LOW DELAY AUDIO COMPRESSION USING PREDICTIVE CODING

Gerald Schuller and Aki Härmä

Media Signal Processing Research, Agere Systems,
Murray Hill, NJ 07974, USA;
{schuller,harma}@agere.com

ABSTRACT

A low delay audio coding scheme for communications applications is proposed. Its compression ratio is comparable to current state-of-the-art audio coding schemes, but with a much lower delay. The source of delay in conventional audio coding are the filters for the subband coding, and the block switching of the filter bank. The block switching leads to high peaks in bit-rate which necessitates a large bit rate buffer to smooth the bit rate for a transmission channel. To avoid or reduce these delays, we replace the subband coding by predictive coding, and the hard switching of the filter bank by soft switching of the predictors. The overall delay becomes 6 ms at 32 kHz sampling rate. A subjective listening test with bit-rates around 64 kb/s for mono signals shows that the new scheme has a comparable quality to a conventional state-of-the-art coder (PAC).

1. INTRODUCTION

The goal of our scheme is to provide a coding or compression scheme for high quality audio communications. Applications can be high quality teleconferencing or musicians playing together over long distances. Communications applications require a low round-trip time, RTT, which includes encoding/decoding delays of the encoder/decoder, and transmission delays. Based on listening test results we propose that the encoding/decoding delay should not exceed 10 ms.

Conventional audio coding uses the principle of subband coding. In the encoder an analysis filter bank is used to decompose the audio signal into subbands. The subband signals are quantized and coded. The quantization step size is controlled by a psycho-acoustic model such that the quantization distortions remain below the masked threshold.

The goal of a high compression ratio in perceptual coding has historically led to the use filter banks with many bands, usually switchable between 1024 and 128 bands to avoid pre-echoes. The large number of bands contributes to a high encoding/decoding delay. The delay of the coder depends on the filter bank size, the size of the look-ahead block for mode switch decisions, and buffering for constant bit-rate channels. For coders like MPEG2/4 AAC or PAC, the delays caused by the first two factors are 2047 and 576 samples respectively. The delay caused by buffering could be a few thousand samples due to the high bit-rate peaks usually associated with the 128 band mode. This can easily add up to more than 100 ms at 32 kHz sampling rate.

The MPEG-4 low delay coder [1] obtains a lower delay by using a lower number of subbands (480), has no window switching which avoids the look ahead and leads to reduced fluctuations in the bit rate, so that smaller buffers can be used. Without a buffer it has an encoding/decoding delay of 960 samples, or 30 ms at 32kHz sampling rate. The problem with the reduced number of subbands is a reduced coding efficiency, leading to higher bit rates or lower quality. At the same time the 30 ms delay is above our targeted delay.

2. ROUND-TRIP TIME

Coding delay is important in full-duplex applications while in broadcasting and storage applications the algorithmic delay can be arbitrarily high. In full-duplex applications like teleconferencing the problems are related to the echoes bouncing back from the far-end and unnatural delays in response times in interaction between parties. The latter may be a problem in a conversation application if the *round-trip time*, RTT, from near-end to a far-end and back exceeds 90 ms [2]. Reflections are not a problem if the acoustic round-trip path can be eliminated, e.g., by using a combination of headphones and close-talk microphones. Otherwise, it is necessary to use Acoustic Echo Cancellation, AEC, to attenuate return-path reflections. A distinct reflection arriving at listeners ears 40-50 ms after the direct sound is called *echo*. At low RTTs, the perceived effect is colorization of a signal.

Figure 1 shows required attenuation for a single reflection in three different experiments. The dashed curve is from [3] and represents the round-trip attenuation at *acceptable level* for telephone speech. The dashed-dotted curve shows the corresponding ITU-T G.131 recommendation. The solid curve in Fig. 1 shows our measurement data averaged over widely used high-quality audio test material including *Castanets, Suzanne Vega, female and male speakers, and flute* at the sampling rate of 32 kHz. Six experienced listeners evaluated the attenuation for a single reflection at the threshold of audibility of an echo or colorization. The test was based on the method of adjustment. The results are qualitatively in line with earlier results but, as expected, show significantly higher requirements for attenuation especially at high delays. The dip for the very low delays can be explained by the fact that very low delays (which are not realistic in most applications) result in a comb-filter like effect, which becomes noticeable.

