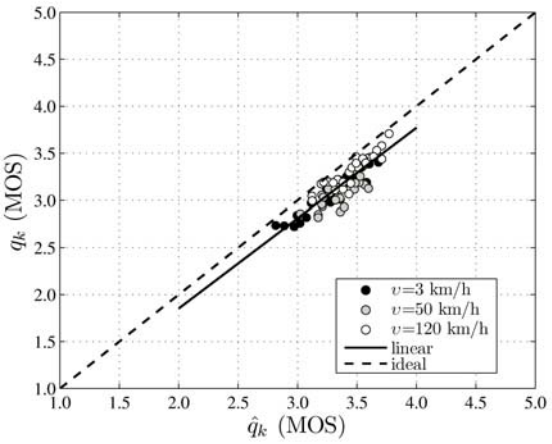


(c)

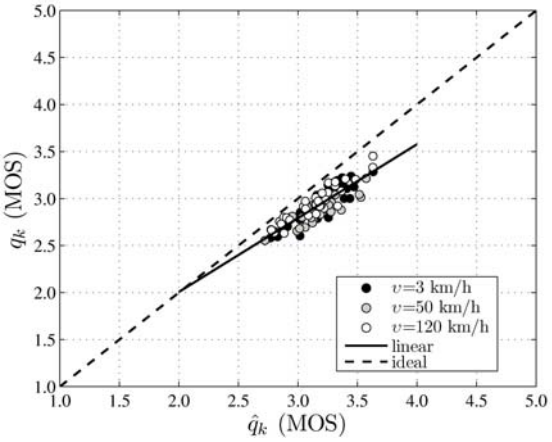
Fig. 6. Scatter diagrams of the actual quality q_k against the estimated quality $\hat{q}_k$ for training data with codec rate R = 4.75 kb/s for (a) *FER* = 1% (b) *FER* = 3%, and (c) *FER* = 5%.



(a)

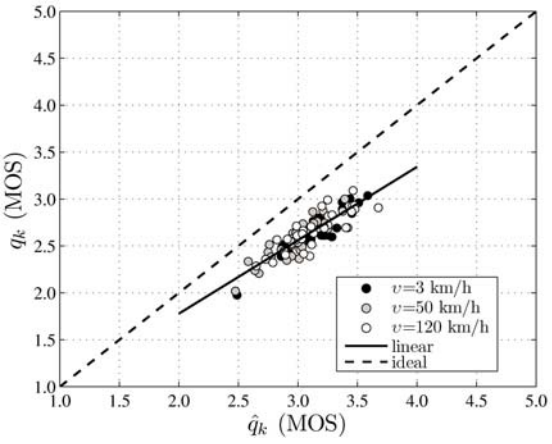


(b)

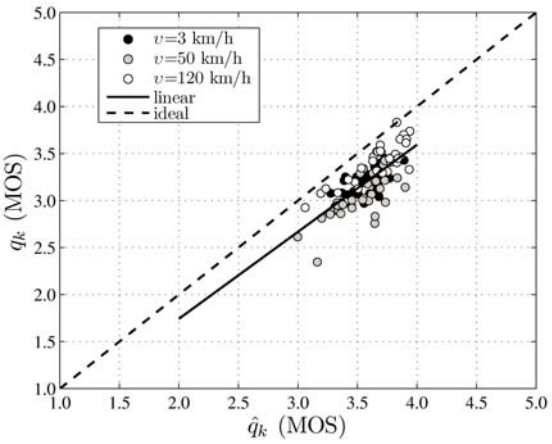


(c)

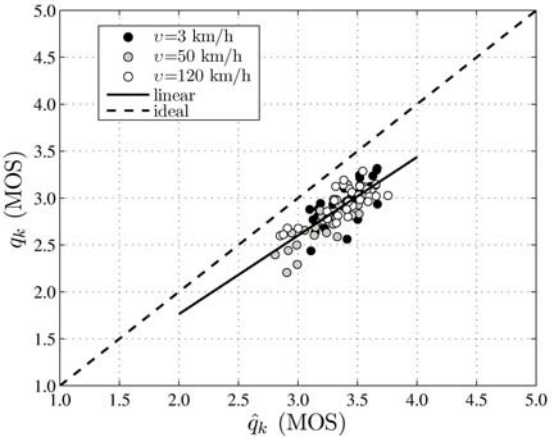
Fig. 7. Scatter diagrams of the actual quality q_k against the estimated quality $\hat{q}_k$ for training data with codec rate R = 7.40 kb/s for (a) *FER* = 1% (b) *FER* = 3%, and (c) *FER* = 5%.



(a)

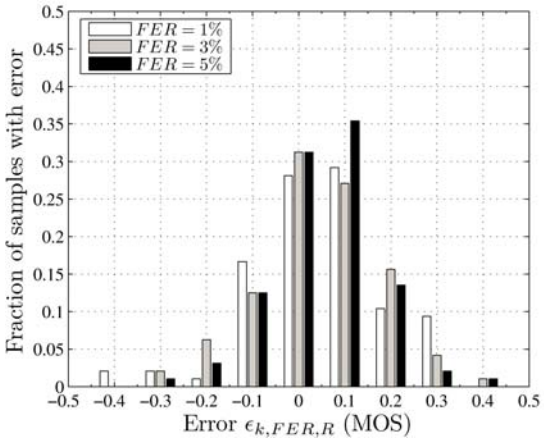


(b)

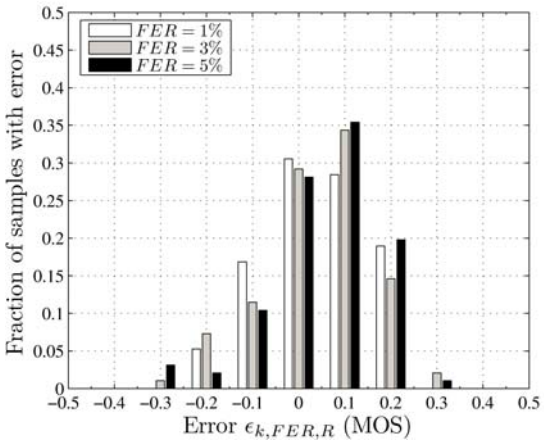


(c)

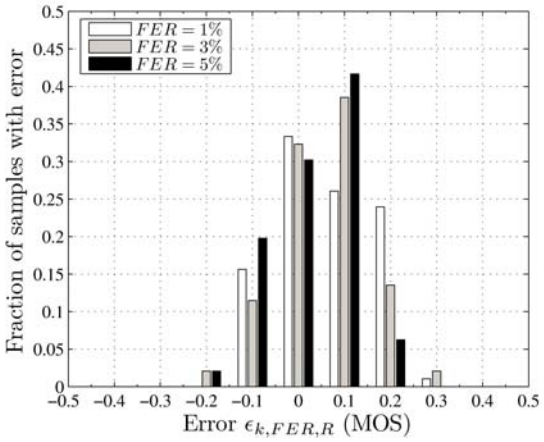
Fig. 8. Scatter diagrams of the actual quality q_k against the estimated quality $\hat{q}_k$ for training data with codec rate R = 12.20 kb/s for (a) *FER* = 1% (b) *FER* = 3%, and (c) *FER* = 5%.



(a)



(b)



(c)

Fig. 9. The distributions of the estimation errors $\varepsilon_{FER;R}$ for codec rates (a) R = 4.75 kb/s, (b) R = 7.40 kb/s, and (c) R =12.20 kb/s.