Corrupted Speech Data Considered Useful

Florian Hammer, Peter Reichl, Tomas Nordström

Telecommunications Research Center Vienna (FTW)
{hammer, reichl, nordstrom}@ftw.at

Extended 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 methods 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.

In general, we identify four major components that contribute to the speech quality as perceived by the user: network mechanisms and parameters, signal processing techniques, like speech coding or packet loss concealment, issues concerning the user terminal, and aspects with regard to the user him- or herself. Our approach uses the combination of signal processing and networking techniques, exploring their joint influence on the perceived speech quality. In particular, we deal with the avoidance of voice packet losses caused by bit errors on a wireline or wireless link.

The IP/UDP/RTP protocol stack is widely used for real-time communications over the Internet. Normally, all packets that are damaged by bit errors are dropped by routers or gateways due to incorrect checksums. In order to provide some flexibility concerning the coverage of the UDP checksum, UDP-lite [6] has been developed. This modified UDP protocol allows for a UDP checksum calculation covering an arbitrary length of its payload. Hence, corrupted speech data can be used for speech signal recovery. Using the UDP-lite approach for our work, we compare three strategies of dealing with bit errors within IP-packets (cf table 1):

- *Strategy 1* represents the usual IP/UDP/RTP packet transport. The UDP-lite checksum covers the entire packet (576 bits), thus the packet is dropped if any of its bits is damaged.
- In *strategy 2*, a packet is only dropped if the header information is damaged. The speech decoder is notified, if perceptually sensitive bits are damaged, but the speech data bits are available for speech frame recovery.
- In *strategy 3*, all of the damaged data is used without notifying the speech

codec about errors, and packets are only dropped if the header information is damaged.

For our simulations, we have chosen the 12.2 kbps mode of the 3GPP/ETSI adaptive multi-rate (AMR) speech codec [2]. This codec features an internal packet loss concealment algorithm that compensates for lost speech frames. Regarding bit errors, uneven level protection distinguishes between the perceptual impact of the speech data bits. "Class A bits" are most sensitive, whereas "Class C bits" are least sensitive to errors. Following the recommendation of the AMR error concealment specification [3], the "SPEECH_BAD" flag should be set at the receiver in the case of a lost speech frame, and if class A data has been corrupted in order to spawn the internal error concealment.

	Strategy 1	Strategy 2	Strategy 3
Header corrupted	drop packet	drop packet	drop packet
"Class A" data corrupted	drop packet	notify decoder	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.

The dropping strategies described above have been simulated in MatLab. Phonetically rich speech samples from the Speechdat-AT database [1], which has been collected at our institute, are coded using the ETSI implementation of the AMR codec [4]. Voice data packetization is simulated according to the bandwidth efficient payload mode as defined in [7], using a single speech frame per packet with respect to the strong delay constraints concerning conversational speech transmission. Bitstreams are generated according to the dropping strategies for a series of bit error rates in the range of 10^{-5} to 10^{-3}. After decoding these bitstreams, the resulting speech quality is estimated by the PESQ-algorithm [5] which provides acceptable accuracy for the degradation factors in the investigated strategies [5].

Figure 1 shows a typical example of our simulation results. Besides the qualitative observation that the use of damaged speech data improves the overall speech quality, we can state an improvement of almost half a PESQ mean opinion score (MOS) value at higher bit rates compared to normal IP-transmission. This behavior can be observed throughout our simulation results and demonstrates that the use of all available data without indicating errors provides better performance than the use of information concerning corrupted sensitive data. Additionally, informal listening tests confirm this conclusion.

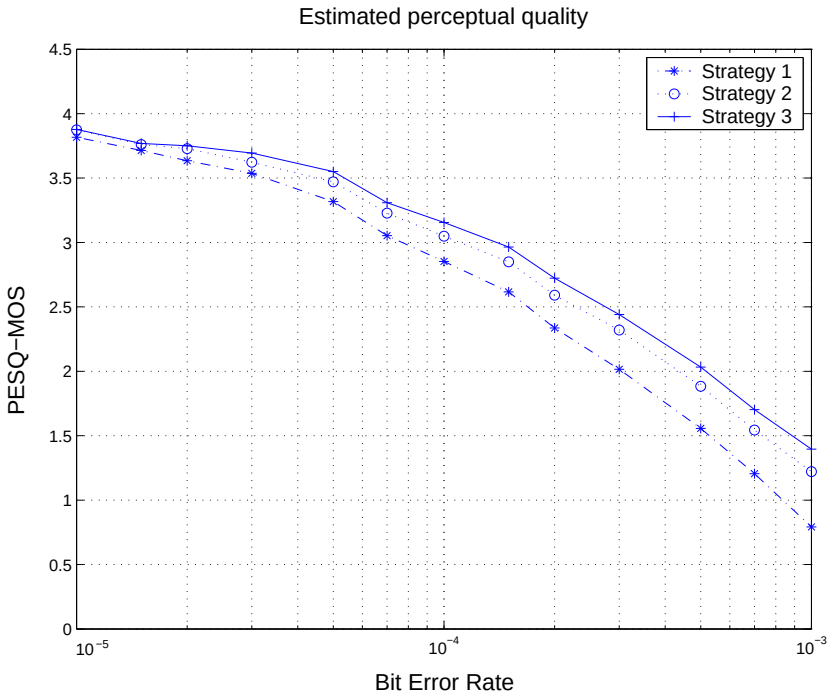

FIGURE 1: Estimated perceived speech quality vs. bit error rate

References

[1] M. Baum, G. Erbach, and G. Kubin. Speechdat-AT: A telephone speech database for Austrian German. In *Proc. LREC Workshop Very Large Telephone Databases (XLDB)*, 2000.

[2] 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*, June 2002.

[3] 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*, June 2002.

[4] European Telecommunications Standards Institute. Universal mobile telecommunications system (UMTS); ANSI-C code for the floating-point Adaptive Multi-Rate (AMR) speech codec (3GPP TS 26.104 version 5.0.0 Release 5). *ETSI TS 126 104 v5.0.0*, June 2002.

[5] 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*, February 2001.

[6] L.-A. Larzon, M. Degermark, and S. Pink. UDP lite for real time multimedia applications. Technical Report HPL-IRI-1999-001, HP Labs, April 1999.

[7] J. Sjoberg, M. Westerlund, A. Lakaniemi, and 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*, June 2002.