BACKWARD ADAPTIVE WARPED LATTICE FOR WIDEBAND STEREO CODING

Aki Härmä, Unto K. Laine, and Matti Karjalainen
Helsinki University of Technology
Laboratory of Acoustics and Audio Signal Processing
P. O. Box 3000, 02015, Espoo, Finland
e-mail: Aki.Harma@hut.fi

ABSTRACT

In this paper an extremely low delay perceptual audio codec is presented. The codec is based on warped linear prediction which inherently utilizes auditory frequency resolution and frequency masking characteristics of hearing. In the current version of the codec the coding delay is the minimum. This is achieved using backward adaptive lattice methods where waveform modeling is completely based on already transmitted data. Coding technique is applied separately to the two channels but the quantization processes are unified to gain more bit rate reduction.

1 INTRODUCTION

The authors have been working with warped linear prediction (WLP) [1, 2, 3, 4] and warped filters and used them in several applications [5]. The warped techniques are beneficial in many audio applications because they have a clear connection to the characteristics of the ear, i.e., the frequency resolution at all the stages of processing is very close to the frequency resolution of hearing.

Backward adaptive lattice is a popular technique in speech coding, e.g., in [6]. The technique has not been applied to warped predictive coding before due to lack of a warped IIR lattice filter. In [7], an implementation of a warped IIR lattice was found and hence it is now possible to apply several efficient techniques used in conventional linear predictive coding (LPC) to warped linear predictive (WLP) coding.

An advantage of backward adaptive coding is that it allows the coding delay to be arbitrarily low. The experimental codec introduced in this paper works on a *sample-in sample-out* basis, i.e., the theoretical coding delay of the encoder-decoder system is 1-2 sample periods. Coding delay is of importance in all bidirectional communication systems such as teleconferencing and in many conceivable systems based on shared acoustical virtual reality, e.g, an Internet orchestra in a virtual space. Obviously, in the latter case, the buffering needed for the Internet is an important factor in the total transmission delay. However, a low or practically missing coding delay is a better starting point in minimizing the total transmission delay than a coding delay of 32 ms which occurs in most of the current audio codecs, e.g., MPEG-2 AAC.

2 WARPED LINEAR PREDICTION

The method of warped linear prediction in speech coding dates back to a paper by Strube [8]. Later, it was applied to speech coding in [9] and to audio coding by the current authors in [2] and [3].

The technique is based on the use of an allpass filter chain. The phase function of a first order allpass (AP) element determines a frequency mapping [10]. By using a suitable value of the warping parameter, $\lambda = 0.723$ at 44.1 kHz sampling rate [11], in the transfer function of an AP element given by

$$H(z) = \frac{z^{-1} - \lambda}{1 - \lambda z^{-1}} \tag{1}$$

the mapping is very close to the mapping from frequency domain to a frequency scale of hearing [1]. This means that low frequencies are processed with higher accuracy than high frequency components. Autocorrelation method of linear prediction may be applied to the outputs of an AP chain to produce the coefficients of an FIR-type[1] warped filter used in encoder and its inverse filter, i.e, a warped IIR-filter used in decoder. Filters where the unit delays of a conventional filter are replaced by first order allpass filters are called warped filters. The relation between a conventional filter and a warped filter is determined by a bilinear conformal mapping from an unit disc onto another unit disc.

It was shown in [3] that WLP is a potential core for a perceptual audio codec because, due to the auditory frequency resolution of the system, certain aspects of perceptual audio coding occur automatically in a warped linear predictive codec without need of a separate auditory model. In [2] it was shown that a warped linear predictive codec outperforms a corresponding non-warped codec in the way the quantization noise is masked in

[1] Abbreviations FIR, and IIR appear in this paper only to illustrate the structural similarity with conventional filters. A warped filter has always an infinite impulse response

