A NEW AUDIO CODER USING A WARPED LINEAR PREDICTION MODEL AND THE WAVELET TRANSFORM

Daryl Ning and Mohamed Deriche

Signal Processing Research Centre
School of Electrical and Electronic Systems Engineering
Queensland University of Technology
GPO Box 2434 Brisbane Qld 4001

ABSTRACT

In this paper, we present results for a wavelet transform (WT) excited warped linear prediction (WLP) audio coder. In contrast to conventional LP, WLP allows for the control of frequency resolution to closely match the response of the human auditory system. The structure of the system is similar to the transform–coded excitation techniques used in wideband speech coding, where LP has been replaced with WLP. Quantisation of the wavelet coefficients is aided by a psychoacoustic model to minimise the perceptually significant noise due to quantisation error. For monophonic signals sampled at 44.1 kHz, the coder achieves near transparent quality for a variety of speech and music signals at an average bit-rate of 64 kb/s. When compared to MPEG layer III at the same bit-rate, the coder delivers superior quality. The power of the proposed coder resides in its easy scalability to lower bitrates.

1. INTRODUCTION

CD quality audio is sampled at 44.1 kHz and encoded with 16 bits/sample PCM, resulting in a large bitrate of 705 kb/s per channel. Such a high bitrate makes storage or transmission of the raw signal very expensive. Because of this, audio coding schemes, (such as the ISO/IEC MPEG [1]), have been designed to reduce the overall bitrate. Although being lossy, these compression schemes, are able to achieve near transparent signal quality at approximately 64 kb/s per channel. Such techniques are generally based upon transform coding, which is well known for its ability to encode audio at high bit-rates. At bit-rates below 16 kb/s, however, quality begins to suffer, especially for speech signals. In general, coders developed to operate at these lower bit-rates are designed primarily with speech in mind. Such coders, eg. ITU-T G.728 and G.729, are based on linear predictive (LP) techniques. These coders are able to obtain excellent speech quality with relatively few bits by modelling the vocal tract using an all pole filter. Unfortunately, these coders don't perform quite so well for high quality audio. Currently there are no universal coders able to provide good quality for speech at low bit-rates, as well as provide high quality reproduction of CD quality audio at higher bit-rates. If an LP based coder can give good results for high quality audio, it can potentially be scaled to encode both speech and audio with good quality over a range of bit-rates.

In this paper, we propose a hybrid audio coder based on warped linear predictive (WLP) techniques. We call the coder *hybrid*, because we use both transform coding and LP techniques. After the initial WLP on the analysis frame, the excitation is transformed into the wavelet domain and coded using perceptual criteria. The coder is currently configured for high quality audio, however it can potentially be modified for lower bit-rates by using CELP coding techniques. In a previous paper [2], a similarly structured coder was presented giving near transparent quality at bit-rates around 90 kb/s. This coder, however, used conventional LP analysis. Here we have chosen to use the WLP technique proposed by Harma [3] to achieve a lower bit-rate.

The major purpose of this on-going research was to determine whether the LP filtering technique could be used to transparently encode high quality audio at bit-rates around 64 kb/s and lower. Among several previous approaches, using both conventional and warped methods, none (to our knowledge) have achieved this goal with 44.1 kHz sampled input signals. Harma and Laine [4] developed a wideband audio coder using WLP based on the G.728 CELP coder. Preliminary tests indicated near transparent audio quality at 84 kb/s. The focus, however, was on low delay coding and not on bit rate reduction. In [5], the excitation was encoded using subband techniques adopted from MPEG Layer II. Quality at 56 kb/s was quoted to be comparable to that of MPEG Layer II at 64 kb/s, however, MPEG Layer II at 64 kb/s is far from being transparent.

In section 2 we discuss the structure of the coder, focusing on the WLP analysis, the DWT, and bit allocation. Section 3 provides information about quantisation, while section 4 presents the results of informal listening tests as

