Presented at the IEEE International Symposium on Circuits and Systems – Monterey, CA, June 1998.

ENHANCING THE PERFORMANCE OF SUBBAND AUDIO CODERS FOR SPEECH SIGNALS

Henrique Malvar

Microsoft Research
One Microsoft Way
Redmond, Washington 98052, USA

ABSTRACT

Transform or subband audio coders can deliver high quality reconstruction at rates around two bits per sample. Most quantization strategies take into account masking properties of the human ear to make the quantization noise less noticeable. In this paper we describe a new coder in which we extend such quantization strategies by incorporating run-length and arithmetic encoders. They lead to improved performance for quasi-periodic signals, including speech. The quantization tables are computed from only a few parameters, allowing for a high degree of adaptability without increasing quantization table storage. To improve the performance for transient signals, the coder uses a nonuniform modulated lapped biorthogonal transform with variable resolution without input window switching. Experimental results show that the coder can be used for good quality signal reproduction at rates close to one bit per sample, and quasi-transparent reproduction at two bits per sample.

1. INTRODUCTION

Transform or subband coders are employed in many modern audio coding standards [1], usually at bit rates of 32 kbps and above, and at 2 bits/sample or more. At low rates, around and below 1 bit/sample, speech codecs such as G.729 and G.723.1 are used in teleconferencing applications. Such codecs rely on explicit speech production models, and so their performance degrades rapidly with other signals such as multiple speakers, noisy environments and specially music signals [2].

With modem speeds having increased by almost a factor of two over the last few years, many applications may afford as much as 8–12 kbps for narrowband (3.4 kHz bandwidth) audio, and maybe higher rates for higher fidelity material. That raises and interest on coders that are more robust to signal variations, at rates similar to or a bit higher than G.729, for example.

In this paper we present a transform coder that can operate at rates down to 1 bit/sample (e.g. 8 kbps at 8kHz sampling) with reasonable quality. To improve the performance under clean speech conditions, we use a run-length and arithmetic encoder, which improves the encoding of the periodic spectral structure of voiced speech.

2. CODER STRUCTURE

A simplified block diagram of the proposed encoder is shown in Figure 1. Overlapping blocks of the input signal $x(n)$ are transformed into the frequency domain via a nonuniform modulated lapped biorthogonal transform (NMLBT) [1]. The NMLBT is essentially a modulated lapped transform (MLT) [4] with different analysis and synthesis windows, in which the high-frequency subbands are combined for better time resolution. Depending on the signal spectrum, the combination of high-frequency subbands may be switched on or off, and a one-bit flag is sent as side information to the decoder. The NMLBT analysis window is not modified, as discussed in Section 5.

The transform coefficients $X(k)$ are quantized by uniform quantizers, as shown in Figure 1. Uniform quantizers are very close to being optimal, in a rate-distortion sense, if their outputs are entropy coded. Vector quantization (VQ) could be employed, as proposed in [5], but the gains in performance are minor, compared to our adaptive run-length & arithmetic encoder. Although the TwinVQ of [5] has a reduced complexity, it is still significantly more complex than scalar quantization.

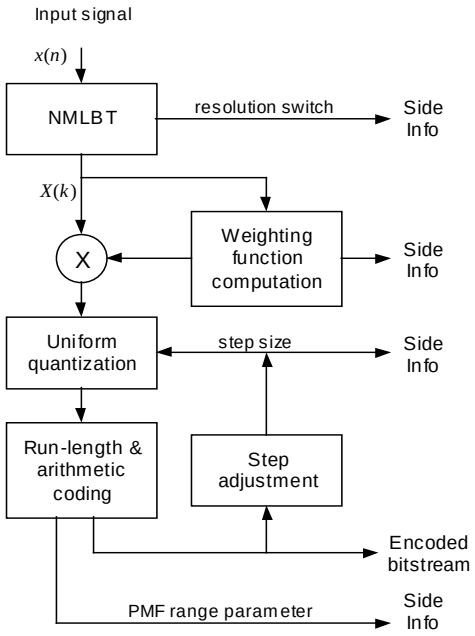

Figure 1. Simplified block diagram of the proposed speech and audio encoder.