The data suggest that a reduction of 10 ms in the round-trip delay corresponds to a 3-4 dB drop in the requirements

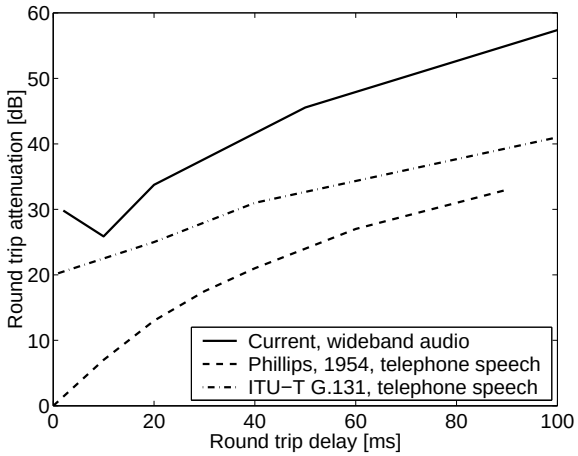

Figure 1: Required attenuation in a round trip loop.

for echo cancellation. For example, if the algorithmic coding delay is diminished from 20 to 6 ms, the requirements for AEC are down by more that 10 dB. The required high attenuation for echoes (RTT $>$ 50 ms) is very difficult to achieve. Therefore, a 25 ms one-way delay is a realistic upper margin for echo-free audio communications. The speed of light in an optical fiber is approximately 200 km/ms. Hence, one may roughly estimate that a 1 ms decrease in algorithmic coding delay corresponds to a 100 km increase in the range of echo-free communications. This all suggest that the coding delay should be below about 10 ms which is also close to the recommendations for low-delay speech coding [4].

3. NEW APPROACH

Since it is difficult to obtain our desired delay with subband coding, we will replace it with predictive coding. Theoretically predictive coding leads to the same coding efficiency as transform coding [5], but at a much lower delay. The main problem of this approach for audio coding is that psycho-acoustic models are based on a subband decomposition of the audio signal, hence there is no direct way to apply the output of a psycho-acoustic model to predictive coding. To solve this problem we separate the application of psycho-acoustics (the irrelevance reduction) from the redundancy reduction, so that we have 2 separate units [6]. The input of our psycho-acoustic model still consists of subband signals from an analysis filter bank. But since the irrelevance reduction unit is not constrained by coding efficiency, we can choose the number of subbands smaller. We chose 128 uniform bands and found it gives sufficient frequency and also time resolution for time and frequency masking effects. The output of the psycho-acoustic model is the masking threshold for each subband. We then view this output as a power spectrum, compute an auto-correlation function of it and finally linear predictive coefficients. These coefficients are used in a linear predictive structure, which we call pre-filter. Its effect is a normalization of the audio signal to its masking threshold. More quantitatively, for every consecutive block of 128 input samples we compute the masking threshold $M(f)$ (dependent on frequency f).

The pre-filters transfer function $H(f)$ should satisfy

$$H(f) = \frac{1}{|M(f)|} \tag{1}$$

The order of the pre-filter is K and its output $x(n)$ is related to its input $s(n)$ through

$$x(n) = s(n) - \sum_{k=1}^{K} a_k s(n-k) \ . \tag{2}$$

The inverse DFT of $|M(f)|^2$ gives the auto-correlation function $r_{mm}(n)$. Then the filter coefficients a_k are obtained by solving the linear equation system [7]

$$\sum_{k=0}^{K-1} r_{mm}(|k-n|)a_k = r_{mm}(n+1) \ , \quad 0 \le n < K \ . \tag{3}$$

We found that a 12th order frequency-warped filter is sufficient to model the masking threshold. To avoid artifacts from interpolating coefficients from one block to the next we implement the filter in a lattice structure [6]. The post-filter in the decoder has a frequency response which is the inverse of the pre-filter. Hence we need to transmit the shape of the frequency response of the pre-filter to the decoder as side information. We use a parameterization with line spectral frequencies [6] for that purpose. This side information is the analog of the scale factors in conventional audio coding.

