Concealment of Lost Speech Packets Using Adaptive Packetization

H. Sanneck
GMD Fokus / GloNe
Kaiserin-Augusta-Allee 31
D-10589 Berlin, Germany
sanneck@fokus.gmd.de

Abstract

Long-term correlation within a speech signal is usually exploited to achieve higher compression ratios (e.g. RPE-LTP coders [6]). In this paper we aim to use the long-term correlation to influence the packetization interval of a voice stream at the sender before sending it over a lossy packet-switched network. If a packet is lost, the receiver can conceal the loss of information by using adjacent signal segments of which (due to the pre-processing/packetization at the sender) a certain similarity to the lost segment can be assumed. Subjective test results show that the "Adaptive Packetization / Concealment" scheme (AP/C) can alleviate significantly the impact of isolated packet losses to speech quality.

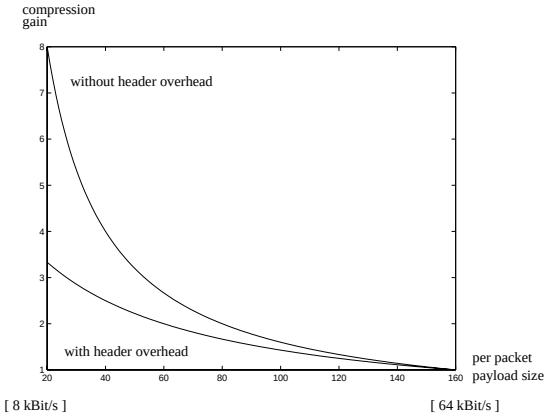

Figure 1. Relative compression gain (20ms speech (sampled at 8 kHz) per packet, 40 octets per-packet header overhead)

1. Motivation

To tackle the problem of lost packets (which for speech transmission means annoying signal drop-outs and a possibly disrupted conversation), different techniques have been proposed, which can be divided as follows:

- Loss avoidance at the application level (adaptation [3], layered coding/multicasting [9], [19])
- Loss avoidance at the network level (reservation [4], buffer management [16])
- Loss reconstruction (redundancy mechanisms [7], [12], [14], [16])
- Loss alleviation (interleaving [8], [10], concealment [5], [15])

Considering a scenario with the presence of numerous, low-bandwidth speech flows in the *Internet*, most of the approaches have scalability limitations, because they either introduce high per-flow state overhead in Internet routers (reservation), data overhead (redundancy mechanisms) or delay overhead (interleaving, receiver-based concealment). Due to the nature of the currently standardized speech codecs, also adaptation of the coder output bitrate to the network conditions, as well as a layered distribution of the coder output signal often is not feasible.

Increasing the *compression* of the speech signal leads to a reduction in the overall amount of data to be sent over the network, i.e. the payload per packet is reduced. Yet the number of packets stays the same (when maintaining the packetization time interval / playout delay), thus inducing the same per-packet processing cost as before in the network. Additionally, the high per-packet RTP/UDP/IP header overhead diminishes the gain of highly compressed speech, as can be seen in Fig. 1 (relative to PCM speech quantized with 8 Bit). On the contrary, highly compressed speech is very sensitive to packet loss (due to coder/decoder state synchronization).

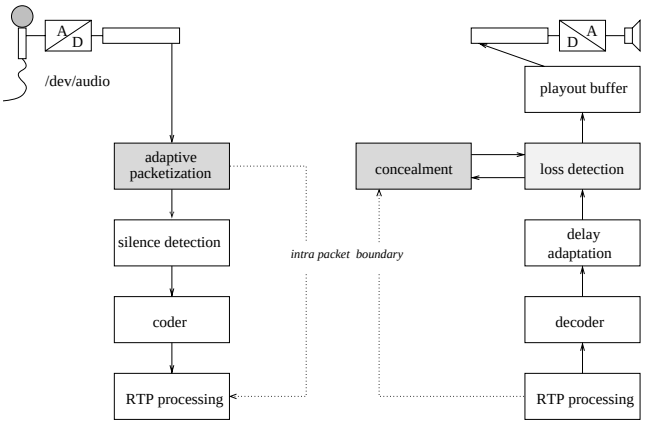

Figure 2. Structure of an AP/C enhanced audio tool

We propose a scheme called *Adaptive Packetization and Concealment (AP/C)*, which uses variable size packets to increase the loss resiliency (previously, in [17] and [18], variable size packetization has only been proposed for compression). Fig. 2 shows the basic structure of the sending and receiving entities of an audiotool with the to-be-added functional blocks shaded. Note that in principle, any speech coder able to operate on variable size frames can be used, as the signal is analyzed before the coder and concealed (when lost) after the decoder.

At the sender, auto-correlation of the signal is used for pitch period estimation. Then, two audio "chunks" of estimated pitch period length are packed into one packet. This results in small packets being sent for voiced speech, large packets sent for speech classified as "unvoiced" (which includes noise and silence).

When loss is detected at the receiver, adjacent speech "chunks" (of the previous and the following) packet are reused. Only a simple sample rate conversion needs to be performed on those chunks to scale them to the needed length and subsequently fill the gap caused by the lost packet.