An optimal rate allocation rule for minimum distortion at any given bit rate would assign the same step size for the subband/transform coefficients, generating white quantization noise. This leads to a maximum signal-to-noise ratio (SNR), but not the best perceptual quality [6]. The "weighting function computation" block in Figure 1 replaces $X(k)$ by $X(k)/w(k)$, prior to quantization, for $k = 0, 1, \ldots, M–1$, where M is the number of subbands, usually a power of two between 256 and 1024. At the decoder we weigh the reconstructed transform coefficients

by $\hat{X}(k) \leftarrow \hat{X}(k)w(k)$. Thus, the quantization noise will follow the spectrum defined by the weighting function $w(k)$. In the next section we address the computation of $w(k)$. The quantized transform coefficients are entropy encoded by run-length and arithmetic coders, as described in Section 4.

The operation of the decoder can be easily inferred from Figure 1. Besides the encoded bits corresponding to the quantized transform coefficients, the decoder needs the side information shown in Figure 1, so it can determine the entropy decoding tables, the quantization step size, the weighting function $w(k)$, and the single/multi-resolution flag for the inverse NMLBT.

3. SPECTRAL WEIGHTING

Ideally, the weighting function $w(k)$ should follow the auditory masking threshold curve for a given input spectrum $\{X(k)\}$. The masking threshold can be computed in a Bark scale (a quasi-logarithmic scale that approximates the critical bands of the human ear) as described in [8],[9]. At high coding rates, e.g. 3 bits per sample, the resulting quantization noise can be below the quantization threshold for all Bark subbands, resulting in a perceptually transparent reconstruction, i.e., the decoded signal is indistinguishable from the original.

At lower rates, e.g. 1 bit/sample, it is not possible to hide all quantization noise under the masking thresholds. In that case, we may not want to raise the quantization noise above the masking threshold by the same dB amount in all subbands, since low-frequency unmasked noise is usually more objectionable. Therefore, assuming $w_{MT}(k)$ is the weighting determined from the masking thresholds, our coder uses the weights

$$w(k) = \left[w_{MT}(k)\right]^{\alpha} \tag{1}$$

where α is a parameter that can be varied from 0.5 at low rates to 1 at high rates. This is similar to the noise spectral coloring used for determining the excitation in CELP coders, except that they use a fractional power of the input spectrum, whereas we use a fractional power of the masking thresholds.

The amount of side information for representing the $w(k)$'s depends on the sampling frequency, f_S. For f_S = 8 kHz, we need 17 Bark spectrum values, and for f_S = 44.1 kHz we need 25. Assuming an inter-band spreading into higher subbands of –10 dB per Bark [9] and differential encoding with 2.5 dB precision, we need about 2 bits per Bark coefficient.

4. RUN-LENGTH & ARITHMETIC CODING

Transform coders usually perform better with complex signals such as music, because of the higher masking levels associated with such signals. With clean speech, transform coders operating at low bit rates may not be able to reproduce the fine harmonic structure. With voiced speech and at rates around 1 bit/sample, the quantization step size is large enough that most transform coefficients are quantized to zero, except for the harmonics of the fundamental vocal tract frequency. Therefore, we can achieve better rates than those predicted by first-order entropy by simply using run-length coding.

Calling $q(k)$ the quantization index corresponding to each $X(k)$, i.e. $\hat{X}(k) = \delta q(k)$, where δ is the quantization step size, our entropy encoder uses the following alphabet:

Quantized value $q(k)$	Symbol
$-A, -A+1, \ldots, A$	$0,1, \ldots, 2A$
Run of $R_{\min}$ zeros	$2A+1$
Run of $R_{\min}+1$ zeros	$2A+2$
:	:
Run of $R_{\max}$ zeros	$2A+1+R_{\max}-R_{\min}$

Table 1. Codewords for the entropy encoder.

A is the maximum quantized index, which varies for each block. The symbols are encoded via an arithmetic encoder, which uses the following parametric probability distribution:

$$\Pr(s=m) = \begin{cases} \beta_1\left[\exp\left(-d_L\left(|m-A|^{0.9}-1\right)\right)+0.01\right], & m \le 2A, m \ne A \\ 0.25, \quad m = A \quad (\text{or } q=0) \\ \beta_2, \quad 2A+2 \le m < 2A+4 \\ \beta_2\left[\exp\left(-d_R\left(|m-2A-4|^{0.8}-1\right)\right)+0.01\right], & m \ge 2A+4 \end{cases} \tag{2}$$

where A, d_L and d_R are parameters, and β_1 and β_2 are computed such that the probabilities in (2) add to one.