0-7803-7402-9/02/$17.00 ©2002 IEEE

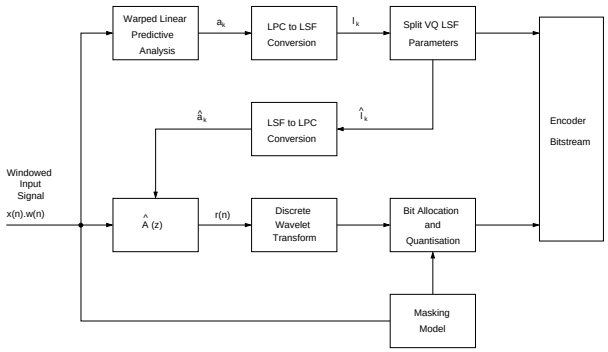

Fig. 1. Proposed Audio Encoder

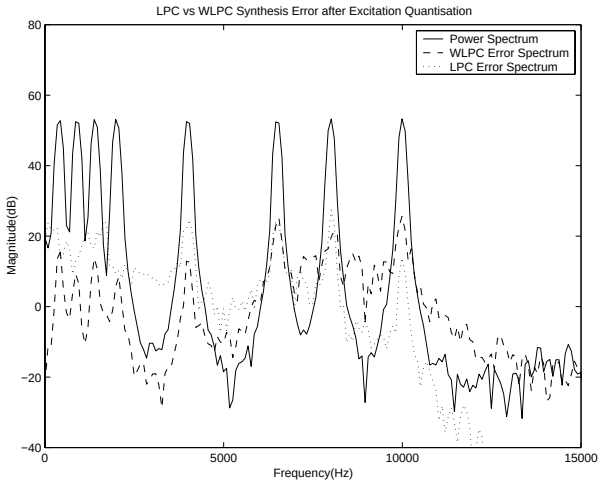

Fig. 2. Comparison of synthesis error using LPC and WLPC

well as a discussion.

2. PROPOSED ENCODER

The block diagram of the proposed encoder is illustrated in figure 1. The audio signal is processed using frame sizes of 1024 samples or 23 ms for 44.1 kHz sampled audio signals. Each frame, $x(n)$, overlaps the previous frame by 32 samples, and is multiplied by a window, $w(n)$, before processing. $w(n)$ is a rectangular window with a raised sine window roll-off in the overlap regions. Such a window is required to reduce blocking artifacts at frame boundaries which can cause clicking in the reconstructed signal.

2.1. Warped Linear Predictive Analysis

In conventional LPC, a sample of a signal, $x(n)$, is estimated from a weighted sum of previous samples. This leads to an LPC error filter given by $A_c(z) = 1 - \sum_{k=1}^{P} a_k z^{-k}$, where a_k are the LPC parameters or filter coefficients. In WLPC, the unit delays are replaced by first order all pass filters of the form:

$$D(z) = \frac{z^{-1} - \lambda}{1 - \lambda z^{-1}}. \quad (1)$$

The corresponding WLPC error filter becomes

$$A(z) = 1 - \sum_{k=1}^{P} a_k D(z)^k, \quad (2)$$

where the a_k's are now the WLPC parameters. In contrast to conventional LPC, WLPC allows for the control of frequency resolution to closely match the response of the human auditory system [3]. Positive values of λ result in a longer group delay for low frequencies and shorter group delay for high frequencies. Consequently, a better resolution is obtained at lower frequencies at the expense of poorer resolution at high frequencies. This tradeoff is beneficial in audio coding due to the similar nature of the human auditory system.

To give a simple example, the WLPC and LPC parameters of a synthetic signal were estimated. The signal consisted of 8 single tones. The residual (prediction error) was then quantised and passed through the corresponding synthesis filters to reconstruct the original signal. Figure 2 plots the spectrum of the original signal, the error from LPC synthesis, and the error from WLPC synthesis. It is clear that at low frequencies, the higher resolution of WLPC shapes the error spectrum more closely under the signal power spectrum than LPC. Similarly at higher frequencies, the coarser resolution of WLPC results in increased spreading compared to LPC.

At the encoder, a 16th order WLP analysis is performed on the windowed input frame. From experiments conducted by Harma [3], an order of 40 would be ideal, however 16 was chosen as a tradeoff between quality and bit-rate. The estimated parameters are converted to line spectral frequencies (LSF) for more efficient quantisation. After vector quantising the LSF parameters, the reverse conversion is performed to obtain the approximated WLPC parameters. Using (2) we obtain the approximated WLPC error filter, $\hat{A}(z)$.

The windowed input signal is filtered with $\hat{A}(z)$ to produce the residual signal, $r(n)$. This residual signal becomes the excitation signal to the AR filter, $1/\hat{A}(z)$, at the decoder.

