Corrupted Speech Data Considered Useful

Florian Hammer[1]*, Peter Reichl*[1]*, Tomas Nordström*[1]*, Gernot Kubin*[2]

[1]Telecommunications Research Center Vienna (ftw.)
[2]Signal Processing and Speech Communication Lab (SPSC),
University of Technology Graz, Austria
{hammer, reichl, nordstrom}@ftw.at, gernot.kubin@TUGraz.at

Abstract

The success of Voice-over-IP technology will crucially depend on the provision of a certain level of perceptual speech quality. In this paper we demonstrate that keeping as much corrupted speech data as possible results in a major step towards this goal. To this end, we explore the combination of two key factors which influence the speech quality as perceived by the user: networking mechanisms and signal processing algorithms. We present three strategies of dealing with voice packets that have been damaged by bit errors. The performance of these strategies is assessed using the standardized PESQ-algorithm. Results of extensive simulations give a detailed insight into the behavior of the different strategies, and thus prove that corrupted speech data in fact have to be considered useful.

1. Introduction

The speech-quality perceived by the user is the ultimate indicator for the transmission quality of today's telecommunication systems. Significant efforts have been made in the areas of signal processing and networking on improving speech transmission algorithms and networking technologies, respectively. However, these areas are usually explored separately, e.g., by improving speech coding algorithms and packet loss concealment methods in the signal processing field, and by developing better queueing mechanisms in the networking area.

In this paper, we investigate the combination of such methods regarding their influence on the perceived speech-quality. This combination is used to deal with the avoidance of voice packet losses caused by bit errors on a wireline (e.g., xDSL) or wireless (e.g., UMTS) link by keeping damaged speech data bits.

The paper is structured as follows. First, we will present some recently developed techniques, which we have used for our investigations: the UDP-Lite protocol, AMR speech coding, and robust header compression. Making use of UDP-Lite, we define alternative strategies for voice packet transport over IP. We then present the environment in which we have simulated these strategies for an error-prone link. Finally, we discuss the simulation results and draw conclusions.

2. Techniques for Improving VoIP Speech-Quality

The popularity of using the Internet for multimedia communications rose in the mid-1990s. Before that time, the Internet has mainly been used for data transmission. The increasing demand for reliable IP-transport of multimedia content spawned an increasing interest of researchers and developers in alternative ways of dealing with Internet Quality of Service (QoS) parameters, e.g., packet loss, as presented in this section.

2.1. UDP-Lite

The User Datagram Protocol (UDP) is commonly used for real-time voice communications over the Internet due to the low protocol processing delays required. UDP's lack of mechanisms for reliable transfer leads to packet drops on links with high bit error rates, such as wireless links, as a result of wrong checksums. But speech codecs that have been developed for such an environment could make efficient use of the damaged data if it would not be dropped. Thus, the perceptual speech-quality could be increased. Bearing this in mind, Larzon *et al.* [1] have developed UDP-Lite, i.e. a modified UDP which allows for a UDP checksum calculation that covers an arbitrary length of the payload. Thus, corrupted speech data can be used for the speech decoding process.

2.2. AMR speech coding

The Adaptive Multi-Rate (AMR) speech codec has been standardized for the Universal Mobile Telecommunications System (UMTS) [2]. The codec's flexibility and robustness allows also for deployment in packet-switched networks, such as the Internet. Like most of today's speech codecs, the AMR codec features an internal frame loss concealment algorithm and provides unequal error protection (UEP). Regarding UEP, the speech bits of each frame are ordered by their impact

on the perceptual speech quality. In particular, the bits are divided into classes A, B, and C, of which class A contains the perceptually most sensitive bits.

The Real-time Transport Protocol (RTP) payload format for AMR speech frames is defined in [3]. In general, the complete payload consists of a payload header, a payload table of contents, and speech data of one or more speech frames.
For our simulations, we decided to use only one speech frame per packet in order to keep the packetization delay and the amount of potential data losses low. We have chosen the "bandwidth efficient mode", in which the RTP payload consists of 10 bits of payload header, the number of speech data bits corresponding to the chosen AMR bitrate, and padding bits as to make the payload octet-aligned. Figure 1 presents this payload definition. The bit distribution between the classes A, B, and C accords with the standard frame structure for the 12.2 kb/s mode [4].

