AN ERROR ROBUST ULTRA LOW DELAY AUDIO CODER USING AN MA PREDICTION MODEL

Stefan Wabnik[1]

[1]Fraunhofer IDMT
Institute for Digital Media Technology
Ehrenbergstr. 31, 98693 Ilmenau, Germany
wbk@idmt.fraunhofer.de

Gerald Schuller[1,2]*, Ferenc Kraemer*[2]

[2]Technical University of Ilmenau
Institute for Media Technology
Helmholtzplatz 2, 98693 Ilmenau, Germany
shl,kraefc@idmt.fraunhofer.de

ABSTRACT

This paper compares two prediction structures for predictive perceptual audio coding in the context of the Ultra Low Delay (ULD) coding scheme. One structure is based on the commonly used AR signal model, leading to an IIR predictor in the decoder. The other structure is based on an MA signal model, leading to an FIR predictor in the decoder. We find that the AR-based predictor has a slightly better performance in case of an undisturbed transmission channel, but the MA-based predictor has a much better performance in case of transmission errors. For a Bit Error Rate (BER) of 1.0e-5, the perceptual quality of the proposed MA model predictor achieves a mean Objective Difference Grade (ODG) of -0.66 ODG whereas the AR model predictor only reaches -3.42 ODG.

Index Terms— Low Delay Audio coding, Linear predictive coding, Moving average processes, Autoregressive processes, Robustness.

1. INTRODUCTION

The perceptually controlled Ultra Low Delay (ULD) audio coding scheme uses predictive coding to reduce signal redundancy. Compared to sub-band coding, predictive coding yields similar coding gain, but has much lower algorithmic encoding/decoding delay [1].

The decision for a certain predictor structure, as well as for forward or backward adaptive coefficient updates, always implies certain advantages and shortcomings. With forward adaptive coefficient update, for example, the transmission of the filter weights is necessary which in turn leads to an increase in bit rate and therefore limits the model order. Backward adaptive predictive coding, on the other hand, is sensitive to transmission errors. An Auto Regressive (AR) source model for the signal to encode leads to an Infinite Impulse Response (IIR) structure for the predictor in the decoder. This is often preferred to a Moving Average (MA) model [2, 3], since many natural signals, like sinusoidal tones for instance, are better synthesized by the AR source model [4, 5]. For random access of the transmission as well as for transmission errors, a reset of the predictor states in both encoder and decoder is used in case of an AR model. To maintain sufficient prediction performance, the time interval between the resets is chosen much larger than the order of the prediction filter. However, in the presence of transmission error, a larger reset interval can also lead to larger amount of corrupted data. To minimize the effect of transmission errors, additional concealment techniques [6, 7, 8] could be applied. With higher error rates, the sensitivity of the IIR structure to transmission errors is a problem, especially when the frequency of errors becomes similar to the frequency of resets. The MA model, on the other side, has the inherent property of an only limited error propagation, so that the performance in case of transmission errors is expected to be higher compared to the AR model.

The goal for this paper is to find a prediction model suitable for the ULD coding scheme for both undisturbed and disturbed transmission. As an approach, we investigate and evaluate the performance of these two predictor versions in the context of our ULD codec, for the case of no transmission errors and with transmission errors.

The rest of the paper is organized as follows: section 2 gives an overview over the ULD encoder and decoder structure, section 3 describes the AR and MA modeling used in the prediction stage of the ULD encoder, section 4 gives experimental results on the behavior of the ULD coding scheme using either AR or MA modeling for disturbed and undisturbed transmission, and Section 5 gives some conclusions.

2. ULTRA LOW DELAY CODING SCHEME

The Ultra Low Delay Audio Coder achieves a total encoding / decoding delay of 5.33 to 8 milliseconds with sampling frequencies from 32 kHz to 48 kHz [9][10]. The scheme achieves bit rates in the range of 80 to 96 kbit/s @ 32 kHz sampling rate.

