ACTA ACUSTICA UNITED WITH ACUSTICA
Vol. 90 (2004) 1052 – 1060

Corrupted Speech Data Considered Useful: Improving Perceived Speech Quality of VoIP over Error-Prone Channels

Florian Hammer, Peter Reichl, Tomas Nordström
Telecommunications Research Center Vienna FTW, Donau-City-Str. 1, 1220 Wien, Austria.
{hammer, reichl, nordstrom}@ftw.at

Gernot Kubin
Signal Processing and Speech Comm. Lab (SPSC), Univ. of Technol. Graz, Austria. gernot.kubin@TUGraz.at

Summary
The provisioning of an appropriate level of perceptual speech quality is crucial for the successful deployment of Voice over the Internet Protocol (VoIP). Today's heterogeneous multimedia networks include links that introduce bit errors into the voice data stream. These errors are detected by the IP packet transport protocol and result in packet losses which eventually degrade the speech quality. However, modern speech coding algorithms can either conceal packet losses or tolerate corrupted packets. In this paper, we investigate to which extent it makes sense to keep corrupted speech data for the special case of uniformly distributed bit errors. We simulate different transport strategies that allow the incorporation of damaged speech data into the speech decoding process. The results from an instrumental speech quality evaluation show that keeping as much damaged data as possible leads to superior performance with regard to the perceptual speech quality compared to dropping packets and using packet loss concealment.

PACS no. 43.71.Gv, 43.72.Kb

1. Introduction

The advent of the Internet in the 1990s has launched an increasing interest in packet-based telephony. First packet voice transport experiments have been accomplished in the mid-1970s, but it took about 20 years to introduce an application to the public [1]. Besides the existence of a well established circuit-switched telephone network, one of the major reasons for the slow evolution of Internet telephony may be that the Internet as such has primarily been designed to support the transmission of non-interactive, non-realtime data. In contrast, interactive applications like telephony require reliable *and* in-time data delivery, otherwise no user would accept the service. Therefore, the network providers need to maintain a certain level of quality of service (QoS) [2, 3].

Compared to the public (circuit-) switched telephone network (PSTN), the "packet-nature" of VoIP exhibits transmission impairments of its own, like packet loss, packet delay, and packet delay jitter. Packet loss results from one of the major obstacles within an IP network, i.e. congestion: if too many users send lots of data at once, router queues become overloaded, and packets need to be dropped. In addition, time-varying traffic causes variations of the packet delay, the so-called jitter, which increases the probability that packets cannot be incorporated into the speech reconstruction process because they have not been received in time. Adaptive buffers can alleviate this problem by buffering packets and delaying their playout time to compensate for the varying network delay [4, 5], while however, the absolute end-to-end delay must be limited to allow a fluent conversation. For a more detailed description of VoIP speech impairments we refer to [6].

In this paper, we are interested in a further impairment of the voice packet stream that occurs mainly in the so-called access network, connecting the user's fixed or mobile terminal to an IP backbone network. Here, even in state-of-the-art broadband access technologies like digital subscriber lines (DSL, wireline) or the universal mobile telecommunication system (UMTS, wireless), transmission impairments introduce bit errors. In fact, the amount of errors represented by the *bit error rate* (BER) serves as an indicator for the quality of the transmission channel. It is important to note that losing only one bit within a packet may have dramatic consequences, as the IP voice packet transport network is designed to simply drop erroneous packets, which results in the loss of the entire information within such packets.

In case of data transmission, the transmission control protocol (TCP, [7]) running on top of the Internet Protocol (IP, [8]) cares for this problem by employing a retransmis-

Received 8 November 2003,
accepted 18 February 2004.

sion mechanism, but at the cost of increased transmission delay. In order to avoid this effect and to meet the real-time constraints, VoIP is based on the user datagram protocol (UDP, [9]) which does not retransmit lost packets.

Speech decoders deal with the packet loss problem by substituting a lost speech entity according to a packet loss concealment (PLC) algorithm [10], e.g., by repeating the last received packet. The loss concealment allows to decrease the perceptual impact caused by the loss of information, but anyhow degrades the speech quality. On the other hand, modern speech codecs can tolerate a certain amount of damaged (but nevertheless delivered) data, especially if the speech bits are ordered according to their perceptual sensitivity and only less important bits are damaged.

This paper presents a performance evaluation of such mechanisms. We have simulated and compared the performance of traditional and modified transport schemes as introduced in [11], where the latter either employ selected parts or even all of the damaged data. The perceptual speech quality resulting from this alternative approach has been evaluated with instrumental quality measurement methods, where "instrumental" refers to the fact that the quality is estimated by computer algorithms instead of being assessed by test persons[1]. In this way, we compare the modified transport schemes with traditional VoIP transport and show that keeping all of the damaged data results in superior performance.

The remainder of the paper is structured as follows: In section 2, we present the techniques that we have used for simulating the incorporation of damaged speech data and for the evaluation of the resulting perceptual speech quality. Furthermore, we briefly review related work. In section 3, we specify three strategies that utilize the techniques for VoIP transport over error-prone links as introduced above. Section 4 presents the framework of the environment in which the transport strategies have been simulated. Our results are presented and discussed in section 5. Finally, we draw conclusions from our work in section 6.

2. Background

In this section, we explain the techniques that facilitate error-tolerant VoIP transport. Based on these techniques, we will propose various strategies for transmitting voice data over error-prone links in section 3. First, we present UDP-Lite, a UDP modification allowing bit errors in the payload. We then introduce the adaptive multi-rate (AMR) speech coding algorithm with its ability to substitute lost packets and to distinguish bits concerning their perceptual sensitivity. Robust header compression can save bandwidth by reducing the huge amount of header information resulting from the IP real-time transmission protocol stack. Then, we present the instrumental speech quality measurement methods we have applied to compare our transmission strategies in terms of perceptual quality. Finally, we briefly review some related work concerning this topic.