The paper is structured as follows: section 2 presents the adaptive packetization employed by the sender. In section 3, the concealment algorithm used by the receiver is described. Section 4 gives results of a subjective test with AP/C. Implications of the proposed scheme with regard to additional delay, etc. are discussed in section 5. Section 6 concludes the paper.

2. Sender algorithm

The part of the sender algorithm interfacing to the audio device copies PCM samples from the audio device to its input buffer (Fig. 3). Pitch period estimation is done by auto-correlation and short-term energy measurement of an input segment of $2T_{max}$ samples (T_{max} being the correlation window size). The result is the value $p(c)$ (c being the number of the found segment, which we call *"chunk"*) reflecting the periodicity for voiced speech (note that only a reliable detection of *periodicity* and changes in periodicity are necessary; the exactness of the pitch period value itself is not crucial). For unvoiced speech, the algorithm typically picks a value close to T_{max} (Fig. 4/Fig. 5). Then the input buffer pointer is moved by $p(c)$ samples (thus constituting a "chunk"), c is incremented and if necessary new audio samples are fetched from the audio device.

Another routine (which may run in parallel and should be integrated with the *silence detection* function) provides a simple check for speech transitions:

$$\Delta p = |p(c) - p(c-1)| > \Delta T$$
and either
$p(c) < T_u$ and $p(c-1) \geq T_u$ (unvoiced $\rightarrow$ voiced: **uv**)
$p(c) \geq T_u$ and $p(c-1) < T_u$ (voiced $\rightarrow$ unvoiced: **vu**)

where ΔT and T_u are pre-configured, fixed bounds.

To alleviate the incurred header overhead, which would be prohibitive for IP-based transport if every chunk is sent in one packet, two consecutive chunks are associated to one packet (see Figures 4 and 5)

However, if a *vu* transition has been detected, the "transition chunk" is partitioned into two parts (8a/b in Fig. 4) with $p(c_a)$ set to $p(c-1)$ and $p(c_b) = p(c) - p(c_a)$ ($p(c)$ being the original chunk size). Note that if $c \ mod \ 2 = 0$, the chunk $c-1$ (no. 7 in Fig. 4) is sent as a packet containing just one chunk.

When a *uv* transition has taken place, *backward* correlation of the current chunk with the previous one (no. 3 in Fig. 5) is tested as it may already contain voiced data (due to the forward auto-correlation calculation). If true, again the previous chunk is partitioned with $p(c_b - 1) = p_{backward}(c-1)$ and $p(c_a - 1) = p(c-1) - p(c_b - 1)$ ($p_{backward}$ is the result of the backward correlation). Note that the above procedure can only be performed if $c \ mod \ 2 = 0$, otherwise the previous chunk has already been sent in a packet (a solution to this problem would be to retain always two unvoiced chunks and check if the third contains a transition, however the gain in speech quality when concealing would