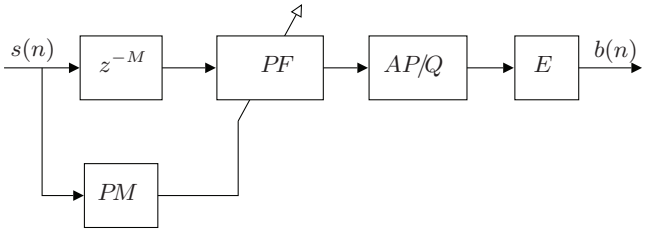

Fig. 1. Design of the ULD encoder: PM perceptual model, AP/Q adaptive prediction and quantization, E entropy coding.

In the encoder (see Fig.1), the input signal $s(n)$ is filtered with a pre-filter PF which is controlled by a perceptual model PM. It estimates the masking threshold causing an algorithmic delay of M=256 input samples. This estimate is used to calculate filter coefficients such that the pre-filter normalizes the input signal with respect to the masking threshold. Compared to the input signal, the pre-filtered signal is much smaller in magnitude. The pre-filtered signal is adaptively predicted and quantized in block AP/Q. The indexed quanti-

zation steps are entropy coded in E, and the output $b(n)$ is sent to the decoder.

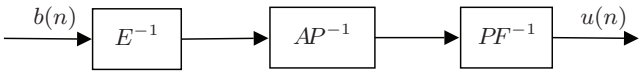

Fig. 2. Design of the ULD Decoder: : E^{-1} inverse entropy coding, AP/Q^{-1} inverse adaptive prediction, PF^{-1} post-filter.

The decoder (Fig.2) contains an entropy decoder E^{-1}, followed by the predictive decoder structure AP^{-1} and a post-filter PF^{-1}. The post-filter transfer function, which is the inverse of the pre-filter transfer function, has a frequency response like the masking threshold. The quantization noise added in the encoder is filtered by the post-filter and thus shapes it like the masking threshold. Hence, the output signal $u(n)$ is a superposition of the encoder input signal and the perceptually shaped quantization noise.

The following section 3 will give a detailed description of both the adaptive prediction and quantization (block AP/Q) in the encoder and inverse adaptive prediction (blockAP^{-1}) in the decoder. Both imply a certain source model for the pre-filtered signal.

3. SOURCE MODELS

In the design of the ULD coding scheme, two structural decisions had to be made, namely to choose a model for the pre- and post-filter and a model choice for the prediction.

Application of the pre- / post-filter maps the signal space onto a perceptually weighted signal space, and a good model fit would certainly boost the performance of the overall coding scheme. The model chosen for the pre- / post-filter combination is an AR synthesis model, i.e the post-filter is a warped lattice-based IIR-like structure and the pre-filter its inverse. This modeling will *not* be under further investigation for this paper. What we will be investigating is the second structural decision mentioned above, namely the modeling of the pre-filtered signal in the prediction stage. The naming of the signal model in this paper is based on the model used in the synthesis predictor in the decoder.

3.1. AR Source Model

This subsection describes the AR modeling of the prediction stage in the ULD coding scheme. In Fig.3, the encoder structure of the prediction stage *AP/Q* is shown. The quantize operator $Q\{\cdot\}$ maps the difference sequence $d(n) = x(n) - p(n)$ to the index sequence $i(n)$. This index sequence is mapped to the quantized difference sequence $\tilde{d}(n)$ using $Q^{-1}\{\cdot\}$. The summation $u(n) = \tilde{d}(n)+p(n)$, which is the quantized version of $x(n)$, is the input sequence to the adaptive predictor $W_n(z)$.

The decoder is depicted in Fig.4. The index sequence $i(n)$ is mapped to the quantized difference sequence $\tilde{d}(n)$, to which the predicted sequence $p(n)$ is added. The sequence $u(n)$ is the quantized and decoded version of the pre-filter encoder sequence $x(n)$.

The N time variant prediction coefficients $\mathbf{w}_n$ are updated from past decoded samples $\mathbf{u}_n$ using the NLMS algorithm with constant step size μ and regularization parameter δ:

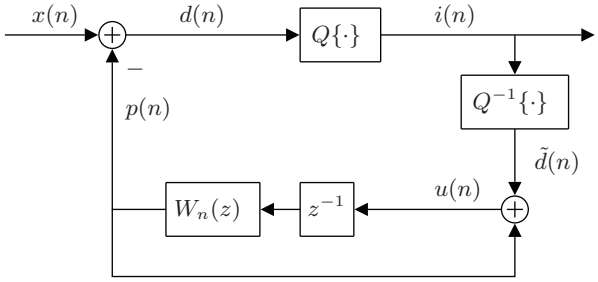

Fig. 3. AR model, encoder.