[1] However, such methods depend on information obtained from subjective listening tests. Thus, we avoid the term "objective" measurement which is widely used in the literature (see, e.g., [12]).

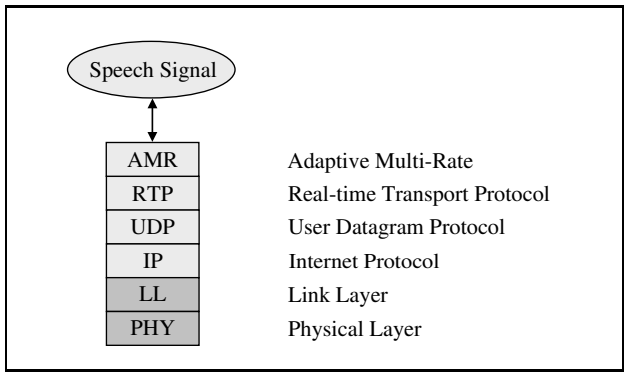

Figure 1. Layer Model.

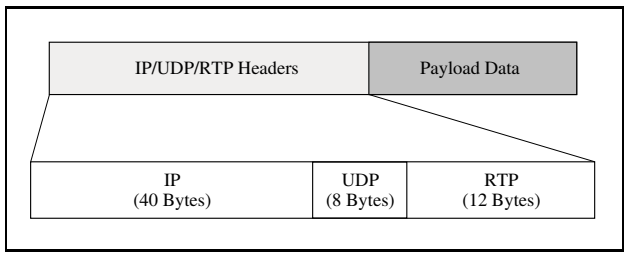

Figure 2. IP/UDP/RTP packet structure.

For convenience, the layer model we use is depicted in Figure 1. PHY and LL represent the physical layer and link layer, respectively. These lower layers handle the physical transmission of data either over a wire or radio link. In this paper, we are concerned about the layers above the link layer.

2.1. UDP-Lite

IP-telephony is based on the real-time transport protocol (RTP, [13]) and the user datagram protocol (UDP, [9]).

The structure of a voice over IP packet is illustrated in Figure 2. Note that the 20 Bytes of IP header contain amongst others a length field that indicates the total length of the packet, and a checksum that may be used to detect errors in the IP header itself, but *not* in the IP payload (the carried data).

In contrast, UDP protects both header *and* payload by calculating and adding a checksum to each of the packets at the sender side. Thus, routers can detect bit errors by recalculating the checksum and comparing it with the original. A difference between these numbers indicates that one or more bits in the packet have been corrupted, and as a consequence, the packet is stopped from being forwarded. Furthermore, UDP does not retransmit a packet if it got lost along its way to the receiver because, for real-time traffic, there is simply no time to wait for a retransmitted packet. Therefore, any packet corrupted by bit errors gets lost.

For a more detailed illustration, Figure 3 shows the header of UDP containing the source and destination port

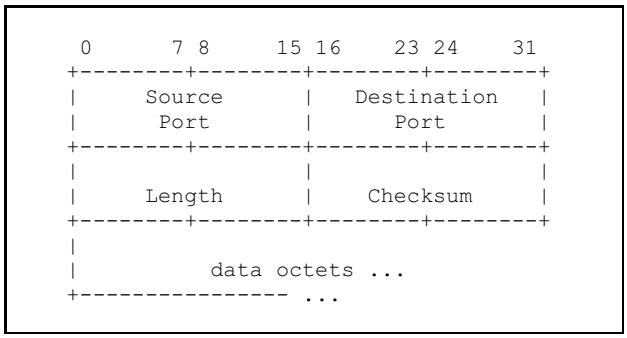

Figure 3. UDP Header Format [9].

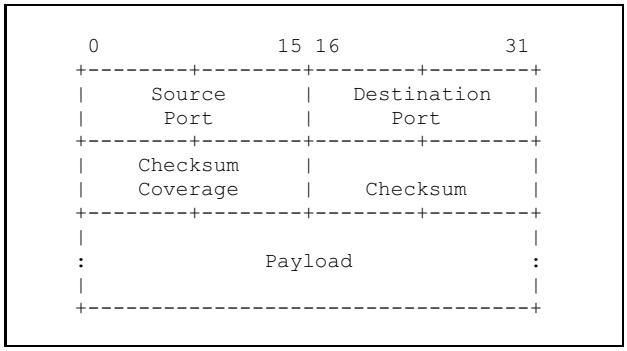

Figure 4. UDP-Lite Header Format [14].

numbers, the packet length (including the 8 UDP header bytes), and a checksum that is calculated over both header and payload data.

The functionality of the length and checksum fields has been slightly varied by Larzon et al. [14]. Their proposal, the so-called UDP-Lite, allows for checksums that cover the payload only partially. To this end, the length field is substituted by a field that defines the checksum coverage size, as depicted in Figure 4. Therefore, only the first part of the payload is covered by the checksum, whereas bit errors are allowed towards the tail end of the payload, assuming that the link layer supports the forwarding of damaged information.

2.2. Adaptive Multi-Rate Speech Coding

The IP/UDP-Lite/RTP protocol stack provides the transmission of speech data over the Internet (cf. Figure 1). In this section, we will describe important features of the adaptive multi-rate (AMR) speech codec which transforms the speech signal into a set of data frames and vice versa. Moreover, the AMR codec is especially suited for our investigations because the Internet Engineering Task Force (IETF) has defined an RTP payload format that allows for employment in an all-IP system.

The AMR speech codec [15] was originally developed for the (circuit-switched) global system for mobile communications (GSM) and has then been chosen as a mandatory codec for third generation (3G) cellular systems [16]. Speech signals, sampled at 8 kHz, are processed in frames of 20 ms, and coded to bitrates ranging from 4.75 to 12.2 kbps. Thus, in a circuit-switched mobile communication system, the codec can adapt its bitrate and the corresponding error protection according to the quality of the wireless transmission channel. The worse the channel quality, the lower the bitrate chosen, and the higher the respective error protection.

