Voice Quality Evaluation for Wireless Transmission with ROHC

Stephan Rein*
Dept. of Electrical Eng.
Technical University Berlin
Germany
email: rein@asu.edu

Frank H.P. Fitzek
acticom GmbH
Berlin
Germany
email: fitzek@acticom.de

Martin Reisslein†
Dept. of Electrical Eng.
Arizona State University
Tempe, AZ, USA
email: reisslein@asu.edu

ABSTRACT
Robust Header Compression (ROHC) has recently been proposed to reduce the large protocol header overhead when transmitting voice and other continous media over RTP/UDP/IP in wireless networks. In this paper we evaluate the transmission of Global System Mobile telecommunications (GSM) encoded voice with ROHC over a wireless link. We evaluate the bandwidth savings and the voice qualities using a wide array of objective voice quality metrics including SNR metrics, spectral distance metrics, as well as parametric distance metrics. We find that for a wide range of loss probabilities on the wireless link, ROHC roughly cuts the bandwidth required for voice transmissions in half. In addition, our extensive voice quality evaluations indicate that ROHC improves the voice quality compared to transmissions without ROHC, especially for large bit error probabilities on the wireless link. The improvment increases exponentially from about 0.075 for an error probability of 10^{-4} to 0.36 for an error probability of 10^{-3} on the 5-point Mean Opinion Score.

KEY WORDS
Internet Protocol, Mean Opinion Score, Mobile Multimedia, Robust Header Compression, Voice Quality

1 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). One major problem with the RTP/UDP/IP protocol architecture is the large overhead, which affects the limited bandwidth of mobile channels. A low bitrate speech application can result in IP packets with a ratio of 30 bytes of payload to 60 bytes of overhead. Recently, RObust Header Comression (ROHC) [1] has been proposed to compress the protocol headers for packet transmission over a wireless link.

In this paper we evaluate ROHC for the packetized transmission of voice 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 segmental Signal to Noise (SNR) ratios, 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.

We find that for a wide range of bit error probabilities on the wireless link, ROHC reduces the protocol overhead for voice transmission with IPv4 by approximatly 85%, which reduces the bandwidth required for a typical voice transmission by about 47%. We find that 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 0.025 for an error probability of $10^{-4.5}$ and reaches 0.175 and 0.36 as the error probability increases to $10^{-3.5}$ and 10^{-3}. We also find that ROHC slightly increases the jitter for small error probabilities and slighly reduces the jitter for large error probabilities.

1.1 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 and 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 does not reflect the 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-

[*] Currently visiting Arizona State University.

[†] Corresponding Author: M. Reisslein, Dept. of Electrical Eng., Goldwater Ctr., MC 7206, Tempe, AZ 85287-7206, USA, phone: (480)965-8593, fax: (480)965-8325, email: reisslein@asu.edu, web: http://www.eas.asu.edu/˜mre

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.

2 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. The IPv6 version requires a 40 bytes 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 typically 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 originally layer generated headers, see [2] for details.

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

$$gain_{\max} = \frac{headersize}{headersize + payload}. \qquad (1)$$

With an 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 bitrate audio streams, where the header size is typically larger than the payload.

3 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) taken from the European Broadcasting Union [4]. The files, given in the wave file mono format, are at first down sampled to 8 kHz and then transferred to the communication system shown in Figure 1. On the sender's side the wave file is GSM encoded (using the encoder [5]). The coded file consisting of 33 byte 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 [6] have revealed that uncorrelated bit errors give a good approximation at 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}. We repeat each experiment numerous times with independent bit errors to obtain 95% confidence intervals that are smaller than 10% of the corresponding sample mean.

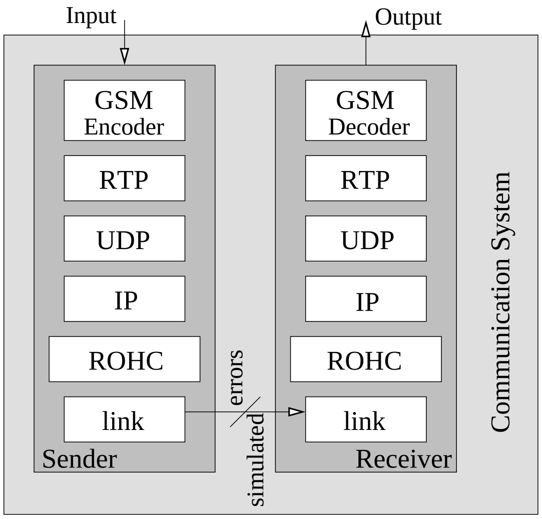

Figure 1. Communication System: Two Linux PCs, ROHC is optional.