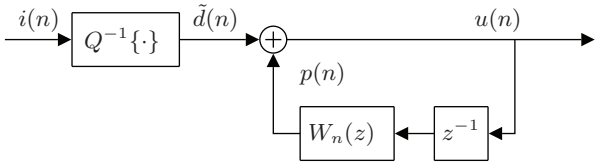

Fig. 4. AR model, decoder.

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \frac{\mu \mathbf{u}(n)}{\delta + \mathbf{u}(n)^T \mathbf{u}(n)} \tilde{d}(n) \tag{1a}$$

$$\mathbf{w}(n) := [w_1, w_2, \ldots, w_N]^T \tag{1b}$$

$$\mathbf{u}(n) := [u(n-1), u(n-2), \ldots, u(n-N)]^T. \tag{1c}$$

For the analysis of the prediction stage behavior in the decoder after a transmission error, we start with the assumption that one single index value $i_e(n)$ at time instant n is disturbed in the index sequence, where e stands for an erroneous sample. After mapping $i_e(n)$ with $Q^{-1}\{\cdot\}$, the erroneous value $\tilde{d}_e(n)$ disturbs $u_e(n) = p(n) + \tilde{d}_e(n)$. Additionally, the update $\mathbf{w}_e(n+1)$ of the NLMS predictor and $\mathbf{u}_e(n+1)$ will be disturbed, too. The disturbance in $\mathbf{w}_e(n+1)$ in turn produces an additional erroneous sample $p_e(n+1)$ which in turn generates disturbed $u_e(n+1)$, $\mathbf{u}_e(n+2)$ and $\mathbf{w}_e(n+2)$. Since the synthesis system has an infinite impulse response, the disturbances can make the system unstable. Thus, to recover from a transmission error, the predictor states in both encoder and decoder have to be reset to predefined values from time to time. This procedure generally lowers the prediction gain, so resets should occur as sparsely as possible.

3.2. MA Source Model

This subsection describes the proposed MA modeling of the prediction stage in the ULD coding scheme (see Fig.5). The difference of pre-filtered input sequence $x(n)$ and the predicted sequence $p(n)$ forms the prediction residual sequence $d(n)$. This sequence is mapped to an index sequence $i(n)$ via $Q\{\cdot\}$, which is the output of the prediction stage. The Mapping $Q^{-1}\{\cdot\}$ generates the quantized prediction error sequence $\tilde{d}(n)$, which is the input sequence to the adaptive predictor $W_n(z)$.

In the decoder (see Fig.6), the quantized prediction residual $\tilde{d}(n)$ is generated using $i(n)$ and $Q^{-1}\{\cdot\}$. With $\tilde{d}(n)$ and predictor

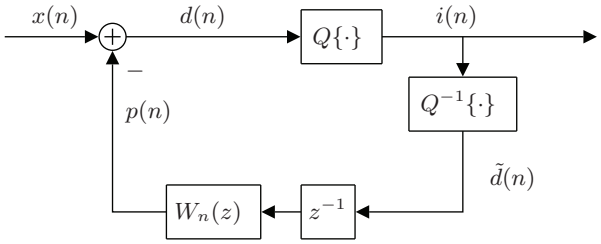

Fig. 5. MA model, encoder.

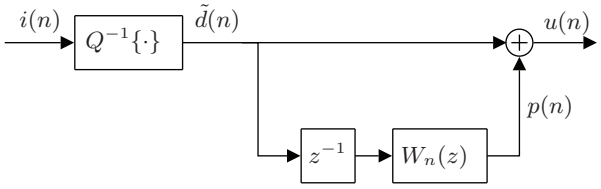

Fig. 6. MA model, decoder.

$W_n(z)$, the predicted sequence $p(n)$ is generated. Addition of $\tilde{d}(n)$ and $p(n)$ generates the output sequence $u(n)$.