Like most of today's speech codecs, the AMR codec features an internal packet loss concealment (PLC) method, discontinuous transmission, and unequal error protection (UEP). Thus, it provides the flexibility and robustness needed for deployment in packet-based networks. For our explorations, we are mainly interested in the AMR codec's capability of providing UEP, and in its PLC algorithm.

Unequal error protection is provided at the coder's side by ordering the speech data bits of a frame according to their perceptual importance. The importance levels are referred to as class A (most sensitive), class B, and class C (least sensitive). If an entire speech frame is lost or if A-bits are corrupted during the transmission, it is recommended to forget about the corrupted packet and to use the internal PLC algorithm [17] instead. Otherwise, the damaged B/C-bits may be used.

This bit ordering feature is fundamental for the construction of our strategies for error tolerant speech data transport. For our purposes, we have chosen the 12.2 kbps mode of the AMR which produces 244 speech bits per frame. These bits are divided into 81 A-bits, 103 B-bits, and 60 C-bits ([18], cf. figure 5).

The *packet loss concealment* algorithm [17] works as follows. If a speech frame has been lost, the PLC algorithm substitutes this frame by utilizing adapted speech parameters of the previous frames. In principle, the gain of the previous speech frame is gradually decreased, and the frequency parameters are shifted towards the overall mean of the previous frames. It is important to note that the AMR codec has an internal state that includes the samples required for long-term and short-term prediction, and a memory for predictive quantizers. However, aside from missing speech information, packet losses may lead to the de-synchronization of the encoder and the decoder which results in error propagation. In other words, the decoder needs some time to recover from the lost data.

The standard-compliant transport of AMR frames over IP has been defined by the IETF by specifying corresponding RTP payload formats [19]. An RTP payload format consists of the RTP payload header, payload table of contents, and payload data. Payloads may contain one or more speech frames. For our simulations, we have chosen the "bandwidth efficient mode" payload format (one frame per packet) that is illustrated in Figure 5. The H and T fields represent the payload header and TOC (table of contents) field, respectively, and sum up to 10 bits.

2.3. Robust Header Compression

IP/UDP(-Lite)/RTP transport of speech data results in a major drawback regarding the transmission efficiency. The protocol headers in total form a 320 bits large cluster (20

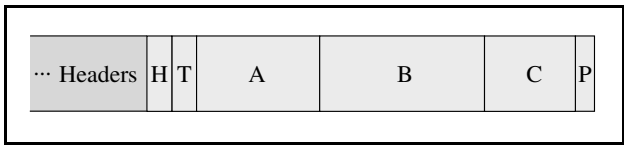

Figure 5. AMR RTP payload format: Bandwidth efficient mode [19]. H: RTP payload header (4 bits), T: RTP payload table of contents (6 bits), A: 81 Class A speech bits, B: 103 Class B speech bits, C: 60 Class C speech bits, P: 2 Padding bits.

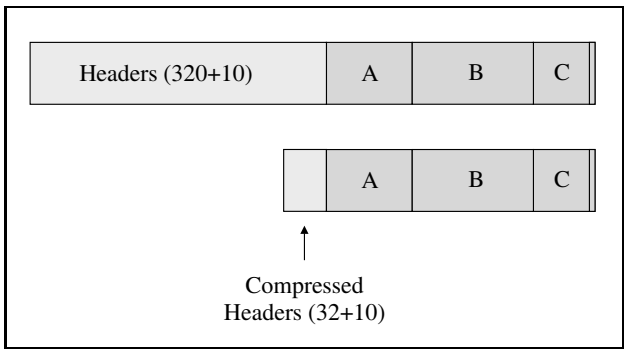

Figure 6. Efficiency of robust header compression.

Bytes IP, 8 Bytes UDP, 12 Bytes RTP) of administrative overhead. Assuming that a packet carries only one speech frame and contains 256 actual payload bits, this overhead comprises more than half of the total packet size. Hence, the majority of packets are lost due to bit errors in the headers when sent over an error-prone serial link.

Robust Header Compression (ROHC, [20]) resolves this problem by utilizing redundancy between header fields within the header and in particular between consecutive packets belonging to the same packet stream. In this way, the overhead can be reduced to a minimum. The term "robust" expresses that the scheme tolerates loss and residual errors on the link over which header compression takes place without losing additional packets or introducing additional errors in decompressed headers. ROHC profiles for UDP-Lite are defined in [21].

In our simulations, we reduce the header size from 40 to 4 Bytes, i.e. 10% of its original size. Figure 6 illustrates the efficiency of the compression. Note that the 10 bits of RTP payload header and TOC are additionally included in the headers.

2.4. Speech Quality Evaluation

Instead of quality of service as characterized by technical parameters, users are first of all concerned about the quality of service as perceived by themselves, because they want to communicate in a comfortable way without having to care about the underlying technology.

Perceived quality is primarily measured in a subjective way. To this end, test persons rate the quality of the media either in listening-only or conversational tests. The subjective assessment of speech quality is addressed in ITU-T Recommendation P.800 [22]. Absolute Category Rating

Table I. Absolute Category Rating scale ([22]).

Speech quality	Score
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

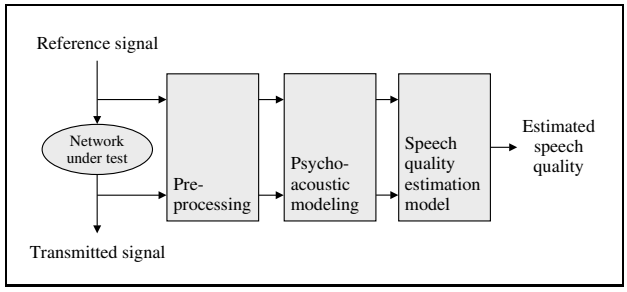

Figure 7. "Intrusive" Instrumental Perceptual Speech Quality Assessment.