Figure 1: *AMR-RTP payload format ("bandwidth efficient mode").*

HDR ... RTP payload header (4 bits)
TOC ... 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

Due to the features described above, the AMR codec is an excellent candidate for our kind of investigations.

2.3. Robust Header Compression

RTP/UDP-Lite/IP voice packet transport has a major drawback with regard to the efficiency of the transmission. The administrative headers of the protocols sum up to 40 Bytes (20 Bytes for IP, 8 Bytes for UDP-Lite, and 12 Bytes for RTP). That is more than the amount of data of a 20 ms speech frame, which is 32 Bytes (=10+244+2 bits) including the data added by the RTP payload format as demonstrated above. This overhead causes lots of packets to be dropped at links with high bit error rates, such as cellular links. Hence, robust header compression (ROHC) has been developed. This mechanism reduces the 40 Bytes header information to 2 or 4 Bytes by utilizing the redundancy between header fields both within the same header but in particular between consecutive packets belonging to the same packet stream. ROHC profiles for UDP-Lite are defined in [5].

3. Strategies

We investigate the impact of bit errors on the perceived speech-quality for different VoIP transmission strategies that are based on RTP/UDP-Lite/IP transport of AMR speech frames. These strategies are presented in Figure 2 and Table 1. As a "reference", strategy 1 represents traditional RTP/UDP/IP transport. Hence, the UDP-Lite checksum covers the entire UDP-Lite payload, and a packet is dropped if a bit error occurs. Strategy 2 limits the checksum coverage to the headers of UDP-Lite, RTP and the RTP payload, and the class A speech bits. Thus, packets with damaged class B/C bits can be saved for the speech decoding process. In strategy 3, none of the speech data is covered by the checksum, so all of it can be used for the reconstruction of the speech signal. In any strategy, the header of IPv4 is protected by its own checksum.
Based on these strategies, we have also simulated the use of robust header compression, since it provides the facility to efficiently utilize low-speed serial links.

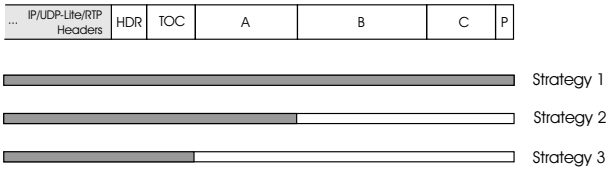

Figure 2: *UDP-Lite checksum coverage.*

Case	Strategy 1	Strategy 2	Strategy 3
Header corrupted	drop packet	drop packet	drop packet
"Class A" data corrupted	drop packet	drop packet	keep data
"Class B/C" data corrupted	drop packet	keep data	keep data
UDP-Lite checksum coverage [bits]	576	411	330

Table 1: *Packet drop strategies and corresponding UDP-Lite checksum coverage.*

4. Simulations

4.1. Simulation Environment

Our simulation environment, depicted in Figure 3, consists of the following parts: a speech database, AMR speech encoding and decoding, and a module simulating the strategies described above for different bit error rates. Finally, we assess the resulting perceptual speech-quality.

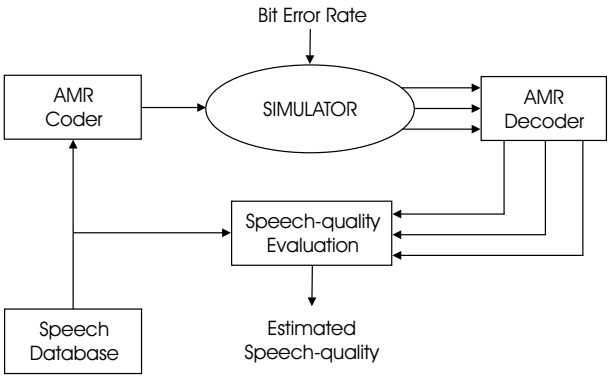

Figure 3: *Simulation environment.*

The PHONDAT speech database [6] contains phonetically rich German sentences, and therefore provides the ideal material for our speech quality studies. We have chosen 12 sentences pairs spoken by 4 talkers (2 female and 2 male).