The probability distribution function (PDF) above is close to a two-sided exponential (Laplacian) for the quantized values part, except that $\Pr(s = A)$ (i.e. $q = 0$) is reduced to 0.25, and close to exponential for the run lengths, but with a plateau on the first three points. The decaying parameters are usually set to d_L = 0.4 and d_R = 0.3, and the run-length range parameters are set to $R_{\min}$ = 4 and $R_{\max} = M/4$. We found that the model in (2) leads to bit rates within about 5% of the optimal ones (corresponding to the measured PDFs) for a large variety of speech and audio signals.

For a given quantized transform block $X(k)$, the parameter A is computed as $A = \max\{q(k)\}$, and that parameter is sent to the decoder as side information (A can be coarsely quantized, e.g. 32 values in a nonlinear scale). The decoder can then compute the PDF to be used in the arithmetic decoder, since all other parameters are fixed. When a constant bit rate is desired, the block "step adjustment" in Figure 1 adjusts the step size iteratively until the bit budget for the block is attained.

5. SWITCHING TIME RESOLUTIONS WITHOUT SWITCHING WINDOWS

The run-length and arithmetic encoder of the previous section helps to alleviate one of the problems of transform/subband coders: the reproduction of voiced speech. Another well-known problem with transform coders is that the number of subbands M has to be large enough to provide adequate frequency resolution, which usually leads to block sizes in the 30–60 ms range. That leads to a poor response to transient signals, with noise

patterns that last the entire block, including the so-called pre-echo [7].

During such transient signals a fine frequency resolution is not needed, and therefore one way to alleviate the problem is to use a smaller M for such sounds [5],[8]. Switching the block size for a modulated lapped transform is not difficult but may introduce additional encoding delay. An alternative approach is to use a hierarchical transform [4] or a tree-structured filter bank, similar to a discrete wavelet transform. Such decomposition achieves a nonuniform subband structure, with small block sizes for the high-frequency subbands and large block sizes for the low-frequency subbands. Hierarchical (or cascaded) transforms have a perfect time-domain separation across blocks, but a poor frequency-domain separation. For example, if a QMF filter bank is followed by a MLTs on the subbands, the subbands residing near the QMF transition bands may have stopband rejections as low as 10 dB, a problem that also happens with tree-structured transforms [4].

One alternative is to increase the time-domain resolution by merging subbands, leading to new subbands with the same frequency but different time localizations [3]. That is the idea behind the frequency-varying MLTs suggested in [10]. The construction in [10], cascading MLTs with smaller IMLTs, does not lead as much time-domain separation as constructions based on the biorthogonal MLTs, as described in [3]. By simply adding and subtracting two consecutive subbands, it is possible to generate two news subbands whose filters have effectively half the time width, in the form

$$\begin{aligned} X'(2r) &= X(2r) + X(2r+1) \\ X'(2r+1) &= X(2r) - X(2r+1) \end{aligned} \tag{3}$$

The main advantage of this approach is that new subbands signals with narrower time resolution can be computed after the MLT of the input signal has been computed. Therefore, there is no need to switch the MLT window functions or block size. To improve the time resolution by a factor of four, i.e., to generate impulse responses with effective widths of a quarter block size, we can use the construction

$$\begin{bmatrix} X'(4r) \\ X'(4r+1) \\ X'(4r+2) \\ X'(4r+3) \end{bmatrix} = \begin{bmatrix} a & a & a & a \\ b & c & -c & -b \\ -b & c & c & -b \\ -a & a & -a & a \end{bmatrix} \begin{bmatrix} X(4r) \\ X(4r+1) \\ X(4r+2) \\ X(4r+3) \end{bmatrix} \tag{4}$$

where $a = 0.5412$, $b = \sqrt{1/2}$, $c = a^2$, $r = M_0, M_0+1, \ldots$, and M_0 usually set to $M/16$. Figure 2 shows plots of the synthesis basis functions corresponding to the construction in (4); we see that the time separation is not perfect, but it does lead to a reduction of error spreading for transient signals [3].