(ACR) is the most common rating method for speech quality listening tests. It is based on the listening-quality scale shown in Table I. The quantity evaluated from the score averaged over the complete set of test persons is called Mean Opinion Score (MOS).

Currently, a lot of research effort is invested in developing algorithms that derive an instrumental measure of the perceived quality, often referred to as "objective" measure. So-called "intrusive" instrumental speech quality assessment algorithms compare a degraded speech signal with its undistorted reference in the perceptual domain, and estimate the corresponding speech quality. This principle is shown in Figure 7. In comparison, "non-intrusive" instrumental assessment methods do not require a reference signal, but estimate the perceptual speech quality by measuring network parameters.

In our experiments, we evaluate the perceptual quality of the speech samples resulting from our simulations by using ITU-T Rec. P.862 "Perceptual Evaluation of Speech Quality" (PESQ, [23]) and the "Telecommunication Objective Speech Quality Assessment" (TOSQA, [24]). Compared to PESQ, an important feature of TOSQA is a modification of the reference signal by utilizing an estimated transfer function of the spectral distortions. Therefore, some of the effects of linear distortions can be balanced.

The quality assessment methods introduced above require reference speech samples. For our investigations, we have chosen the PHONDAT speech sample database [25] that contains phonetically rich German sentences recorded in studio-quality (16 bit/16 kHz). The results are also applicable to English speech samples. We have selected 12 sentence pairs spoken by 4 talkers (2 female and 2 male). The speech samples were down-sampled to 8 kHz, modified-IRS [26] filtered and normalized to an active

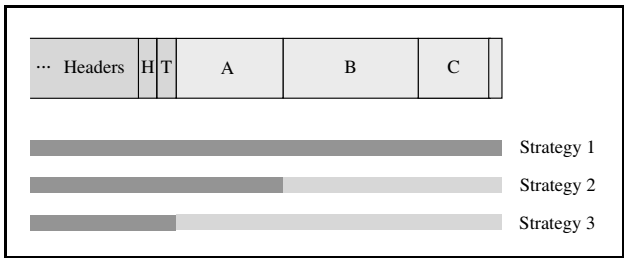

Figure 8. UDP-Lite checksum coverage.

Table II. Packet drop strategies.

Corrupted part of packet	Strategy 1	2	3
Header	drop	drop	drop
A-bits	drop	drop	keep
B/C-bits	drop	keep	keep

Table III. UDP-Lite checksum coverage and coverage degrees.

Strategy	No ROHC [bits]	α	ROHC [bits]	α
1	576	1	288	1
2	411	0.71	123	0.43
3	330	0.57	42	0.15

speech level of -26 dBov (units of dB relative to overload, [27]).

2.5. Related Work

After presenting the technical background on which we base our investigations, we give a brief overview of related work.

The application of UDP-Lite for video transmission over wireless links has been explored by Singh et al. [28]. In that work, GSM radio frame error traces have been collected in a cellular IP testbed, and have then been used to simulate the transmission of video streams over a wireless link. Compared to traditional UDP, the use of UDP-Lite provides 26% less end-to-end delay, constant inter-arrival time of the packets, slightly higher throughput, and 50% less packet losses. The perceptual quality is claimed to be significantly higher, but neither subjective nor instrumental quality assessment has been accomplished.

In addition to packet loss concealment, forward error correction (FEC) can be used to compensate for packet losses [10]. At the cost of bandwidth and delay, either Reed-Solomon (RS) block coded data [29] or low bit-rate redundancy data (LBR), i.e. a low quality version of the same speech signal, are added as redundant information within one of the following voice packets or in a separate packet. Jiang [30] shows that LBR performs worse, with regard to the perceptual speech quality, than the use of FEC in terms of RS-codes.

However, in our study we aim to investigate the impact of bit errors on the transmitted speech data, so we do not facilitate any additional FEC or channel coding for our simulations, with the exception of perceptual bit ordering.

3. Alternative Strategies for VoIP Transport

This section introduces the strategies following [11]. We explore the impact of using erroneous speech data on the perceived speech quality by defining three strategies which handle corrupted packets in different ways. The strategies are based on IP/UDP-Lite/RTP transport of AMR speech frames facilitating different UDP-Lite checksum coverage which is illustrated in Figure 8 and Table II. In addition, we apply ROHC to the headers.

We define the strategies as follows:

- *Strategy 1* simply corresponds to traditional IP transport, hence the UDP-Lite checksum covers the entire UDP payload. If any data is corrupted, the packet is lost and substituted by the receiver's PLC. Including the traditional transport method into the simulations constitutes a reference with regard to the speech quality performance.
- In accordance with the AMR standard, *strategy 2* permits B- and C-bits to be faulty, but detects errors within the header and the class A bits. Thus, a reasonable amount of packets with erroneous B- and C-bits can be saved.
- *Strategy 3* exhibits the most tolerant behavior. All of the payload data are allowed to be corrupted, consequently a packet is only dropped when the header is corrupted. All of the corrupted speech data can be incorporated in the reconstruction of the speech signal.

Under any strategy, the IPv4 header is protected by its own checksum.

In order to further characterize the strategies, we introduce the *coverage degree* α as a parameter that corresponds to the relation of the checksum coverage N' to the total packet length N (including the headers),

$$\alpha = \frac{N'}{N}. \tag{1}$$

Hence, a coverage degree of $\alpha = 0$ means that none of the data is covered by the checksum, and a coverage degree of $\alpha = 1$ indicates that the entire packet is covered by the checksum. Note that with smaller coverage degree α, fewer packets are discarded due to bit errors (cf. section 4.2).

Table II summarizes the properties of the three proposed strategies and Table III provides the corresponding values of the coverage degree α. Note for the extreme case of strategy 3, the use of ROHC may reduce the coverage degree down to 0.15.