Each sample has been coded using the AMR-codec's 12.2 kb/s mode, and fed to the simulator module, which is described in section 4.2. For each of the three transmission strategies, a degraded bitstream is generated. After decoding these bitstreams, the quality of the degraded speech samples is assessed using PESQ (see section 4.3). This procedure has been repeated 40 times for each speech sample. The next section describes the simulator in detail.

4.2. Bit Error Model

The heart of the simulator is a simple network layer bit error generation model which corresponds to the behavior of a binary symmetric channel in wireless communications, and an additive white Gaussian noise channel in wireline transmission. For each packet, the number of bit errors X within the actual packet is determined by a binomial distribution:

$$X \sim B(N, p) \tag{1}$$

where N represents the packet size [Bits] and p represents the bit error rate (BER). The location of the bit errors L_X within the packet is then calculated using a uniform distribution, as shown in Equation 2. Each location must only occur once in a frame.

$$L_X \sim \lfloor N * U(0,1) + 1 \rfloor \tag{2}$$

We have chosen bit error rates ranging from 10^{-5} to 10^{-3}. The relation between the bit error rate and packet loss for our model is shown in Table 2. The packet losses that can be obtained by ROHC are given in Table 3.

	Packet Loss Rate [%]		
BER	Strategy 1	Strategy 2	Strategy 3
10^{-5}	0.57	0.40	0.32
10^{-4}	5.71	4.10	3.29
10^{-3}	43.64	33.52	27.95

Table 2: *Relation between bit error rates and packet loss rates for AMR voice packets at 12.2 kb/s without ROHC (576 bits/packet).*

	Packet Loss Rate [%]		
BER	Strategy 1	Strategy 2	Strategy 3
10^{-5}	0.29	0.12	0.04
10^{-4}	2.88	1.24	0.42
10^{-3}	25.01	11.59	4.09

Table 3: *Relation between bit error rates and packet loss rates for AMR voice packets at 12.2 kb/s using ROHC (288 bits/packet).*

4.3. Speech-Quality Evaluation

For the user, the perceived speech-quality is the only measure for the quality of a voice connection. Subjective speech-quality assessment, i.e., using human test persons, is costly and time consuming. Hence, objective methods have been developed, which compare a speech signal degraded by some network with the original signal (reference). ITU-T Rec. P.862 "Perceptual Evaluation of Speech-Quality" (PESQ) [7] defines such a method. The quality estimated by PESQ corresponds to the average user perception of the speech sample under assessment (PESQ-Mean Opinion Score, PESQ-MOS). The PESQ-algorithm provides acceptable accuracy for packet loss concealment methods and transmission errors. Thus, we have chosen it for our studies.

5. Results

Tables 2 and 3 present significant decreases in packet loss for strategies 2 and 3, compared to strategy 1. The relations between the packet loss rates of a certain bit error rate correspond with the relation of the number of bits covered by the UDP-Lite checksum.

The simulation results are shown in Figures 4 and 5 for the non-compressed-header case and ROHC, respectively. The perceived speech-quality (PESQ-MOS) is plotted as a function of the bit error rate, which is scaled logarithmically. Independently of the strategy and whether ROHC is used or not, the PESQ-MOS values roughly coincide at BERs of 10^{-5}. In general, we can observe that using all damaged speech data (strategy 3) performs better than dropping a packet if its class A bits are corrupted (strategy 2).