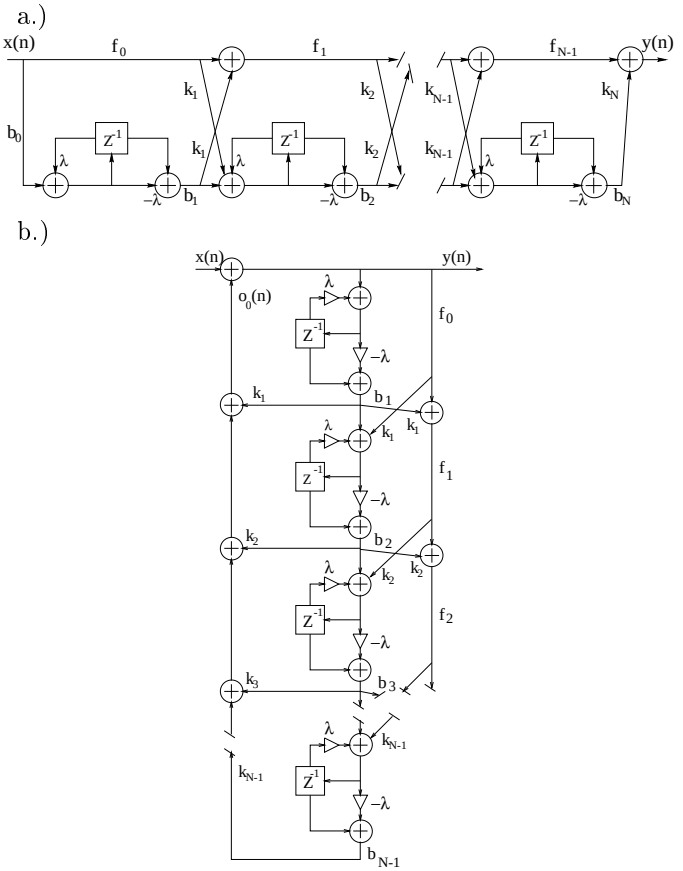

Figure 1: a.) A warped FIR lattice b.) A warped IIR-type lattice

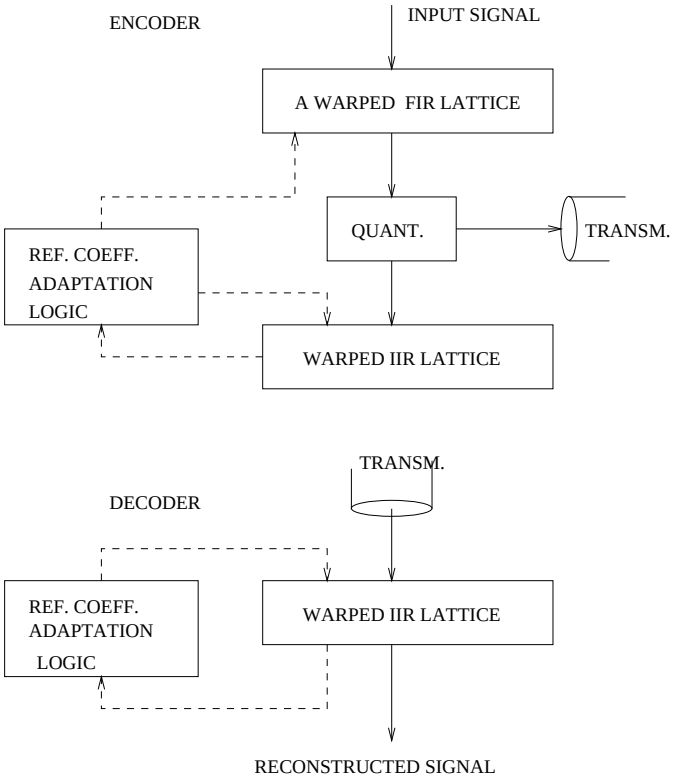

Figure 2: The general block diagram for one channel (right or left) of the codec: encoder, decoder, and the transmission line as a tube. For the second channel a similar structure is used. The quantizers of the two channels are linked together.

frequency. However, the warped technique is advantageous only in wideband coding. For example, if the sampling rate is 8 kHz, the perceptually motivated value for $\lambda \approx 0.4$ and therefore the effect of warping is weaker.

3 WARPED LATTICE FILTERS

As mentioned above a warped filter is obtained by replacing the unit delays of a conventional filter structure by first order allpass filters. However, this leads to problems in implementing recursive filters because an AP element is delay-free and therefore the resulting system would include delay-free recursive loops. A technique to transform a non-realizable IIR-type filter to a new directly realizable structure with a new set of coefficients exists [12].

A warped FIR-type lattice filter is shown in Fig. 1a. The warped IIR-type lattice shown in Fig. 1b is a new tool first introduced in [7]. Even though the filter contains several delay-free recursive paths it may be implemented without any modifications to the structure or the coefficients of the system using a two-step procedure [7].

4 THE STRUCTURE OF THE CODEC

The general structure used in coding the right channel is shown in Fig. 2. A similar structure is used for the left channel. However, the quantization of the two channels is performed in the same process.

The coefficients are updated using an *exponential window* backward adaptive lattice algorithm introduced in [6]. The adaptive lattice algorithm is also known as the *gradient adaptive lattice algorithm* (GAL)[13]. Recently, the algorithm was applied to an adaptive *Laguerre*- lattice filter in [14]. A *Laguerre*-lattice filter is essentially the same as the warped FIR-type lattice filter used in the current paper.