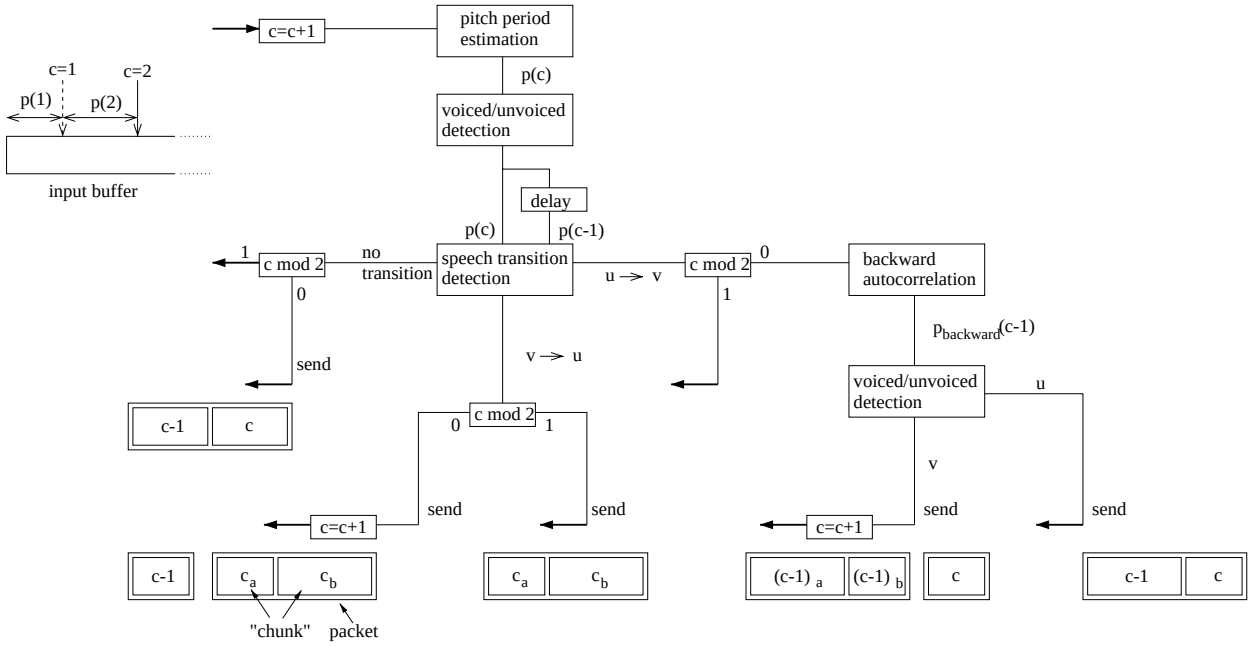

Figure 3. AP/C sender algorithm

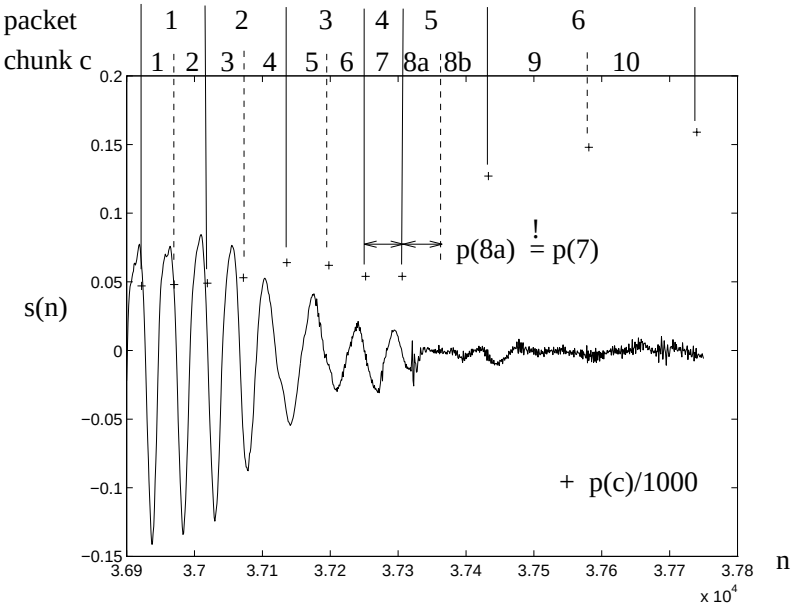

Figure 4. AP/C sender operation: transition voiced → unvoiced; $s(n)$: time domain signal, n: sample number

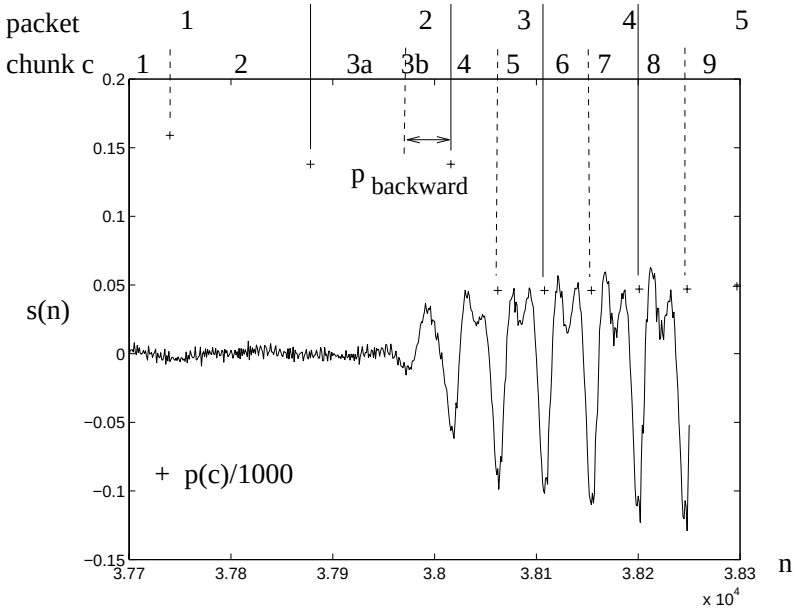

Figure 5. AP/C sender operation: transition unvoiced → voiced

not justify the incurred additional delay).

With the above algorithm "more important" (voiced) speech is sent in smaller packets and thus the resulting loss impact/distortion (assuming that the network's loss probability is independent of the packet size) is less significant than using fixed size packets of the same average length, even without concealment. With our scheme, the packet size is now adaptive to the measured pitch period. Fig. 6 shows this dependency for four different speakers. The mean packet size $\bar{l}$ is approximately twice the mean pitch period $\bar{p}_v$, as the most frequent combination is a packet consisting of two voiced chunks.

Distributions of the packet size for test signals of about $10s$ featuring four different speakers in Fig. 7 show that the parameter settings [1] can accomodate a range of pitches, as their overall shapes are similar to each other. As mentioned above, the most common packets contain two voiced chunks (*vv* packets), as distributions are centered around a value that is twice the mean pitch period (i.e. the mean of voiced chunks).

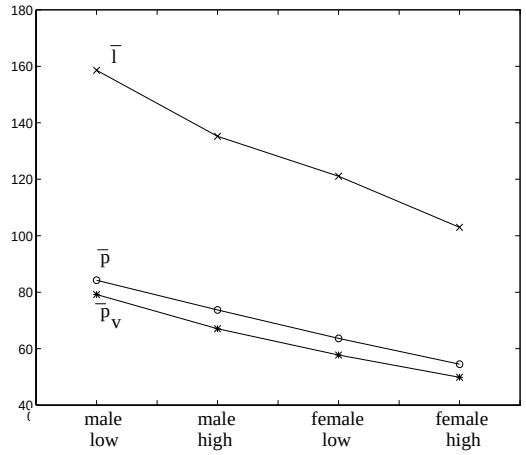

Figure 6. Dependency of the mean packet size $\bar{l}$ on the mean chunk size $\bar{p}$ / mean pitch period $\bar{p}_v$

[1] $T_{min} = 30$ (start offset point of the auto-correlation); $T_u = 120$; $T_{max} = 160$ samples. Note that the packet size extends from sending a single voiced chunk ($l \geq T_{min}$) to sending two unvoiced chunks ($l \leq 2T_{max}$).

