PACKET LOSS CONCEALMENT IN PREDICTIVE AUDIO CODING

Stefan Wabnik, Gerald Schuller, Jens Hirschfeld, Uli Kraemer

Fraunhofer Institute for Digital Media Technology IDMT
Langewiesener Strasse 22, 98693 Ilmenau, Germany
shl@idmt.fraunhofer.de

ABSTRACT

In this paper we present and evaluate several concealment strategies for packet losses in the context of a low delay predictive audio coder. Our goal is to minimize the audible impact of a packet loss. The problem is that the predictive coder is backward adaptive, hence depending on past values. There is a predictor reset, but to increase coding efficiency, the distance between two resets is several hundred packets. Hence, not only the lost packet itself cannot be reconstructed, but the transmitted data up to the next reset does not result in an exact reconstruction of the audio signal. Our approach is to try to use as much information as possible from the still available data until the next reset, and to reconstruct an audio signal such that distortions are least objectionable. To compare different approaches, we conducted a listening test. The result is that an adaptive reconstruction filter works best in this context.

1. INTRODUCTION

Standard audio coders, like MP3 [1] or MPEG AAC [2], are based on subband coding. As a result of this coding principle and the high number of subbands used, they feature a high encoding/decoding delay. This delay is too high for communications applications or applications where there is a superposition of the direct sound and the encoded/decoded sound (singer on a small stage, feedback channels for musicians, musicians playing together remotely, a mix of wired and coded wireless speakers). For this purpose we developed an audio coder with a very low delay, based on backward adaptive prediction, where the adaptation is based on past samples [3].

There are several concealment strategies in use for the case of packet transmission losses in standard audio coders. Examples are an interpolation of subband values between received blocks, which adds delay, or an interpolation between frequencies if parts of the spectrum are lost [4]. Another way to deal with packet losses is to interpolate the missing samples in the time domain, which also adds delay [5, 6].

These methods cannot be applied to our approach, because it is based on a different principle (backward adaptive predictive coding), and because we do not want to increase our end-to-end delay for concealment strategies. The problem with backward adaptive prediction is that the predictor depends on all past samples. Thus, error-free decoding of the signal is not possible after a transmission error occurs. To overcome this problem, the adaptive predictor is reset about every second. The reset does not appear very often because a reset reduces prediction accuracy and hence compression performance. But this rare reset also leads to a long loss of valid data after a packet loss. We would like to find out how to best make use of the data that is still available before the next reset (the packets are much shorter than the reset interval). For that purpose we describe several possible concealment strategies that use this data to reduce the effect of a packet loss. In order to find out which strategy is the most effective we conducted a listening test as an evaluation.

2. DESCRIPTION OF THE CODER USED

The structure of our Ultra Low Delay coder (ULD) [7] is depicted in Fig. 1. It can be subdivided into three major processing steps, as described in the following sections.

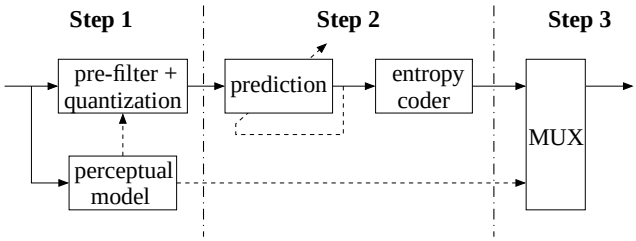

Figure 1: Basic structure of the ULD Encoder.

2.1. Encoder

In step one, the irrelevancy of the signal is reduced. The input signal is filtered by a so called pre-filter, divided by a gain factor and quantized. The pre-filter and the gain factor are controlled by a perceptual model, resulting in a a frequency response inverse to the masking threshold. The pre-filter ”normalizes” the input signal with respect to the masking threshold. This ensures that the quantization error added by the fixed uniform quantizer stays below the masked threshold. The filter coefficients and the gain value form the so-called side information which has to be transmitted to the decoder. This side info is calculated every 128 samples. The output of step one is small in amplitude compared to the input signal, and, together with the side info, represents the psychoacoustically relevant information. In order to access this information in the decoder, the following processing steps have to be invertible.

