Voice Quality Evaluation for Wireless Transmission with ROHC (extended version and software code)

Stephan Rein   Frank Fitzek   Martin Reisslein

Abstract

Robust Header Compression (ROHC) has recently been proposed to reduce the large protocol header overhead when transmitting voice and other continuous media over RTP/UDP/IP in wireless networks. In this paper we evaluate the real-time transmission of GSM encoded voice with ROHC over a wireless link. We first present a tutorial on voice quality evaluation. 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. Besides the impact of ROHC on the voice quality we consider the impact of ROHC on the consumed bandwidth and the delay jitter in the voice signal. We find that for a wide range of error probabilities on the wireless link, ROHC roughly cuts the bandwidth required for the transmission of GSM encoded voice in half. In addition, ROHC improves the voice quality compared to transmissions without ROHC, especially for large bit error probabilities on the wireless link. The improvement reaches 0.26 on the 5-point Mean Opinion Score for a bit error probability of 10^{-3}.

I. INTRODUCTION

While the main service of first and second generation wireless cellular systems has been voice, third generation systems are designed to support a wide range of services, including audio and video applications. This flexibility is achieved by using the Internet protocol (IP) in conjunction with the User Datagram Protocol (UDP) and the Real Time Protocol (RTP). One major problem with the RTP/UDP/IP protocol architecture is the large overhead, which affects the limited bandwidth of wireless channels. A low bit rate speech application can result in IP packets with a ratio of 30 bytes of payload to 60 bytes of overhead. Recently, RObust Header Compression (ROHC) [1] has been proposed to compress the protocol headers for packet transmission over a wireless link.

In this paper we provide an evaluation methodology and performance results for the real-time transmission of voice with ROHC over a wireless link. Our evaluation metrics are the compression gain (reduction in header and total packet size), the voice quality, and the delay jitter. Importantly, we employ a wide array of objective voice quality metrics, including both the traditional and the segmental Signal to

Noise (SNR) ratio, 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 ROHC on the voice quality in terms of the MOS. Our delay jitter measurements do not consider the jitter of the voice packets; instead we consider the jitter within the voice signals, which is closer related to the subjective quality perceived by the user.
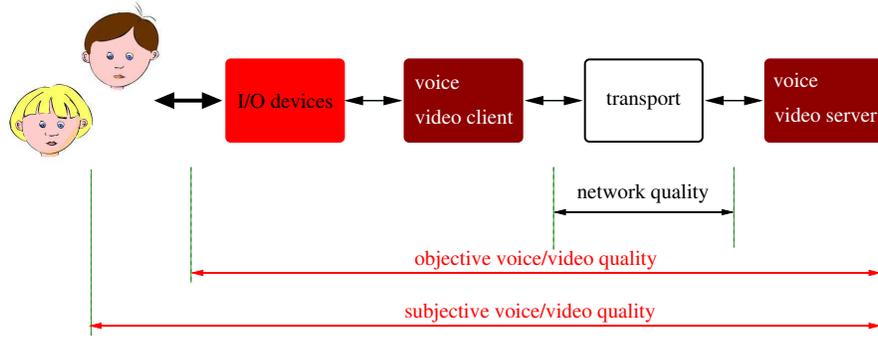

Fig. 1

DIFFERENT PERSPECTIVES ON QUALITY IN ROHC PERFORMANCE EVALUATION.

Generally, when evaluating ROHC one may distinguish between three qualities, namely the network quality, the objective quality, and the subjective quality, as illustrated in Figure 1. While the network quality reflects the provider's perspective, the objective and the subjective quality reflect the customer's perspective. The network quality can be easily measured by network parameters, such as the packet loss rate or the packet delay. The subjective quality is generally more meaningful than the network quality, as it relates directly to the user perceived quality. Assessing the subjective quality, however, is very tedious as it requires listening tests with a large number of test persons. For this reason, objective quality measures that predict the subjective quality are typically employed in the evaluation of voice transmission systems. In this paper we give a tutorial introduction to elementary objective voice quality metrics that allow for computationally efficient and accurate voice quality evaluations without requiring the purchase of specialized software.

We find in our evaluation that for a wide range of bit error probabilities on the wireless link, ROHC reduces the protocol overhead for voice transmission with IPv4 by approximately 85%, which reduces the bandwidth required for a GSM coded voice transmission by about 47%. On top of these bandwidth savings, ROHC improves the voice quality. On the 5-point MOS scale the improvement increases roughly exponentially with the bit error probability. The improvement is about 0.028 for an error probability of $10^{-4.5}$ and reaches 0.134 and 0.264 as the error probability increases to $10^{-3.6}$ and 10^{-3}. We also find that ROHC slightly increases the jitter for small error probabilities and slightly reduces the jitter for large error probabilities.

This paper is organized as follows. In the following subsection we review related work. In Section II we describe the principles and integration of ROHC in the IP protocol stack. In Section III we describe our evaluation methodology. In Section IV we explain how to evaluate the objective voice quality using

an array of metrics ranging from Signal to Noise (SNR) ratio based metrics to spectral and parametric distance metrics which are based on a linear predictive coding (LPC) analysis. In Section V we present our segmental cross correlation (SCC) algorithm for synchronizing the original voice stream with the voice stream after network transport. In Section VI we present our bandwidth reduction, objective voice quality, and delay jitter results for using ROHC. In Section VII we summarize our contributions.

A. Related Work

There exists a large body of literature on the development of header compression schemes for wireless networks and on the evaluation of these schemes in terms of the network metrics of throughput, packet delay, and packet jitter. This literature is comprehensively surveyed in [2]. The impact of header compression on the quality of the transmitted medium (e.g., voice) has received very little attention so far. The only study in this direction that we are aware of is [3]. In [3] the objective speech quality degradation (using the traditional SNR which has only a weak correlation with user perception) is studied for Robust Checksum-based Compression (ROCCO) and the Compressed Real Time Protocol (CRTP), which may be considered as precursors to ROHC. In contrast, in this paper we consider the state-of-the-art ROHC compression scheme and evaluate the voice quality using an array of objective metrics that allow accurate predictions of the subjective voice quality of hearing tests. (The impact of ROHC on the wireless transport of video, whose quality evaluation is significantly different from the voice quality considered here, is examined in a companion paper [4].)

As reviewed in more detail at the beginning of Section IV, objective voice quality evaluation metrics have received significant attention over the past 20 years in the research literature of the acoustics and signal processing community. However, to the best of our knowledge there is no succinct self-contained tutorial available that is readily accessible and usable by the networking engineer or researcher. For this reason we provide a tutorial on objective voice quality evaluation in Section IV of this paper.

II. Overview of Robust Header Compression

A multimedia stream packet composed for an IP network transmission consists of a 20 byte IP header, an 8 byte UDP header, and a 12 byte RTP header, as shown in Figure 2. The IPv6 version requires a 40 byte IP header, so the total header size can sum up to 60 bytes. A speech application generates compressed data at a low bit rate of around 13 kbit/s. Considering a typical payload smaller than 40 bytes, the ratio of header size to payload results in an significant waste of link bandwidth. The ROHC compressor replaces the RTP/UDP/IP overhead by its own, much smaller header. On the receiver side the decompressor transforms the ROHC header into the original protocol layer headers.

In Figure 3, the different header fields of an IP packet are classified in order to show the potential of a compression scheme.

Many IP header fields of a given data flow are static, i.e., do never change. ROHC stores the values of these static header fields as *static context* at the decompressor. More challenging for the compression scheme is the treatment of the changing (dynamic) fields in the IP header. ROHC uses linear functions based on the packets' sequence numbers to derive the values of the dynamic header fields, see [1] for details. The parameters characterizing these linear functions are stored and updated as so-called *full context*
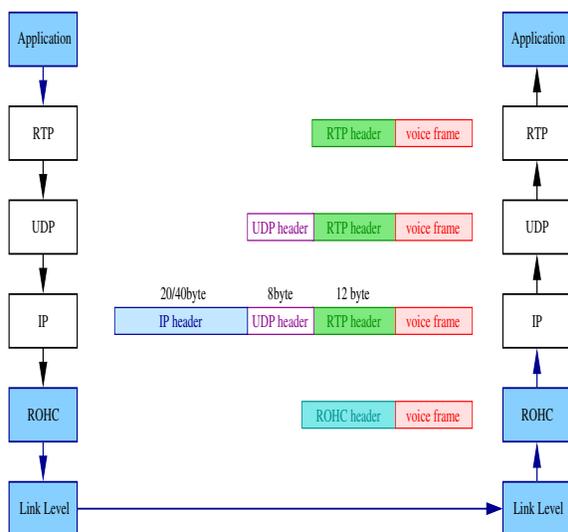
Fig. 2

THE 40 BYTE RTP/UDP/IPv4 HEADER (OR 60 BYTE
RTP/UDP/IPv6 HEADER) IS COMPRESSED TO A SMALLER
ROHC HEADER BY THE ROHC LAYER, WHICH RESIDES
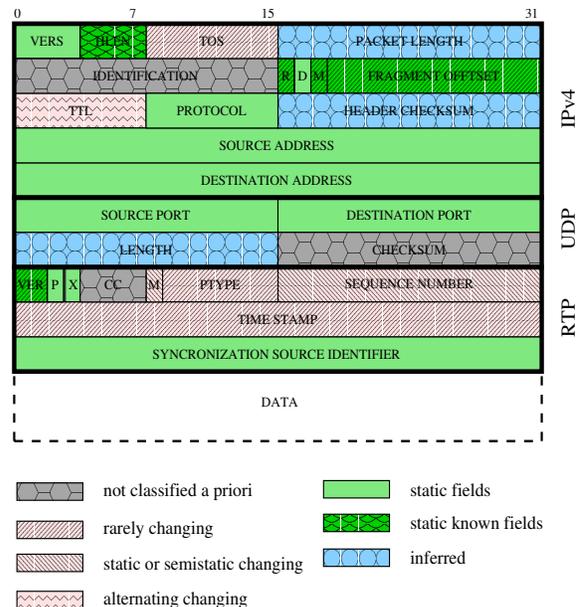BETWEEN IP AND LINK LAYER.

