AN EXPERIMENTAL AUDIO CODEC BASED ON WARPED LINEAR PREDICTION OF COMPLEX VALUED SIGNALS

Aki Härmä, Unto K. Laine, and Matti Karjalainen

Helsinki University of Technology
Laboratory of Acoustics and Audio Signal Processing
Otakaari 5 a, 02150 ESPOO, FINLAND
Aki.Harma@hut.fi

ABSTRACT

Bark-scale warped linear prediction [WLP] is a very potential core for a monophonic perceptual audio codec [2]. In the current paper the WLP scheme is extended for processing complex valued signals (CWLP). Three different methods of converting a stereo signal to one complex valued signal are introduced. The philosophy behind the coding scheme is to integrate some aspects of modern wideband audio coding (e.g. perceptuality and stereo signal processing) into one computational element in order to find a more holistic and economic way of processing.

1. INTRODUCTION

Most of the current wideband audio coding schemes use either transform coding or subband coding. LPC codecs are considered to be inappropriate in coding music and other non-speech audio because LPC is a source model of an excitation and a vocal tract, but an audio signal is typically not produced by a single source. Nevertheless, several audio codecs based on *multipulse LPC* (MPLPC) have been proposed (e.g. [11]). In addition, some hybrid codecs where LPC is combined with subband methods and transform coding do exist.

The codec described in the current article utilizes warped linear prediction where the spectral resolution approximates the frequency representation of the human auditory system [5],[6],[8].

In most of the modern audio subband codecs there is a separate perceptual model controlling bit allocation. The WLP codec is an auditory model and therefore many functions of a separate auditory model are integrated in the core of the codec. For instance, the quantization noise in the reconstructed signal has similar spectral shape to the *masked threshold* of the signal [2]. In addition, the LPC process produces sharper filters to model tonal than non-tonal signal components.

In the current paper, the same codec is used to compress stereo signals. Three different methods of forming the input signal are compared. The emphasis is on analyzing the preservation of spatial information in the reconstructed signal. It is shown that the WLP encoder both *whitens* spectra and *despatializes* the stereo signal. It is sufficient to transmit one signal containing a representation of the temporal fine structure of the original signal and filter coefficients containing both the spectral characteristics and spatial information of the stereo signal.

2. WARPED LINEAR PREDICTION (WLP)

In *forward* linear prediction the estimate $\hat{x}(n)$ for the next value of the sequence $x(n)$ is a linear combination of the previous values. The predictor given as

$$\hat{x}(n) = \sum_{j=1}^{N} \alpha_j^* x(n-j), \tag{1}$$

is obviously equivalent to an FIR-filter. The α-coefficients are the *predictor coefficients* that can be estimated from a signal by using, e.g., the *autocorrelation* method [1].

As it is the case in ordinary real valued LPC-analysis, the coefficients of a complex predictor may be solved from a set of linear equations. The optimal coefficient vector α, in matrix notation, (in MMSE sense) is given by

$$\alpha = \mathbf{R}^{-1}\mathbf{r}, \tag{2}$$

where r is the autocorrelation vector of the sequence and R is the correlation matrix. Since R is *hermitian* or in the case of a real valued sequence *symmetric*, there exist efficient recursive algorithms (e.g. Levinson-Durbin algorithm) to solve the Eq. 1.

For convenience, we consider a complex valued signal $c = x + iy$ as a two-channel signal $[x, y]$. Typically, the two-sided spectrum of such signal is non-symmetric. Let us have a complex valued signal such that it contains the same sinusoidal component on both channels with a small inter-channel delay d. A set of two-sided spectra are represented in Fig. 2.

The signal is *analytical* in positions $d = 2.5$ and 7.5, i.e., the imaginary part is the *Hilbert*-transform of the real part or vice versa. If the phase difference is $d = n * pi$, where $n \in \mathcal{Z}$, the spectrum is symmetric. The frequency response of an estimated filter is non-symmetric as well.

In our case the predicted sample value is not produced from the previous samples as is the case of the conventional LP but from the samples of a *warped signal* [8] [5]. The warped sequence is formed using the outputs of the first order allpass filter chain shown in Fig. 2.