The delay of this stage is only the 128 samples needed for the psycho-acoustic model. A simple rounding operation is used after pre-filtering to quantize the signal. Since the pre-filter normalizes the signal to its masking threshold rounding results in quantization distortions just at the masking threshold. The sequence of integers after this rounding now needs to be encoded in a lossless way for the redundancy reduction. Our goals for the lossless coding unit are again a low delay and at the same time to maintain a high compression ratio.

Current lossless audio coders are typically based on block wise forward prediction. The prediction coefficients for a block are transmitted as overhead, and the residuals are Huffman coded and transmitted. This means there is a delay of at least one block size. Lossless coders are typically intended for file compression, where delay is of no concern, and where the computational complexity is of some importance because it determines the compression time. This means that their compression performance is not optimized. To obtain a low delay we use backward adaptive predictive coding which is also a standard techniques in low-delay speech coding [4]. Backward adaptive prediction has also been used in previously in audio coding, e.g., in backward adaptive warped lattice algorithm proposed in [8]. However, in these cases backward adaptive prediction was used in LPC filters, and in a lossy scheme, while in the current article it is used in lossless compression after the quantizer.

The backward adaptive prediction is implemented using the normalized least means squares (NLMS) algorithm [7]. To obtain a higher compression performance we use soft switching between different predictors [9, 10]. The filter

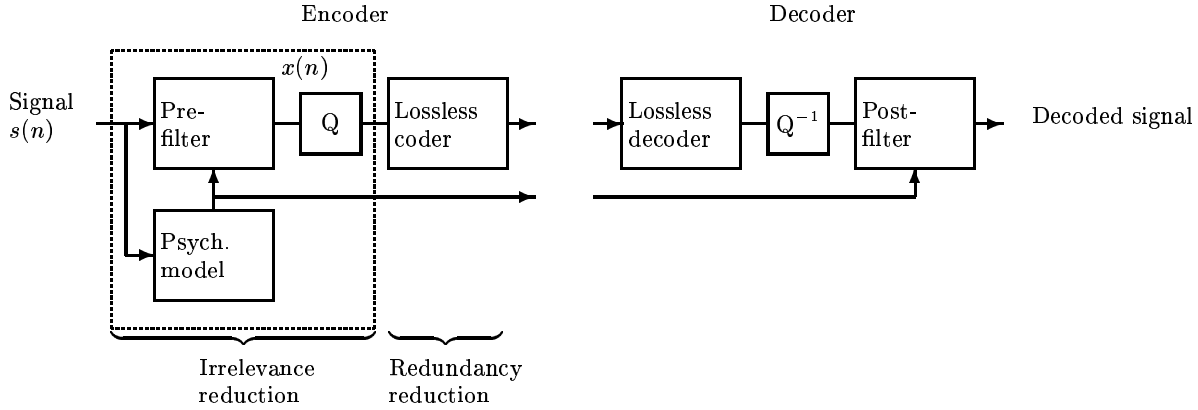

Figure 2: The audio coding scheme with separated irrelevance and redundancy reduction, using a psycho-acoustic pre- and post-filter and lossless compression.

bank in conventional audio coding has 2 modes, one with a high number of bands (typically 1024) but reduced time resolution, and one with a lower number of bands (typically 128) but higher time resolution. The mode depends on the signal and is hard switched (either one or the other). The analog of a filter bank with many bands in subband coding is a predictor with high order in predictive coding. In predictive coding, having more modes is simpler than in subband coding. We found that 3 modes (meaning 3 different predictors) instead of 2 provides an advantage for the compression performance. In predictive coding it is also possible to have *soft* switching instead of hard switching. We found that soft switching also provides an advantage for the compression performance. We implement the soft switching between the 3 different individual predictors as follows. $P_1(n)$, $P_2(n)$, and $P_3(n)$ are the 3 different individual predictors with different order. Then the final predicted value $P(n)$ is a linear combination of the 3 individual predictors, using weights w_i

$$P(n) = \sum_{i=1}^{3} w_i(n) \cdot P_i(n), \;\; w_i(n) \geq 0, \;\; \sum_{i=1}^{3} w_i(n) = 1.$$