Automatic switching of the subband combination matrix in (4) can be done at the encoder by analyzing the input block waveform: if the power levels within the block vary considerably, the combination matrix is turned on. The switching flag is sent to the receiver as side information, so it can use the inverse 4x4 operator to recover the MLT coefficients. An alternative switching strategy that we used in our experiments is to analyze the power distribution among the MLT coefficients X(k) and to switch the combination matrix on when a high-frequency noise-like pattern is detected.

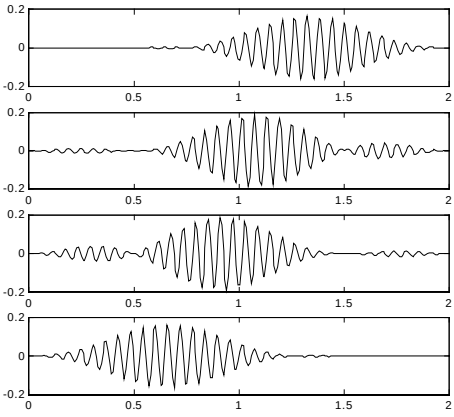

Figure 2. Some of the NMLBT synthesis functions when the combination matrix is switched on. Horizontal scale = block length.

6. EXAMPLES

We have tested the coder describe here with a variety of speech and music signals, with sampling rates varying from 8 to 32 kHz, and bit rates varying from 1 to 2 bits per sample. In one experiment, we used signals sampled at 8 kHz and compared our coder operating 10 kbps (1.25 bits/sample) to the ITU-T standard G.729, which operates at 8 kbps (1bit/sample). We set $\alpha = 0.5$ in (1) and used $M = 256$ subbands, corresponding to a block size of 32 ms and an algorithmic delay of 64 ms.

Simple objective performance measurements for low-rate coders may not necessarily correlate well with subjective measurements, specially at low bit rates. The well-known segmental SNR [6] can be useful in comparing the relative performances of two waveform coders, or as a tool for parameter optimization. However, it can be misleading when applied to a hybrid parametric/waveform coder such as G.729 [2]. So, besides the SNR we also used the segmental noise-to-masking ratio (NMR) [9] as an objective performance measurement.

The NMR measures how many dBs the coding noise is above the auditory masking threshold. NMRs of a few negative dBs means perceptually transparent or near-transparent reproduction, whereas a few positive dBs mean noticeable distortion. We computed the NMR as suggested in [9], by subtracting the power spectra of the original and decoded signals, and comparing it to the masking thresholds. Instead of averaging the NMR over all Bark subbands to compute the NMR for each signal block, we computed the NMR as the worst-case ratio among all Bark subbands, because our ears can easily spot errors that are above the threshold in just a few Bark bands. So, we will refer to it as a peak NMR (PNMR). The segmental PNMR is then computed the average of the block PNMRs.

Table 2 shows the segmental SNR and PNMR for the two coders. The SNR numbers for G.729 indicate that it does not do a good job of preserving the signal waveform, as expected. The SNR numbers for our coder indicate that it still reconstructs reasonably the signal the waveforms, even with the auditory

error weighting. The PNMR numbers correlate much better with informal listening tests. For G.729, they indicate a drop in performance (higher PNMRs) for the singing voice and violin signals. For our coder, the PNMR improves considerably for the violin signal; that is consistent with listening tests.

Signal	Seg. SNR, dB		Seg. PNMR, dB	
	G.729 8 kbps	Proposed 10 kbps	G.729 8 kbps	Proposed 10 kbps
Music – violin	4.0	13	4.2	1.9
Singing voice	3.7	15	2.2	4.7
Clean speech	3.8	13	1.8	4.3
Noisy speech	0.2	10	3.1	4.6
Hands-free speech	2.1	11	2.7	4.8

Table 2. Performance comparison for narrowband coders.

The PNMR values in Table 2 show that our coder operating at 10 kbps has a similar level of quality for music signals as G.729 for speech. The 1.5–2.5 dB difference for clean speech shows that G.729 is better in that case, but the difference is less noticeable in loudspeaker playback with average office noise conditions.

If we remove the run-length encoding component in our coder, the SNRs decrease by as much as 2 dB and the NMRs increase by as much as 1 dB, with a noticeable degradation in quality. For a fixed quantization step size (constant fidelity, variable bit rate), the run-length encoder can reduce the bit rate by about 10% for music signals, and 25% for speech. Therefore, the run-length encoder is quite efficient in encoding the periodicity in voiced speech. That is also indicated in Table 2, where the clean speech signal has a slightly better PSNR than the other speech signals.