0-8186-7919-0/97 $10.00 © 1997 IEEE

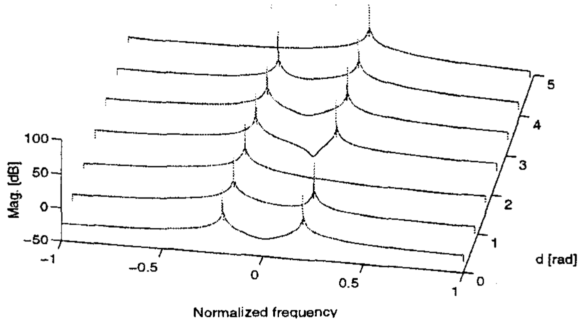

Figure 1: Spectra for a set of complex sinusoidal signals with an interchannel phase difference *d*.

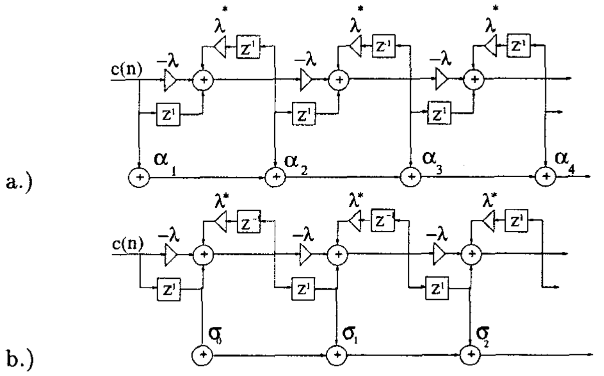

Figure 2: a.) A warped predictor. b.) A modified predictor – suitable for recursive structures.

The value of λ determines the frequency warping of the system (see Fig. 3). If $\lambda = 0$ the chain reduces to a chain of unit delays and thus we have the conventional LPC. If $\lambda = 0.723$ [1] [7] the frequency resolution follows the psychoacoustic *Bark*-scale [9]. If λ is allowed to be complex valued, a variety of different frequency resolutions can be obtained. For example, we may adaptively focus the range of best resolution to the frequency range where it is most needed. However, the act of focusing is, again, one-sided in frequency domain and hence it works optimally only in the case of analytic signals.

3. THE CODEC

The encoder and the decoder based on the warped predictor are illustrated in Fig. 2. Each signal *frame* is processed in two phases. First, the filter parameters are estimated and then the *residual* signal *d* is produced by the inverse filtering process.

Since the predictor in Fig. 2a contains a delayless branch a modified filter structure [4], shown in Fig. 2b is used instead. The σ_i-coefficients of the modified structure may be calculated by employing the following algorithm:

$$S_i = \sum_{k=i}^{N} (-1)^{k-i+1} \lambda^{k-i} \alpha_k \tag{3}$$

$$\begin{aligned} \sigma_1 &= S_1 \\ \sigma_2 &= S_2 + \lambda^* S_1 \\ \sigma_3 &= S_1 + \lambda^* S_2 \\ \sigma_i &= S_i + \lambda^* S_{i-1} \end{aligned} \tag{4}$$

In Eq. 3, coefficients α_i are the coefficients of the original predictor structure in Fig. 2a. The common coefficient g in the modified structure is given by

$$g = 1 + \sum_{k=1}^{N} (-\lambda)^k \alpha_k. \tag{5}$$

3.0.1. Complex valued signals

Three different methods of forming complex valued signals are considered.

i. A stereo-pair $[l, r]$ may be converted to a complex valued signal c simply by $c = r + il = x + iy$. This is called the direct form complex stereo signal.

ii. A *joint stereo* approach results a signal given by $c = r + l + i(r - l) = x + iy$. The components of the signal are orthogonal. The direct form signal is rebuild by $c = (x + y + i(x - y))/2$.