In step two, the redundancy of this signal is reduced using backward adaptive prediction and entropy coding. For backward prediction we use the Normalized LMS (NLMS) [8] . The predictor has a length of 64 taps and is used in a lossless fashion, that is, the input is integer-valued and the predicted value is rounded to integers. This way, the integer input can be exactly reconstructed in the decoder. To provide random access of the transmitted sig-

nal for the decoder, the predictor is reset every 32000 samples, whereas the entropy coder is reset every 128 samples.

Step three multiplexes the output of part two with the side info of the pre-filter. The output of part three is transmitted to the decoder in packets corresponding to a block of 128 input samples.

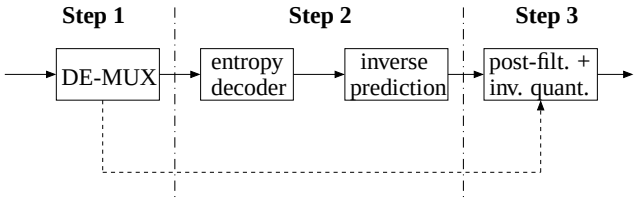

Figure 2: Basic structure of the ULD Decoder.

2.2. Decoder

The three major processing steps of the decoder are shown in Fig. 2.

The first processing step is to demultiplex each received packet. The output of the first step is the entropy coded prediction error signal and the side info.

In the second step, inverse entropy coding produces the prediction error signal. This signal is fed into the inverse predictor. If no transmission error occurred, the output of the predictor is identical to the pre-filtered signal in the encoder.

In step three, the output of step two is multiplied by the gain factor and post-filtered using the transmitted side info. The frequency response of the post-filter is inverse to the pre-filter response, thus resembling the masking threshold. When filtering the pre-filtered and quantized signal with the post-filter, the ”normalization” with respect to the masking threshold is reverted and the added quantization noise is shaped similar to the masking threshold.

3. CONCEALMENT STRATEGIES

The simplest way to deal with a packet loss in the Ultra Low Delay coding scheme is to ignore it, that is, the decoding process continues after inserting 128 zero-valued samples in the output stream. This, however, could result in very annoying full scale colored noise in the output stream. The noise would last up to the next predictor reset. The reason for this behavior is that the predictor in encoder and decoder are not synchronous any more. Even freezing the coefficients of the decoder predictor from just before the packet loss leads to strong artifacts. Since this is not acceptable, alternatives are needed.

3.1. Muting the Predictor Output

The first concealment strategy is to mute the output of the inverse predictor, or equivalently, the input of the post-filter. The muting starts with the first sample in the lost packet and lasts until the next predictor reset. The effect on the output signal is that it is very quickly faded to zero without generating any annoying artifacts. We call this strategy ”muted”.

3.2. Direct Use of the Predictor Error Signal

The second strategy only mutes the post-filter input for the duration of the lost packet and uses the following packets to produce an output signal until the next predictor reset. The strategy takes advantage of the short reset interval of the backward adaptive entropy coder (128 samples, i.e. one reset every packet), that is, for every packet following a packet loss, the prediction error signal is decodable. This signal is then fed directly into the post-filter, thus bypassing the inverse prediction. The output of the post-filter is used as the output signal. We call this strategy ”raw”.

3.3. Fixed Reshaping Filter

The third strategy differs from the second strategy in that the decoded prediction error signal is reshaped by a filter prior to being processed by the post-filter, that is, the predictor is replaced by a fixed reshaping filter until the predictor is reset. The reshaping filter is implemented as a sixth order IIR filter with fixed coefficients. The coefficients were calculated from a mean spectrum of pre-filter signals obtained from coding multi-lingual speech signals. The output of the reshaping filter is fed into the post-filter which generates the output signal. We call this strategy ”fixed”.