Fig. 3

HEADER FIELDS FOR RTP/UDP/IP PACKETS (VERSION 4)
WITH THE APPROPRIATE DYNAMICS.

at the decompressor. We also note that some IP header fields can be derived (inferred) from other header fields at the decompressor and are therefore not transmitted once the static context is established.

The fundamental challenge in header compression for transmission over wireless links is to maintain the correct context at the decompressor in the face of quite frequent bit errors in the received packets. ROHC supports three different modes for maintaining the context in different wireless systems. The *unidirectional mode* is designed for systems without a feedback channel from the decompressor to the compressor, i.e., where the decompressor can not acknowledge the correct receipt of context information. To overcome this limitation, the compressor periodically retransmits the context information. The *bidirectional optimistic mode* and the *bidirectional reliable mode* are designed for systems with a feedback channel from the decompressor to the compressor, i.e., where the decompressor can acknowledge the correct receipt of context information and/or send negative acknowledgements to request the retransmission of context information. With the bidirectional optimistic mode, bit errors in the compressed header are detected with a 3-bit cyclic redundancy check (CRC) code. When the CRC check fails the decompressor generally discards the affected packet and attempts to repair its context either locally or by requesting a context update from the compressor. The reliable mode extends the optimistic mode by a more complex error detection and correction which uses a larger number of coding bits. Our evaluation concentrates on the optimistic mode, since it gives generally the best compression efficiency. Also, with the results of the optimistic mode, it is possible to predict the results for the reliable mode

To assess the maximum compression gain (packet size reduction) with header compression we consider an ideal compression scheme that reduces the header size to zero bytes. Clearly, such an ideal compression

TABLE I

TEST MATERIAL FOR VOICE QUALITY EVALUATION

file	text	language	speaker	duration [sec]	total size [kB]	GSM size [kB]
49.wav	A	English	female	19.15	306.45	31.6
53.wav	B	German	female	16.64	266.21	27.46
54.wav	B	German	male	16.79	268.72	27.72

TABLE II

MEAN OPINION SCORE

MOS	
5	imperceptible
4	just perceptible but not annoying
3	perceptible and slightly annoying
2	annoying but not objectionable
1	very annoying and objectionable

scheme has a compression gain (i.e., reduces the packet size) by

$$gain_{\max} = \frac{header size}{header size + payload}. \tag{1}$$

With a GSM codec generating 33 byte frames, the maximum saving potential is 55% when using IPv4, it grows to 65% when using IPv6. As the overhead is constant, the maximum saving with compression increases as the payload size decreases. Therefore ROHC is well suited for low bit rate voice streams, where the header size is typically larger than the payload.

In the commonly used RTP/UDP/IP protocol suite, ROHC is installed between the network and the link layer. In the third generation Universal Mobile Telecommunication System (UMTS), ROHC compressor and decompressor are part of the UMTS mobile phone and the corresponding UMTS radio network controller (RNC). (Other solutions are possible, but ROHC always resides above the link layer.) The other Internet components do not notice the usage of a compression scheme, but the wireless service provider can take advantage of a significant reduction of the required bandwidth, as demonstrated by our results in Section VI. ROHC requires from the link layer that the packets are sent in a strictly sequential order. Also the packets are not allowed to contain routing information (single hop restriction).

ROHC supports three different modes in order to adapt to different requirements of reliability and channel capacity. The *unidirectional mode* is the least efficient mode as there is no feedback channel from the decompressor to the compressor. To ensure a correct context at the decompressor side, the compressor periodically has to sent context information. The *bidirectional optimistic mode* and the *bidirectional reliable mode* use feedback information sent from the decompressor to the compressor. The feedback allows the compressor to respond to successful or unsuccessful transmissions. The reliable mode extends the optimistic mode by a 7-bit error correction scheme. Our evaluation concentrates on the optimistic mode, since it gives generally the best compression efficiency. Also, with the results of the optimistic mode, it is possible to predict the results for the reliable mode.

III. EVALUATION METHODOLOGY

The ROHC measurements were conducted on a testbed consisting of two Linux machines. The Linux kernels had been enhanced by an ROHC implementation (provided by the acticom GmbH, www.acticom.de). We used three different voice files (track 49, track 53, and track 54) obtained from the European Broadcasting Union [5], as shown in Table I.

The files, given in the wave file mono format, are first down sampled to 8 kHz and then transferred to the communication system shown in Figure 4. On the sender's side the wave file is GSM encoded (using the encoder [6]). The coded file consisting of 33 byte GSM frames, is passed to the RTP/UDP/IP protocol

stack. (The wave file header (44 bytes) is not part of the transmission, because the GSM encoder expects raw audio data.) The RTP/UDP/IP packet finally arrives at the ROHC and link layers. The two Linux machines are connected by an Ethernet network. Recent channel characterization studies [7] have revealed that uncorrelated bit errors give a good approximation of the error process in 3G networks. Consequently, we simulate uncorrelated bit errors on the link layer. We use nine different bit error probabilities ranging from 10^{-6} to 10^{-3}. Figure 5 illustrates how the original and the transferred (and possibly distorted) voice files are employed in our quality evaluation. As the ROHC is optional, the quality evaluation is obtained by a comparison of transmissions with and without ROHC. This comparison answers the question whether
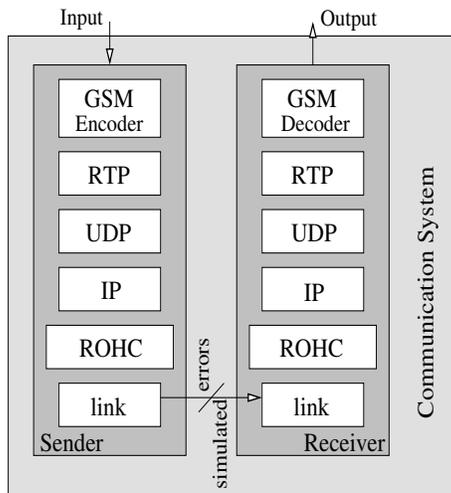

Fig. 4

COMMUNICATION SYSTEM: TWO LINUX PCS, INTERCONNECTED BY AN ETHERNET NETWORK WITH SIMULATED WIRELESS LINK ERRORS. EXPERIMENTS ARE CONDUCTED WITH ROHC AND WITHOUT ROHC.
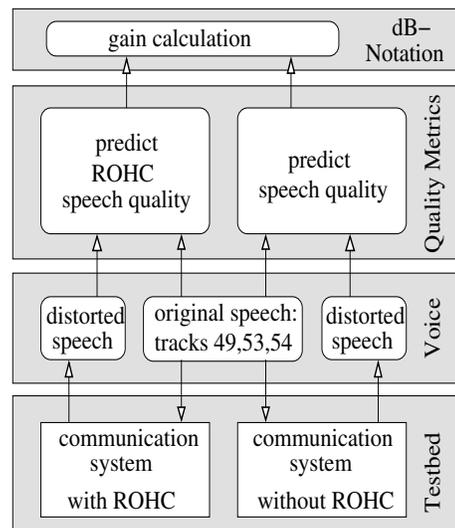

Fig. 5

METHODOLOGY OF ROHC EVALUATION.

the voice quality is deteriorated or improved by using ROHC.

IV. VOICE QUALITY EVALUATION METRICS

Expensive and time consuming speech perception tests with human listeners as detailed in [8] 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 Mean Opinion Score (MOS) scale summarized in Table II. 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.

Generally, there are three classes of objective voice quality evaluation metrics, the network parameter based metrics, the psycho-acoustic metrics, and the elementary metrics. The 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 [9], which is one recent proposal for a parameter based metric. Parameter based metrics, such

TABLE III

CORRELATIONS BETWEEN OBJECTIVE VOICE QUALITY
METRICS AND SUBJECTIVE VOICE QUALITY. THE DISTORTION
TYPES (INDEXED BY THE FOOTNOTE MARKERS 1–8) ARE
GIVEN IN TABLE IV.

Objective Metric	Correlation
(traditional) SNR	$+0.24^1 / + 0.31^2$
segmental SNR	$+0.77^1 / + 0.78^2$
inverse linear unweighted distance	$+0.63^3 / + 0.48^4$
unweighted delta form	$-0.61^3 / - 0.51^5$
log root mean square (RMS)	$(-)^{10}$
Log Area Ratio	$-0.62^3 / - 0.65^4$
Energy Ratio [17]	$-0.59^3 / - 0.61^4$
log likelihood [17]	$-0.49^3 / - 0.48^6$
cepstral distance	$-0.96^7 / - 0.95^8 / - 0.93^9$

TABLE IV

DISTORTION TYPES FOR THE MEASURED CORRELATIONS.
THE DISTORTION TYPES INDEXED BY 1–4 ARE FROM [18].

Index	Distortion
1	waveform coders: 8 types
2	additive– and narrow-band noise
3	coding distortions, controlled distortions, and narrow-band distortions: 23 types
4	waveform coders and controlled distortions in the time and frequency domain, 18 types
5	cellular phone: [19]
6	cellular phone: [20]
7	coding and other non-linear distortions: [21]
8	PCM, ADPCM, G.728, MNRU: [14]
9	noise masking, band pass filtering, echo, and peak clipping: [22]
10	theoretical approach: [23]

as the E–model hold promise for predicting the subjective voice quality [10] but still require extensive refinements and verifications [11].

