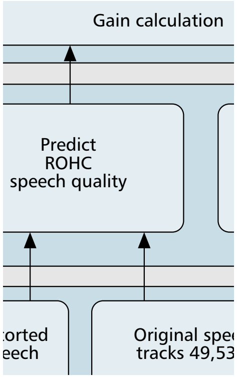
Accepted from Open Call

Voice Quality Evaluation in Wireless Packet Communication Systems: A Tutorial and Performance Results for ROHC

Stephan Rein, Technical University Berlin
Frank H. P. Fitzek, University of Aalborg
Martin Reisslein, Arizona State University



As wireless systems are evolving toward supporting a wide array of services, including the traditional voice service, using packet-switched transport, it becomes increasingly important to assess the impact of packet-switched transport protocols on the voice quality.

Abstract

As wireless systems evolve toward supporting a wide array of services, including traditional voice service, using packet-switched transport, it becomes increasingly important to assess the impact of packet-switched transport protocols on voice quality. In this article we present a tutorial on voice quality evaluation for wireless packet-switched systems. We introduce an evaluation methodology that combines elementary objective voice quality metrics with a frame synchronization mechanism. The methodology allows networking researchers to conduct effective and accurate quality evaluation of packet voice. To illustrate the use of the described evaluation methodology and interpretation of the results, we conduct a case study of the impact of robust header compression (ROHC) on the voice quality achieved with real-time transmission of GSM encoded voice over a wireless link.

Introduction

While the main service of the circuit-switched first- and second-generation wireless cellular systems has been voice, third-generation systems are being designed to support a wide range of services, including audio and video applications. This flexibility is achieved by employing packet-switched transport in conjunction with the Internet Protocol (IP). The development and refinement of packet-based transport over wireless systems has been and continues to be an active area of research and development. As novel communication and networking protocol mechanisms and refinements for wireless packet-switched transport are developed and wireless packet voice systems are deployed, it is important to evaluate the performance of the transport protocol mechanisms and refinements not only in terms of network metrics such as packet loss, delay, and jitter, but also in terms of the subjective quality experienced by voice users. Generally, when evaluating the quality of packet voice one may distinguish between three qualities: the network quality, the objective quality, and the subjective quality, as illustrated in Fig. 1. While the network quality reflects the provider's perspective, objective and subjective quality reflect the customer's perspective. The network quality can be relatively easily measured by network parameters, such as the packet loss rate or packet delay or jitter. Subjective quality is generally more meaningful than network quality, as it relates directly to user-perceived quality. Assessing subjective voice quality, however, is very tedious as it requires listening tests with a large number of test subjects. For this reason, objective quality measures that predict subjective quality are typically employed in the evaluation of voice transmission systems.

In this article we describe an evaluation methodology for the transmission of packet voice over a wireless system. We first give a tutorial introduction to elementary objective voice quality metrics. We then describe an evaluation methodology that allows computationally efficient and accurate voice quality evaluations without requiring specialized software. Our evaluation methodology employs a wide array of objective voice quality metrics, including both traditional and segmented signal-to-noise ratio (SNR), spectral distance metrics, and parametric distance metrics. The considered parametric distance metrics include the cepstral distance metric, which can be transformed into the mean opinion score (MOS), thus enabling us to quantify the effect of a protocol mechanism or refinement on voice quality in terms of the MOS.

We illustrate the use of our evaluation methodology by applying it to the problem of assessing the

A related conference paper appears in the Proceedings of the IASTED International Conference on Internet and Multimedia Systems and Applications (IMSA), *Honolulu, HI, August 2003, pp. 461–66.*

Stephan Rein was with the Communication Systems Group, Technical University Berlin, Germany. He is now with the Electronic Measurement and Diagnostics group. Part of this work was conducted while he was visiting Arizona State University in 2003.

impact of robust header compression (ROHC) on voice quality. In particular, we compare the voice quality achieved in a wireless system without ROHC with that achieved in a wireless system with ROHC. We find in our evaluation that for a wide range of bit error probabilities, ROHC improves the voice quality and at the same time reduces the protocol overhead for voice transmission with IPv4 by approximately 85 percent, which reduces the bandwidth required for a GSM coded voice transmission by about 47 percent.

This article is organized as follows. We describe the overall evaluation setup. We explain how to evaluate the objective voice quality using an array of metrics ranging from SNR-based metrics to spectral and parametric distance metrics based on a linear predictive coding (LPC) analysis. We present the segmental cross correlation (SCC) algorithm for synchronizing the original voice stream with the voice stream after network transport. We apply our evaluation methodology to evaluate the impact of ROHC on voice quality. We then summarize our contributions.

Evaluation Methodology

In this section we give a general overview of the system setup for voice quality evaluation. In an evaluation one is often interested in the change in voice quality caused by a refinement or modification to a basic communication system. To keep the following discussion concrete we consider the addition of ROHC to a standard wireless communication system with the RTP/UPD/IP protocol stack, illustrated in Fig. 2. In this example the basic communication system consists of the sender and receiver protocol stacks containing the RTP, UDP, IP, and link protocol layers, but not the ROHC protocol layer. The modified system consists of the protocol stacks including ROHC, depicted in Fig. 2. We emphasize that the addition of ROHC is only considered as an illustrative example. The evaluation methodology presented here and the following sections can be applied in analogous fashion to other refinements or modifications to the communication or networking protocols or mechanisms.

Our evaluation methodology employs a set of original speech files that consist of a sequence of voice signal samples. In our example evaluations we use tracks 49, 53, and 54 of the sound quality assessment material from the European Broadcasting Union, as illustrated in the center of Fig. 3. The original voice files are fed into the input of the communication system without and with the modification under study; in our illustrative example a real-time voice transmission system with a GSM codec and the RTP/UPD/IP and link layer protocol stack, without and with the added ROHC. Following the transmission over the wireless link, which we simulate in our example evaluation, the voice packets pass up through the protocol stack on the receiver side to the GSM decoder. The GSM decoder decompresses each received GSM frame into a sequence of audio samples, which are the output of the communication system under study. Importantly, the wireless link errors typically result in decoded voice signal samples that differ from the original voice signal samples (i.e., the wireless link errors result in distortion of the speech).