To obtain the 3 different individual predictors we cascaded 3 predictors to obtain a computational efficient structure, and hence call it weighted cascaded LMS (WCLMS) prediction). To obtain the highest compression performance we found a order of 200 for the first predictor of the cascade, order 80 for the second, and order 40 for the last is suitable. Observe that the orders also correspond roughly to the number of subbands used in conventional audio coding. The weights w_i are adjusted such that the weight is higher for an individual predictor with small past prediction error $e_i(n)$. With the assumption of a Laplacian distribution of the prediction error we obtain the weights as [9]

$$w_i(n) = \exp(-c(1-\mu)) \sum_{i=1}^{n-1} |e_i(n-i)| \cdot \mu^{i-1}$$

with tuning parameters which we chose to $c = 2$ and $\mu = 0.9$. Since the input signal $x(n)$ of our lossless coder is integer valued, we use the rounded value of our final predictor $P(n)$ to obtain the prediction error signal $e(n)$,

$$e(n) = [x(n)] - [P(n)]$$

where the square brackets [] denote the rounding operation. This predictor introduces no delay. The prediction error signal $e(n)$ is then entropy coded and transmitted. We use low delay entropy coding schemes. We found that adaptive Huffman coding and arithmetic coding lead to similar or slightly better compression results compared to conventional block based Huffman coding [10]. Our adaptive Huffman coding algorithm leads to a delay of only 17 samples, our Arithmetic coding scheme to a delay of about 100 samples. The entire low delay audio encoder, consisting of the combination of the pre- and post-filter (PPF) with the WCLMS lossless unit then leads to, depending on the adaptive Huffman coding or arithmetic coding, a delay of $128 + 17$ or $128 + 100$, which are both in the order of 200 samples. Since the decoder does not introduce additional delay, this is about 6 ms encoding/decoding delay at 32 kHz sampling rate, if no bit-rate buffering is used. Hence it is even below our targeted delay of 10 ms at 32 kHz.

To give an impression of the performance of the combined system PPF-WCLMS in terms of bit-rate and audio quality, we compare it with a state-of-the-art audio coder, PAC [11], in mono-mode. We use a subjective listening test on a set of 10 test signals. The 10 test signals are chosen from a set of 73 signals by several experienced listeners to be particularly critical (coding artifacts are more pronounced). They consist of speech signals (mspeech, spot), single instruments (tink, castanet, triangle, oboe), music with several instruments (chart, jazz), and mixed speech and music (mixed).

Both coders are used without an output bit-rate buffer, as they could be used for transmission channels with variable bit-rate (e.g. packet networks). Not using a bit-rate buffer is also helping our goal of a low encoding/decoding delay. We set both coders such that they use the same average bit-rate over the length of each individual signal. The adjustment is done such that the bit-rate is not too far from the starting point given by their psycho-acoustic model (for most signals this starting point is quite similar between them), and such that it is between 1.5 and

Signal	Bit/sample
A tink	1.625
B chart	2.0625
C jazz	2.0625
D castanet	2.0625
E harps	1.8437
F mixed	2.375
G mspeech	2.3125
H oboe	1.5937
I spot	2.25
K triangle	1.5937

Table 1: The signals for the subjective comparison test and their bit-rate (in bit per sample) for both coders, at 32 kHz sampling rate.

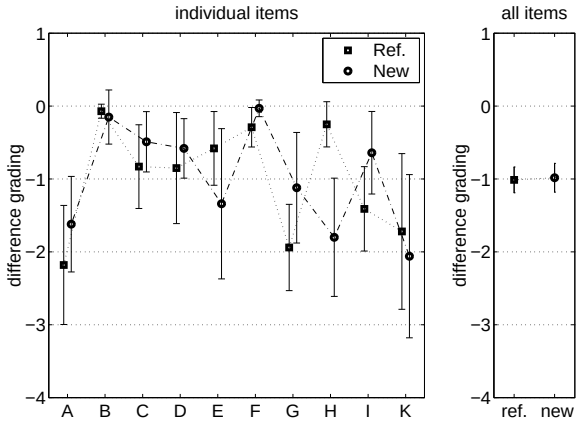

Figure 3: The result of our listening test. "Ref." is the PAC mono coder, "New" is our PPF-WCLMS coder. The vertical bars around each value show the 95 % confidence interval.