In another experiment, we used music signals sampled at 32 kHz to compare the performance of our coder operating at 56 kbps (1.75 bits/sample) to that of the MPEG-2 Layer III standard operating at the same 56 kbps rate. We set $\alpha = 0.5$ in (1) and used $M = 1024$ subbands, corresponding to a block size of 32 ms and an algorithmic delay of 64 ms. The objective performance results for that experiment are shown in Table 3.

Signal	Seg. SNR, dB		Seg. PNMR, dB	
	MPEG-2 Layer III 56 kbps	Proposed 56 kbps	MPEG-2 Layer III 56 kbps	Proposed 56 kbps
Singer + guitar	23	16	1.6	2.4
Soft rock	22	11	1.9	2.7
Mix: classic + soft rock + speech	19	13	1.1	2.6

Table 3. Performance comparison for music coders.

We see that our coder approaches the performance of MPEG-2 Layer III, with PNMRs that are only about 0.5–1.5 dB higher. For both coders the quality of the reconstructed signal is very high; most people under common listening conditions would not distinguish the originals from the encoded ones.

For signals sampled at 16 kHz, we tested our coder at 24 kbps with $M = 512$ (still a block size of 32 ms). The resulting segmental SNRs and PNMRs were around 12–18 dB and PNMRs around 2–3 dB. The PNMR numbers and subjective quality are close to the ITU-T standard G.722 operating at 48 or 56 kbps.

7. CONCLUSION

We presented a subband/transform coder with three distinguishing characteristics: 1) a signal-dependent subband decomposition via a biorthogonal MLT without window switching; 2) encoding via uniform scalar quantization with run-length and arithmetic encoding; and 3) spectral noise shaping via weighted auditory masking functions. Although the run-length encoding improved the performance for clean speech, the encoder still has a better performance with music signals. The main advantage of the encoder is its robustness; it can operate with signals ranging from narrowband speech to high-fidelity music and rates from 10 kbps and above. Another advantage is the robustness to packet losses, because a missing block will affect the reconstruction up to the next block only, i.e., the coder recovers within a one block period. A disadvantage of the coder is a relatively large algorithmic delay of 64 ms (for the block sizes we reported), which may not be an issue for applications such as videoconferencing and Internet audio streaming.

REFERENCES

[1] S. Shlien, "The modulated lapped transform, its time-varying forms, and its applications to audio coding standards," *IEEE Transactions on Speech and Audio Processing*, vol. 5, pp. 359 –366, July 1997.

[2] R. V. Cox and P. Kroon, "Low bit-rate speech coders for multimedia communication," *IEEE Commun. Magazine*, vol.33, pp. 34–41, Dec. 1996.

[3] H. S. Malvar, "Biorthogonal and nonuniform lapped transforms for transform coding with reduced blocking and ringing artifacts," *IEEE Trans. Signal Processing*, vol. 46, no. 4, Apr. 1997.

[4] _____, *Signal Processing with Lapped Transforms*. Norwood, MA: Artech House, 1992.

[5] T. Mirya, N. Iwakami, A. Jin, K. Ikeda, S. Miki, "A design of transform coder for both speech and audio signals at 1 bit/sample," *Proc. IEEE ICASSP*, Munich, Germany, pp. 1371–1374, May 1997.

[6] A. S. Spanias, "Speech coding: a tutorial review," *Proc. IEEE*, vol. 82, pp. 1541–1582, Oct. 1994.

[7] N. S. Jayant, J. D. Johnston, and R. J. Safranek, "Signal compression based on models of human perception," *Proc. IEEE*, vol. 81, pp. 1385–1422, Oct. 1993.

[8] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE J. Selected Areas Commun.*, vol. 6, pp. 314–323, Feb. 1988.

[9] K, Brandenburg and T. Sporer, "NMR and masking flag: evaluation of quality using perceptual criteria," *Proc. 11th Int. AES Conference*, Portland, pp. 169–179, May 1992.

[10] M. V. Purat and P. Noll, "Audio coding with a dynamic wavelet packet decomposition based on frequency-varying modulated lapped transforms," *Proc. IEEE ICASSP*, Atlanta, pp. 1021–1024, May 1996.