In contrast to the AR modeling, this time the N time variant prediction coefficients $\mathbf{w}_n$ are updated from past quantized prediction **residuals** $\tilde{d}(n)$ using the NLMS algorithm with constant step size μ,regularization parameter δ and an additional leakage factor $\alpha, 0 < \alpha \ll 1$:

$$\mathbf{w}(n+1) = (1-\alpha)\mathbf{w}(n) + \frac{\mu \tilde{\mathbf{d}}_n}{\delta + \tilde{\mathbf{d}}(n)^T \tilde{\mathbf{d}}(n)} \tilde{d}(n) \quad (2a)$$

$$\mathbf{w}(n) := [w_1, w_2, \ldots, w_N]^T \quad (2b)$$

$$\tilde{\mathbf{d}}(n) := [\tilde{d}(n-1), \tilde{d}(n-2), \ldots, \tilde{d}(n-N)]^T. \quad (2c)$$

For the analysis of the prediction stage behavior in the decoder in the presence of transmission error, we again assume that one single index value $i_e(n)$ at time instance n is disturbed, where e stands for an erroneous sample. After mapping $i_e(n)$ with $Q^{-1}\{\cdot\}$, the disturbed value $\tilde{d}_e(n)$ directly corrupts the value of $u_e(n) = \tilde{d}_e(n) + p(n)$ on time instant n. For the next N time instances, the erroneous value $\tilde{d}_e(n)$ shifts through $\tilde{\mathbf{d}}(n)$. This disturbance also affects $\mathbf{w}_{n+1}$ through $\mathbf{w}_{n+N}$ and produces erroneous prediction values $p_e(n+1, \ldots, n+N)$, which in turn cause an erroneous output sequence $u_e(k) = \tilde{d}(k) + p_e(k), k = n+1, \ldots, n+N$. After $N+1$ time instances, $\tilde{\mathbf{d}}_n$ does not contain the disturbed sample anymore and thus the second update term in equation 2a produces correct values again. The only disturbance left in the system is the difference $\Delta\mathbf{w}(n+N) = \mathbf{w}_e(n+N) - \mathbf{w}(n+N)$ of the disturbed coefficient vector. Since update equation 2a forms a first order difference system: $\Delta\mathbf{w}_e(k+1) = (1-\alpha) * \Delta\mathbf{w}_e(k), k = n+N, n+N+1, \ldots$, the disturbance $\Delta\mathbf{w}_e(n+N)$ will fade out over time. The decay time of this system depends on the factor alpha.

4. EXPERIMENTAL RESULTS

This section starts with an example illustrating the effects of AR and MA modeling in the context of the ULD coding scheme. We then present experimental results from the evaluation of these modelings in the ULD coding scheme under different test scenarios.

For the illustrative example, we disturbed a single prediction residual value (sample) and plotted the evolving error pattern for the pre-filtered signal in the decoder. In Figure 7, we see three signals: a) the undisturbed pre-filtered signal, b) the error signal in case of MA modeling, and c) the error signal in case of AR modeling. Comparing plot b) and c), the advantage of the MA modeling in the decoder becomes obvious: The AR predictor (Fig.7c) in the decoder stalls after a single sample error until the next reset, while the same error causes only minor deterioration when applying the MA predictor (Fig.7b). Although the time constant of the leakage factor for the MA model may be relatively longthe duration of perceptible errors is actually much less. In Fig.7 b) the disturbance due to a sample error lasts for about $L = 2\,ms$ corresponding to the chosen model order ($N = 64$), but afterwards almost immediately disappears.