2.2. Discrete Wavelet Transform of the Excitation Signal

In order to encode the perceptually significant portion of the excitation signal, we first perform a subband decomposition using a DWT. The analysis uses a cascade of M-band wavelet transforms to divide the spectrum into 14 non-

uniform bands. The subbands are designed to closely resemble the critical band divisions of the human auditory system. Although the critical bandwidths are as small as 100 Hz at low frequencies, the minimum bandwidth is maintained at 345 Hz since any further decomposition results in a greater number of significant sidelobes in the overall magnitude frequency response of the cascaded filterbank.

2.3. DWT Bit Allocation

Before the bit allocation proceeds, a masking threshold is calculated using a technique based closely upon the MPEG psychoacoustic model II described in [1]. Having computed the masking threshold (as a function of frequency) for the input frame, the pool of bits reserved for the DWT coefficients must be judiciously allocated amongst the subbands to minimise the perceived distortion. In this scheme, bits are allocated to an entire group of wavelet coefficients within a subband, hence any quantisation error will have the effect of introducing bandpass filtered noise. We can therefore attempt to shape the overall noise spectrum to lie beneath the masking threshold.

The error between the original signal and the reconstruction is due to the quantisation of the residual signal. Since this is filtered through the AR filter, $1/\hat{A}(z)$, the Fourier transform of the error, $E(e^{jw})$, can be written as

$$E_{all}(e^{j\omega}) = \frac{\hat{R}(e^{j\omega}) - R(e^{j\omega})}{\hat{A}(e^{j\omega})}. \tag{3}$$

The numerator is simply the filterbank error. Since near perfect reconstruction filters are being used, the filterbank error is almost entirely due to the quantisation error of the wavelet coefficients. We assume the noise within the i^{th} subband is white with energy, q_i, given by

$$q_i = \sum |\mathbf{w}_i - \hat{\mathbf{w}}_i|^2, \tag{4}$$

where $\mathbf{w}_i$ and $\hat{\mathbf{w}}_i$ are the wavelet coefficients and quantised wavelet coefficients in subband i respectively. The estimated spectrum of the filterbank error, $\hat{S}_{FB}(e^{j\omega})$, can therefore be written as

$$\hat{S}_{FB}(e^{j\omega}) = \sum_{i=1}^{M} q_i |F_i(\omega)|^2, \tag{5}$$

where M is the number of subbands, and $F_i(\omega)$ is the frequency response of the i^{th} subband filter. The overall noise spectrum, $S_{all}(e^{j\omega})$, can therefore be re-written as

$$\hat{S}_{all}(e^{j\omega}) = \frac{\sum_{i=1}^{M} q_i |F_i(\omega)|^2}{|\hat{A}(e^{j\omega})|^2}. \tag{6}$$

Since we have previously calculated masking thresholds using the masking model, we can now calculate the Noise-to-Masking Ratio (NMR) within each subband to aid us in allocating bits. The procedure is as follows:

1. Calculate $\hat{A}(e^{j\omega})$, and $F_i(\omega) \quad \forall i$
2. Calculate $\hat{S}_{all}(e^{j\omega})$ using (6), (4), and calculations from step 1.
3. Sum the noise within each subband frequency range and calculate a NMR for each subband.
4. For each subband i, repeat steps 2 and 3 assuming one extra bit were added to subband i.
5. Find the subband which gives the greatest increase in its own NMR (due to the extra bit) and allocate one bit to this subband. Calculate bits left for allocation.
6. Repeat steps 3 - 5 until no more bits are available for allocation.

Note that bits are only allocated to unmasked subbands, i.e. those bands with negative NMR's. (The NMR is defined as Signal-to-Noise Ratio – Signal-to-Masking Ratio).

3. QUANTISATION

The LSF parameters are encoded using a split vector quantisation scheme, similar to that proposed by Paliwal and Atal [6]. Each 16 dimensional LSF vector is split into 5 subvectors and coded using a total of 50 bits. The bit per sample ratios are higher for the lower LSF's since the human ear can more accurately resolve lower frequencies.

The wavelet coefficients are first grouped into their respective subbands. For each subband, we divide by a scalefactor to normalise the coefficient values between -1 and 1. Each scalefactor can be represented using 4 bits. The normalised coefficients within each subband are non-uniformly quantised using the number of bits determined by the bit allocation procedure. The bit allocation information is sent as side information. Both the scalefactors and the normalised wavelet coefficients are Huffman coded to reduce the overall bitrate.