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

Table 1. Mean Opinion Score

4 Voice Quality Evaluation Metrics

A very reliable method to evaluate a speech communication system is to perform speech perception tests with human listeners. One reliable scheme to conduct subjective measurements is described by ITU-T Recommendation P.830, the mean opinion score, as shown in Table 1. Subjective measurements are expensive and time-consuming. Therefore, significant effort has been devoted to developing an objective, computer based metric in order to predict the results of a subjective evaluation. The realiability of objective metrics is usually verified by a correlation analysis between the calculated metric and hearing tests among a distorted data base. A fundamental analysis has been conducted by Quackenbush, Barnwell, and Clement [7], who evaluated numerous at that time available objective metrics in the time and frequency domain. We measure the voice quality using elementary, objective metrics, proposed in [7]. We selected the set of metrics to obtain a wide range of distortions. Table 2 gives the correlations to subjective

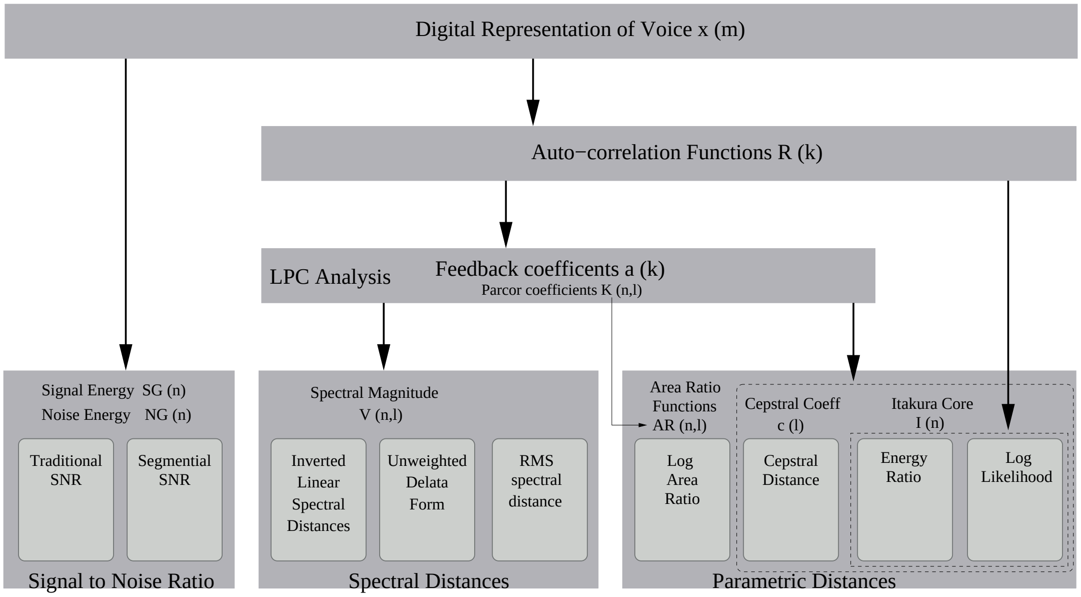

Figure 2. Set of used objective voice quality metrics. The calculations are partially similar, but the metrics have different scopes.