3.4. Adaptive Reshaping Filter

The fourth strategy is a refinement of the third strategy such that the coefficients for the reshaping filter are calculated in a signal-adaptive way. Basis for this calculation is the predictor output signal in the decoder. Starting with each predictor reset, the pre-filter signal is used to estimate the autocorrelation sequence of this signal. Once every packet, twelve filter coefficients are estimated via the Levinson-Durbin recursion using the estimated autocorrelation sequence. As soon as a packet loss occurs, this calculation process is stopped and the last calculated filter coefficients are used for the reshaping filter. We call this strategy ”adaptive”.

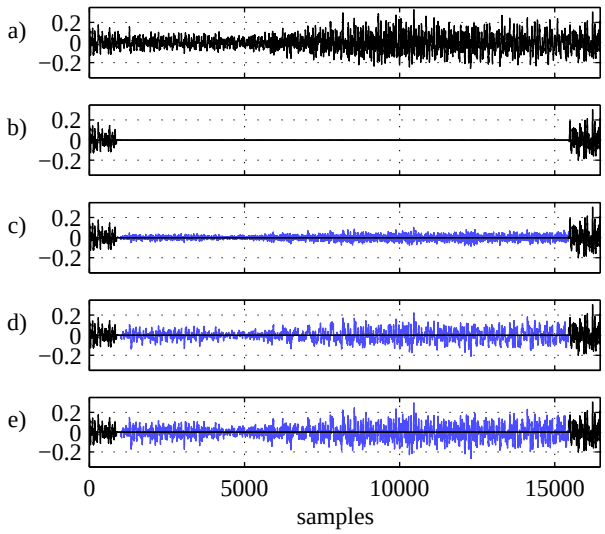

Figure 3: Example of different concealment methods: a) original signal, b) muted, c) raw, d) fixed, e) adaptive.

3.5. Example

Fig.3 illustrates the effects of the different concealment techniques with a segment of the test item sc02 (see Tab.1). The reconstructed parts of the audio signal are colored. Plot a) shows the original signal. The next four plots show the decoded signal after packet loss with the different concealment strategies applied. Plot b) shows the muting strategy ("muted"). In c), the prediction error signal is used as input for the post-filter ("raw"). In plot d), the fixed reconstruction filter is applied to the prediction error signal before it is fed into the post filter ("fixed"). in e), the adaptive reshaping filter is used on the prediction error signal before the post-filter is used ("adaptive").

As can be seen, the reconstructed signal from strategy e) has the closest resemblance to the original signal.

4. EVALUATION

For an evaluation of the four different concealment strategies, we conducted a subjective listening test using a modified MUSHRA test [9] (not using band-limited anchors).

File	Description
es01	Suzanne Vega, solo
es02	male speech, german
es03	female speech, english
sc01	trumpet
sc02	orchestra
sc03	pop music
si01	harpsichord
si02	castanets
si03	pitch pipe
sm01	bagpipe
sm02	glockenspiel
sm03	plucked strings

Table 1: List of the MPEG test files.

The test items were generated from the MPEG test set consisting of 12 audio files (see Tab. 1). In a first step, each file was mixed to mono and encoded with the Ultra Low Delay encoder at sampling frequency of 32 kHz and a constant bit rate of 96 kb/s.

In a second step, a list of lost packets was generated for each file using a random number sequence (see also Tab. 2).

In a third step, the four concealment strategies were applied in the following way: all coded MPEG files were decoded *without* any packet loss, thus forming the reference for the subjective listening test. Then, using the list of lost packets, for each concealment strategy all twelve encoded files were decoded twice. In the first decoding run, all information of the lost packet was removed, including the side info. In the second decoding run, only the entropy coded prediction error signal was lost, whereas the side info was kept. The test items of the second run could help to answer the question whether or not additional error protection for the side info could improve the perceptual quality in case of loss concealment. Altogether, the two decoding runs produced eight test items per MPEG file.