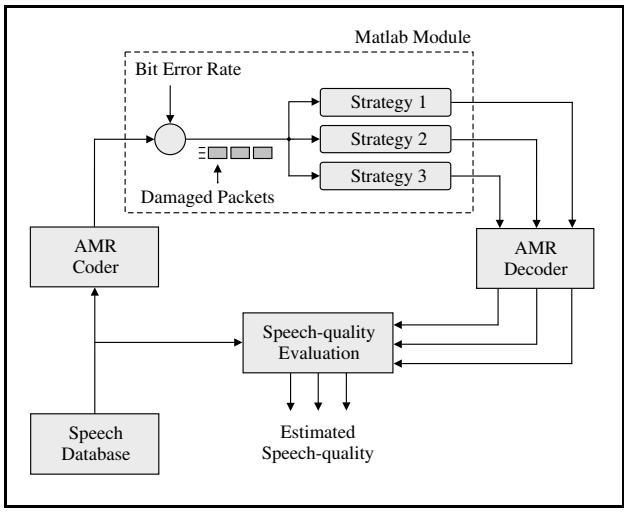

Figure 9. Simulation Environment.

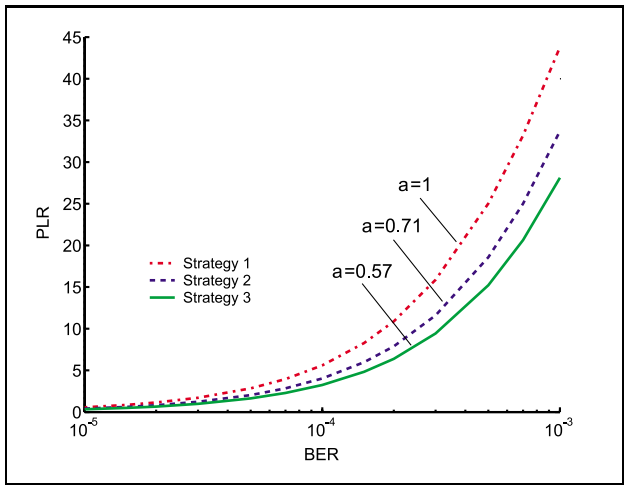

Figure 10. Relation between Packet Loss Rate and Bit Error Rate (Without ROHC).

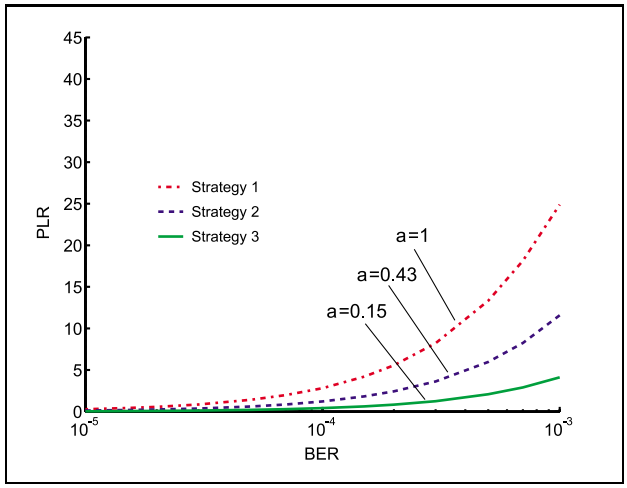

Figure 11. Relation between Packet Loss Rate and Bit Error Rate (Using ROHC).

4. Simulations

4.1. Simulation Environment

The simulation environment, as depicted in Figure 9, represents an example for the interworking between signal processing and networking methods. It contains the following parts: the speech database, AMR speech encoding and decoding, a Matlab [31] module simulating the strategies specified in section 3 for different bit error rates, and the perceptual speech quality assessment unit.

After coding a speech sample, the voice bitstream is processed corresponding to each of the three transport strategies. To obtain a good resolution of the area of decreasing speech quality, we have chosen 13 bit error rates between 10^{-5} and 10^{-3}. The three bitstreams are then decoded, and instrumental measurements estimate the perceptual quality of the degraded speech samples. This procedure is repeated 24 times per bit error rate per speech sample in order to get good average values of the resulting speech quality.

4.2. Bit Error Model

As already mentioned in the introduction, digital transmission of data over wireline or wireless access networks can result in a certain amount of bit errors. The amount and the distribution of the bit errors can be controlled by channel coding. In this study, we assume the special case that the channel coding at the physical link provides uniformly distributed bit errors. We further assume that the lower system layers provide support for UDP-Lite by forwarding erroneous data to the upper layers.

Based on these assumptions, the number of bit errors X that occur in an actual packet is calculated using the binomial distribution

$$X \sim B(N, p), \tag{2}$$

where N represents the packet size [bits] and p represents the bit error rate.

The location of the erroneous bits within the packet is then uniformly distributed over the packet.

To be able to present the effects of keeping damaged data in detail, we deal with bit error rates ranging from 10^{-5} to 10^{-3} which can be expected for wireless channels. For DSL, the BER is typically controlled at 10^{-7}. However, our choice of bit error rates might be relevant for "customized" wireline techniques like "Channelized Voice over DSL".

At this range of the bit error rate, the behaviors of the strategies highly affect the amount of packets lost due to bit errors. The packet loss rate, PLR, depends on the bit error rate p according to

$$PLR(p) = 1 - (1 - p)^{\alpha N}, \tag{3}$$

where α represents the checksum coverage degree as defined in Equation (1). Figures 10 and 11 depict the packet loss relations among the three strategies without and with ROHC, respectively. The graphs illustrate that the loss rate is substantially reduced by compressing the header and by reducing the checksum length.