The psycho-acoustic metrics transform the voice signals to a reduced representation to retain only the 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 the evaluation of non-waveform preserving speech coding algorithms. These coding algorithms perform waveform distortions that are perceptually not significant. Various complex metrics have been developed and refined over the last decade. These include the Bark spectral distance [12], the measuring normalizing blocks (MNB) technique [13] [14], and the PESQ measure [15] [16], 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 the subjective voice quality. The elementary metrics have generally smaller correlations with the subjective voice quality than the highly complex psycho-acoustic metrics and do not provide the perception modeling that is needed for psycho-acoustic coder algorithm development. The elementary metrics, however, do represent a good engineering trade-off for networking researchers in that they allow for fairly detailed conclusions about the voice quality while having low computational complexity. We also note that in our evaluation methodology, as illustrated in Figure 5, we focus on system modification in the networking domain (e.g., the introduction of ROHC). Both, the unmodified system (without ROHC) and the modified system (with ROHC) employ the same voice codec and thus experience approximately the same voice codec distortions. Our evaluation is focused on the impact of the modification in the networking domain on the voice quality (and is not designed to evaluate voice codec distortions).

We have selected the elementary metrics listed in Table III for our evaluations. 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 data base. Table III gives the distortion types that the various objective metrics have been examined for and the resulting correlations to subjective hearing tests. The larger the magnitude of the correlation, the better the prediction of the subjective voice quality. We note

that the traditional SNR has a poor correlation performance. However, we include it because it is often considered as a purely objective quality metric. The RMS spectral distance is included because in [23],
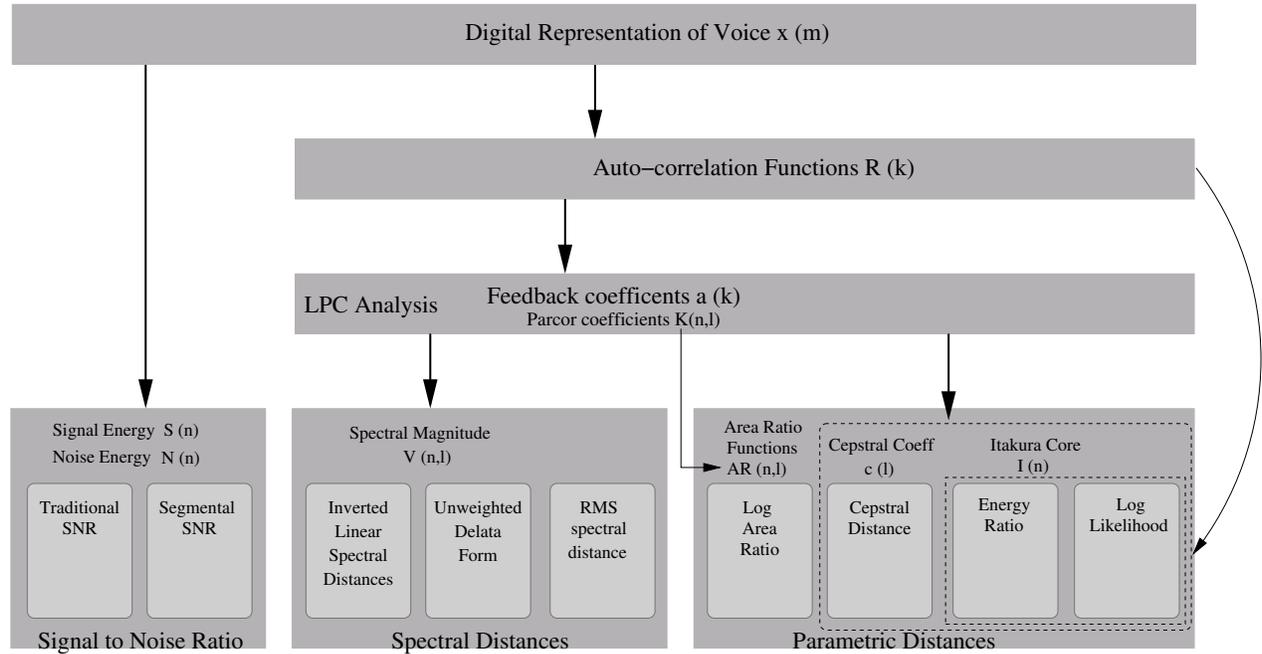

Fig. 6

SET OF USED OBJECTIVE VOICE QUALITY METRICS. THE CALCULATIONS ARE PARTIALLY SIMILAR, BUT THE METRICS COVER DIFFERENT TYPES OF DISTORTIONS.

it is shown that it is a very meaningful measure for speech perception, as it can be physically interpreted and efficiently computed. We note that the cepstral distance achieves the best correlation performance for its respective distortion types. In addition to this good correlation performance, the cepstral distance has the distinctive property that its values can be transformed to the predicted mean opinion score (MOS) by a publicly available mapping. (We note that mappings from other objective metrics to a subjective score are generally possible, however, we are not aware of any such mapping being publicly available.) As illustrated in Figure 6, many metrics use the same coefficients and are similarly calculated. However, their performance differs among different types of distortions, as verified in [14], [18], [19]–[22].

We close this general overview of voice quality evaluation metrics by noting that we have chosen the elementary metrics in Table III as they represent a sensible engineering approach for our networking study. The chosen elementary metrics have good correlations with the subjective voice quality and thus allow for meaningful conclusions about the voice quality. At the same time the chosen metrics are computationally efficient and do not require costly proprietary software (in fact we make our evaluation software source code publicly available: http://www.eas.asu.edu/~mre). In order to cover a reasonably wide range of distortion types we selected a set of elementary metrics (see Table III), which as we shall demonstrate are highly correlated to the cepstral distance (and thus to the MOS) for the considered wireless voice transmission. To synchronize the received voice stream after packet based transport we developed a low complexity, yet effective segmental cross correlation (SCC) algorithm, see Section V.

TABLE V

NOTATION OF OBJECTIVE MEASURES

n	frame index
N	total number of frames in voice file
m	sample index
M	number of samples in a frame; $M = constant = 160$
$x_{n,\phi}(m)$	amplitude of sample m in frame n of the undistorted signal
$x_{n,d}(m)$	amplitude of sample m in frame n of the distorted signal
$F(n)$	distortion index for frame n
D	total quality of file calculated by a metric

A. Notation

For the calculation of the objective quality metrics a given uncompressed voice signal is broken into frames of 20 msec duration. These 20 msec frames are introduced for the voice quality evaluation in accordance with the human voice recognition. 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, $n = 1, \ldots, N$. (Note that with a sample rate of 8 KHz a 20 msec frame contains $M = 160$ samples, each 16 bits worth of uncompressed voice data. These 320 bytes of voice data are typically compressed into one 33 byte GSM frame.) 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 (after network transport). 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. The signal energy $S(n)$ and the noise energy $N(n)$ of frame n are given by

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

and

$$N(n) = \sum_{m=1}^{M} [x_{n,d}(m) - x_{n,\phi}(m)]^2. \tag{3}$$

Each metric gives a distortion index $F(n)$ for a given frame n. 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:

$$D = \frac{1}{N} \sum_{n=1}^{N} F(n). \tag{4}$$

A slightly more complex approach may weigh the distortion indices of the individual frames by the corresponding signal energies, but this weighting has typically negligible impact on the total quality. Equation (4) is only used with the spectral and parametric measures, because the SNR metrics give directly the total quality. Table V summarizes the notation of our objective measures.

B. SNR Measures

The *traditional* SNR and the *segmental (short–time or framed)* SNR are given by

$$D_{\text{trad}} = 10 \cdot \log_{10} \frac{\sum_{n=1}^{N} S(n)}{\sum_{n=1}^{N} N(n)} \tag{5}$$

and

$$D_{\text{seg}} = 10 \cdot \log_{10} \sum_{n=1}^{N} \frac{S(n)}{N(n)}.$$ (6)

Whereas the segmental SNR is based on a frame by frame calculation, the classical SNR is calculated across the entire sequence of N frames. As a result, the classical SNR aggregates the signal energy in the entire file and relates this aggregate signal energy to the aggregate noise energy. In contrast, the segmental SNR relates the signal energy of each individual frame to the noise energy of the frame. This finer granularity relates more meaningfully to the perception of the voice file.

C. Spectral Distances

The spectral distance metrics are based on the so-called *spectral magnitude* $V(n, l)$, which is provided in Appendix C. Based on the spectral magnitude, the frame distortion indices for the inverse linear unweighted distance, the unweighted delta form, and the log root-mean-square distance are given by

$$F_{inv}(n) = \left[\frac{1}{L} \sum_{l=0}^{L-1} \left[\frac{1}{b + |V_\phi(n,l) - V_d(n,l)|} \right]^p \right]^{1/p},$$ (7)

$$F_\delta(n) = \left[\frac{1}{L} \sum_{l=0}^{L-1} |V_\phi(n,l)^\delta - V_d(n,l)^\delta|^q \right]^{1/q},$$ (8)

and

$$F_{rms}(n) = \sqrt{\frac{1}{L} \sum_{l=0}^{L-1} \left[\log_{10} \frac{V_\Phi(n,l)}{V_d(n,l)} \right]^2},$$ (9)

where L is usually set to 128. In [18], the constants $p = 8$, $q = 1$, and $\delta = 0.2$ have been heuristically verified to achieve good performance.