The listening test was performed in a quiet office environment. The group of the nine test listeners consisted of expert and non-expert listeners. Before the subjects started with the listening test, they had the possibility to listen to a test set. During the test, the grading of the test items is performed using a scale from 0 to 100, corresponding from "bad" to "excellent".

File	length in sec.	no. of lost packets
es01	10.736	7
es02	8.600	5
es03	7.608	4
sc01	10.972	7
sc02	12.736	8
sc03	11.556	8
si01	7.996	5
si02	7.728	4
si03	27.888	12
sm01	11.152	5
sm02	10.096	6
sm03	13.988	8

Table 2: List of the test item lengths and number of lost packets per item.

For the modified MUSHRA test, a laptop was used for collecting the ratings and playback of the test items via external DA-converter and STAX amplifier / headphones.

5. RESULTS

The results of our subjective listening test are depicted in Fig. 4. For each of the MPEG test files (es01, ..., sm03) as well as for the mean of all test files (all items), the gradings for the eight concealment strategies and for the hidden reference are given as nine different items with mean and 95%-confidence interval. The hidden reference is depicted as item 1 ("hidden_reference"). Items 2 to 5 give the results for the strategies "adaptive", "fixed", "raw" and "muted" in case of lost side info during packet loss (prefix "all_"). Items 6 to 9 give the results for the strategies "adaptive", "fixed", "raw" and "muted" for continuing side info transmission during packet loss (prefix "audio_"). As long as the confidence intervals overlap, there is no statistical significant difference in the grading.

For all MPEG files, the "adaptive" strategy was rated best. For all files except for "pitch pipe" (sc03), "muted" is significantly worse than any other strategy tested. When the strategy "adaptive" is compared to "raw", there are statistically significant differences observable between individual concealment techniques for "orchestra" (sc02), "pop music" (sc03), "castanets" (si02) and "all items". For all these files, "adaptive" is significantly better than "raw". For the "castanets" (si02), the "adaptive" strategy was also significantly better than the "fixed" strategy. For all files, no statistical significant difference is observable between no side info transmission during packet loss (prefix "all_") and continuing side info transmission(prefix "audio_").

6. CONCLUSIONS

The first significant result of our listening test is that muting the prediction error signal until the next reset is the worst strategy. For every signal tested the results are significantly below the other strategies. The second important result is that the adaptive reshaping filter for the prediction error is significantly better than using