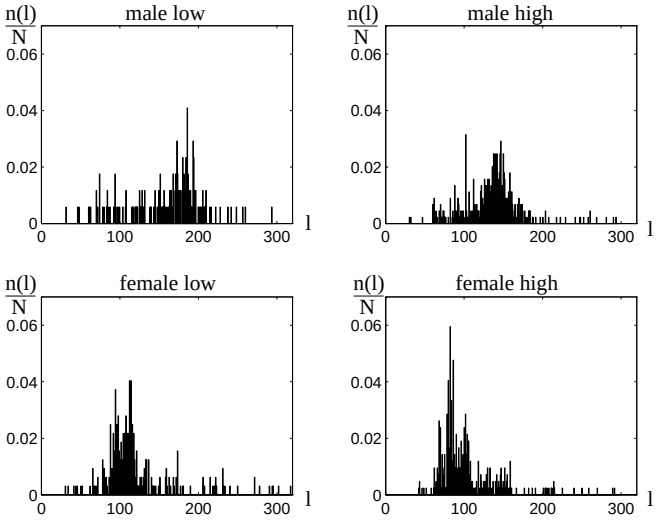

Figure 7. Normalized packet size frequency distributions for four different speakers; $n(l)$: number of packets of size l, N: overall number of packets

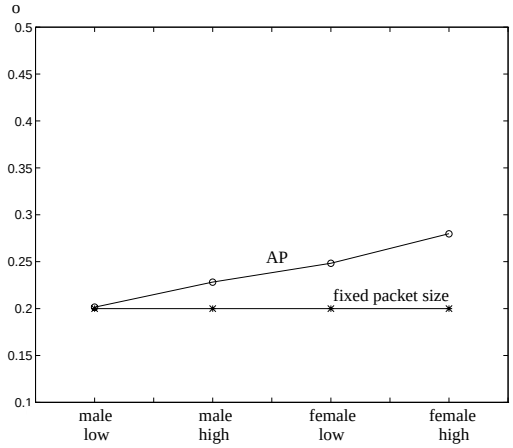

Figure 8. Relative cumulated header overhead o for AP and fixed packet size (160 octets) assuming 40 octets per-packet overhead for four different speakers

Fig. 8 shows the resulting relative packet header overhead for different speakers. The overhead is comparable to a typical parameter setting in IP networks (160 octets payload [$= 20ms$ $\mu-$law PCM audio] in an IP/UDP/RTP packet [20+8+12 octets header]), yet increases with increasing mean pitch period.

To support a possible concealment operation it is necessary to transmit the intra-packet boundary between two chunks as additional information in the packet itself and the following packet. That amounts to two octets of "redundancy" for every packet, that could e.g. be transmitted by the proposed redundant encoding scheme for RTP ([11]).

3. Receiver algorithm

At the receiver, packet loss is detected by means of RTP sequence numbers (and timestamps when using silence detection), taking into account the current playout delay (when late packets have to be assumed as lost). Due to the pre-processing at the sender, the receiver can assume that the chunks of a lost packet resemble the adjacent chunks. The adjacent chunks (c_{12} and c_{31} in Fig. 9) are resampled in the time domain by a factor of $k = c_{12}/c_{21}$ and $k = c_{31}/c_{22}$ for the left and right adjacent packet respectively. This is done to match the lost chunk sizes, which are given by the packet length and the transmitted intra packet boundary[2]. Resampling is done using a linear interpolator (as in [20]). The conversion factor k is linearly varied throughout the signal segment. This enables a replacement signal with a correct phase, thus avoiding discontinuities in the concealed signal leading to distortions, while maintaining the original pitch frequency at both edges. Then these chunks are copied into the outbut buffer as a replacement for the lost packet. No time-scale adjustment ([15]) is necessary as the chunk sizes are small. Because the sizes of the lost and the adjacent chunk most probably only differ slightly for either voiced or unvoiced speech (and thus the respective spectra), no specific distortion caused by the operation can be observed. Fig. 10 shows the concealment operation in the time domain.

However, *transitions* in the signal might lead to extreme expansion/compression operations. Table 1 lists the possible cases. v_a, u_a are voiced/unvoiced *available* chunks, v_L, u_L are voiced/unvoiced *lost* chunks which are relevant for the case. A $u(u|v)$ packet is a packet where the second chunk contains an unvoiced/voiced transition that was not recognized by the sender algorithm (see section 2). To avoid extreme expansion/compression an upper bound f_{max}

[2] Further study is needed, how good an *estimation* of the intra-packet boundaries would perform.