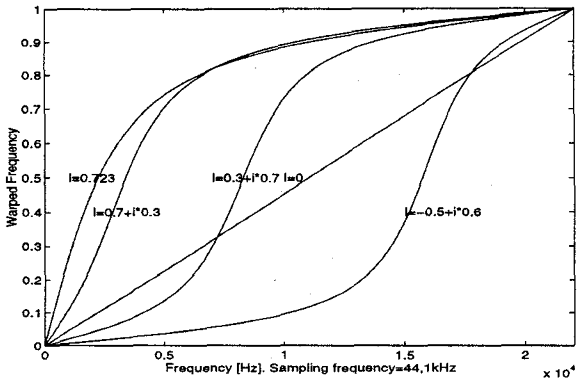

Figure 3: The bilinear frequency mapping from the frequency scale to the warped frequency scale as a function of the coefficient λ. The steepest part of a curve represents the range of highest frequency resolution.

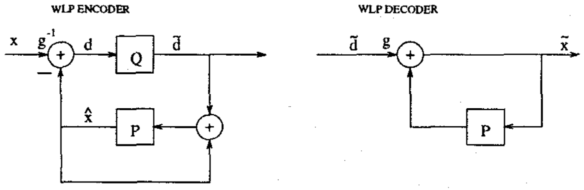

Figure 4: General structures of the encoder and the decoder.

[1] For the sample rate of 44.1 kHz

iii. By using *Hilbert*-transform $\mathcal{H}$, it is possible to constuct an *analytic signal* with one-sided spectrum. In a complex signal given by

$$c = r + i\mathcal{H}[r] + l - i\mathcal{H}[l] = r + l + i(\mathcal{H}[r-l]) = x + iy,$$

the right channel is mapped to the negative frequencies and the left to the positive frequencies. The direct form signal is given by

$$c = (x + \mathcal{H}[y] + i(x - \mathcal{H}[y]))/2.$$

There are basically two different quantization geometries: cartesian and polar quantization. In cartesian quantization $\mathcal{C}$ the real and imaginary parts of a signal are quantized separately so that the reconstructed residual in the decoder is $\hat{c} = \mathcal{Q}[x] + i\mathcal{Q}[y]$. In polar quantization $\mathcal{P}$ the magnitude A and the angle ϕ of a complex number are quantized so that $\hat{c} = \mathcal{Q}[A]e^{-i\mathcal{Q}[\phi]}$.

In the current paper, the quantization grid is typically linear and $\mathcal{C}$ is used. In most cases, there is no significant difference between the coding results in $\mathcal{C}$ and $\mathcal{P}$.

To summarize, the input signal is a complex signal produced by one of the methods. The estimation process gives complex valued inverse filter coefficients α_i. The σ_i coefficients are calculated and a complex valued *residual* signal is obtained by applying the filter to the original signal. The residual and the coefficients are quantized and transmitted to the receiver where the signal is rebuild. Coefficient quantization is not discussed in the current paper.

4. NUMERICAL SIMULATIONS

Due to quantization, the channels leak in the case of method i (Fig. 5) and also in using method ii. This results from the fact that, in both methods, in the spectra of the complex valued signals the channels are mixed so that components belonging to originally separate channels may be very close to each other.

However, in iii the channels are always separated (Fig. 5.) For instance, a component that appears only in the left channel is only in the righthand side of the two-sided spectrum of that complex signal.

Leakage between the channels may change the spatial impression of the stereo signal. Therefore, the method iii may be the best candidate.

In a stereo signal the noise is masked if it is masked in frequency, as in the case of monophonic signals, and also spatially. Lowest *Binaural Masking Level Difference* [BMLD] is achieved if *Interaural Time Difference* [ITD] and *Interaural Level Difference* [ILD] of the signal and the quantization noise N are equal [10], i.e., the direction of noise is the same as the direction of the signal. In the current paper, ILD is the difference between total energies given by $ILD = 20 * lg(\sum_k l_k^2 / \sum_k r_k^2)$. $ILD(f)$ is a level difference as a function of frequency. In Fig. 6 we have ITD-histograms (interchannel cross-correlation functions) for the noise produced in the codec with methods i, ii and iii, respectively. In N_i and N_{ii} there is a strong correlation corresponding to the same $ITD = 0.23\ ms$ as the signal, but in N_{ii}, due to leakage, there is also an annoying correlation at $ITD_{ii} = -0.23ms$. In N_{iii}, as one could expect, the noise channels are not correlated and therefore, in terms of ITD, the noise is uniformly spread in the space.