4. RESULTS AND DISCUSSION

To evaluate the proposed codec, double blind tests were used. For each test source, a pair of signals was presented to the listener. The listener was told that either one, both, or none of the two samples may be the result of compression. The listener was able to listen to the two samples as many times as he/she wished. The listener was then asked to decide which of the two samples had the better overall quality. A "not sure" answer was allowed. All test material were taken from the European Broadcasting Union (EBU) Sound Quality Assessment Material (SQAM) CD. Each track is sampled at 44.1 kHz and varies in length between 6 and 17 seconds. The six tracks used comprised of a female voice,

Audio Signal	Likelihood of Listener Preferring Original Signal Over WLPC Encoded Signal	Comments
Castanets	0.56	Near Transparent
Pop Music	0.67	Original Preferred
Female Speech	0.49	Transparent
Acoustic Guitar	0.61	Original Preferred
Male Speech	0.55	Near Transparent
Piano	0.57	Near Transparent

Table 1. Listening Test Results: Original vs WLPC Encoder.

Audio Signal	Likelihood of Listener Preferring MP3 Signal Over WLPC Encoded Signal	Preferred Coder
Castanets	0.46	WLPC
Pop Music	0.39	WLPC
Female Speech	0.46	WLPC
Acoustic Guitar	0.44	WLPC
Male Speech	0.48	Similar
Piano	0.48	Similar

Table 2. Listening Test Results: MPEG Layer 3 vs WLPC Encoder.

male voice, piano, pop music, acoustic guitar, and castanets. 18 listeners were used to evaluate the performance of the proposed WLPC codec compared to the original source, and the MPEG layer III codec. These results are summarised in tables 1 and 2. Both the WLPC and MPEG codecs operated at 64 kb/s.

When compared to the original source, 4 out of the 6 WLPC encoded signals gave transparent or near transparent quality. The pop music piece and acoustic guitar did not. Given that a number of the listeners actually played the guitar, and were therefore accustomed to its sound, the result for this signal did not come as a surprise. The pop music piece was found to sound slightly muffled by some listeners, which may be due to an over generous psychoacoustic model. Although the higher frequencies were calculated to be masked, they did actually contribute to the sound for some of the listeners. When compared to the MPEG layer III codec, the WLPC codec gave superior or similar quality for all signals.

A further reduction in bit-rate (say 10%) could be expected with optimisation of the present scheme. Current work, however, is targeted at modifying the proposed coder to be scalable. Bit-rates are expected to range between 8 and 80 kbps. At low rates, a CELP–like coding structure will be used as a core. At higher rates (above 16 kbps), techniques from the current scheme will be incorporated for higher quality coding of audio.

5. CONCLUSION

In this paper, we have presented results on high quality audio coding using a hybrid wavelet-WLPC scheme. Previously (to our knowledge), no LPC based scheme has provided transparent coding of CD quality audio at bit-rates around 64 kb/s. Informal listening tests indicate that this structure is capable of delivering near transparent quality for a range of audio signals at 64 kb/s. Given these results, we are currently modifying the existing scheme in an attempt to develop a scalable coder capable of delivering excellent quality for audio signals, as well as delivering good speech quality at low bit-rates ($<$16 kb/s).

6. REFERENCES

[1] ISO/IEC, *International Standard ISO/IEC 11172-3. Information Technology – Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s – Part 3: Audio*, 1992.

[2] Simon Boland and Mohamed Deriche, "Hybrid LPC and discrete wavelet transform audio coding with a novel bit allocation algorithm," *ICASSP*, vol. 6, pp. 3657–60, 1998.

[3] Aki Harma, Matti Karjalainen, Lauri Savioja, Vesa Valimaki, Unto Laine, and Jyri Huopaniemi, "Frequency warped signal processing for audio applications," *Journal of the Audio Engineering Society*, vol. 48, no. 11, pp. 1011–1031, Nov. 2000.

[4] Aki Harma and Unto Laine, "Warped low delay CELP for wideband audio coding," *AES 17th International Conference on High Quality Audio Coding*, pp. 207–215, 1999.

[5] Yu Rongshan and Ko Chi Chung, "High qualty audio coding using a novel hybrid WLP-subband coding algorithm," *5th International Symposium on Signal Processing and its Applications*, vol. 1, no. 483–486, 1999.

[6] Kuldip Paliwal and Bishnu Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame," *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 1, pp. 3–14, Jan. 1993.