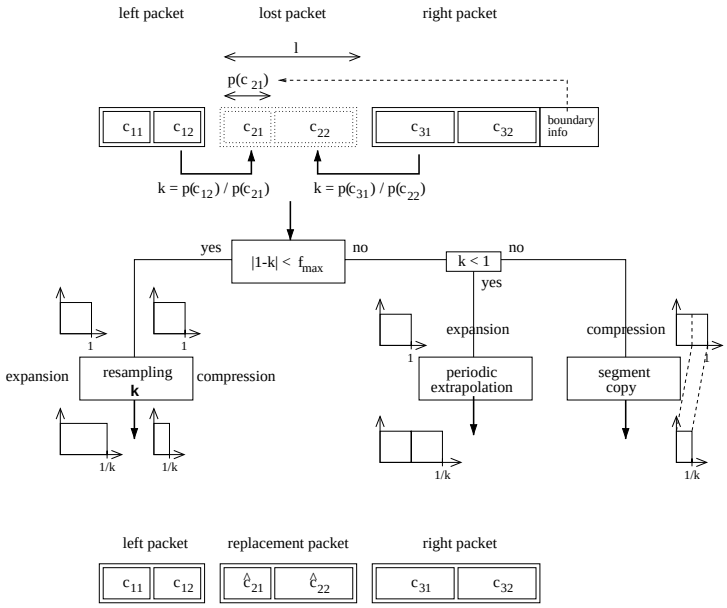

Figure 9. AP/C receiver operation

packet left	packet lost	packet right	expansion (exp.) compression (comp.)
$v\|u_a$	$u_L u$		$u_a \ll u_L \rightarrow$ exp.
	uu_L	$u_a\|v$	$u_a \ll u_L \rightarrow$ exp.
uu_a	$u_L\|v$		$u_a \gg u_L \rightarrow$ comp.
	$v\|u_L$	$u_a u$	$u_a \gg u_L \rightarrow$ comp.
	$u(u\|v)_L$	$v_a v$	$v_a \ll (u\|v)_L \rightarrow$ exp.
$u(u\|v)_a$	$v_L v$		$(u\|v)_a \gg v_L \rightarrow$ comp.

Table 1. Concealment of/with packets containing speech transitions leading to high expansion or compression

for the resampling has been introduced (Fig. 9): $|1-k| \stackrel{!}{<} f_{max}$. We used $f_{max} = 50\%$. If the bound is exceeded when compressing, adjacent samples of the relevant length are taken and inserted in the gap ("segment copy" in Fig. 9). An audible discontinuity which might occur can be avoided by overlap-adding the concealment chunk with the adjacent ones. High expansions are avoided by repeating a chunk until the necessary length is achieved ("periodic extrapolation" in Fig. 9) and then again overlap-adding it.

4. Subjective Test Results

To evaluate the properties and performance of AP/C a subjective test was carried out. Test signals were the four signals (with different speakers) also used in the previous objective analysis (PCM 16 bit linear, sampled at 8 kHz). The new technique was compared against silence substitution and the simple receiver-based concealment algorithm "Pitch Waveform Replication" (PWR, [15]), which is the only one able to operate under very high loss rates (isolated losses). With PWR, one pitch period found in the packet preceding the missing one is repeated throughout the loss gap.

Primary goal of the test was to assess the performance in the presence of numerous, yet isolated losses. Internet measurements (e.g. [2] and [21]) have shown that the probability to loose one packet is relatively high, however drops significantly for the loss of several consecutive packets. Loss *bursts* need to be alleviated by a combination of concealment/FEC (for losses caused by bit errors, e.g. on wireless links, and congestion) and queue management in the routers (for losses caused by congestion).

We therefore modelled equally distributed random packet losses and introduced only isolated losses (to allow *complete concealment* and thus a relative evaluation of the algorithms) with the following function:

$$P_i(i) = \begin{cases} 0 = P_i(i|i-1): & \text{packet i-1 lost,} \tilde{P}_i(i) = P_L \\ 1 = \tilde{P}_i(i|i-1): & \text{packet i-2 lost,} \tilde{P}_i(i) = 0 \\ P_L = 2\bar{P}: & \text{otherwise,} \tilde{P}_i(i) = 0 \end{cases}$$

where $P_i(i)$ denotes the probability of the loss of packet i and $P_i(i|i-n)$ the conditional probability to loose the ith packet if the $(i-n)$th packet was lost. $\tilde{P}_i$ can force