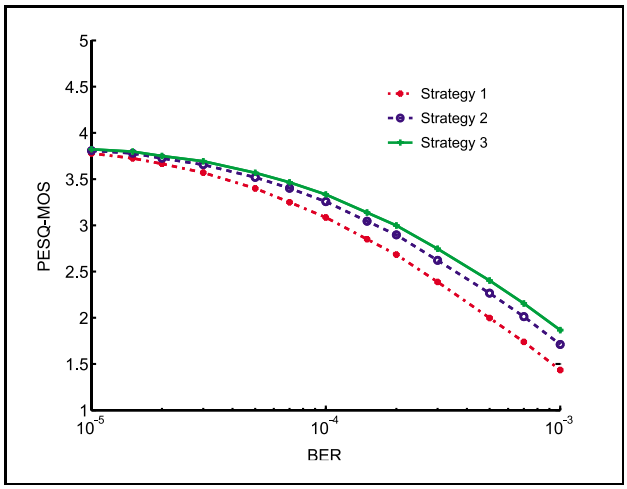

Figure 12. PESQ-MOS vs. BER: Without ROHC.

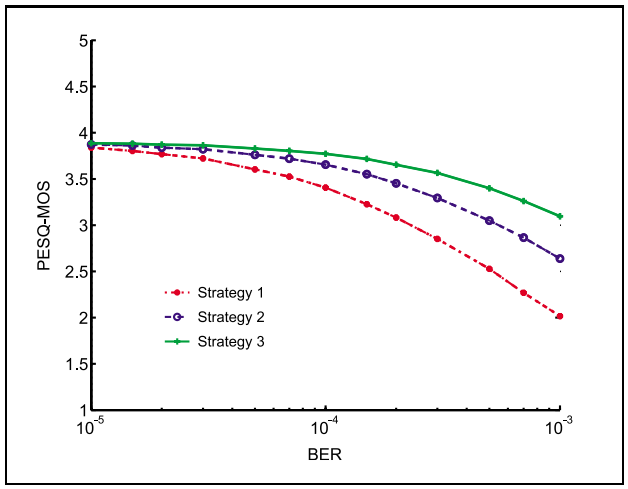

Figure 13. PESQ-MOS vs. BER: Using ROHC.

5. Results and Discussion

5.1. Estimated Perceived Speech Quality vs. Bit Error Rate

At first, we compare the performance of the strategies with regard to the perceptual speech quality estimated by PESQ as a function of the bit error rate. We have chosen PESQ as the main tool for the quality evaluation, since it is widely used and has been standardized by the ITU-T. The results for the non-header compressed case are shown in Figure 12. The differences in quality are noticeable for strategies 2 and 3 compared to strategy 1. Strategy 3 performs best, although the average improvement compared to strategy 2 is only marginal. The standard deviations of the speech quality MOS estimates are around 0.14, 0.25, and 0.19 at bit error rates of 10^{-5}, 10^{-4}, and 10^{-3}, respectively, for all strategies.

In less than 1% of the test cases strategy 2 performs better than strategy 3 for the non-header compressed case. However, this behavior can only be observed at very low bit error rates.

When the packet header is compressed, strategy 3 significantly outperforms strategy 2. Figure 13 shows that at a bit error rate of 10^{-3}, strategy 3 results in a perceptual quality that is half a MOS point higher compared to strategy 2, and an increase of more than one MOS point compared to strategy 1. The standard deviations of the MOS estimates for strategies 1/2/3 are 0.25/0.20/0.14 for a bit error rate of 10^{-4}, and 0.21/0.24/0.23 for a bit error rate of 10^{-3}. Similar to the non-header compressed case, strategy 2 performs better than strategy 3 at low bit error rates in only 0.7% of the test cases. This underlines the consistent trend of the results.

As a result, we conclude that applying the packet loss concealment in case of erroneous A-bits performs worse than keeping them for decoding. This result may reflect the fact that employing corrupted data saves a considerable amount of packets from being dropped (cf. section 4.2). As the codecs maintain an internal state, they need some time to recover from a lost packet. Employing damaged speech data introduces artifacts but avoids error propagation. We conclude that for a certain bit rate, a damaged speech data packet is of significantly higher "perceptual value" than its substitution by the loss concealment. We regard this conclusion to be one of the central results of our investigations.

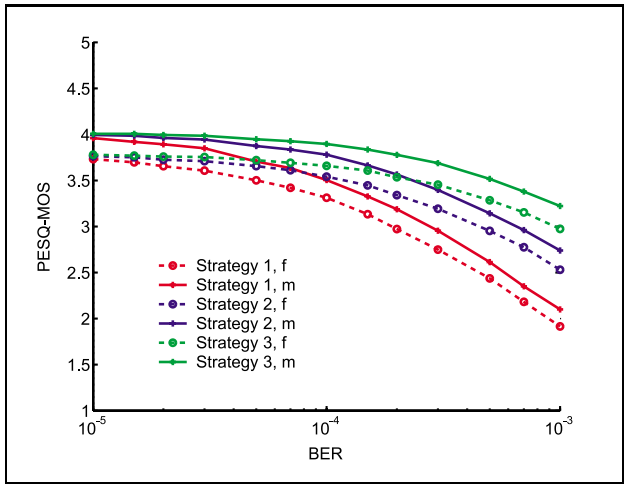

Figure 14. Gender dependency of the speech quality estimated by PESQ when ROHC is used.

5.2. Gender Dependency

The dependency of the estimated speech quality on the gender of the talkers is shown for the header compressed case in Figures 14 and 15 for PESQ and TOSQA, respectively.

The PESQ results indicate a significant difference between samples of female and male speakers. At low bit error rates, male voices are rated about 0.23 MOS points higher than female voices. For strategies 1 and 2, this difference decreases with increasing bit error rate, while for strategy 3 it slightly increases. At 10^{-3}, the differences are 0.18, 0.21, and 0.25 for strategies 1, 2, and 3, respectively.