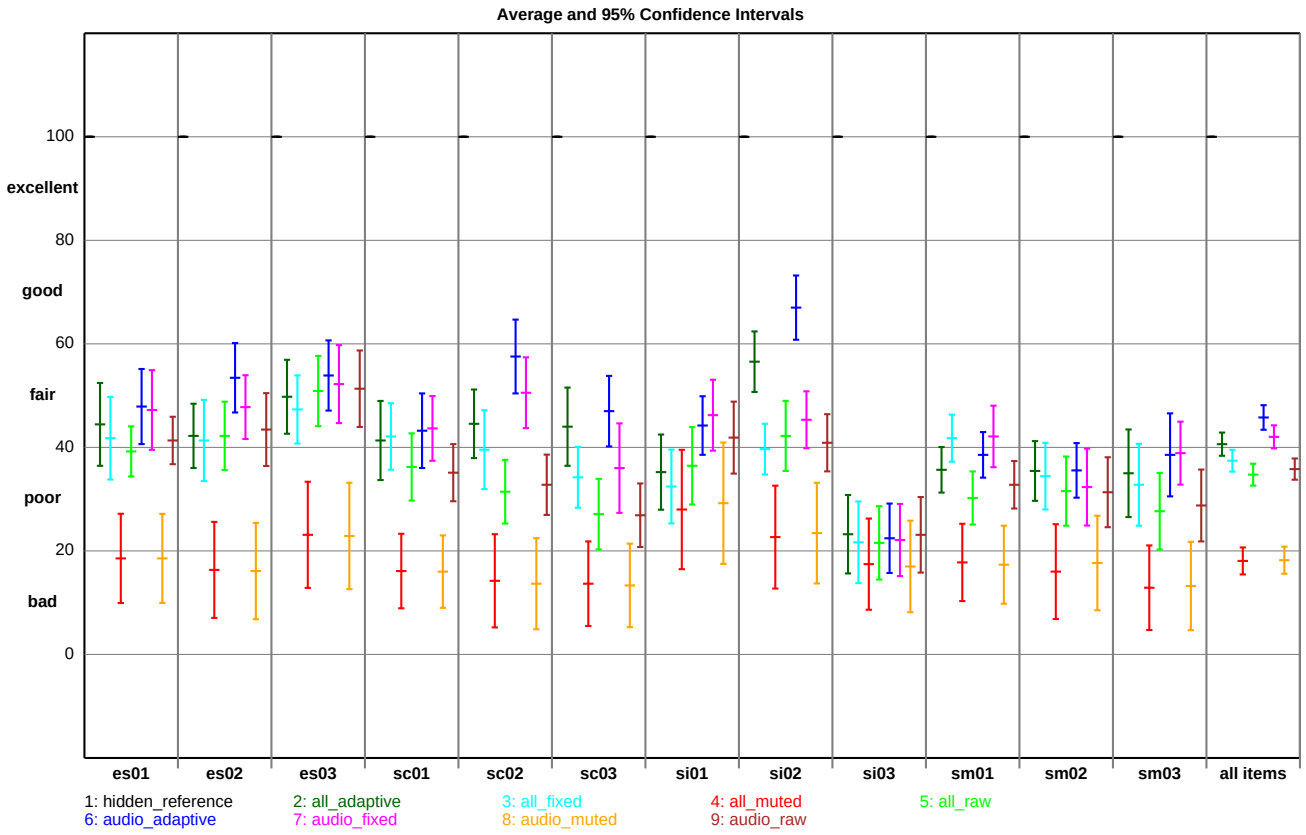

Figure 4: Results of the subjective listening test.

the unfiltered prediction error. The adaptive strategy is found to be significantly better than the fixed strategy for one MPEG file. For all other files, the average for the adaptive strategy is better than for the fixed strategy, although not statistically significant. The difference between the two cases of available side information for the psycho-acoustic post-filter and loss of this side information is not statistically significant, which was surprising. In conclusion, the best strategy is the adaptive reshaping of the prediction error, and an extra loss protection for the side information may not be worthwhile.

7. REFERENCES

[1] *Generic Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s.* ISO/IEC JTC1/SC29/WG11 (MPEG), 1993, international Standard ISO/IEC IS 11172-3.

[2] *Generic Coding of Moving Pictures and Associated Audio: Advanced Audio Coding.* ISO/IEC JTC1/SC29/WG11 (MPEG), 1997, international Standard ISO/IEC IS 13818-7.

[3] G. Schuller and A. Härmä, "Low delay audio compression using predictive compression," *IEEE International conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2002, Orlando, FL, USA.

[4] P. Lauber and R. Sperrschneider, "Error concealment for compressed digital audio," *111th AES Convention*, Sept. 2001, New York.

[5] C. Perkins, O. Hodson, and V. Hardman, "A survey of packet loss recovery techniques for streaming audio," *IEEE Network*, vol. 12, no. 5, pp. 42–48, Sept./Oct. 1998.

[6] W.-T. Liao, J.-C. Chen, and M.-S. Chen, "Adaptive recovery techniques for real-time audio streams," *Proceedings of the IEEE Infocom 2001, Anchorage, Alaska, USA*, Apr. 2001.

[7] U. Kraemer, G. Schuller, S. Wabnik, J. Klier, and J. Hirschfeld, "Ultra low delay audio coding with constant bit rate," *117th AES Convention*, Oct. 2004, San Francisco, CA.

[8] S. Haykin, *Adaptive Filter Theory*, 4th ed. Upper Saddle River, New Jersey: Prentice Hall, 2002.

[9] "Method for the subjective assessment of intermediate quality levels of coding system," ITU-R BS.1534-1, January 2003.