D. Parametric Distances

Parametric distances use transformations of the linear predictive coding (LPC) coefficients, which are presented in Appendix B. We consider three classes of parametric distance measures,

1) the *log area ratio* measure,
2) the *energy ratio*/*log likelihood* measure, and
3) the LPC cepstral distance measure,

which are defined in [18]. The first two classes are evaluated in [18] while the third class is evaluated in [21], [22], [14]. The log area ratio is more efficient than a spectral distance measure, because it requires a lower computation time and gives a comparable correlation to the spectral distances.

The energy ratio/log likelihood measures are computationally even less demanding, but give lower correlations. The log likelihood measure was one of the first objective metrics to measure the voice quality. As these measures are strongly related to the Itakura likelihood ratio distance measure [17], Figure 6 categorizes this class by the Itakura core.

Kitawaki et al. [21] compared elementary objective speech quality measures for voiceband codecs. The cepstral distance had the best correspondence to the mean opinion score among all objective measures

studied. These results are confirmed by Wu and Pols [22], who estimated a correlation of 0.926 for the LPC cepstral distance measure with the mean opinion score. This correlation performance has been further verified for waveform preserving codecs and for 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 [14]. We use the results of [21] to predict the mean opinion score from the cepstral distance.

1) Log Area Ratio Measure: From the LPC analysis, the lth *Parcor* coefficient $K_{d/\phi}(n,l)$ for the distorted/undistorted frame n is calculated, as given in (28) in Appendix B. Let $AR(n,l)$ denote the *area ratio* function:

$$AR_{d/\phi}(n,l) = \frac{1 + K_{d/\phi}(n,l)}{1 - K_{d/\phi}(n,l)}. \tag{10}$$

The *log area ratio* measure is given by

$$F_{log}(n) = \left\{ \frac{1}{10} \sum_{l=1}^{10} \left| 20 \cdot \log_{10} \frac{AR_d(n,l)}{AR_\phi(n,l)} \right|^p \right\}^{1/p}, \tag{11}$$

where the constant $p = 1$ achieves a good performance [18].

2) Energy Ratio and Log Likelihood Measure: Let $\mathbf{R}_\phi(n)$ denote the autocorrelation matrix of frame n, given in (35), and $\vec{a}_{d/\phi}(n)$ denote a vector of all LPC coefficients, given in (36). The core $I(n)$ of the second class of parametric distance measures is given by

$$I(n) = \frac{\vec{a}_d^T(n) \cdot \mathbf{R}_\phi(n) \cdot \vec{a}_d(n)}{\vec{a}_\phi^T(n) \cdot \mathbf{R}_\phi(n) \cdot \vec{a}_\phi(n)}. \tag{12}$$

The *energy ratio* is calculated by

$$F_{en}(n) = |I(n)|^{\delta/2}, \tag{13}$$

where the best correspondence to subjective quality was obtained for $\delta = 0.5$. The *log likelihood* measure is given by

$$F_{like}(n) = 10 \cdot \log_{10}\{I(n)\}. \tag{14}$$

3) Cepstral Distance: The *cepstral* distance measure calculates the difference in the shape of the original and the distorted spectrum. It is based on LPC–derived cepstral coefficients $c_{n,d/\phi}(l)$, see Appendix E. The cepstral distance is defined in [21] as

$$F_{cep}(n) = \frac{10}{\log_e(10)} \left[2 \sum_{l=1}^{L} [c_{n,\phi}(l) - c_{n,d}(l)]^2 \right]^{\frac{1}{2}}, \tag{15}$$

where L is the number of cepstral coefficients, which we choose equal to the order of the LPC analysis from which the predictor coefficients $a(i)$ are derived.

V. Segmental Cross Correlation algorithm (SCC)

We transfer voice over a communication system. Thereby, voice frames may (i) be completely lost, (ii) experience varying delays, or (iii) suffer voice signal distortions due to bit errors. The objective voice quality is based on a comparison between the received (distorted) and the original (reference) voice streams, which need to be synchronized for the comparison. There are generally two types of approaches to synchronize the streams: (i) packet based approaches, and (ii) voice signal based approaches. Packet
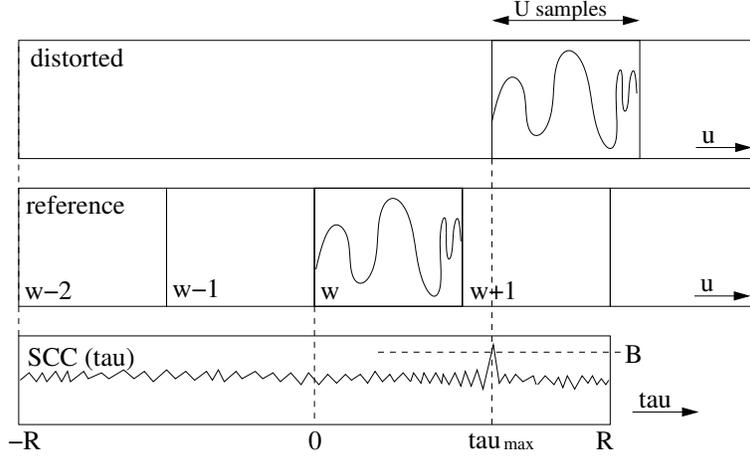
Fig. 7

PRINCIPLE OF SCC: FOR EVERY FRAME w OF THE REFERENCE FILE, A FRAME OF THE DISTORTED FILE IS MATCHED.

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, see for instance [24], and adjusting the playout time of the voice frame. Voice signal based approaches, on the other hand, employ signal correlation techniques to align frames in the distorted and reference streams, see for instance [15]. For signal based synchronization we have developed the segmental cross correlation (SCC) algorithm, which we outline in this section and employ in our measurements reported in Section VI. We note that the voice quality evaluation methodology presented in the previous section can be employed both in conjunction with packet based synchronization and signal based synchronization.

For the synchronization the reference file is divided into consecutive synchronization frames of U samples each. 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. More formally, let $x_{w,\phi}(u)$, $u = 1, \ldots, U$, 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 segmental cross correlation function

$$SCC_w(\tau) = \frac{\sum_{u=1}^{U}[x_{w,\phi}(u) - \bar{x}_{w,\phi}] \cdot [(x_d(u + (w-1)U + \tau) - \bar{x}_d(w,\tau)]}{\sqrt{\sum_{u=1}^{U}[x_{w,\phi}(u) - \bar{x}_{w,\phi}]^2}\sqrt{\sum_{u=1}^{U}[x_d(u + (w-1)U - \tau) - \bar{x}_d(w,\tau)]^2}}, \qquad (16)$$

where we denote

$$\bar{x}_d(w,\tau) = \frac{1}{U}\sum_{u=1}^{U} x_d(u + (w-1)U + \tau). \qquad (17)$$

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

$$\tau_{\max}(w) = \arg \max_{-R \leq \tau \leq R} SCC_w(\tau). \qquad (18)$$

If this maximum cross correlation is larger than a threshold then the displacement estimate (match) is accepted, otherwise the search range is increased. If an acceptable match is not found for the increased

TABLE VI

PARAMETERS OF THE SEGMENTAL CROSS CORRELATION ALGORITHM

threshold value for a sufficient correlation	$B = 0.3$
number of samples for one frame	$U = 4000$
range of search area	$R = 200$ samples
range of augmented search area	$R' = 4000$ samples

search range, then the synchronization fails for this distorted file (replication of the experiment). Such a failure is usually due to failures in the measurement set-up. For the subsequent frames w, $w \geq 2$, the cross correlation is initially evaluated for the search range $\tau_{\max}(w-1) - R \leq \tau \leq \tau_{\max}(w-1) + R$, i.e., the search range is adaptively shifted according to the displacement of the preceding frame $w - 1$. The parameters synchronization frame length U, initial search range R (and policy for increasing the search range), and acceptance threshold represent trade-offs between computational effort and likelihood of successful synchronization (see Table VI). A successful synchronization is determined by a sufficient *correlation value*, i.e., a sufficient mean value of all SCC frame correlations (this value is calculated by our evaluation software, see source file AudioMeter.cpp, output file quality.dat, variable "correlation") of a synchronized voice file. Visual comparisons of the synchronized voice file and the reference file revealed that a correlation value, which is significantly larger than the acceptance threshold value B, indicates a successful synchronization, thus allowing a meaningful evaluation by the elementary voice quality metrics. For the experiments reported in the following two sections we have typically successfully synchronized the voice files of 94 % of the experiments. The remaining 6% of the experiments did not give valid voice files for synchronization due to problems with the measurement set-up, such as operating system problems and hardware failures. As detailed shortly, many independent replications are conducted for each experiment to obtain statistical confidence levels on all results.

We note that the computation time of the SCC algorithm can be reduced by using the well known Fast Fourier Transform (FFT) algorithm. Due to space constraints we only roughly describe its application for time synchronization and refer the interested reader to [25] for details. The SCC algorithm is based on a cross correlation in the time domain. The reference signal $x_{w,\phi}(\cdot)$ and the distorted signal $x_d(\cdot)$ are transformed to the frequency domain using the FFT algorithm: $X_\phi(k) = FFT[x_{w,\phi}(\cdot)]$ and $X_d(k) = FFT[x_d(\cdot)]$. In the frequency domain the component-wise product $Z(k) = X_\phi(k) \cdot X_d(k)$ is calculated and retransformed to the time domain using the inverse FFT algorithm. The result is equivalent to $SCC_w(\tau)$, in (16), thus allowing directly for the determination of $\tau_{\max}$.