The communication system both without and with the considered refinement or modification gives rise to distorted speech at the output. To assess the impact of the system modification on voice quality we need to compare the speech distortions from the two systems in a meaningful manner. Toward this end we predict the subjective speech quality for the communication system without and with the modification. In particular, we employ the objective voice quality metrics detailed in the next section to predict the subjective speech quality. As a final step we compare the predicted subjective speech qualities to calculate the gain in voice quality, as also detailed in the next section.

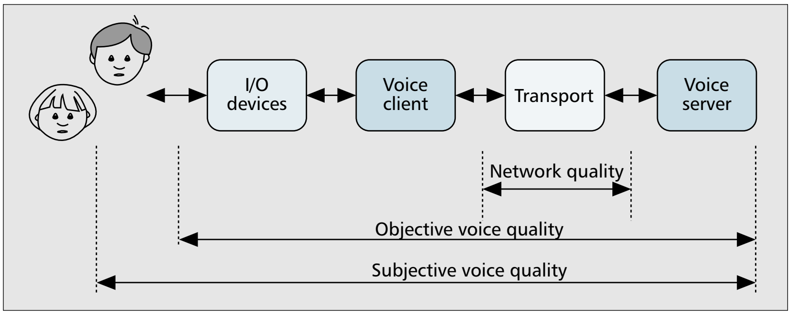


Figure 1. *Different perspectives on quality in performance evaluation of packet voice.*

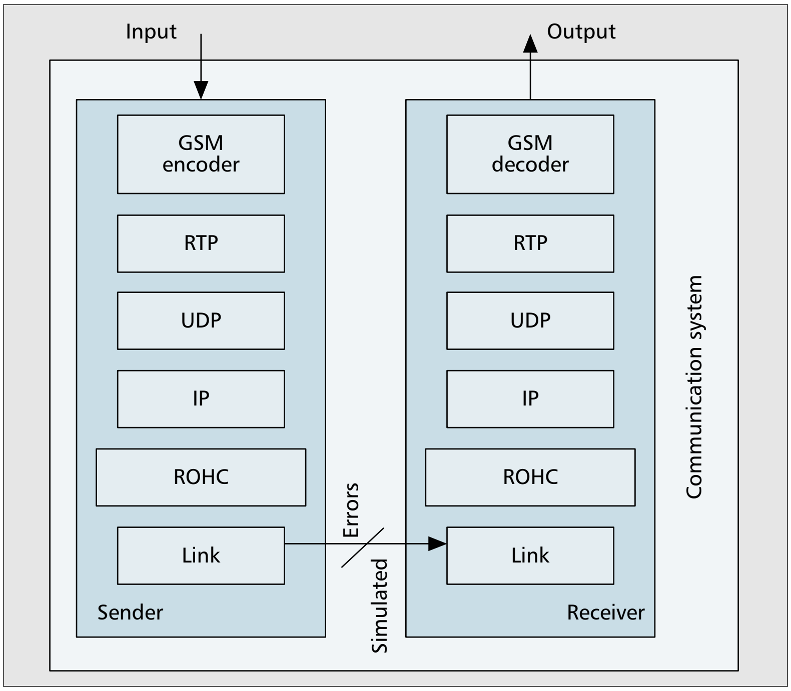


Figure 2. *The protocol stack of a typical wireless packet voice communication system. As an example system modification we consider the impact of ROHC.*

Notation

Before we proceed to the voice quality evaluation we introduce the following basic notation for voice signal samples. For calculation of the objective quality metrics a given sequence of voice signal samples is broken into analysis frames of 20 ms duration, which are introduced for the voice quality evaluation in accordance with the temporal resolution of the human ear. Let N denote the total number of frames in a given voice file. Let M denote the total number of samples in a given frame n, $n = 1, \ldots, N$, and note that with a typical sample rate of 8 kHz an analysis frame contains $M = 160$ samples. Let m, $m = 1, \ldots, M$ index the individual samples within a given frame. Throughout we denote ϕ for the undistorted signal and d for the distorted signal (from the output of the communication system). Let $x_{n,\phi}(m)$ denote the amplitude of sample m in frame n of the undistorted voice signal, and let $x_{n,d}(m)$ refer to the distorted sample.

Voice Quality Evaluation

Expensive and time consuming speech perception tests with human listeners as detailed in International Telecommunication Union — Telecommunication Standardization Sector (ITU-T) Recommendation P.800.1 are required to reliably obtain the subjective voice quality achieved by a communication system. The subjective voice quality is typically given on the 5-point MOS scale, which ranges from 5 (excellent) to 1 (bad). To avoid the expense and effort required for subjective voice quality evaluation, significant effort has been devoted to developing objective computer-based metrics that predict the results of a subjective evaluation [1].

Overview of Objective Voice Quality Metrics

Generally, there are three classes of objective voice quality evaluation metrics: network-parameter-based metrics, psycho-acoustic metrics, and elementary metrics. Parameter-based metrics do not consider the actual voice signal. Instead, these metrics sum impairment factors that characterize the individual components of the communication system. The packet loss and delay in a packet voice system, for instance, are translated into impairment factors according to provisional translation tables in the ITU-E-model, which is one recent proposal for a parameter-based metric. Parameter-based metrics such as the E-model hold promise for predicting subjective voice quality but still require extensive refinements and verifications.

Psycho-acoustic metrics transform voice signals to a reduced representation to retain only perceptually significant aspects. These metrics aim to predict the subjective quality over a wide range of voice signal distortions, allowing for the development as well as evaluation of non-waveform-preserving speech coding algorithms. These coding algorithms perform waveform distortions that are perceptually insignificant. Various complex metrics have been developed and refined over the last decade. These include the Bark spectral distance, the measuring normalizing blocks (MNB) technique [2], and the PESQ measure [3], which was recently standardized by ITU-T as Recommendation P.862.