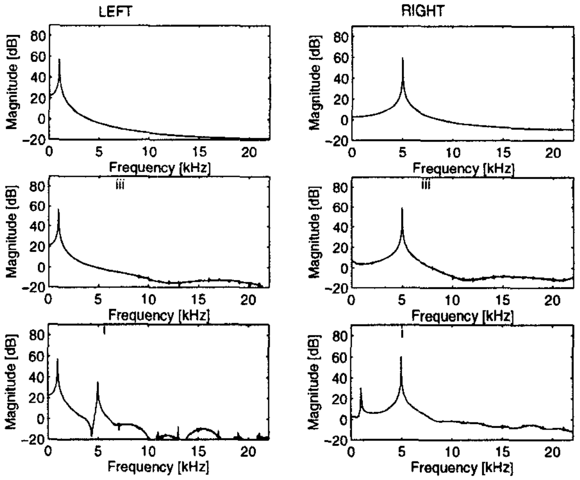

Figure 5: Top: A stereo signal containing two tones. Middle: reconstructed signal (method iii and 2-bit $\mathcal{C}$) Bottom:(method i and 2-bit $\mathcal{C}$)

Preliminary listening tests show that if the only difference between the two channels is ITD, the method i might be the best choice. However, in binaural listening ITD works as a cue only at low frequencies. Experiments with ILD show (Fig. 7) that ILDs of the original signal (symbol '+') are transmitted to the quantization noise in iii, but in i the level differencies vary randomly. The ii appears to be the worst method.

The whitening process of LPC has an especially interesting feature in the case of the method iii. In estimating the parameters of a two-sided spectrum it, by the same price, estimates the $ILD(f)$ of the stereo signal. Therefore, the inverse filter both whitens the spectra and *despatializes* the stereo signal. As a result, if the estimation process were successful, the residual is a white noise signal with two almost identical channels $[l, r]$. Hence, a significant amount of bit-rate reduction is achieved by transmitting only one, i.e., real valued, residual channel. As a first guess, the mean of the channels was used, i.e, $s = (x + y)/2$. For some typical music material the quality of the reproduced signal is almost as good as in the case of two residual streams. However, for some material the apparent direction of source is changed, e.g., a sound source in the left hemisphere may move slightly towards the center.

5. DISCUSSION

Complex valued WLP appears to be a potential starting point in developing a new stereo audio codec. However, the current version of the codec is structurally a simple *schoolbook* example of a LPC. The parameters are estimated in a fixed rectangular window without overlapping and there is no interpolation between the coefficients of adjacent frames. Thus, there are disturbing transients at frame boundaries. Quantization method is direct linear quantization with frame-based scalefactors. It is quite prob-

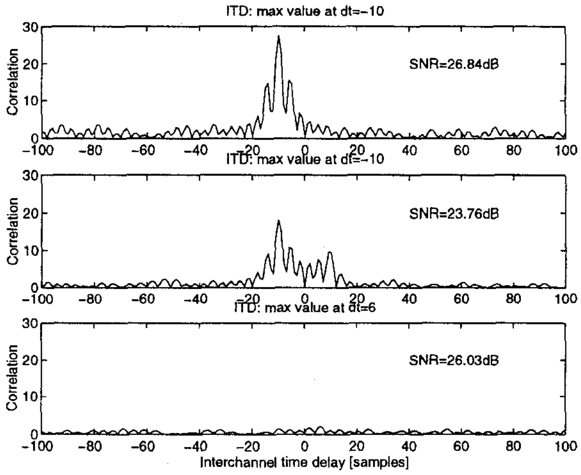

Figure 6: Interchannel cross-correlation functions for quantization noise in the case of a narrow band of noise with an ITD of 10 samples (0.23 ms if $f_s = 44.1kHz$). Top:i Middle:ii Bottom:iii

able, that an adaptive quantization or a form of vector quantization might be a better choice. The usage of proper error feedback techniques would enhance the performance of codec. There are several efficient variants of linear prediction (e.g. CELP or Multi-Pulse LP (MPLPC)) that could be adopted to the WLP scheme. Recently, MPLPC has been applied to audio coding with promising results [11],[12].