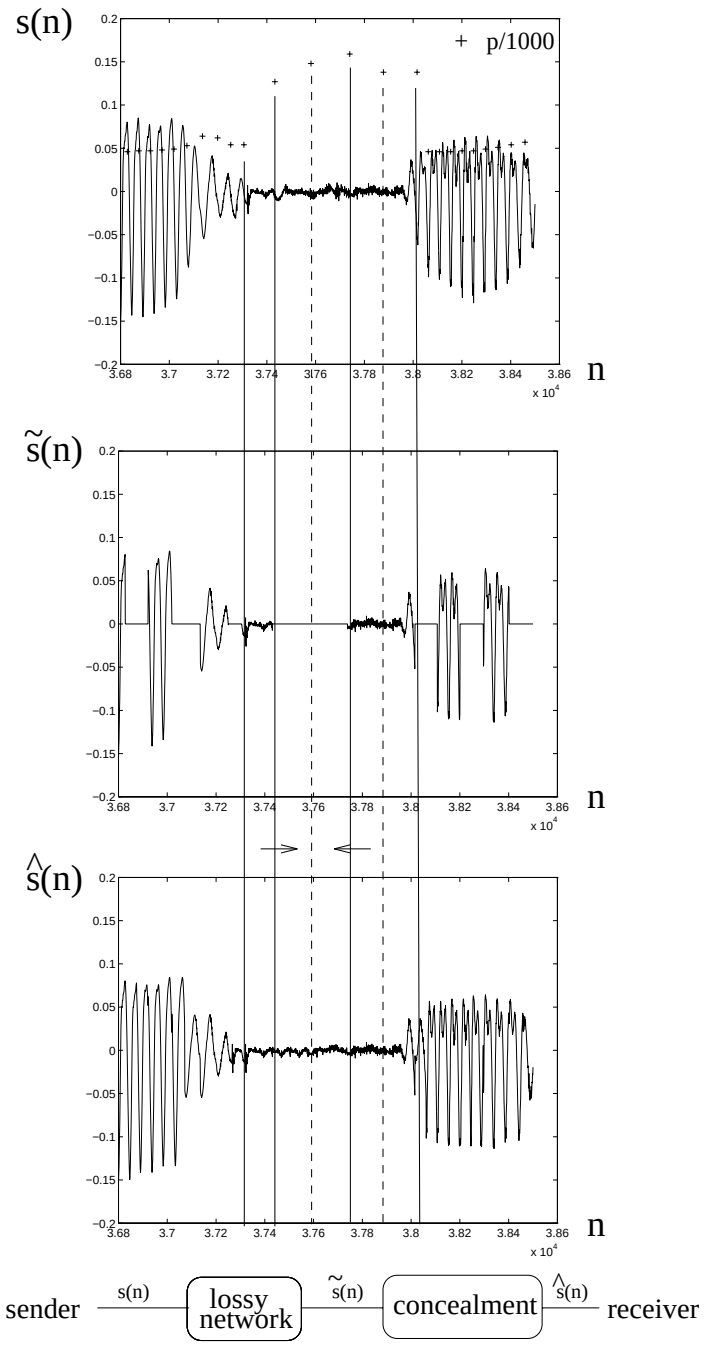

Figure 10. Concealment of a distorted signal (50% loss)

a loss on the next packet if the previous packet was lost. $\bar{P} \leq 0.5$ is the mean loss rate that needs to be simulated for the signal. Note that $\bar{P} = 0.5$ results in a deterministic loss of every second packet.

Thirteen non-expert listeners evaluated the overall quality of 40 test conditions[3] on a five-category scale (Mean Opinion Score: MOS). Before testing started, an "Anchoring" procedure took place, where the quality range (Original = 5, "Worst Case" (WC) signal[4] = 1) was introduced.

Figures 11/12 show the MOS values for the four different speakers (male low/high, female low/high). As loss values we give the actual *sample loss rate* instead of the packet loss rate, as we deal with variable size packets. It can be seen that for all speakers AP/C leads to a significant enhancement in speech quality compared to the "silence substitution" case, which is maintained also for higher loss rates. However for speakers (female) with a higher pitch frequencies, the relative performance (distance between "silence substitution" and "AP/C") decreases. A reason for this is the chosen start offset point T_{min} (= 30 samples) of the auto correlation computation, which constitutes a lower bound on the chunk/packet size to avoid excessive packet header overhead, but also limits the accurateness of the periodicity measurement (note the small distance between the peak of the packet size distribution and the lower bound in Fig. 7 for "female high"). Additionally, female speakers receive relatively high MOS values for the worst case signal ($> \ 1.5$). This is due to the adaptive packetization: a higher number of shorter gaps are introduced (compared to fixed size packetization with the same loss rate) which are less perceivable. The PWR algorithm performs well for loss rates of about 20% (cf. [15]), however speech quality drops significantly for higher loss rates, as the specific distortions introduced by that algorithm become significant. Standard deviations of MOS values for all but two of the forty test conditions are below 1.

Objective measurements are clearly inappropriate for PWR (no aim at mathematical approximation of the missing signal segments). AP/C is not a *reconstruction* scheme as well, however the adaptive packetization and subsequent resampling should perform better concerning mathematical correctness. Calculated overall SNR values for PWR (for the examples which figure below) are always below those for the distorted signal. SNR values for AP/C are always above those for the distorted signal and at least 4dB higher than for PWR. This confirms our conjecture, yet conclusions about speech quality should be only based on the subjective test results.

[3] 4 speakers $\times$ (3 algorithms $\times$ 3 loss rates + original)

[4] In this test we used the unconcealed 50% loss signal.

5. Discussion

The additional delay introduced in the current implementation consists of

- time interval corresponding to the length of the buffered speech segment needed for the sender processing (auto-correlation computation) of the second chunk minus the actual size of the second chunk (as this belongs to the "conventional" packetization interval to create a packet) : $T_{max} \leq d_S \leq 2T_{max} - T_{min}$.
- time corresponding to one packet length after a loss was detected at the receiver ($T_{min} \leq d_R \leq 2T_{max}$)
- time needed for computations d_C

The computational complexity is low at sender and very low at the receiver as only simple operations (auto-correlation, sample rate conversion) have to be performed (thus $d_C \ll d_S + d_R$). This makes the scheme well suited for multicast environments with low-end receivers. As shown in Fig. 8, the additional packet header overhead is even for the highest pitch voice below 10%, which is comparable to adding a very low bitrate additional source coding to reconstruct isolated losses ([7]). Because of the dependency on the pitch period, AP/C is aimed at speech transmission only.