The GAL algorithm is given below:

Initialize: $f_0(n) = b_0(n) = y(n)$

Repeat for $l = 1, 2, \cdots, N$

$$c_l(n) = \gamma c_l(n-1) + f_l(n) b_l(n)$$

$$d_l(n) = \gamma d_l(n-1) + 0.5 * (f_l^2(n) + b_l^2(n))$$

$$k_l(n+1) = c_l(n)/d_l(n)$$

In the algorithm f_l and b_l are the forward and backward prediction error signals and k_l are the reflection

coefficients of the lth stage of the lattice filter (see Fig. 1.) The coefficients k_l are updated at each sample. The adaptation parameter γ determines the length of the exponential window function. For steady state signals a suitable value is $\gamma = 0.998$. In simulations discussed below the number of the lattice stages was fixed to $N = 40$.

To avoid numerical problems in the cases where the amplitude of the residual signal becomes very small the value of $d_l(n)$ must be given a lower limit D_{ll}. Suitable lower limit is $0.001 < D_{ll} < 0.1$.

Since the adaptation of the coefficients is based on the decoded data there is no need to transmit filter coefficients. The same algorithm resides at encoder and decoder and both are driven by the same quantized residual signal.

In the current version of the codec a simple Jayant's one-word memory backward adaptive quantizer [15] is used. The quantizer is a 3-bit quantizer, i.e., there are 8 available step sizes per sample. For a stereo signal this means that there are $8 * 8 = 64$ possible step size combinations per a *stereo sample*. Since some adjacent step size combinations are highly infrequent, the number of combinations may be reduced. This is done so that for a given previously transmitted step size combination there is only a limited set of combinations available for the current sample. The most frequently occurring step size combinations for each possible previous combination were found by running the codec with a wide collection of speech and music material so that all combinations were allowed. In preliminary listening tests it was found that no deterioration of quality occurred if the combinations per a stereo sample were restricted to 32 instead of 64. This gives a total bit rate of 2.5 bits/sample, i.e., for a stereo signal at 44.1 kHz sampling rate this is 220.5 kbits/s. The selection of the step size combination is based on finding the available combination for which the sum of squared quantization errors of the two channels is the smallest.

5 ADAPTATION TO TRANSIENTS

The codec works well with smoothly varying signals but it often fails with sharp transients. This is an obvious result from the fact that coding is completely based on already transmitted data. In some cases, e.g., in test sequence 6 (see Table 1) the beginning of a note loses some of its brightness. With sequences 7 and 8 the codec produces clearly audible artifacts and the reconstructed signal may suddenly turn into a complete disaster. If the dimension of the filter is too high, e.g., $N > 50$, or the adaptation coefficient is too small the filter may enter an unstable state after almost any sudden change in the signal. This usually produces extraneous high level tonal components which may remain in force over several hundred milliseconds after the onset of a transient.

For the speech sequence 2, the output of the codec has some growling noise at low frequencies which is obviously a result from the strong periodic pitch pattern to which the codec can not adapt to. A conventional technique in speech coding is to use long term prediction to model periodic pitch pattern of speech. The same method may also work with some music signals. However, the long term predictor has not been implemented in the codec represented in this paper.

The mistuning of the codec is indicated by the rapid increase in energy of the prediction error signals, i.e., the sum of the values of d_l given by

$$E(n) = \sum_{l=1}^{N} d_l(n). \tag{2}$$

Since $E(n)$ is available at both ends of the codec, it is possible to react to the transients simultaneously in the encoder and the decoder.

One way to enhance the performance of the codec is to make the parameter γ adaptive so that its value depends on the value of $E(n)$. This is done in the current codec so that

$$\gamma = \begin{cases} \gamma_{sw}, & \text{if } E(n) > E_{max} \\ \gamma_{lw}, & \text{otherwise} \end{cases} \tag{3}$$

where γ_{sw} is a small value corresponding to a short window function and γ_{lw} is a larger value used when the prediction error energy is below certain value E_{max}. Typically, $\gamma_{sw} = 0.98$ and $\gamma_{lw} = 0.998$. This technique works with some signals but it does not help in the case of, e.g., the test sequence 7 (*castanets*), where at least some of the percussions are severely deteoriated.

6 SIMULATIONS

The codec was tested with widely used test sequences from the MPEG community. The test sequences are listed in Table 1. An informal *three stimulus hidden reference listening test* was performed where subjects were allowed to switch between the sequences A, B, C at any time and give B or C a grade continuously from 1 to 5 with usual attributes, i.e., 5 is *imperceptible*, 4 is *just perceptible but not annoying*, etc. The test procedure was implemented as a combination of C-code and a MATLAB-based user interface on an Unix workstation with high-quality audio output. The computer program is well in line with recommendations on listening tests but due to the tight schedule it was not possible to perform detailed listening tests in our standard listening room. Instead, the test was performed very quickly in an ordinary office room using headphones and six untrained listeners. The variance in grading for most of the test sequences is small but for some of the signals (average grade in parenthesis) the grade varied among the subjects significantly.