In contrast, quality evaluation using TOSQA results in marginally better quality for female voices for all strategies at low bit error rates. Additionally, at high bit er-

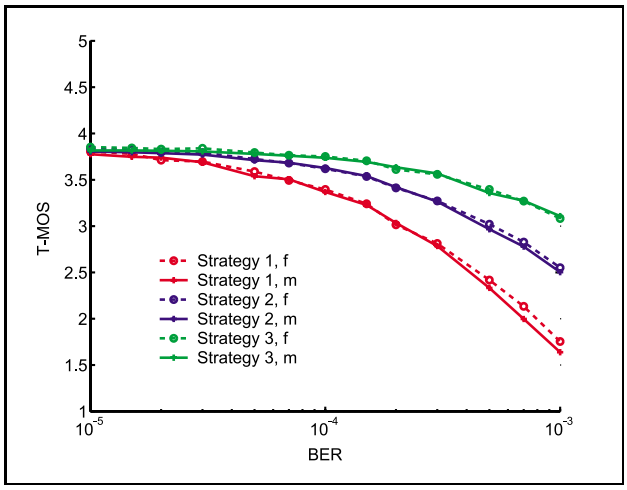

Figure 15. Gender dependency of the speech quality estimated by TOSQA when ROHC is used.

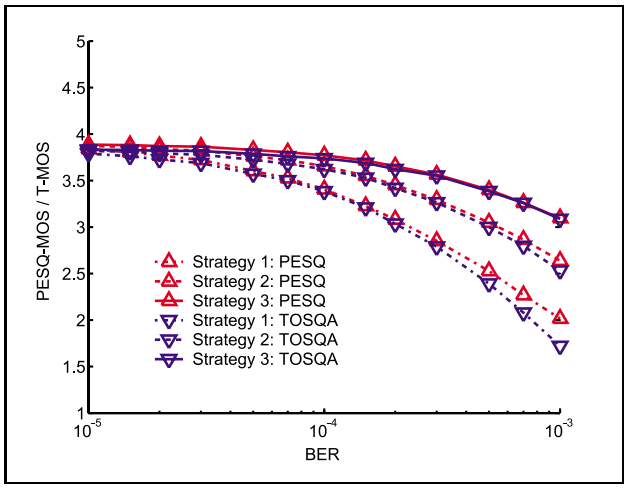

Figure 16. A comparison of PESQ and TOSQA results (Using ROHC).

ror rates, female voices are rated slightly better than male voices. This small difference decreases with an increasing amount of damaged data that is incorporated into the speech decoding process.

The difference between the PESQ and TOSQA evaluation results may have two reasons:

- The *speech coding algorithm* results in different speech quality for female and male speakers.
- The *instrumental speech quality assessment methods* behave in different ways.

From subjective tests conducted at AT&T for AMR characterization[2] [32] we imply that there is no significant talker dependency for the AMR codec.

A closer look at the input filter responses at the preprocessing stages of PESQ and TOSQA (cf. Figure 7) offers a possible explanation for its different behavior. PESQ cuts the signal energy below 250 Hz and applies an IRS receive filter [33] to the input signal. In comparison, TOSQA uses an input frequency response that is based on acoustic handset measurements [34]. This frequency response is, especially at the lower frequencies, more bandlimited than the frequency response of the PESQ input filter.

The different input filter responses may be the main reason for the difference of the results regarding the gender of the talkers. However, this issue is out of the scope of this paper, and needs to be investigated in more detail.

5.3. PESQ vs. TOSQA

As a final result, we present a comparison of the mean speech quality estimates given by PESQ and TOSQA. The results of both methods are given in Figure 16 for the header compressed case. The major observation is that for strategies 1 and 2, TOSQA estimates a lower MOS compared to PESQ at high bit error rates. At a BER of 10^{-3}, the differences in estimated quality are 0.3 and 0.1 for strategies 1 and 2, respectively. On the contrary, PESQ as well as TOSQA provide equal quality results at higher bit error rates for strategy 3.

As we can observe from Figures 14 and 15, the difference between the results of PESQ and TOSQA seems to be caused mainly by the different ratings obtained for speech samples of male talkers.

In any case, we may conclude that the TOSQA results approve the trend, indicated by the PESQ results, that the use of all corrupted data results in superior speech quality.

6. Conclusions

In this paper, we have simulated a VoIP system making use of speech data that have been corrupted due to bit errors. We have distinguished traditional VoIP transport which drops damaged packets, the use of corrupted data that is perceptually less sensitive, and the incorporation of all available erroneous data into the speech decoding process.

The results of an instrumental perceptual speech quality evaluation clearly indicate that keeping all damaged speech packets in combination with robust header compression results in superior performance compared to dropping the damaged packets and utilizing the packet loss concealment algorithm at the receiver. It is especially remarkable that the dropping of packets which contain perceptually sensitive corrupted speech bits does not yield a gain in quality compared to the use of all erroneous data. Thus, in fact all corrupted speech data have to be considered useful.

As future work, the instrumental speech quality assessment results concerning strategies 2 and 3 remain to be evaluated subjectively using comparison category rating [22]. In order to enhance the scope with regard to the bit error distribution, different levels of bit error bursts will be studied in the future. Finally, testing the performance of the Internet Low Bit-rate Codec (iLBC,[35]) as an alternative to the AMR codec will also be interesting because the iLBC includes an improved packet loss concealment method which reduces the impact of error propagation due to the codec state.

[2] As well as our simulations, these experiments were based on the speech material including 2 female and 2 male talkers.

Acknowledgement

Part of this work has been funded under the Austrian government's Kplus Competence Center Program.

The authors would like to thank Jens Berger and Paolo Usai for their valuable input and the useful discussions regarding the gender dependency of the instrumental quality assesment methods and the AMR, respectively. Furthermore, we thank T-Systems Nova for providing TOSQA.

References

[1] H. Schulzrinne: Converging on internet telephony. IEEE Internet Computing **3** (1999) 40–43.