Elementary objective voice quality metrics rely on low-complexity signal processing techniques to predict subjective voice quality. Elementary metrics generally have smaller correlations with subjective voice quality than highly complex psycho-acoustic metrics and do not provide the perception modeling needed for psycho-acoustic coder algorithm development. Elementary metrics, however, do represent a good engineering trade-off for communication and networking system researchers and developers in that they allow for fairly detailed conclusions about voice quality while having low computational complexity. We also note that in our evaluation methodology, as illustrated in Fig. 3, we focus on system modification in the networking domain (e.g., the introduction of ROHC). Both, the unmodified (without ROHC) and modified (with ROHC) systems employ the same voice codec and thus experience approximately the same voice codec distortions. Our evaluation methodology is focused on the impact of the modification in the communication or networking system on voice quality, and is not designed to evaluate voice codec distortions.

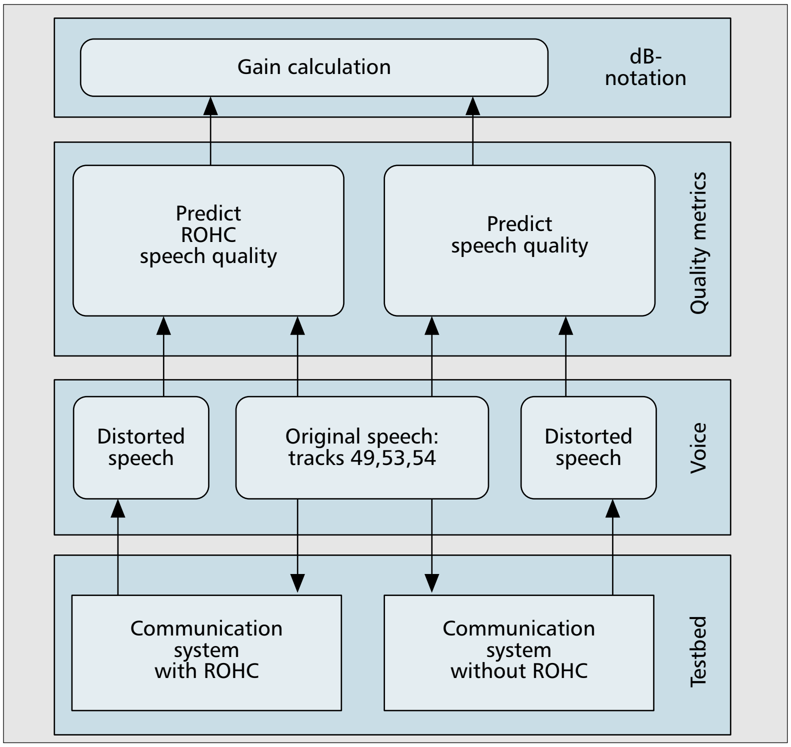

Gain calculation
dB-notation
Predict ROHC speech quality
Predict speech quality
Quality metrics
Distorted speech
Original speech: tracks 49,53,54
Distorted speech
Voice
Communication system with ROHC
Communication system without ROHC
Testbed

Figure 3. *Methodology for assessing the impact of a system modification, the addition of ROHC in the considered example. The distorted speech from the system with and without the modification is compared with the original speech signal to determine the speech quality with and without the refinement. The two speech qualities are then compared to determine the quality gain achieved by the system modification.*

Evaluation Methodology Based on Elementary Objective Metrics

We have selected the elementary metrics listed in Table 1 for our evaluation methodology. The reliability of objective voice quality metrics is usually verified by a correlation analysis between the calculated objective metric and subjective hearing tests among a distorted database. Table 1 gives the distortion types for which the various objective metrics have been examined and the resulting correlations to subjective hearing tests. The larger the magnitude of correlation, the better the prediction of subjective voice quality. We note that the traditional SNR has poor correlation performance. However, we include it because it is often considered a purely objective quality metric. The traditional SNR aggregates the signal energy in the entire file and relates this aggregate signal energy to the aggregate noise energy. Thereby soft and loud voice analysis frames are not equally weighted. More formally, the signal energy $S(n)$ and noise energy $N(n)$ of frame n are given by

$$S(n)=\sum_{m=1}^{M} x_{n,\phi}^{2}(m) \tag{1}$$

and

$$N(n)=\sum_{m=1}^{M}\left[x_{n,d}(m)-x_{n,\phi}(m)\right]^{2}. \tag{2}$$

The traditional SNR is given by

$$D_{\text{trad}}=10\cdot\log_{10}\frac{\sum_{n=1}^{N}S(n)}{\sum_{n=1}^{N}N(n)}. \tag{3}$$

In contrast, the segmental (short-time or framed) SNR relates the signal energy of each individual frame to the noise energy of the corresponding frame, formally,

$$D_{\text{seg}}=10\cdot\log_{10}\sum_{n=1}^{N}\frac{S(n)}{N(n)}. \tag{4}$$

This finer granularity relates more meaningfully to the perception of the voice file.

Spectral distances measure the distortions of the frequency amplitudes (see [8] for details) and represent meaningful speech recognition features over a wide range of voice signal distortion types. The inverse linear unweighted and unweighted delta form spectral distances revealed superior performance among all spectral distances in [4]. The root mean square (RMS) spectral distance is included because in [9] it is shown that it is a very meaningful measure for speech perception, as it can be physically interpreted and efficiently computed.