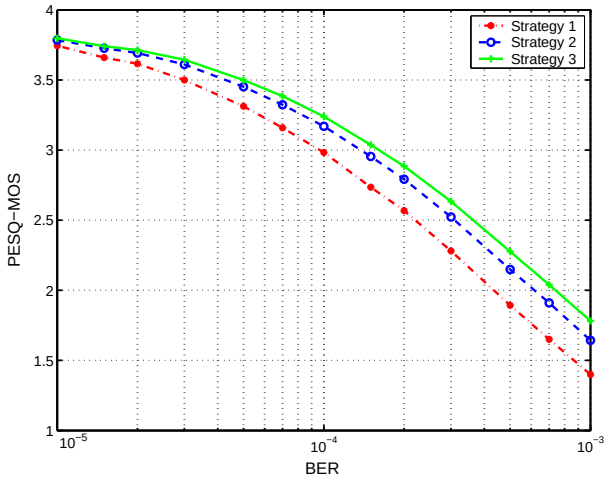

Figure 4: *Estimated perceived speech-quality vs. bit error rate (no header compression).*

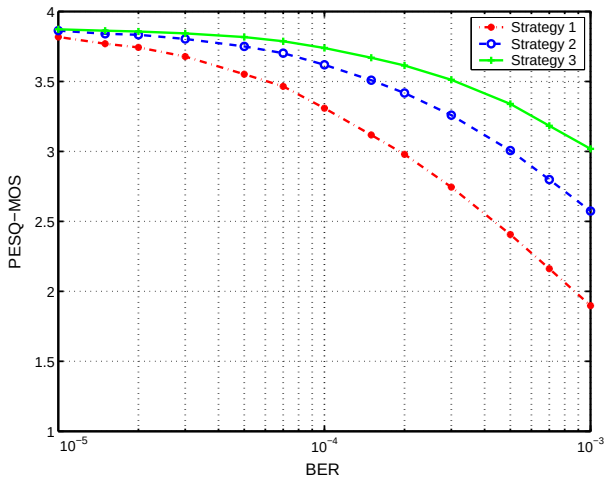

Figure 5: *Estimated perceived speech-quality vs. bit error rate using ROHC.*

Moreover, the use of ROHC results in a significant increase in speech quality. Again, strategy 3 yields best performance, resulting in a faily stable PESQ-MOS score of about 3.8 up to a bit rate of 10^{-4} which then decreases to about 3.0 at a bit error rate of 10^{-3}. In comparison, strategies 1 and 2 reach speech-quality scores of 1.9 and 2.6, respectively. Contrary to the non-header-compressed case, where the PESQ-MOS scores seem to decrease in parallel with the BER, ROHC causes the speech-quality values of the three strategies to spread with increasing BER.

We identify the reduction of packet losses due to keeping erroneous bits as the main cause for the increased speech-quality in our simulation scenario. In general, note that using all damaged speech data (strategy 3) dominates the performance of packet loss concealment in case of corrupted class A bits throughout the investigated BER range. Therefore, in our scenario the distinction of bit relevance appears to be useless.

6. Conclusions

This paper has dealt with the question whether keeping packets containing erroneous speech data can improve the speech-quality perceived by the user. We have presented networking and signal processing techniques that seemed suitable for this purpose. Based on these techniques, we have introduced strategies, by which speech data damaged by bit errors can be saved from being dropped. Simulations with and without robust header compression gives an insight into the behavior of the strategies.

As a general conclusion, a VoIP system should be designed to make use of as much damaged data as possible in order to keep the packet loss rate as low as possible. Thus, bit errors within the speech data have perceptually less impact on the speech-quality than lost speech frames, hence corrupted speech data have to be considered useful.

7. References

[1] Larzon, L.-A., et al., “The UDP-Lite Protocol”, Internet Draft, Internet Engineering Task Force, Dec. 2002. Work in Progress.

[2] 3rd Generation Partnership Project, “AMR speech Codec; General description”, TS 26.071 version 5.0.0, June 2002.

[3] Sjöberg, J., et al., “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”, RFC 3267, Internet Engineering Task Force, June 2002.

[4] 3rd Generation Partnership Project, “AMR speech Codec; Frame Structure”, TS 26.101 version 5.0.0, June 2002.

[5] Pelletier, G., “RObust Header Compression (ROHC): Profiles for UDP-Lite”, Internet Draft, Internet Engineering Task Force, Jan. 2003. Work in Progress.

[6] Bavarian Archive for Speech Signals (BAS), “PhonDat 1 Corpus”, `http://www.bas.uni-muenchen.de/Bas/BasPD1eng.html`.

[7] 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 Rec. P.862, Feb. 2001.