Metric	Correlation	Reference
(traditional) SNR	$+0.24/+0.31$	[7]
segmental SNR	$+0.77/+0.78$	[7]
inv. linear distance	$+0.63/+0.48$	[7]
unw. delta form	$-0.61/-0.51$	[7], [8]
root mean square	*theor. approach*	[9]
Log Area Ratio	$-0.62/-0.65$	[7]
Energy Ratio	$-0.59/-0.61$	[7], [10]
log likelihood	$-0.49/-0.48$	[7], [11], [10]
Cepstral distance	-0.93	[12], [13]

Table 2. Performance of elementary objective metrics. Correlations are given for different distortion types, see [14] for details.

hearing tests. The traditional SNR has a very poor performance. However, it is included because of its relevance in terms of purely objective voice quality. The 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. As illustrated in Figure 2, many metrics use the same coefficients and are similarly calculated. However, their performance differs among different types of distortions, as verified in [7], [8], [11], and [12]. Due to space constraints we have to keep this overview of the voice quality metrics necessarily short, we refer the interested reader to [14] for more details on these metrics and their calculation. We close this section by noting that one could employ the very complex (and expensive) PESQ [15] measure for the voice quality evaluation. Instead, we selected the objective metrics in Table 2, which have also good correlations with the subjective voice quality (especially the cepstral distance) and do not require proprietary software (and thus allow for replication of our experiments; in fact we plan to make our evaluation software publicly available to the research community).

5 Segmental Cross Correlation algorithm (SCCA)

We transfer voice over a communication system. Thereby, parts of the voice file can be delayed, other parts can be lost. The objective quality is based on a comparison between the distorted and the reference file. To synchronize these files, we have developed a synchronization algorithm in the time domain, the *segmental cross correlation algorithm* (SCCA). For every frame n of the reference file a frame $\hat{n}$ in the distorted file has to be found. The frames $\hat{n}$ are stringed and finally form the *reconstructed* file, which is synchronized to the reference file.

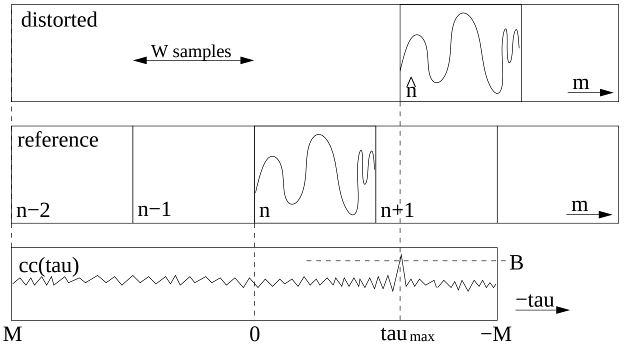

Figure 3. Principle of SCCA: For every frame n of the reference file, a frame $\hat{n}$ of the distorted file is matched.

Let M devote the maximally allowed displacement in samples between the distorted and the undistorted frames, each of the length W samples. To match a frame $\hat{n}$ to a frame n, we use a cross correlation function $CC_n(\tau)$ proposed in [16] to estimate a $\tau_{\max}$, the actual displacement between the frames $\hat{n}$ and n:

$$\tau_{max}(n) = \max\ CC_n(\tau),\ -M \leq \tau \leq M. \qquad (2)$$

Figure 3 illustrates that with (2), for every τ a correlation is calculated, where $\tau = \tau_{\max}$ attains the maximum correlation. This calculation is repeated for the entire sequence of N frames of a voice file.

6 Evaluation Results

In this section we give an overview of our extensive evaluations of voice transmission with ROHC. Due to space constraints we present here only a representative sample of our results and refer the interested reader to [14] for more details.

6.1 ROHC network gain

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

$$header\ gain = 1 - \left(\frac{size\ ROHC\ header}{size\ uncompressed\ header}\right). \qquad (3)$$

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. Now the compression gain for the total RTP/UDP/IP packet with a payload of 33 bytes can be 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 Eqn. (1). Next we address the question whether this significant reduction in consumed bandwidth affects the voice quality.

6.2 Voice quality gain

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	$D - D_{ROHC}$

Table 3. Gain definitions for different metrics.

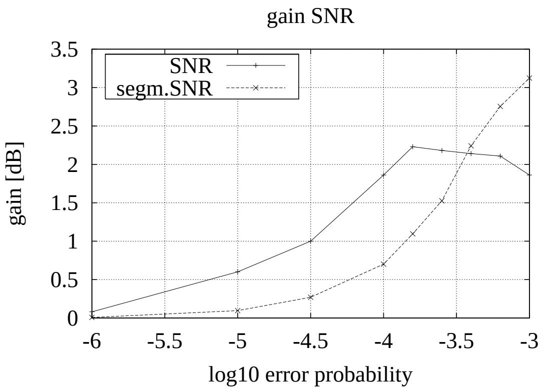

Figure 4. ROHC gain for SNR measures.