Parametric distances use transformations of the linear predictive coding (LPC) coefficients, which are standard signal descriptors in signal processing. We consider three classes of parametric distance measures:

- The *log area ratio* measure
- The *energy ratio/log likelihood* measure
- The LPC *cepstral* distance measure

These three classes of measures allow comparisons of the spectra without calculating computationally demanding Fourier transformations. In signal communications the cepstral distance is a widely employed reference measure for calculating the difference in shape of the original and distorted spectra. Its general applicability for speech quality evaluation was discovered by Kitawaki *et al.* [6], who compared elementary objective speech quality measures for voiceband codecs. The cepstral distance revealed the best correspondence to the MOS of all objective measures studied. These results are confirmed by Wu and Pols [7], who estimated a correlation of 0.926 for the LPC cepstral distance measure with the MOS. This correlation performance has been further verified for waveform preserving codecs and the MNRU, which is one of the most common reference conditions for subjective and objective voice quality assessments, as part of the recent study by Voran [2]. Because of its widely verified correlation performance to subjective hearing tests, we use the results of the fundamental study [6] to predict the MOS from the cepstral distance, as detailed later.

As illustrated in Fig. 4, many metrics use the same coefficients and are similarly calculated. Thus, our approach represents a framework of voice quality metrics allowing computationally effective voice quality evaluation. Each metric gives a distortion index $F(n)$ for a given frame n, as detailed in [8]. The total quality D of a given distorted voice file with respect to the corresponding undistorted file is typically obtained by averaging the individual distortion indices,

$$D=\frac{1}{N}\sum_{n=1}^{N}F(n). \tag{5}$$

A slightly more complex approach may weigh the distortion indices of the individual frames by the corresponding signal energies, but this weighting typically has negligible impact on the total quality. Equation 5 is only used with the spectral and parametric measures, because the SNR metrics directly give the total quality.

Objective metric	Correlation
(Traditional) SNR	+0.24[1]/+ 0.31[2]
Segmental SNR	+0.77[1]/+ 0.78[2]
Spectral distances	
Inverse linear unweighted distance	+0.63[3]/+ 0.48[4]
Unweighted delta form	–0.61[3]
Log root mean square	Theoretical approach
Parametric distances	
Log area ratio	–0.62[3]/ –0.65[4]
Energy ratio	–0.59[3]/–0.61[4]
Log likelihood	–0.49[3]/ –0.48[5]
Cepstral distance	–0.96[6]/–0.95[7]/–0.93[8]

[1] Waveform coders: 8 types: [4]; [2] additive and narrowband noise: [4]; [3] coding distortions, controlled distortions, and narrowband distortions (23 types): [4]; [4] waveform coders and controlled distortions (18 types): [4]; [5] cellular phone: [5]; [6] coding and other nonlinear distortions: [6]; [7] PCM, ADPCM, G.728, MNRU: [2]; [8] noise masking, bandpass filtering, echo, and peak clipping: [7].

Table 1. *Correlations between objective voice quality metrics and subjective voice quality. The distortion types are indexed by footnotes 1–8.*

Evaluation of Voice Quality Gain

To evaluate the impact of a communication system modification such as the addition of ROHC on voice quality we obtain the total quality both without the modification (denoted D) and with the modification (denoted D_{ROHC} for the considered addition of ROHC) for the objective quality metrics described above. For ease of evaluating the voice quality improvement (gain) achieved by the system modification under study we define the gain metrics in dB in Table 2. (The right half of the table containing the mapping function can be ignored for now.) Positive gains indicate improved voice quality while negative gains indicate deteriorated voice quality. Note from Table 1 that the SNR and inverse linear spectral distance have positive correlations with subjective voice quality (i.e., $D_{ROHC} \geq D$ indicates higher voice quality). All other metrics have a negative correlation with subjective voice quality; thus, $D_{ROHC} \leq D$ indicates improved voice quality. For metrics that involve a logarithm (i.e., SNR, segmental SNR, RMS distance, log area ratio, log likelihood) we define the gain in dB as the difference of the metric values. For the inverse linear spectral distance and unweighted delta spectral distance (which do not employ a logarithm) we use the standard dB formula to obtain the dB gain. For the energy ratio we use 10 as the multiplicative factor (and a power of 4 to compensate for the power of 1/4 in the metric definition [8]) in the gain definition to make it comparable to the closely related log likelihood. We note that we adopt these dB gain definitions to facilitate comparison of the results of the different metrics and also note that other definitions are possible.

Voice Quality Gain on MOS Scale

Finally, in order to asses the impact on user perception we study the impact of the system modification on the subjective 5-point MOS scale. We transform the values of the cepstral distance to the predicted MOS using the mapping verified in [6]. Let D_{cep} denote the voice quality calculated by the cepstral distance. The MOS value is given by

$$MOS = 3.56 - 0.8 \cdot D_{cep} + 0.04 \cdot D_{cep}^2. \qquad (6)$$

We note that the absolute MOS values obtained with this mapping need to be interpreted with caution; however, the relative difference in the MOS between a base system and a modified system is meaningful [7]. In the context of our illustrative example, we define the MOS gain for the addition of ROHC to the communication system as

$$MOS_{gain} = MOS_{wROHC} - MOS_{w/oROHC}. \qquad (7)$$

We close this tutorial on voice quality evaluation by noting that we have chosen the elementary metrics in Table I as they represent a sensible engineering approach for the evaluation of communication or networking system modifications. The chosen elementary metrics have good correlations with the subjective voice quality and thus allow for meaningful conclusions about voice quality. At the same time the chosen metrics are computationally efficient and do not require proprietary software (in fact we make our evaluation software source code publicly available [8]). In order to cover a reasonably wide range of distortion types we selected a set of elementary metrics (Table 1). We provide a technique for verifying the correlation of the other elementary metrics to the cepstral distance (and thus to the MOS).