We finally note that PESQ, which requires the purchase of proprietary software (with a cost on the order of $ 10.000, see http://www.pesq.org), performs highly complex algorithms in the time and frequency domain [15] and gives generally better synchronization performance than our low complexity SCC algorithm (for which we make the source code publicly available: http://www.eas.asu.edu/~mre). However, the SCC algorithm does allow for meaningful delay jitter measurements in the received voice signal, as presented in Section VI-C and synchronizes the voice signals to allow for the objective voice quality evaluations presented in Section VI-B.

VI. Performance Results for ROHC

In this section we give an overview of our evaluations of voice transmission with ROHC. We first evaluate the compression performance (bandwidth reduction) achieved by ROHC. Next, we employ the objective voice quality metrics explained in Section IV to assess the impact of ROHC on the voice quality. Finally, we evaluate the impact of ROHC on the jitter in the voice signal.

Throughout this evaluation study we consider the three voice tracks (files) given in Table I. For each track we conduct many experiments, each with statistically independent bit errors, to obtain 95 % confidence intervals less than 10 % of the corresponding sample mean for all performance metrics. For additional statistically reliability we then average the results for the three tracks.

A. ROHC Bandwidth Compression

We measure the header compression gain (i.e., reduction of header size) achieved by ROHC, which is calculated for each track as

$$header\ gain = 1 - \left(\frac{size\ ROHC\ header}{size\ uncompressed\ header} \right). \tag{19}$$

We found that the header compression gain is 84.7 % for all tracks for the entire range of considered error probabilities from 10^{-6} to 10^{-3}. With IPv4 this implies that the header size is in the long run average reduced from 40 to approximately 6 bytes. The compression gain for the total RTP/UDP/IP packet with a payload of 33 bytes is calculated as

$$total\ gain = 1 - \left(\frac{(6+33)\ bytes}{(40+33)\ bytes} \right) = 0.47.$$

This actual compression gain of 47% for the total IP packet is close to the maximum gain of 55%, obtained from Equation (1). Next we address the question whether this significant reduction in consumed bandwidth affects the voice quality.

B. Voice Quality Evaluation of ROHC

To evaluate the impact of ROHC on the voice quality we obtain the total quality both without ROHC (denoted by D) and with ROHC (denoted by D_{ROHC}) for the objective quality metrics described in Section IV. For ease of evaluating the voice quality improvement (gain) achieved by ROHC we define the gain metrics in decibel (dB) in Table VII. Positive gains indicate an improved voice quality while

TABLE VII

Gain definitions for different metrics.

metric	gain [dB]
SNR	$D_{ROHC} - D$
segm. SNR	$D_{ROHC} - D$
inv. lin. spectral dist.	$20 \cdot \log \left(D_{ROHC}/D \right)$
unw. delta spectral dist.	$20 \cdot \log \left(D/D_{ROHC} \right)$
RMS distance	$D - D_{ROHC}$
log area ratio	$D - D_{ROHC}$
energy ratio	$10 \cdot \log \left(D/D_{ROHC} \right)^4$
log likelihood	$D - D_{ROHC}$

negative gains indicate a deteriorated voice quality. Note from Table III that the SNR and the inverse linear spectral distance have positive correlations with the subjective voice quality, i.e., $D_{ROHC} \geq D$ indicates a higher voice quality. All other metrics have a negative correlation with the subjective voice quality, thus $D_{ROHC} \leq D$ indicates an 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 the 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 multiplicative factor (and a power of 4 to compensate for the power of $\frac{1}{4}$ in the metric definition) 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 the comparison of the results of the different metrics and also note that other definitions are possible.

1) Voice Quality Gain Results: In Figures 8, 9, and 10, we plot the gain (in dB) 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 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 two and three decibels 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.
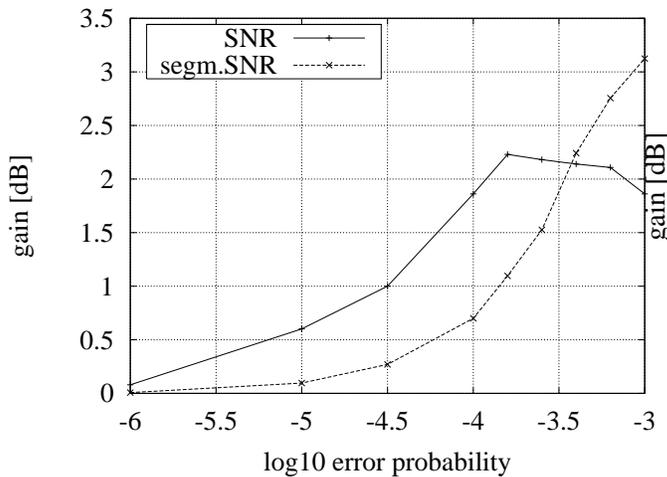

Fig. 8

GAIN IN OBJECTIVE VOICE QUALITY WITH ROHC FOR SNR MEASURES AS A FUNCTION OF BIT ERROR PROBABILITY.
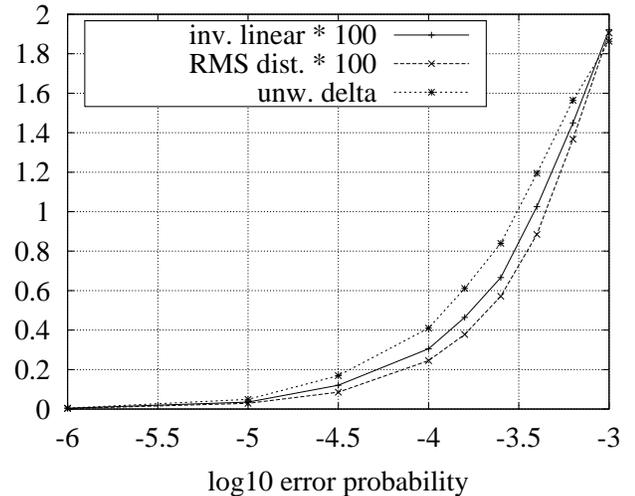
Fig. 9

GAIN IN OBJECTIVE VOICE QUALITY WITH ROHC FOR SPECTRAL DISTANCES AS A FUNCTION OF BIT ERROR PROBABILITY.
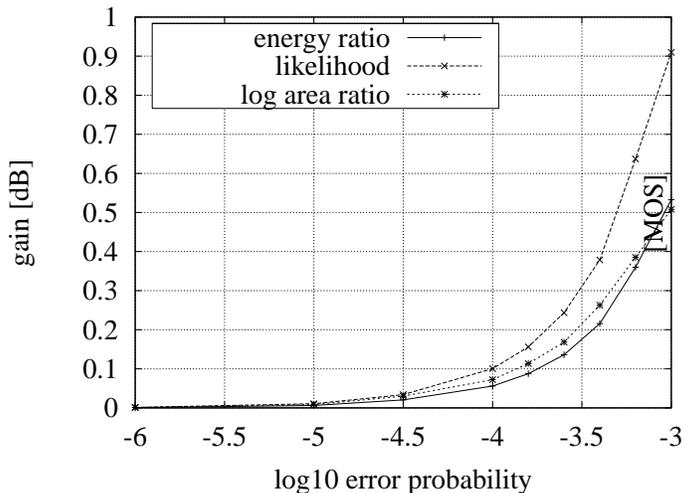
Fig. 10

GAIN IN OBJECTIVE VOICE QUALITY WITH ROHC FOR PARAMETRIC DISTANCES AS A FUNCTION OF BIT ERROR PROBABILITY.
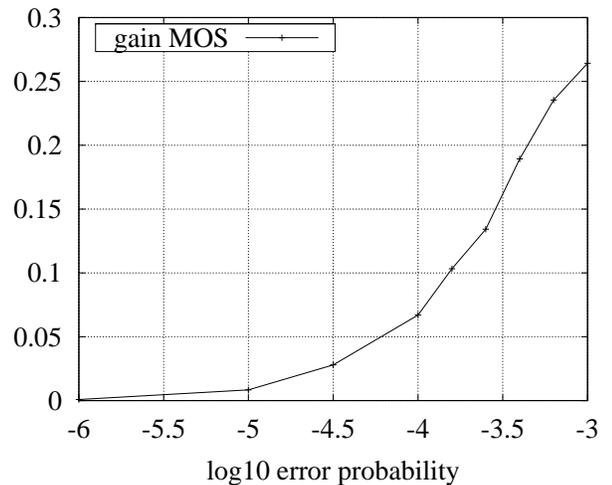
Fig. 11

GAIN IN VOICE QUALITY WITH ROHC IN TERMS OF MEAN OPINION SCORE AS A FUNCTION OF BIT ERROR PROBABILITY.

Overall, these results indicate that the voice quality does not suffer from header compression. On the contrary, it is improved, especially for large bit error probabilities on the wireless link. Note that these gain values in dB represent the improvement in terms of objective voice quality and not in terms of user perception.

In order to asses the impact on the user perception we now investigate the improvements on the subjective 5 point MOS scale (see Table II). We transform the values of the cepstral distance to the predicted mean opinion score (MOS), using the mapping verified in [21]. Let D_{cep} devote the voice quality calculated by the cepstral distance. The MOS value is given by

$$MOS = 3.56 - 0.8 \cdot D_{cep} + 0.04 \cdot D_{cep}^2. \tag{20}$$

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 two differently processed versions of the voice file is meaningful [22]. Hence, we define the MOS gain for ROHC as

$$MOS_{\mathrm{gain}} = MOS_{\mathrm{wROHC}} - MOS_{\mathrm{w/oROHC}}. \tag{21}$$

As shown in Figure 11, the predicted gain for ROHC in terms of the MOS increases roughly exponentially with increasing error probability and reaches 0.26 for error probabilities of 10^{-3}.