7 DISCUSSION

The codec works well with smoothly varying signals but the performance is unsatisfactory in the cases of sharp

Label	Description	Ave. grade
1	Suzanne Vega	3.8
2	German speech	3.9†
3	Trumpet solo and orchestra	4.5
4	Orchestral piece	4.6
5	Pop	(3.8)
6	Harpsichord	(3.7)
7	Castanets	–
8	Glockenspiel	–
9	Plucked Strings	(3.1)†

Table 1: Test sequences and grading. The sequences 7 and 8 were not tested because of the clearly audible artifacts. For the sequences 2 and 9 a lower order filter $N = 20$ was used in order to avoid instability. Parentheses refer to the cases where the average value is somewhat unreliable due to high intersubjective variance.

onsets and clicks. It is possible to remove the instability problem by using some special techniques in cases where the error energy $E(n)$ exceeds certain high value but usually audible errors in those cases can not be avoided. Therefore it is probably necessary to increase the coding delay by using some *forward* adaptive processing. At 44.1 kHz sampling rate a forward buffer of 88 samples means only a 2 ms growth in coding delay but it could enable the use of various methods which might enhance the performance of the codec significantly. This is going to be tested in a future version of this codec. In addition, long term prediction is going to be integrateed into the system to enhance the performance with periodic signals.

A complex-valued extension to the WLP scheme was introduced in [4]. In that, a stereo signal was converted to a representation where the left and the right channel appeared, respectively, on the positive and negative frequencies in the spectrum of the complex-valued signal. An advantage of this representation is that the spectral estimation may be performed with a single complex valued WLP process so that for a given number of filter taps the estimated all-pole model is optimized in respect to the stereo signal. It is possible to use the same technique with the current codec, too. The complex-valued version of the warped backward adaptive codec may be computationally more efficient and it works in some cases even better than the current codec where the two channels are processed separately. However, the conversion from the two channels of a stereo signal to a single complex valued signal can not be made without an additional delay.

8 ACKNOWLEDGEMENT

This work has been partially supported by the Academy of Finland. The authors are grateful to several workers of the Laboratory Acoustics and Audio Signal Processing for taking part in the listening tests.

References

[1] U. K. Laine and M. Karjalainen, "WLP in speech and audio processing," in *ICASSP*, vol. III, (Adelaide), pp. 349–352, 1994.

[2] A. Härmä, U. K. Laine, and M. Karjalainen, "Warped linear prediction in audio coding," in *Proc. of NORSIG'96*, (Helsinki), 1996.

[3] A. Härmä, U. K. Laine, and M. Karjalainen, "WLPAC – a perceptual audio codec in a nutshell," in *AES 102nd Conv. preprint 4420*, (Munich), 1997.

[4] A. Härmä, U. K. Laine, and M. Karjalainen, "An experimental audio codec based on warped linear prediction of complex valued signals," in *Proc. of ICASSP'97*, vol. 1, (Munich), pp. 323–327, 1997.

[5] M. Karjalainen, A. Härmä, J. Huopaniemi, and U. K. Laine, "Warped filters and their audio applications," in *IEEE Workshop on ASPAA 97*, (Mohonk, New York), 1997.

[6] V. Iyengar and P. Kabal, "A low delay 16 kbits/sec speech coder," in *Proc. of ICASSP'88*, pp. 243–246, 1988.

[7] A. Härmä, "Implementation of recursive filters having delay free loops," in *Proc. of ICASSP'98*, (Seattle), 1998.

[8] H. W. Strube, "Linear prediction on a warped frequency scale," *J. of the Acoust. Soc. Am.*, vol. 68, no. 4, pp. 1071–1076, 1980.

[9] K. Koishida, K. Tokuda, T. Kobayashi, and S. Imai, "Celp coding system based on mel-generalized cepstral analysis," in *Proc. of ICSLP'96*, vol. 1, 1996.

[10] A. V. Oppenheim, D. H. Johnson, and K. Steiglitz, "Computation of spectra with unequal resolution using the fast Fourier transform," *Proc. of IEEE*, vol. 59, pp. 299–301, 1971.

[11] J. O. Smith and J. S. Abel, "The Bark bilinear transform," in *Proc. of IEEE ASSP Workshop*, (Mohonk), 1995.

[12] T. Kobayashi, S. Imai, and Y. Fukuda, "Mel-generalized log spectral approximation filter," *Trans. IECE*, vol. 68, pp. 610–611, 1985.

[13] S. Haykin, *Adaptive Filter Theory*. Prentice Hall, 3 ed., 1996.

[14] Z. Fejzo and H. Lev-Ari, "Adaptive Laguerre-lattice filters," *IEEE Tr. on Signal Processing*, vol. 45, no. 12, pp. 3006–3016, 1997.

[15] N. S. Jayant, "Adaptive quantization with one-word memory," *Bell Syst. Tech. J.*, pp. 1119–1144, 1973.