[2] M. A. El-Gendy, A. Bose, K. G. Shin: Evolution of the Internet QoS and support for soft real-time applications. Proc. of the IEEE **91** (2003) 1086–1104.

[3] V. Firoiu, J.-Y. Le Boudec, D. Towsley, Z.-L. Zhang: Theories and models for internet quality of service. Proc. of the IEEE **90** (2002) 1565–1591.

[4] R. Ramjee, J. Kurose, D. Towsley, H. Schulzrinne: Adaptive playout mechanisms for packetized audio applications in wide-area networks. Proc. IEEE INFOCOM, 1994, 680–688.

[5] S. B. Moon, J. Kurose, D. Towsley: Packet audio playout delay adjustment: Performance bounds and algorithms. ACM/Springer Multimedia Systems **6** (1998) 17–28.

[6] A. Raake: Predicting speech quality under random packet loss: Individual impairment and additivity with other network impairments. Appearing in this issue (2004).

[7] J. Postel: Transmission control protocol. RFC 793 (1981).

[8] J. Postel: Internet protocol. RFC 791 (1981).

[9] J. Postel: User datagram protocol. RFC 768 (1980).

[10] C. Perkins, O. Hodson, V. Hardman: A survey of packet loss recovery techniques for streaming audio. IEEE Network **12** (1998) 40–48.

[11] F. Hammer, P. Reichl, T. Nordström, G. Kubin: Corrupted speech data considered useful. Proc. First ISCA International Tutorial and Research Workshop on Auditory Quality of Systems, Mont-Cenis, Germany, 2003.

[12] European Telecommunications Standards Institute: Speech processing, transmission and quality aspects (STQ); Specification and measurment of speech transmission quality; Part 1: Introduction to objective comparison measurement methods for one-way speech quality across networks. ETSI EG 201 377-1 v1.2.1 (2002).

[13] H. Schulzrinne et al.: RTP: A transport protocol for real-time applications. Request for Comments (Standards Track) RFC 3550, Internet Engineering Task Force (2003).

[14] L.-Å. Larzon et al.: The UDP-Lite protocol. IETF Internet Draft (work in progress), draft-ietf-tsvgq-udp-lite-02.txt (2003).

[15] European Telecommunications Standards Institute: Universal mobile telecommunications system (UMTS); AMR speech codec; General description (3GPP TS 26.071 version 5.0.0 Release 5). ETSI TS 126 071 v5.0.0 (2002).

[16] 3rd Generation Partnership Project (3GPP). http://www.3gpp.org/.

[17] European Telecommunications Standards Institute: Universal mobile telecommunications system (UMTS); AMR speech codec; Error concealment of lost frames (3GPP TS 26.091 version 5.0.0 Release 5). ETSI TS 126 091 v5.0.0 (2002).

[18] European Telecommunications Standards Institute: Universal mobile telecommunications system (UMTS); Mandatory speech codec speech processing functions; Adaptive Multi-Rate (AMR) speech codec frame structure (3GPP TS 26.101 version 5.0.0 Release 5). ETSI TS 126 101 v5.0.0 (2002).

[19] J. Sjöberg, M. Westerlund, A. Lakaniemi, Q. Xie: Real-time transport protocol (RTP) payload format and file storage format for the adaptive multi-rate (AMR) and adaptive multi-rate wideband (AMR-wb) audio codecs. Request for Comments (Standards Track) RFC 3267, Internet Engineering Task Force (2002).

[20] C. Bormann et al.: Robust header compression (ROHC): Framework and four profiles: RTP, UDP, ESP, and uncompressed. Request for Comments (Standards Track) RFC 3095, Internet Engineering Task Force (2001).

[21] G. Pelletier: Robust header compression (ROHC): Profiles for UDP lite. Internet Draft, draft-ietf-rohc-udp-lite-01.txt (2003).

[22] International Telecommunication Union: Methods for subjective determination of transmission quality. ITU-T Recommendation P.800 (1996).

[23] International Telecommunication Union: Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. ITU-T Recommendation P.862 (2001).

[24] International Telecommunication Union: TOSQA - telecommunication objective speech quality assessment. ITU-T COM 12-34 (1997).

[25] Bavarian Archive for Speech Signals (BAS): Phondat 1 corpus. http://www.bas.uni-muenchen.de/Bas/BasPD1eng.html.

[26] International Telecommunication Union: Subjective performance assessment of telephone-band and wideband digital codecs. ITU-T Recommendation P.830 (1996).

[27] International Telecommunication Union: Objective measurement of active speech level. ITU-T Recommendation P.56 (1993).

[28] A. Singh, A. Konrad, A. D. Joseph: Performance evaluation of UDP lite for cellular video. Proc. Int. Workshop Network and Operating Systems Support for Digital Audio and Video NOSSDAV, Port Jefferson, NY, 2001.

[29] R. E. Blahut: Theory and practice of error control codes. Addison-Wesley, NY, 1983.

[30] W. Jiang, H. Schulzrinne: Comparison and optimization of packet loss repair methods on VoIP perceived quality under bursty loss. Proc. Int. Workshop Network and Operating Systems Support for Digital Audio and Video NOSSDAV, Miami Beach, FL, 2002.

[31] Mathworks: Matlab reference guide. The MathWorks, Inc., Natick, MA., 1998.

[32] European Telecommunications Standards Institute: AT&T labs AMR characterization phase final report. ETSI SMG11#11, Tdoc 193/99 (1999).

[33] International Telecommunication Union: Specification for an intermediate reference system. ITU-T Recommendation P.48 (1988).

[34] J. Berger: Instrumentelle Verfahren zur Sprachqualitätsschätzung - Modelle auditiver tests. Dissertation. CAU Kiel, 1998.

[35] S. V. Andersen et al.: Internet low bit rate codec. Internet Draft, draft-ietf-avt-ilbc-codec-04.txt (2003).