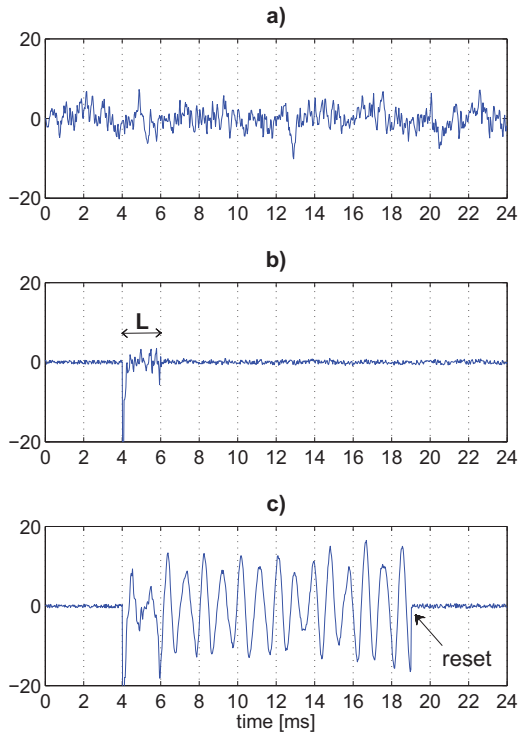

Fig. 7. Exemplary error behavior of the different source models. **a)** Pre-filtered signal ($x(n)$). **b)** MA model: error signal $x(n) - u(n)$ after erroneous predictor input sample ($\tilde{d}_e(n)$) at $4\,ms$. **c)** AR model: error signal $x(n) - u(n)$, error identical to **b)**.

For the evaluation of the ULD coding scheme with both AR and MA prediction models, three different codec versions and three different transmission error scenarios were evaluated

The codec versions under test all used identical pre-/post-filter and entropy encoding/decoding units, but differed in the applied prediction unit and reset distance. Version one (*MA*) incorporated the proposed MA source model approach , version two (AR_{100ms}) and three (AR_{1s}) both utilized the same AR model structure, but with a different reset interval of $100\,ms$ and $1\,s$.

The material to be tested with PEAQ consisted of 12 mono audio

files of the MPEG reference test set, all with a sampling frequency of 32 kHz: es01 (Suzanne Vega), es02 (male speech, German), es03 (female speech, English), sc01 (trumpet), sc02 (orchestra), sc03 (pop music), si01 (cembalo), si02 (castanets), si03 (pitch pipe), harpsichord01 (bagpipe), sm02 (glockenspiel) and sm03 (plucked strings).

The test material was encoded and decoded at a bit rate of $96\,kbps$. For the transmission error simulation, burst errors of a length up to $4\,ms$ with a mean error distance of $111\,ms$ at a BER of 10^{-4} and $1110\,ms$ at a BER of 10^{-5} were simulated. The error patterns only affected the encoded prediction residual after entropy coding.

To obtain a comparison of the perceptual quality, we used the advanced model implementation of the (PEAQ) standard from OPTICOM [11, 12]. The OPTICOM implementation compares the original file with the decoded file and gives a perceptual quality rating based on the five grade impairment scale ranging from 0.0 (no noticeable difference) to -5.0 (very annoying). The trend reflected in the ODGs values was confirmed by informal listening comparisons.

Test File	PEAQ ODG(*AM*)		
	MA	$\boldsymbol{AR}_{100ms}$	$\boldsymbol{AR}_{1s}$
es01	-0.37	-0.38	-0.39
es02	-0.35	-0.36	-0.36
es03	-0.37	-0.39	-0.37
sc01	-0.24	-0.24	-0.25
sc02	-0.27	-0.27	-0.27
sc03	-0.28	-0.29	-0.29
si01	-0.39	-0.39	-0.38
si02	-0.39	-0.42	-0.46
si03	-1.43	-1.08	-0.59
sm01	-0.28	-0.28	-0.30
sm02	-2.10	-1.60	-1.99
sm03	-0.27	-0.27	-0.27
mean	-0.56	-0.50	-0.49

Table 1. ODGs of PEAQ (*Advanced Model*) in the error free case.

For the error free case, the achieved Objective Difference Grades (ODGs) for all three version are listed in Table 1. Observe that the decreased source model fit for the MA model (*MA*) can mainly be noticed in the error free case for test item sm02.