The Segmental Cross Correlation Algorithm

We consider the transfer voice over a communication system. Thereby, voice frames may:
- Be completely lost
- Experience varying delays
- Suffer voice signal distortions due to bit errors

Objective voice quality is based on a comparison between the received (distorted) and original (reference) voice streams, which need to be synchronized for comparison. There are generally two types of approaches to synchronize the streams:
- Packet-based approaches
- Voice-signal-based approaches

Packet-based approaches employ timestamps and sequence numbers (e.g., using RTP) to detect lost packets and varying packet delays, and compensate for these effects by replacing lost packets (i.e., by using interpolation techniques) and adjusting the playout time of the voice frames. Voice-signal-based approaches, on the other hand, employ signal correlation techniques to align frames in the distorted and reference streams (e.g., [3]). For signal-based synchronization we employ the segmental cross correlation (SCC) algorithm, which we outline in this section and use in our case study reported in the following section. We note that the voice quali-

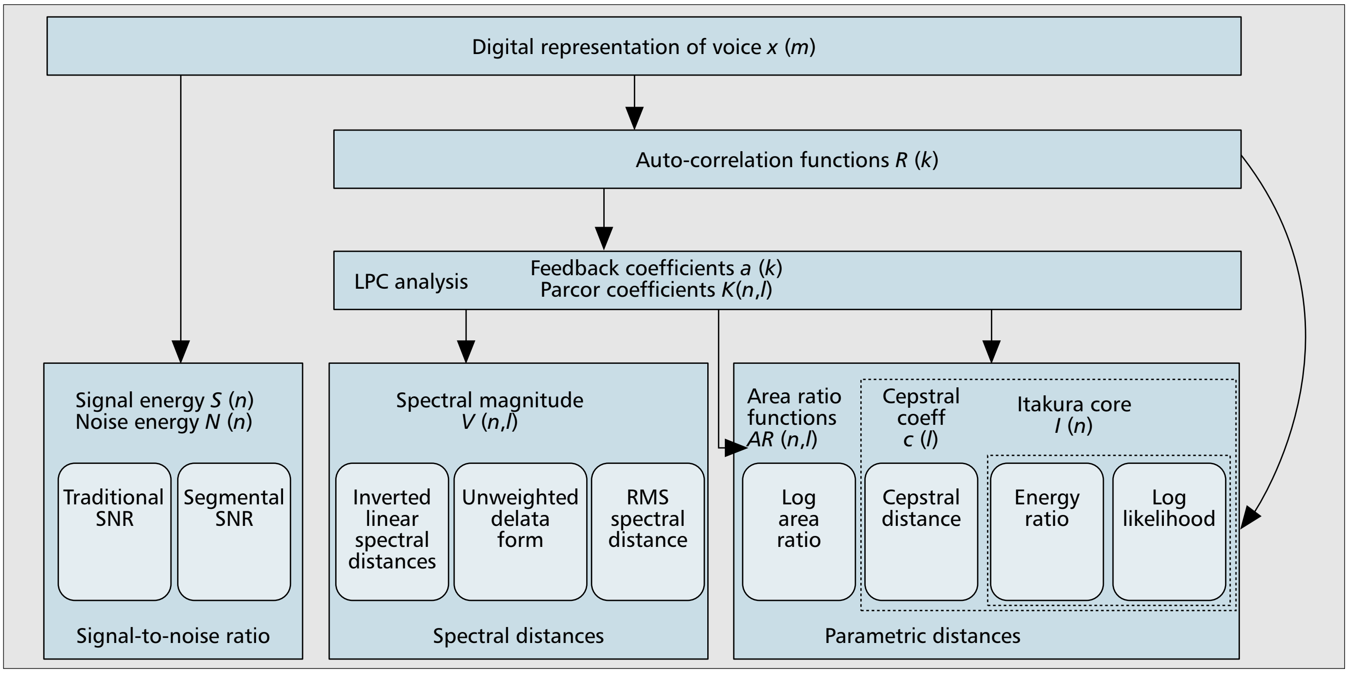


Figure 4. *A framework of the used objective voice quality metrics. The calculations are partially similar, but the metrics cover different types of distortions.*

Metric	Gain [dB]	Mapping function	Symbol
SNR	$D_{ROHC} - D$		
Segmented SNR	$D_{ROHC} - D$		
Inv. lin. spectral distance	$20 \cdot \log (D_{ROHC}/D)$	D_{cep} = -5281.82D + 105.98	◊
Unw. delta spectral distance	$20 \cdot \log (D/D_{ROHC})$	D_{cep} = 17.65D + 0.38	▽
RMS spectral distance	$D - D_{ROHC}$	D_{cep} = 12.89D + 0.44	△
Log area ratio	$D - D_{ROHC}$	D_{cep} = 0.46D + 0.23	°
Energy ratio	$10 \cdot \log (D/D_{ROHC})4$	D_{cep} = 8.17D – 7.40	□
Log likelihood	$D - D_{ROHC}$	D_{cep} = 0.29D + 0.74	×

Table 2. *Gain definitions for different metrics and linear mappings of LPC based metrics* D *to the cepstral distance* $\mathrm{D_{cep}}$*. the symbols are used in the scatter plot Fig. 6.*

ty evaluation methodology presented in the previous section can be employed in conjunction with both packet- and signal-based synchronization.

Due to space constraints we give here only an outline of a simplified version of the SCC algorithm and refer the interested reader to [8] for details on the full algorithm. For synchronization the reference file is divided into consecutive synchronization frames of U samples each. The goal of synchronization is to divide the distorted file into synchronization frames such that a frame in the distorted file matches well with the corresponding frame in the reference file. More formally, let $x_{w,\phi}(u)$, u = 1, ..., U, denote the sample values in synchronization frame w in the reference file. Let $x_d(\cdot)$ denote the sample values in the (unframed) distorted file. The algorithm is based on the normalized SCC function