2.4 bits/sample. Table 1 shows the used test signals and their corresponding bit-rates for both coders. We use a three-alternative hidden-reference test, as described by the ITU [12]. For the evaluation the ITU five-grade impairment scale is used. We had 5 expert listeners in our test. The listening test is conducted in a sound proof booth, with a Linux workstation with a high quality sound board and STAX Lambda Pro headphones. The results are displayed in Fig. 3, where the difference grading is the difference between the grade a subject gives for the original and for the encoded/decoded signal. The circles show the average grading for the PPF-WCLMS coder, the squares the averages for the PAC coder. It can be seen that for most signals there is no statistically significant difference in the evaluation of the two coders. Also, on average there is no statistically significant difference between the two coders. Recall that PPF-WCLMS has a delay of about 200 samples (6 ms)compared with 2047 sample filter bank delay plus 576 sample look ahead for window switching, or about 3000 sample (100 ms) delay for PAC. Hence we conclude that we can indeed significantly reduce our encoding/decoding delay without sacrificing quality or compression performance compared to traditional audio coders.

4. CONCLUSIONS

We constructed a high quality audio coder with a low encoding/decoding delay for communications purposes. Its structure is analog to conventional audio coders, but we use predictive coding instead of subband coding, and soft switching instead of hard switching. The cost for the lower delay is an increased complexity of our approach compared to conventional subband based audio coders. Our approach results in a considerably reduced delay of about 6 ms at 32 kHz sampling rate, compared to 100 ms or more for conventional audio coding. A subjective listening showed that the audio quality delivered by our coder is comparable to a conventional coder (PAC).

REFERENCES

[1] E. Allamanche, R. Geiger, J. Herre T. Sporer, "MPEG-4 Low Delay Audio Coding based on the AAC Codec", 106th AES Convention, Munich, Germany, May,1999.

[2] N. Kitawaki and K. Itoh, "Pure delay effects on speech quality in telecommunications," *IEEE J. Sel. Areas in Comm.*, vol. 9, pp. 586–593, May 1991.

[3] G. M. Phillips, "Echo and its effects on the telephone user," *Bell Laboratories Record*, vol. 32, pp. 281–284, August 1954.

[4] J.-H. Chen, R. V. Cox, Y.-C. Lin, N. Jayant, and M. J. Melchner, "A low-delay CELP coder for the CCITT 16 kb/s speech coding standard," *IEEE J. Sel Areas in Comm.*, vol. 10, pp. 830–849, June 1992.

[5] N.S. Jayant, P. Noll, "Digital Coding of Waveforms", Prentice Hall, Englewood Cliffs, New Jersey, 1984.

[6] B. Edler and G. Schuller, "Audio Coding Using a Psychoacoustic Pre- and Post-Filter", ICASSP 2000, Istanbul, Turkey, pp. II-881:884.

[7] S. S. Haykin (1999). *Adaptive Filter Theory.* Englewood Cliffs, N.J. : Prentice Hall.

[8] A. Härmä, U. K. Laine, and M. Karjalainen, "Backward adaptive warped lattice for wideband stereo coding," in *Proc. of EUSIPCO'98*, (Greece), 1998.

[9] G. Schuller, B. Yu, D. Huang, "Lossless coding of audio signals using cascaded prediction", ICASSP 2001, Salt Lake City, Utah, May 2001

[10] S. Dorward, D. Huang, S. A. Savari, G. Schuller, B. Yu, "Low Delay Perceptually Lossless Coding of Audio Signals", Data Compression Conference, Snowbird, UT, March 2001, pp. 312-320

[11] V. Madisetti, D. B. Williams, eds., "The Digital Signal Processing Handbook", Chapter 42, D. Sinha et al., "The Perceptual Audio Coder (PAC)", CRC Press, Boca Raton, Fl., 1998.

[12] ITU-R, "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems", Rec. ITU-R BS.1116-1, Geneva, 1997