Test File	PEAQ ODG(*AM*)					
	MA		$\boldsymbol{AR}_{100ms}$		$\boldsymbol{AR}_{1s}$	
	BER		BER		BER	
	10^{-5}	10^{-4}	10^{-5}	10^{-4}	10^{-5}	10^{-4}
es01	-0.44	-1.21	-3.51	-3.98	-3.58	-3.98
es02	-0.44	-0.90	-2.70	-3.98	-2.62	-3.98
es03	-0.42	-1.03	-3.02	-3.98	-3.02	-3.98
sc01	-0.34	-1.05	-3.96	-3.97	-3.96	-3.97
sc02	-0.31	-0.63	-3.54	-3.98	-3.52	-3.98
sc03	-0.34	-0.67	-2.47	-3.98	-2.77	-3.98
si01	-0.44	-1.09	-3.71	-3.97	-3.72	-3.97
si02	-0.51	-0.86	-2.52	-3.98	-2.61	-3.98
si03	-1.73	-3.20	-3.95	-3.97	-3.95	-3.97
sm01	-0.35	-1.18	-3.97	-3.97	-3.96	-3.97
sm02	-2.25	-3.10	-3.91	-3.97	-3.85	-3.97
sm03	-0.35	-0.77	-3.80	-3.97	-3.80	-3.97
mean	-0.66	-1.31	-3.42	-3.97	-3.45	-3.97
mean var	0.04	0.28	0.45	0.00	0.44	0.00

Table 2. Mean ODGs of PEAQ (*Advanced Model*) over 100 realizations for BERs of 10^{-5} and 10^{-4}. *mean var*: Mean variance of the deduced ODGs.

Table 2 shows the mean ODG values in the error case with a BER of 10^{-5} and 10^{-4} calculated from 100 different error pattern realizations. In general the *MA* version outperforms the AR version, especially for non harmonic signals. For version AR_{100ms} and AR_{1s}, going from the error free case to a BER of 10^{-4} leads to a reduction of the ODG of about 3.5, whereas the degradation for the *MA* version is only in the order of 0.7, except for si03 and sm02.

5. CONCLUSIONS

We compared two different approaches for predictive audio coding. One has a predictor with an underlying signal model which is a good fit for most audio signals (the AR signal model). Hence it has a potentially higher coding gain. The other has a predictor with an underlying signal model, which is not the best fit for many audio signals (the MA signal model), but which leads to a more robust coding scheme in case of transmission errors. Our comparisons showed that the penalty using the MA signal model is small compared to the gain in performance in the case of disturbed transmission. With a Bit Error Rate of 10^{-5} the resulting decoded audio quality is hardly affected for the MA case, with a reduction of the ODG value of only about 0.17, compared to about 3.5 for the AR case.

6. REFERENCES

[1] M. Lutzky, G. Schuller, M. Gayer, U. Krämer, and S. Wabnik, "A Guideline to Audio Codec Delay," *116th AES Convention*, May 2004.

[2] Ali H. Sayed, *Fundamentals of Adaptive Filtering*, Wiley-IEEE Press, 2003.

[3] ITU-R G.722, "7 khz (Wide Band) Audio-Coding within 64 kbit/s," Nov. 1988.

[4] S. Haykin, *Adaptive Filter Theory*, Prentice-Hall, 4th edition, 2001.

[5] N. S. Jayant and P. Noll, *Digital Coding of Waveforms - Principles and Applications to Speech and Video*, Prentice Hall, Englewood Cliffs, New Jersey, 1984.

[6] G. Schuller, S. Wabnik, U. Krämer, and J. Hirschfeld, "Concealment Strategies fot the Ultra Low Delay Audio Coding Scheme," Tech. Rep., Fraunhofer IDMT, dept. of audio applications, audio coding team, Jun. 2004.

[7] S. Wabnik, G. Schuller, J. Hirschfeld, and U. Krämer, "Packet Loss Concealment in Predictive Audio Coding," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 2005.

[8] C. Perkins, O. Hodson, and V. Hardman, "A Survey of Packet Loss Recovery Techniques for Streaming Audio," *IEEE Network Mag.*, vol. 12, no. 5, pp. 40–48, 1998.

[9] B. Edler, C. Faller, and G. Schuller, "Perceptual Audio Coding Using a Time-Varying Linear Pre- and Post-Filter," *109th AES convention*, Sep. 2000, Los Angeles, CA, USA.

[10] 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.

[11] ITU-R BS.1387-1, "Method for Objective Measurements of Perceived Audio Quality," Nov. 2001.

[12] Official website of OPTICOM, "http://www.opticom.de/," Last checked: 09/29/08.