$$SCC_w(\tau) = \frac{\sum_{u=1}^{U}\left[x_{w,\phi}(u) - \bar{x}_{w,\phi}\right] \cdot \begin{bmatrix} (x_d(u+(w-1)U+\tau) \\ -\bar{x}_d(w,\tau) \end{bmatrix}}{\sqrt{\sum_{u=1}^{U}\begin{bmatrix} x_{w,\phi}(u) \\ -\bar{x}_{w,\phi} \end{bmatrix}^2} \sqrt{\sum_{u=1}^{U}\begin{bmatrix} (x_d(u+(w-1)U \\ -\tau) - \bar{x}_d(w,\tau) \end{bmatrix}^2}}, \tag{8}$$

where we denote

$$\bar{x}_d(w,\tau) = \frac{1}{U}\sum_{u=1}^{U} x_d(u+(w-1)U+\tau).$$

For the first frame, w = 1, in a file the cross correlation is initially evaluated for a search range $0 \le \tau \le R$. The displacement between the frame in the reference file and the distorted file is tentatively estimated as the displacement that attains the maximum correlation, ithat is,

$$\tau_{\max}(w) = \arg \max_{0 \le \tau \le R} SCC_w(\tau). \tag{9}$$

If this maximum cross correlation is larger than a threshold, the displacement estimate (match) is accepted; otherwise, the search range is increased. For the subsequent frames w, $w \ge 2$, the cross correlation is initially evaluated for the search range $\tau_{\max}(w-1) - R \le \tau \le \tau_{\max}(w-1) + R$; that is, the search range is adaptively shifted according to the displacement of the preceding frame w – 1, as detailed in [8] where we also provide fast Fourier transform techniques to reduce the computation time of the SCC algorithm.

We finally note that both the elementary and psycho-acoustic voice quality metrics described earlier generally do not include synchronization and therefore cannot be used to directly evaluate the received voice signal after packetized transport. The main innovation of PESQ [3] over previous perceptual metrics is the signal-based synchronization of the received voice signal. PESQ, which requires the purchase of proprietary software, performs highly complex algorithms in the time and frequency domain [3] and may give better synchronization performance than the SCC algorithm. However, the SCC algorithm, which has low complexity and for which we make the source code publicly available [8], does allow for meaningful delay jitter measurements in the received voice signal, as presented in [8], and synchronizes the voice signals to allow for the objective voice quality evaluations presented in the next section.

Case Study: The Impact of ROHC on Voice Quality

The purpose of the case study presented in this section is to illustrate the use of the evaluation metrics presented in the preceding section and give an example of how to interpret the results obtained from an evaluation. In this case study we examine the impact of adding ROHC [10] to a basic wireless packet voice communication system, as illustrated in Fig. 2. ROHC was recently developed to reduce the overhead due to protocol headers, which typically result in voice packets consisting of 30 bytes of compressed voice data and 40 bytes of RTP/UDP/IP headers. ROHC exploits redundancies between the headers in successive packets of a given voice flow to compress the protocol headers.

We note that the impact of header compression on voice quality has received very little attention so far. The only study in this direction of which we are aware is [11]. In [11] objective speech quality degradation is studied (using the traditional SNR which has only a weak correlation with user perception) for robust checksum-based compression (ROCCO) and the Compressed Real-Time Protocol (CRTP), which may be considered precursors to ROHC. In contrast, in this case study we consider the state of-the-art ROHC compression scheme and evaluate voice quality using our evaluation methodology which employs an array of objective metrics that allows for accurate prediction of the subjective voice quality of hearing tests. We give here only a brief overview of our evaluations of voice

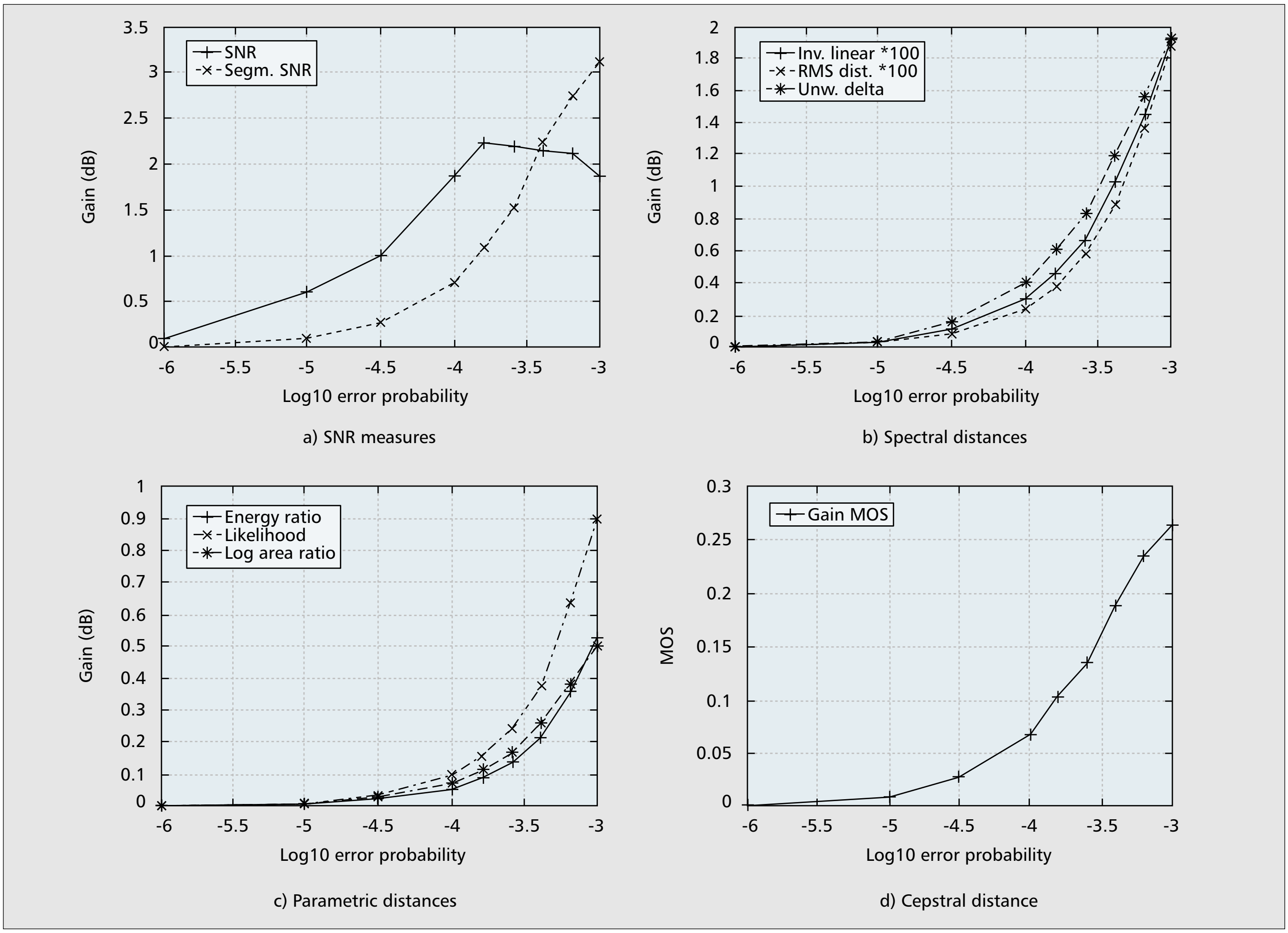
The goal of the synchronization is to divide the distorted file into synchronization frames such that a frame in the distorted file matches well with the corresponding frame in the reference file.