2) Relationship between Quality Metrics: Generally, in objective voice quality evaluation it is advisable to consider a variety of metrics since each individual metric (including our key metric, the cepstral distance) has been evaluated for a limited set of distortions, see Table III. We therefore examine now the correlations

TABLE VIII

LINEAR MAPPINGS OF OTHER LPC BASED METRICS D TO THE CEPSTRAL DISTANCE D_{cep}. THE SYMBOLS ARE USED IN THE SCATTER PLOT FIGURE 12.

Metric	Mapping function	Symbol
inv. lin. spectral dist.	$D_{cep} = -5281.818D + 105.982$	◇
unw. delta form	$D_{cep} = 17.6542D + 0.37997$	▽
RMS spectral dist.	$D_{cep} = 12.8911D + 0.4383$	△
log area ratio	$D_{cep} = 0.46107D + 0.23373$	○
energy ratio	$D_{cep} = 8.1716D - 7.404$	□
likelihood	$D_{cep} = 0.2867D + 0.7428$	×

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} without ROHC. The resulting linear mappings are reported in Table VIII. 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 Figure 12. 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 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.

C. Delay Jitter Results

The voice quality metrics considered in the preceding section do not capture the signal delays. Therefore, we investigate the delay, or more precisely, the delay variation (jitter) separately in this section. Recall that we employ our SCC algorithm to perform delay corrections to the received (distorted) voice signal before evaluating the voice quality metrics. The amount of these delay corrections gives the delay jitter within the voice signal.

We examine both the delay jitter histogram and the standard deviation of the delay jitter. Figure 13 shows a typical histogram of delay jitter for the bit error probability 10^{-3}. Each bar represents a delay jitter range of 5 msec. (The bars of ROHC are slightly thinner for graphical reasons.) Figure 14 depicts the ROHC gain for jitter (i.e., reduction in delay standard variation). For the bit error probabilities 10^{-3} to $10^{-3.4}$ there is a gain between 0 and 10 msec for the average of all tracks. For the other error probabilities there is a loss of around 5 msec. Track 54 is mainly responsible for the loss, for all other tracks ROHC mostly causes a gain. Overall, our results indicate that ROHC does not significantly deteriorate the delay jitter. Note that — in contrast to the widely studied packet delay jitter with ROHC — throughout this section we have considered the delay jitter in the received voice signal, which is closer related to the user's perception.

VII. CONCLUSIONS

In this paper we have provided a tutorial on an evaluating methodology for transmitting voice with robust header compression over a wireless link. Our methodology employs elementary objective voice quality
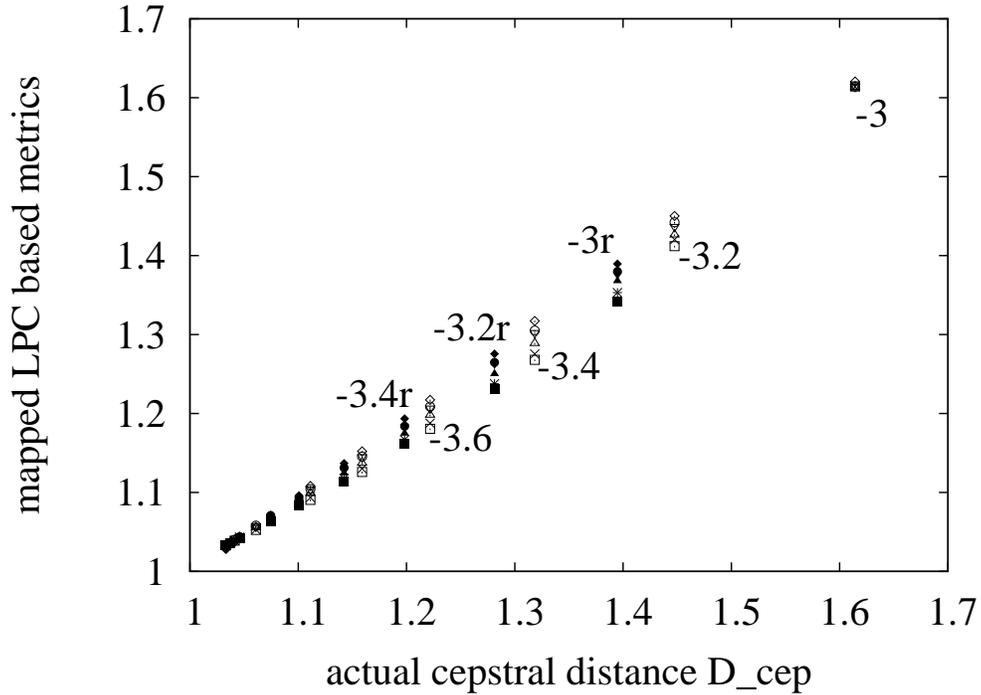
Fig. 12

SCATTER PLOT OF CEPSTRAL DISTANCE OBTAINED FROM LINEAR MAPPINGS OF OTHER LPC BASED METRICS AS A
FUNCTION OF ACTUAL CEPSTRAL DISTANCE.

metrics which predict the subjective voice quality with good reliability. In addition, our methodology employs a segmental cross correlation (SCC) algorithm to synchronize the received (distorted) voice signal with the original (reference) signal. This synchronization makes the elementary objective voice quality metrics robust and usable for the evaluation of modern packet voice communication systems. We note that both elementary and psycho-acoustic voice quality metrics do generally not include synchronization and are therefore not directly applicable to packet voice. The main innovation of PESQ [15] [16] over previous perceptual metrics is the synchronization of the voice signals. By combining the SCC synchronization with elementary objective voice quality metrics we provide an alternative evaluation methodology for packet voice quality. Our tutorial makes the objective voice quality metrics and the SCC algorithm readily accessible and employable by networking researchers to evaluate similar voice communication systems.

Our evaluations of RObust Header Compression (ROHC) indicate that with ROHC the header size is reduced by approximately 85%, which for the considered GSM encoded voice with 33 byte GSM frames cuts the total bandwidth required for the voice transmission almost in half. (This reduction of the total bandwidth is expected to be even larger for lower bit rate encoders with smaller voice frames.) Our extensive voice quality evaluations, which employ the presented objective voice quality metrics indicate
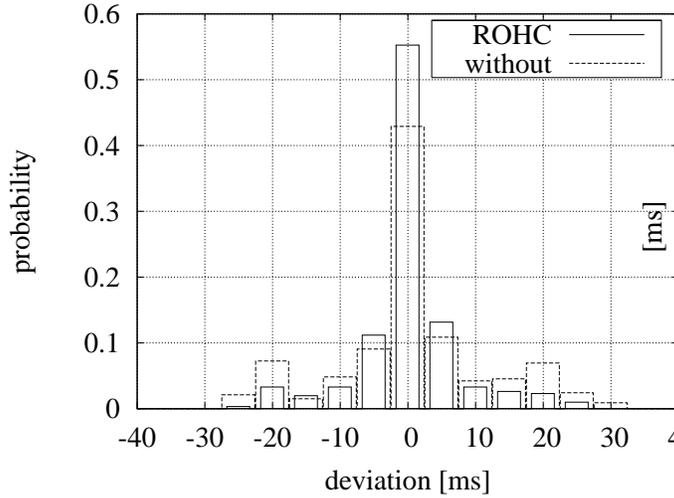
Fig. 13

TYPICAL DELAY JITTER HISTOGRAM FOR A TRANSMISSION
WITH AND WITHOUT ROHC. THE PROBABILITY OF A DELAY
BETWEEN -2.5 AND $+2.5$ MSEC IS HIGHER FOR ROHC
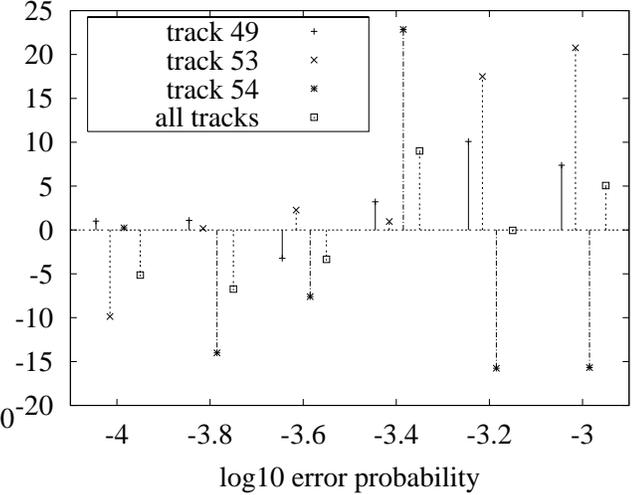TRANSMISSIONS.

Fig. 14

JITTER GAIN FOR ROHC: NEGATIVE GAIN (I.E., LARGER
JITTER) FOR ERROR PROBABILITIES $10^{-4} \ldots 10^{-3.6}$,
POSITIVE GAIN (I.E., SMALLER JITTER) FOR $10^{-3.4} \ldots 10^{-3}$.

that this enormous reduction in used bandwidth does not deteriorate the voice quality. On the contrary, the voice quality is improved by ROHC. All of the considered parametric and spectral distances indicate improvements in the objective voice quality. In addition, the cepstral distance predicts a subjective quality improvement of 0.26 on the 5-point Mean Opinion Score (MOS) for a wireless bit error probability of 10^{-3}. Our phase timing measurements indicate that ROHC does not significantly deteriorate the delay jitter in the voice signal. One explanation for the improved voice quality is that the smaller packets with ROHC are more resistant against wireless link errors. Overall, we note that even if the voice quality improvements with ROHC are moderate and barely perceivable in many practical settings (with ambient noise), the compression gain of ROHC promises remarkable benefit for wireless service providers. The number of 3rd generation mobile cell phone users could nearly be doubled by employing ROHC without allocating more link bandwidth.