As shown above, the estimation process extracts spectral information of the both channels and the *ILD(f)* of the channels. What is left in the two-channel residual is a representation of the temporal fine structure of the signals and the *ITD*. *ITD* is a set of time shifts between the residual signals. It is a feasible task to model *ITD* of the residuals, because the residual signal is almost white noise.

The bit-rate of the current version of the codec for 44.1 kHz stereo material is approximately 128 $kbits/s$, but the quality of the reproduced signal is in most cases worse than CD-quality. Fortunately, there are many known methods to enhance the performance of the codec and reduce the gross bit-rate.

The frequency representation in the codec, due to the characteristics of the 1st order allpass filters, follows very closely the psychoacoustic Bark-scale. Unfortunately, the critical band rate scale (Bark-scale) is based on incorrect assumptions about the characteristics of auditory perception [13]. Hence, the resolution in the current codec is insufficient at low frequencies. The future work will involve the usage of ERB-resolution warping in the coding process.

6. ACKNOWLEGEMENT

This work has been partially supported by the Foundation for Finnish Inventions (Keksintösäätiö) and the Academy of Finland. The authors are grateful to Jyri Huopaniemi for providing data and Matlab-tools for HRTF (Head Related Transfer Function) -processing.

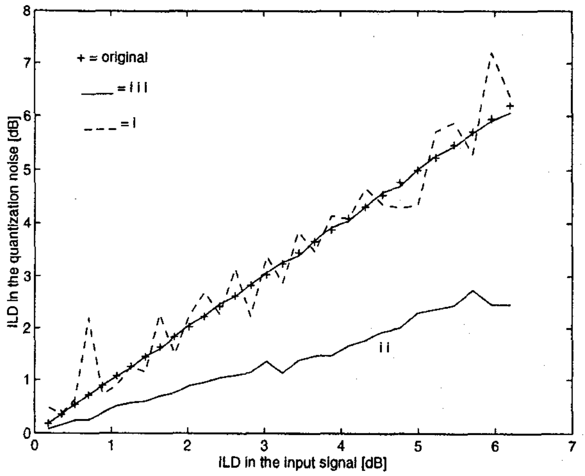

Figure 7: ILD of the original signal ('+'). Solid curves represent the the ILD's of pure quantization noise in methods i, ii and iii.

7. REFERENCES

[1] Haykin S., Adaptive Filter Theory, Prentice-Hall inc., New Jersey, 1996.

[2] Härmä A., Laine U. K., Karjalainen M., Warped Linear Prediction (WLP) in Audio Coding, Proc. NORSIG-96, Espoo, Finland, 1996.

[3] Jayant N. S., Noll P., Digital coding of waveforms, Prentice-Hall inc., New Jersey, 1984.

[4] Karjalainen M., Härmä A. Laine U., Realizable Warped IIR Filter Structures, Proc. NORSIG-96, Helsinki, 1996.

[5] Laine U. K., Karjalainen M., Altosaar T., "WLP in speech and audio processing", Proc. of ICASSP-94, Adelaide, South Australia, III pp.349-352, 1994.

[6] Laine U. K., Generalized linear prediction based on analytic signals, Proc. of ICASSP-95, Detroit, II, pp. 1701-1704, 1995.

[7] Smith J. O., Abel J. S., The Bark Bilinear Transform, Proc. of IEEE ASSP Workshop, Mohonk, New Paltz, 1995.

[8] Strube H. W., Linear prediction on a warped frequency scale, JASA, 68, 4, pp.1071-1076, 1980.

[9] Zwicker E., Fastl H., Psychoacoustics, Springer-Verlag, Berlin, Germany, 1990.

[10] Blauert J., *Spatial Hearing*, The MIT Press, Massachusetts, p. 494, 1997.

[11] Singhal S., High Quality Audio Coding Using Multipulse LPC, Proc. Of ICASSP'90, 2, pp. 1101-1104, 1990.

[12] Chang W., Wang C., A masking-Threshold-Adapted Weighting Filter for Excitation Search, IEEE Trans. on Speech and Audio Processing, Vol 4, 2, 1996.

[13] Sek A., Moore B.C.J., The critical modulation frequency and its relationship to auditory filtering at low frequencies, *JASA* 95, 5, pp.2606-2615, 1994.