A deployment problem (for Internet speech transmission in general) is that modern speech coders are designed to operate on (small) fixed size audio frames (e.g. $F = 10ms$ for G.729 [1], $F = 30ms$ for GSM). However, we plan to integrate a fragmentation algorithm which would allow to use those codecs, i.e. chunks are always of size $kF \geq p(c)$ (k being a positive integer), resulting in overlapping chunks and additional alignment information needing to be transmitted. Subjective tests have been performed with PCM samples, this carries the implicit assumption that the speech immediately after the gap is decoded properly. However modern speech coders rely on synchronization of coder and decoder, which is lost during a packet loss gap ([13]) thus the decoding is worse after the gap due to previous coder state loss.

Backwards compatibility to existing audio tools is ensured, as most tools can receive properly variable length PCM packets (and then mix them into their output buffer), however specific delay adaptation algorithm might need to be modified.

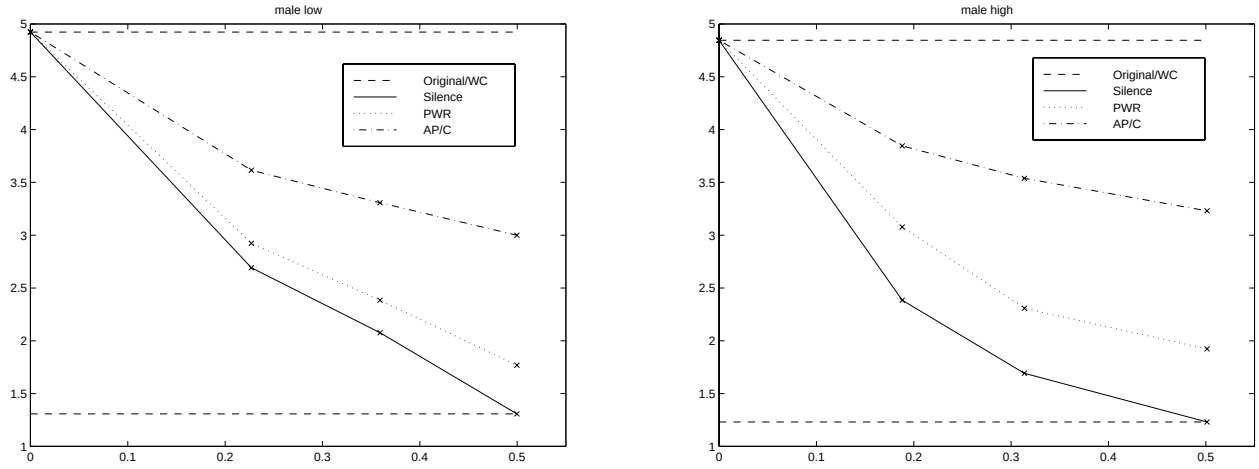

Figure 11. MOS as a function of sample loss rate for speakers ’male low’ and ’male high’

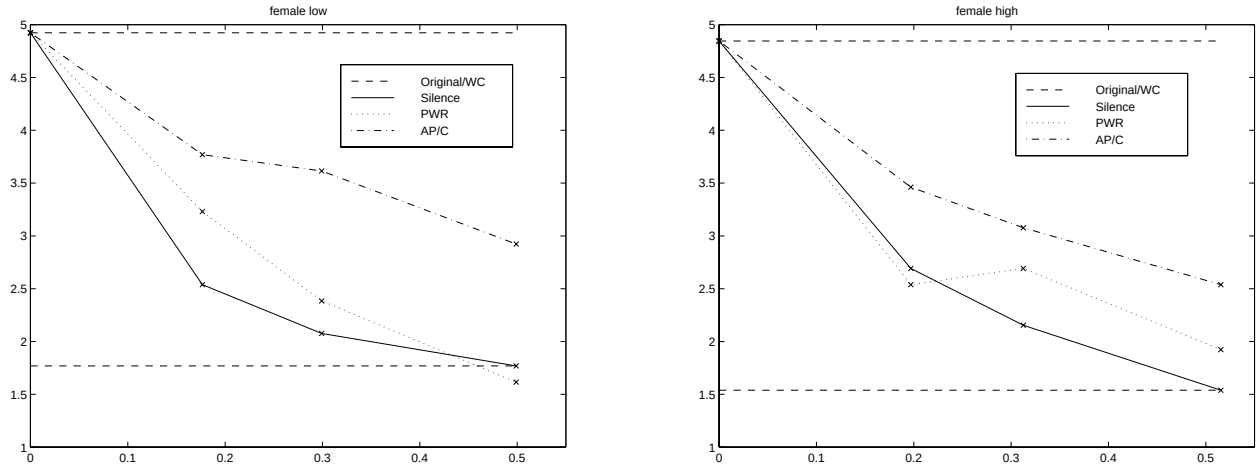

Figure 12. MOS as a function of sample loss rate for speakers ’female low’ and ’female high’

6. Conclusions

A technique for the concealment of lost speech packets has been presented. The core idea of preprocessing a speech signal at the sender to support possible concealment operations at the receiver has proven to be successful. It results in an inherent adaptation of the network to the speech signal, as predefined portions of the signal ("chunks" assembled to packets) are dropped under congestion. From the perspective of the network, the presented application level scheme could be complemented by influencing loss patterns at congested routers (queue management), thus also supporting more fairness between flows by avoiding bursty losses within one voice flow.