transmission with ROHC and refer the interested reader to [8] for a more extensive evaluation.

In Figs. 5 a–c, we plot the voice quality gain from the addition of ROHC (in dB) for the objective metrics described earlier as a function of the logarithm with base 10 of the bit error probability on the wireless link. We observe that all metrics indicate an increasing positive gain with larger bit error probabilities. As an exception, the gain for the traditional SNR decreases for bit error probabilities above $10^{-3.8}$. Because of the unequal weighing of soft and loud frames, the traditional SNR reveals here its worse granularity. The SNR measures indicate a gain between 2 and 3 dB for link error probabilities in the $10^{-3.4}$ to 10^{-3} range. Similarly, the spectral distances indicate gains between 0.02 and 2 dB for link error probabilities of 10^{-3}, and the parametric distances give gains between 0.5 and 1 dB. Overall, these results indicate that voice quality does not suffer from the addition of ROHC to the base system. On the contrary, it is improved, especially for large bit error probabilities on the wireless link.

Next, we consider the change in voice quality due to the addition of ROHC on the 5-point MOS scale, using the mapping from the cepstral distance to the MOS given in Eq. 7. In Fig. 5d we plot the gain in voice quality from the addition of ROHC in terms of MOS as a function of the bit error probability. We observe that the gain in MOS increases roughly exponentially with increasing error probability and reaches 0.26 for error probabilities of 10^{-3}.

Relationship between Quality Metrics — Generally, in objective voice quality evaluation it is advisable to consider a variety of metrics since each individual metric (including the cepstral distance used to evaluate MOS_{gain}) has been evaluated for a limited set of distortions (Table 1). We therefore now describe a technique for examining the correlations between the total objective quality D_{cep} obtained with the cepstral distance and the corresponding quality D obtained with the other individual LPC analysis-based metrics. We examine these correlations by means of a scatter plot, which is generated as follows. We express the qualities D of the other LPC-based metrics as a linear function of the cepstral distance quality D_{cep}. We determine the slope and offset of these linear functions by considering the D and D_{cep} obtained for the bit error probabilities of 10^{-6} and 10^{-3} for the base system (without ROHC). The resulting linear mappings are reported in Table 2. Next, we plot the D_{cep} obtained by these linear mappings as a function of the actual measured D_{cep}, resulting in the scatter plot in Fig. 6. In the plot the filled (shaded) symbols correspond to the qualities with ROHC. The unfilled symbols correspond to the qualities without ROHC. We observe that the points are fairly closely scattered



a) SNR measures

b) Spectral distances

c) Parametric distances

d) Cepstral distance

■ **Figure 5**. *Gain in voice quality with addition of ROHC as a function of bit error probability: a) SNR measures; b) spectral distances; c) parametric distances; d) cepstral distance.*

around a straight line with slope one. This indicates that there is a high correlation between the total qualities D obtained with the considered LPC-based metrics, and the total quality D_{cep} obtained with the cepstral distance.

Conclusions

In this article we provide a tutorial on a methodology for evaluating voice quality in wireless packet voice systems. Our methodology employs elementary objective voice quality metrics that predict subjective voice quality with good reliability. In addition, we have provided a segmental cross correlation algorithm for voice-signal-based synchronization of the received (distorted) voice signal with the original (reference) signal. Our tutorial makes the objective voice quality metrics and SCC algorithm readily accessible and employable by networking researchers to evaluate novel protocol mechanisms and refinements for wireless voice communication and networking systems.

We have applied our evaluation methodology to assess the impact of ROHC on a wireless voice communication system. We have found that the addition of ROHC improves voice quality. The improvement reaches 0.26 on the 5-point MOS for a wireless bit error probability of 10^{-3}. This result in conjunction with the result that ROHC cuts the total bandwidth required for voice transmission almost in half [8] indicates that by adding ROHC, the number of third-generation mobile cell phone users could nearly be doubled without allocating more link bandwidth and without compromising voice quality.

Acknowledgment

This work is supported in part by the National Science Foundation under Grant No. Career ANI-0133252 and Grant No. ANI-0136774, and the state of Arizona through the IT301 initiative.

We are grateful for interactions with Professor Thomas Sikora of Technical University Berlin, Germany, throughout this work. We are also indebted to Patrick Seeling of Arizona State University (formerly with acticom GmbH), who helped with setting up the experimental testbed and scripting of the measurement experiments.

We thank acticom GmbH for providing the acticom ROHC source code.