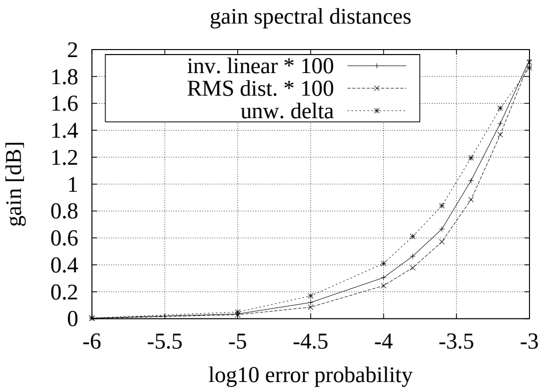

Figure 5. ROHC gain for spectral distances.

As we analyzed three different tracks, we calculated the mean value of all tracks and thereby obtained a better statistical reliability. (Note that the individual track measurements were repeated many times with independent bit errors to obtain 95% confidence intervals less than 10% of the corresponding sample mean.) We use a decibel notation to simplify the analysis of the quality gain achieved by ROHC. Table 3 gives the gain definitions used for Figures 4, 5, and 6. All metrics show an increasing gain with higher error rates. As an exeption, the gain for the traditional SNR slightly decreases for error rates higher than $10^{-3.8}$. 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. Overall, these results indicate that the voice quality does not suffer from header compression, on the contrary, it is improved, especially for high link error probabilities. Note that these gain values in dB represent the improvement in terms of objective voice quality and not in terms of

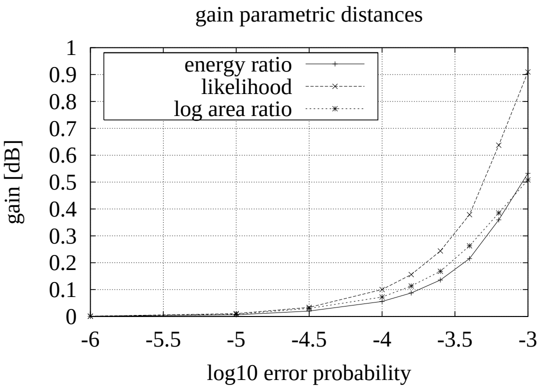

Figure 6. ROHC gain for parametric distances.

user perception. To asses the impact on the user perception we investigate the improvements on the subjective 5 point scale (Table 1) next.

We transform the values of the cepstral distance to the predicted mean opinion score (MOS), using the mapping verified in [12]. Let c devote the voice quality calculated by the cepstral distance. The MOS value is given by

$$MOS = 4.0013 - 2.7227 \cdot c + 0.35735 \cdot c^2. \qquad (4)$$

We define the MOS gain for ROHC as

$$MOS_{gain} = MOS_{\mathrm{w}ROHC} - MOS_{\mathrm{w}/oROHC}. \qquad (5)$$

As shown in Figure 7, the predicted gain for ROHC in terms of the MOS increases roughly exponentially with increasing error probability and reaches 0.36 for error probabilities of 10^{-3}.

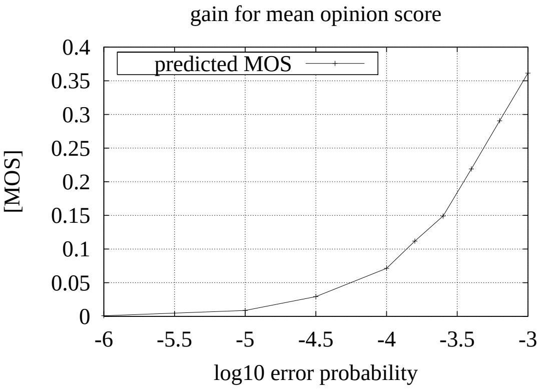

Figure 7. Predicted ROHC gain for mean opinion score.

6.3 Delay analysis

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 SCCA 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. (To capture the entire range of jitters we set the maximum total time window of the SCCA algorithm to the voice file length in the experiments reported in this section.)