For future work, better speech classification/ processing algorithms should be explored in conjunction with AP/C, yet always taking into account the compromise of quality and computational complexity. An important task is to assure efficient support for existing frame-based codecs. Finally, an integrated scheme, aligning coding, packetization and loss recovery functionalities should be developed.

7. Acknowledgements

This work was funded in part by the BMBF (German Ministry of Education and Research), the DFN (German Research Network) and in part by the EEC within the ACTS project AC012 MULTICUBE.

The work benefitted from discussions and participation in the subjective test from members of the GloNe (Global Networking) research group at GMD Fokus and the Telecommunications Department at TU Berlin.

References

[1] Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear-prediction (CS-ACELP). ITU-T Recommendation G.729, March 1996.

[2] J.-C. Bolot, H. Crépin, and A. Garcia. Analysis of audio packet loss in the Internet. In *Proceedings of the 5th International Workshop on Network and Operating System Support for Digital Audio and Video*, pages 163–174, Durham, NH, April 1995.

[3] J.-C. Bolot and A. Garcia. Control mechanisms for packet audio in the Internet. In *Proceedings IEEE Infocom '96*, pages 232–239, San Francisco, CA, April 1996.

[4] R. Braden, D. Clark, and S. Shenker. Integrated services in the Internet architecture: an overview. RFC, IETF, 1994. ftp://ftp.nordu.net/rfc/rfc1633.txt.

[5] K. Clüver. *Rekonstruktion fehlender Signalblöcke bei blockorientierter Sprachübertragung (Reconstruction of missing signal blocks for block-orientated voice transmission)*. PhD thesis, Telecommunications Department, Technical University of Berlin, January 1998.

[6] J. Degener. GSM 06.10 lossy speech compression. Documentation, TU Berlin, KBS, October 1996. http://kbs.cs.tu-berlin.de/ jutta/toast.html.

[7] V. Hardman, M. Sasse, M. Handley, and A. Watson. Reliable audio for use over the Internet. In *Proceedings Inet 95*, http://info.isoc.org/HMP/PAPER/070/abst.html, 1995.

[8] N. Jayant and S. Christensen. Effects of packet losses in waveform coded speech and improvements due to an odd-even sample-interpolation procedure. *IEEE Transactions on Communications*, COM-29(2):101–109, February 1981.

[9] S. McCanne, V. Jacobson, and M. Vetterli. Receiver-driven layered multicast. In *Proceedings ACM SIGCOMM '96*, pages 117–130, Stanford, CA, September 1996.

[10] C. Perkins. Options for the repair of streaming media. Internet Draft, IETF Audio-Video Transport Group, August 1997. ftp://ftp.nordu.net/internet-drafts/draft-ietf-avt-info-repair-00.txt.

[11] C. Perkins. RTP payload for redundant audio data. RFC, IETF Audio-Video Transport Group, September 1997. ftp://ftp.nordu.net/rfc/rfc2198.txt.

[12] M. Podolsky, C. Romer, and S. McCanne. Simulation of FEC-based error control for packet audio on the Internet. In *Proceedings IEEE Infocom*, pages 48–52, San Francisco, CA, March 1998.

[13] J. Rosenberg. G. 729 error recovery for Internet Telephony. Project report, Columbia University, 1997.

[14] J. Rosenberg and H. Schulzrinne. An A/V profile extension for generic forward error correction in RTP. Internet Draft, IETF Audio-Video Transport Group, July 1997. ftp://ftp.nordu.net/internet-drafts/draft-ietf-avt-fec-00.txt.

[15] H. Sanneck, A. Stenger, K. B. Younes, and B. Girod. A new technique for audio packet loss concealment. In *Proceedings IEEE Global Internet 1996 (Jon Crowcroft and Henning Schulzrinne, eds.)*, pages 48–52, London, England, November 1996.

[16] N. Shacham and P. McKenney. Packet recovery in high-speed networks using coding and buffer management. In *Proceedings ACM SIGCOMM '90*, pages 124–131, San Francisco, CA, June 1990.

[17] R. Steele and F. Benjamin. Variable-length packetization of μ-law PCM speech. *AT&T Technical Journal*, 64:1271–1292, July-August 1985.

[18] R. Steele and P. Fortune. An adaptive packetization strategy for A-law PCM speech. In *Conference Record of the International Conference on Communications (ICC)*, pages 941–945 (29.6), Chicago, IL, June 1985.

[19] T. Turletti, S. F. Parisis, and J.-C. Bolot. Experiments with a layered transmission scheme over the Internet. Research report 3296, INRIA, November 1997.

[20] R. Valenzuela and C. Animalu. A new voice packet reconstruction technique. In *Proceedings ICASSP*, pages 1334–1336, May 1989.

[21] M. Yajnik, J. Kurose, and D. Towsley. Packet loss correlation in the MBone multicast network. In *Proceedings IEEE Global Internet 1996 (Jon Crowcroft and Henning Schulzrinne, eds.)*, pages 94–99, London, England, November 1996.