References

[1] ETSI EG 201 377-1 V1.2.1 (2002-12), "Speech Processing, Transmission and Quality Aspects (STQ); Specification and Measurement of Speech Transmission Quality; Part 1: Introduction to Objective Comparison Measurement Methods for Oneway Speech Quality Across Networks," Dec. 2002.

[2] S. Voran, "Objective Estimation of Perceived Speech Quality, Part II: Evaluation of the Measuring Normalizing Block Technique," *IEEE Trans. Speech and Audio Proc.*, vol. 7, July 1999, pp. 383–90.

[3] A. W. Rix *et al.*, "Perceptual Evaluation of Speech Quality (PESQ), the New ITU Standard for End-to-End Speech Quality Assessment; Part I — Time-Delay Compensation," *J. Audio Eng. Soc.*, Oct. 2002, pp. 755–64.

[4] S. Quackenbush, T. Barnwell III, and M. Clements, *Objective Measures of Speech Quality*, Prentice Hall, 1988.

[5] K. Lam *et al.*, "Objective Speech Quality Measure for Cellular Phone," *Proc. IEEE Int'l. Conf. Acoustics, Speech, and Signal Proc.*, vol. 1, Atlanta, GA, May 1996, pp. 487–90.

[6] N. Kitawaki, H. Nagabuchi, and K. Itoh, "Objective Quality Evaluation for Low-Bit-Rate Speech Coding Systems," *IEEE JSAC*, vol. 6, no. 2, Feb. 1988, pp. 242–48.

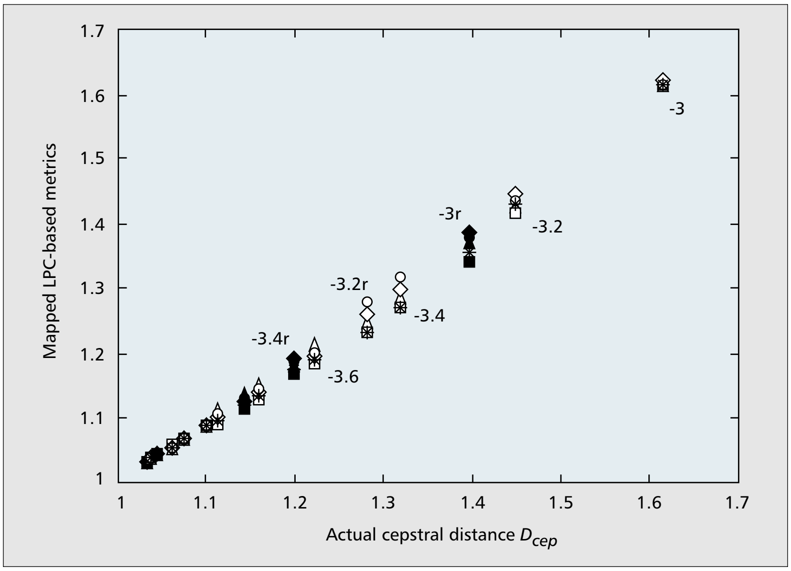


Figure 6. *Scatter plot of cepstral distance obtained from linear mappings of other LPC-based metrics as a function of actual cepstral distance.*

[7] S. Wu and L. Pols, "A Distance Measure for Objective Quality Evaluation of Speech Communication Channels Using also Dynamic Spectral Features," *Proc. Inst. of Phonetic Sci. Amsterdam*, vol. 20, 1996, pp. 27–42.

[8] S. Rein, F. Fitzek, and M. Reisslein, "Voice Quality Evaluation for Wireless Transmission with ROHC (Extended Version and Evaluation Software Source Code)," Dept. of Elec. Eng., AZ State Univ., tech. rep., May 2004, http://www.fulton.asu.edu/˜mre

[9] A. Gray and J. Markel, "Distance Measures for Speech Processing," *IEEE Trans. Acoustics, Speech, and Sig. Proc.*, vol. 24, no. 5, Oct. 1976, pp. 380–91.

[10] C. Bormann *et al.*, "RObust Header Compression (ROHC): Framework and Four Profiles: RTP, UDP, ESP, and Uncompressed," July 2001.

[11] A. Cellatoglu *et al.*, "Robust Header Compression for Real-Time Services in Cellular Networks," *Proc. 2nd Int'l. Conf. 3G Mobile Commun. Tech.*, London, UK, Mar. 2001, pp. 124–28.

Biographies

STEPHAN REIN (stephan.rein@tu-berlin.de) studied electrical engineering at the Technical University of Aachen, Germany, and Technical University Berlin (TUB), Germany. He received a Dipl.-Ing. degree in electrical engineering from TUB in 2003. From March 2003 to October 2003 he visited the multimedia networking group in the Department of Electrical Engineering at Arizona State University, Tempe. He is currently pursuing a Ph.D. degree at the Institute for Energy and Automation Technology, TUB. His current research interests include digital signal processing with emphasis on wavelet applications for automotive security systems.

FRANK H. P. FITZEK is an associate professor in the Department of Communication Technology, University of Aalborg, Denmark, heading the Future Vision group. He received his diploma (Dipl.-Ing.) degree in electrical engineering from the University of Technology RWTH Aachen, Germany, in 1997 and his Ph.D. (Dr.-Ing.) in electrical engineering from TUB in 2002. He co-founded the startup company acticom GmbH in Berlin in 1999. In 2002 he was an adjunct professor at the University of Ferrara, Italy.

MARTIN REISSLEIN (reisslein@asu.edu) is an assistant professor in the Department of Electrical Engineering at Arizona State University, Tempe. He received his Ph.D. in systems engineering from the University of Pennsylvania in 1998. From July 1998 through October 2000 he was a scientist with the German National Research Center for Information Technology (GMD FOKUS), Berlin and a lecturer at TUB. He maintains an extensive library of video traces for network performance evaluation, including frame size traces of MPEG-4 and H.263 encoded video, at http://trace.eas.asu.edu.