APPENDIX

APPENDIX A: WINDOW FUNCTION

As detailed in the following appendices, the considered objective quality metrics rely on autocorrelations in the speech signal. For such autocorrelation based metrics it is generally beneficial to avoid sharp discontinuities in the time domain by multiplying the voice frames with a window function. Such *windowing* generally reduces the prediction error. We use an M-point Hamming window defined by

$$w(m) = 0.54 - 0.46 \cdot \cos \frac{2\pi m}{M - 1}, \qquad m = 1, \ldots, M \tag{22}$$

and illustrated in Figure 15. The windowed voice signal is obtained by

$$x_{n,d/\phi}(m) = x_{d/\phi}(m + (n - 1)M) \cdot w(m), \tag{23}$$

where $x_{d/\phi}(\cdot)$ denotes the "unwindowed" and "unframed" voice signal. The "framed" and "windowed" signal $x_{n,d/\phi}(m)$ is used for all the calculations detailed in the following appendices. (We note that only the SNR metrics defined in (5) and (6) use the "unwindowed" but "framed" voice signal.)
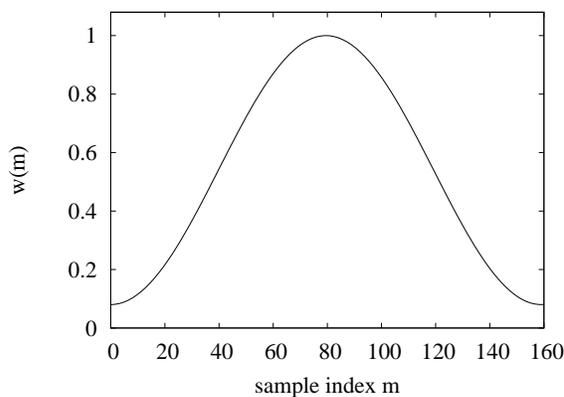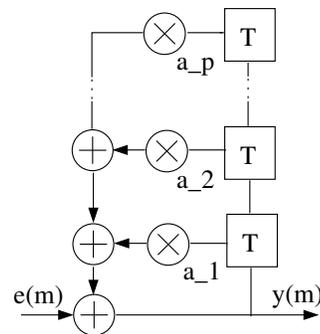

Fig. 15

HAMMING WINDOW

Fig. 16

LPC FILTER: A PURELY RECURSIVE, DIGITAL FILTER. $e(m)$ AND $y(m)$ ARE SAMPLES IN THE TIME DOMAIN.

APPENDIX B: LINEAR PREDICTIVE CODING (LPC) ANALYSIS

The spectral and parametric distances are based on a *linear predictive coding* (LPC) analysis, which gives the feedback coefficients $a(i)$. The coefficients $a(i)$ allow the prediction of a speech sample $y(m)$ by a linear, weighted sum of its p previous values. The *predicted* speech sample $\hat{y}(m)$ is given as

$$\hat{y}(m) = \sum_{i=1}^{p} a(i) \cdot y(m - i), \tag{24}$$

where p is called the LPC *model order*. Figure 16 illustrates the LPC model, a recursive digital filter with the input $e(m)$ and the output $y(m)$. The relationship between output and input is given by

$$y(m) = e(m) + \sum_{i=1}^{p} a(i) \cdot y(m - i), \tag{25}$$

where $e(m)$ is called the *prediction error*. In order to approximate the speech sample $y(m)$ by the *predicted* sample $\hat{y}(m)$, the LPC analysis minimizes the mean squared error $e(m)$:

$$\min \sum_{m=1}^{M} e^2(m). \tag{26}$$

From (26), a linear system of equations is derived. This system can be solved by the so-called *Levinson-Durbin* recursion (LDR). With the help of the LDR, the p feedback coefficients $a(i)$ for each frame n are computed by the following set of equations. Let $R_n(k)$ denote a set of autocorrelation functions, which can be calculated for the distorted or the undistorted signal:

$$R_{n,d/\phi}(k) = \sum_{\forall m} x_{n,d/\phi}(m) \cdot x_{n,d/\phi}(m+k), \text{ for } 0 \leq k \leq p. \tag{27}$$

From now on, the distinction d/ϕ is left out to simplify the notation. The feedback coefficients refer to either the distorted or the undistorted signal. The principle of calculating the feedback coefficients of an LPC model with the order p is to calculate the coefficients for all lower model orders. Without loss of generality we consider the order $p = 10$, which is typically used for voice quality evaluations. We define $a^{(n)}(i)$ to be the ith feedback coefficient of an LPC model with the order n. For $i = n$, the coefficients $a^{(n)}(n)$ are calculated as

$$a^{(n)}(n) = \frac{R_n(n) - \sum_{i=1}^{n-1} a^{(n-1)}(i) \cdot R_n(n-i)}{E(n-1)} \quad \langle =: -K(n)\rangle \tag{28}$$

with

$$E(n) = \{1 - [a^{(n)}(n)]^2\} \cdot E(n-1), \qquad n = [0, \ldots, 9] \tag{29}$$

and the initial conditions

$$E(0) = R_n(0), \qquad a^{(1)}(1) = \frac{R_n(1)}{R_n(0)}.$$

The negative coefficients $a^{(n)}(n)$ equal the so-called Parcor coefficients $K(n)$, in particularly $K(n) = -a^{(n)}(n)$, which are used for calculating the log area ratio metric. If the coefficient number i does not equal the model order n ($i \neq n$), the feedback coefficients are given by

$$a^{(n)}(i) = a^{(n-1)}(i) - a^{(n)}(n) \cdot a^{(n-1)}(n-i), \qquad n > i. \tag{30}$$

When the desired model order p ($= 10$) is obtained, the feedback coefficients are given by

$$a(i) = a^{(p)}(i). \tag{31}$$

We have listed a set of formulas to compute the feedback coefficients. As many of these formulas are recursive, the question of a possible computation order concerning the coefficients has to be answered. Based on (28), (29), and (30), we developed the following computation scheme, which can be used for the implementation:

$$\begin{matrix} a^{(1)}(1) & & & \\ a^{(2)}(2) & a^{(2)}(1) & & \\ a^{(3)}(3) & a^{(3)}(2) & a^{(3)}(1) & \\ \vdots & \vdots & & \ddots \\ a^{(p)}(p) & a^{(p)}(p-1) & \ldots & a^{(p)}(1) \end{matrix} \tag{32}$$

This scheme describes the computation order of the coefficients $a^{(n)}(i)$. The calculation has to be performed in the order $a^{(1)}(1), a^{(2)}(2), a^{(2)}(1), \ldots, a^{(p)}(1)$. The last line reflects the actual feedback coefficients with the model order p.

APPENDIX C: SPECTRAL MAGNITUDE

Let $G(n, d/\phi)$ denote the gain factor

$$G(n, d/\phi) = \left[R_{n,d/\phi}(0) - \sum_{k=1}^{10} a_{n,d/\phi}(k) \cdot R_{n,d/\phi}(k) \right]^{1/2}, \tag{33}$$

where $R_n(k)$ is given by (27). The spectral distance measures all contain a spectral, frame related magnitude

$$V_{d/\phi}(n, l) = \left| \frac{G(n, d/\phi)}{1 - \sum_{k=1}^{10} a_{n,d/\phi}(k) \cdot e^{-jk\frac{\pi l}{128}}} \right|. \tag{34}$$

Equations (33) and (34) assume a model order of 10. The gain factor is typically set to one, as the overall level does not influence the perception [18]. The ten feedback coefficients $a(k)$ are calculated by the LPC analysis.

APPENDIX D: AUTOCORRELATION MATRIX $\mathbf{R}$ AND LPC VECTOR $\vec{a}$

Let $\mathbf{R}_\phi(n)$ denote the *autocorrelation* matrix

$$\mathbf{R}_\phi(n) = \begin{bmatrix} R_{n,\phi}(0) & R_{n,\phi}(1) & \ldots & R_{n,\phi}(10) \\ R_{n,\phi}(1) & R_{n,\phi}(0) & \ldots & R_{n,\phi}(9) \\ \vdots & \vdots & \ddots & \vdots \\ R_{n,\phi}(10) & R_{n,\phi}(9) & \ldots & R_{n,\phi}(0) \end{bmatrix} \tag{35}$$

and $\vec{a}_{d/\phi}(n)$ devote the LPC vector

$$\vec{a}_{d/\phi}(n) = \begin{bmatrix} 1 \\ -a_{n,d/\phi}(1) \\ -a_{n,d/\phi}(2) \\ \vdots \\ -a_{n,d/\phi}(10) \end{bmatrix}. \tag{36}$$

In (35), $R_{n,\phi}(k)$ is defined by (27). In (36), $a_{n,d/\phi}(k)$ are the LPC coefficients and are derived from (31).

APPENDIX E: CEPSTRAL COEFFICIENTS

The LPC–derived cepstral coefficients are given by

$$c_{n,d/\phi}(l) = a_{n,d/\phi}(l) + \frac{1}{l} \sum_{k=1}^{l-1} [l - k] \cdot c_{n,d/\phi}(l - k) \cdot a_{n,d/\phi}(k), \qquad \text{for } 2 \leq l \leq L \tag{37}$$

with the conditions

$$a_{n,d/\phi}(0) = 1, \qquad a_{n,d/\phi}(k) = 0 \text{ for } k > p$$
$$\text{and } c_{n,d/\phi}(0) = 0, \qquad c_{n,d/\phi}(1) = a_{n,d/\phi}(1). \tag{38}$$

L is the number of cepstral coefficients, which we choose equal to the order of the LPC analysis from which the p predictor coefficients $a(i)$ are derived.