We examine both the delay jitter histogram and the standard deviation of the delay jitter. Figure 8 shows a typical histogram of delay jitter for the 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 9 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 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 detoriate 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.

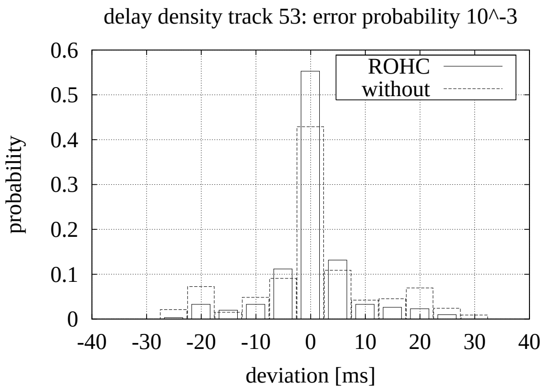

Figure 8. Typical delay jitter histogram for a transmission with and without ROHC. The probability of a delay between -2.5 and $+2.5$ msec is higher for ROHC transmissions.

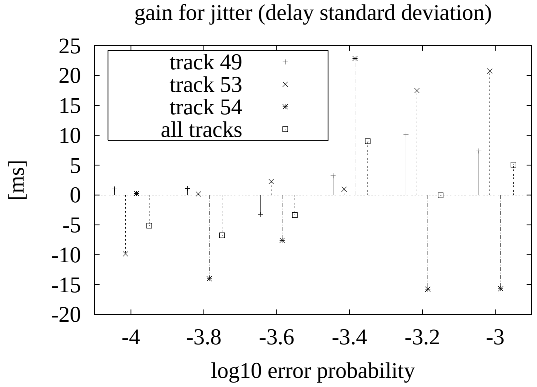

Figure 9. Jitter gain for ROHC: negative gain for error probabilities $10^{-4} \ldots 10^{-3.6}$, positive gain for $10^{-3.4}$ and 10^{-3}.

7 Conclusions

We have examined the performance of RObust Header Compression (ROHC), an IP header compression scheme, for voice applications in 3rd generation mobile networks. Our evaluations of the compression gain indicate that with ROHC the header size is reduced by 85% and the total IP packet size is almost cut in half. As our extensive voice quality evaluations indicate, this enormous reduction in used bandwidth does not detoriate 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 objetive voice quality. In addition, the cepstral distance predicts a subjective quality improvement of 0.36 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 detoriate 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.

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, *Wireless Internet: Technologies and Applications*. CRC Press, 2004, ch. Header Compression Schemes for Wireless Internet Access, to be published.

[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] G. Waters, *Sound Quality Assessment Material — Recordings For Subjective Tests: User's Handbook for the EBU – SQAM Compact Disk*, European Broadcasting Union (EBU), 1988. [Online]. Available: http://www.ebu.ch/tech_32/tech_t3253.pdf

[5] J. Degener and C. Bormann, "GSM 06.10 lossy speech compression," 1994. [Online]. Available: http://kbs.cs.tu-berlin.de/∼jutta/toast.html

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

[11] K. Lam, O. Au, C. Chan, K. Hui, and S. Lau, "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.

[12] S. Wu and L. Pols, "A distance measure for objective quality evaluation of speech communication channels using also dynamic spectral features," Institute of Phonetic Sciences, University of Amsterdam, Tech. Rep., 1996.

[13] N. Kitawaki, K. Itoh, M. Honda, and K. Kakehi, "Comparison of objective speech quality measures for voiceband codecs," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '82)*, vol. 7, May 1982, pp. 1000–1003.

[14] S. Rein, F. Fitzek, and M. Reisslein, "Voice quality evaluation for wireless transmission with ROHC (extended version)," Dept. of Electrical Eng., Arizona State University, Tech. Rep., May 2003, available at `http://www.eas.asu.edu/˜mre`.

[15] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ) — a new method for speech quality assessment of telephone networks and codecs," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '01)*, vol. 2, Salt Lake City, UT, 2001, pp. 749–752.

[16] P. Bourke, "Cross correlation, autocorrelation, 2d pattern identification," Aug. 1996. [Online]. Available: http://astronomy.swin.edu.au/∼pbourke/analysis/correlate