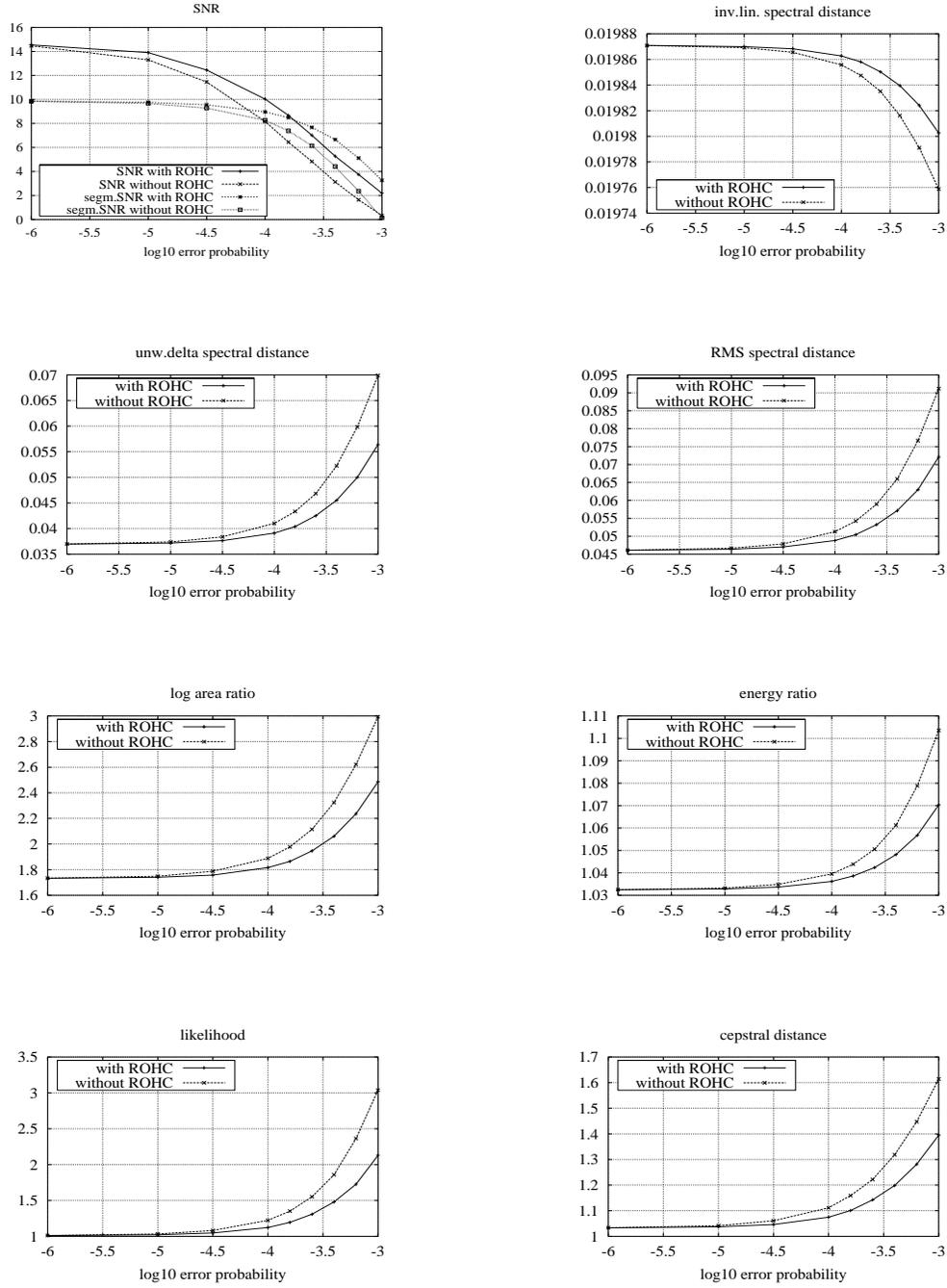APPENDIX F: NUMERIC VALUES OF OBJECTIVE QUALITY METRICS

Fig. 17

NUMERIC VALUES OF OBJECTIVE QUALITY MEASURES FOR ALL TRACKS.

REFERENCES

[1] C. Bormann, C. Burmeister, M. Degermark, H. Fukushima, H. Hannu, L.-E. Jonsson, R. Hakenberg, T. Koren, K. Le, Z. Liu, A. Martensson, A. Miyazaki, K. Svanbro, T. Wiebke, T. Yoshimura, and H. Zheng, "RObust Header Compression (ROHC): Framework and four profiles: RTP, UDP, ESP, and uncompressed," July 2001.

[2] F. Fitzek, P. Seeling, and M. Reisslein, "Header compression schemes for wireless internet access," in *Wireless Internet: Technologies and Applications*, A. Salkintzis and A. Poularikas, Eds. CRC Press, 2004, preprint available at `http://www.eas.asu.edu/~mre`.

[3] A. Cellatoglu, S. Fabri, S. Worrall, A. Sadka, and A. Kondoz, "Robust header compression for real–time services in cellular networks," in *Proceedings of the Second International Conference on 3G Mobile Communication Technologies*, London, UK, Mar. 2001, pp. 124–128.

[4] F. Fitzek, S. Hendrata, P. Seeling, and M. Reisslein, "Video quality evaluation for wireless transmission with robust header compression," acticom GmbH, Tech. Rep., July 2003, available at `http://www.fitzek.net/publication.html` and `http://www.eas.asu.edu/~mre`.

[5] G. Waters, "Sound quality assessment material — recordings for subjective tests: User's handbook for the ebu – sqam compact disk," European Broadcasting Union (EBU), Tech. Rep., 1988, available at `http://www.ebu.ch/tech_32/tech_t3253.pdf`.

[6] J. Degener and C. Bormann, "GSM 06.10 lossy speech compression," 1994, available at `http://kbs.cs.tu-berlin.de/~jutta/toast.html`.

[7] M. Rossi, A. Philippini, and M. Zorzi, "Link error characteristics of dedicated (DCH) and common (CCH) UMTS channels," Universita di Ferrara, FUTURE Group, Italy, Tech. Rep., July 2003.

[8] ITU-T Recommendation P.800.1, "Mean opinion score (MOS) terminology," March 2003.

[9] ITU-T Recommendation G.107, "The E-model, a computional model for use in transmission planning," May 2000.

[10] T. A. Hall, "Objective speech quality measure for internet telephony," in *Proceedings of SPIE Voice over IP VoIP Technology*, vol. 4522, Denver, CO, July 2001, pp. 128–136.

[11] A. Estepa, R. Estepa, and J. Vozmediano, "On the suitability of the E-model to VoIP networks," in *Proceedings of the IEEE International Symposium on Computers and Communications (ISCC)*, Taormina, Italy, July 2002, pp. 511–516.

[12] S. Wang, A. Seley, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE Journal on Selected Areas in Communications*, vol. 10, no. 5, pp. 819–829, June 1992.

[13] S. Voran, "Objective estimation of perceived speech quality, Part I: Development of the measuring normalizing block technique," *IEEE Transactions on Speech and Audio Processing*, vol. 7, pp. 371–382, July 1999.

[14] ——, "Objective estimation of perceived speech quality, Part II: Evaluation of the measuring normalizing block technique," *IEEE Transactions on Speech and Audio Processing*, vol. 7, pp. 383–390, July 1999.

[15] A. W. Rix, M. Hollier, A. Hekstra, and J. Beerends, "Perceptual evaluation of speech quality (PESQ), the new ITU standard for end-to-end speech quality assessment. Part I – time-delay compensation," *Journal of the Audio Engineering Society*, pp. 755–764, Oct. 2002.

[16] J. B. Beerends, A. P. Hekstra, A. W. Rix, and M. P. Hollier, "Perceptual evaluation of speech quality (PESQ), the new ITU standard for end-to-end speech quality assessment. Part II – psychoacoustic model," *Journal of the Audio Engineering Society*, pp. 765–778, Oct. 2002.

[17] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, no. 1, pp. 67–72, Feb. 1975.

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

[19] K. Lam, O. Au, C. Chan, K. Hui, and S. Lau, "Objective speech measure for chinese in wireless environment," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-95)*, vol. 1, Detroit, MI, May 1995, pp. 277–280.

[20] ——, "Objective speech quality measure for cellular phone," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-96)*, vol. 1, Atlanta, GA, May 1996, pp. 487–490.

[21] N. Kitawaki, H. Nagabuchi, and K. Itoh, "Objective quality evaluation for low-bit-rate speech coding systems," in *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 2, Feb. 1988, pp. 242–248.

[22] S. Wu and L. Pols, "A distance measure for objective quality evaluation of speech communication channels using also dynamic spectral features," in *Proceedings of the Institute of Phonetic Sciences Amsterdam (IFA)*, vol. 20, 1996, pp. 27–42.

[23] A. Gray and J. Markel, "Distance measures for speech processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 5, pp. 380–391, Oct. 1976.

[24] C. Perkins, O. Hodson, and V. Hardman, "A survey of packet loss recovery techniques for streaming audio," *IEEE Network*, vol. 12, no. 5, pp. 40–48, September/October 1998.

[25] E. O. Brigham, *The Fast Fourier Transform and its Applications*. Prentice–Hall, 1988.