Coding Overcomplete Representations of Audio using the MCLT

Byung-Jun Yoon

California Institute of Technology
1200 E. California Blvd.
Pasadena, CA 91125, USA

Henrique S. Malvar

Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA

Abstract

We propose a system for audio coding using the modulated complex lapped transform (MCLT). In general, it is difficult to encode signals using overcomplete representations without avoiding a penalty in rate-distortion performance. We show that the penalty can be significantly reduced for MCLT-based representations, without the need for iterative methods of sparsity reduction. We achieve that via a magnitude-phase polar quantization and the use of magnitude and phase prediction. Compared to systems based on quantization of orthogonal representations such as the modulated lapped transform (MLT), the new system allows for reduced warbling artifacts and more precise computation of frequency-domain auditory masking functions.

1. Introduction

Most modern audio compression systems use a frequency-domain approach [1]. The main reason is that when short audio blocks (say, 20 ms) are mapped to the frequency domain, for most blocks a large fraction of the signal energy is concentrated in relatively few frequency components, a necessary first step to achieve good compression. The mapping from time to frequency domain is usually performed by the modulated lapped transform (MLT, also known as the modified discrete cosine transform – MDCT) [1], an overlapping orthogonal transform that allows for smooth signal reconstruction even after heavy quantization of the transform coefficients, without discontinuities across block boundaries (blocking artifacts) [2].

One disadvantage of the MLT is that it is not a shift-invariant representation [3], [4], that is, if the signal is shifted by a small amount (say $1/8^{th}$ of a block), the transform coefficients will change significantly. In fact, just like with wavelet decompositions, there is no overlapping transform or filter bank that can be both shift invariant and orthogonal [5]. In particular, consider that the audio signal is composed of a single sinusoid of constant frequency and amplitude; then the MLT coefficients vary from block to block. Therefore, if they are quantized, the reconstructed audio will be a modulated sinusoid [6]. When all harmonic components of a more complex audio signal suffer from these modulations, "warbling" artifacts can be heard in the reconstructed signal.

Such modulation artifacts can be significantly reduced if we replace the MLT by a transform that supports a magnitude-phase representation, such as the modulated complex lapped transform (MCLT) [3]. However, the MCLT is an overcomplete (or

oversampled) transform by a factor of two, because it maps a block with M new real-valued signal samples into M complex-valued transform coefficients, which can significantly hurt compression performance.

In this paper we discuss strategies for encoding audio signals represented in the MCLT domain, and propose an encoder that significantly reduces the rate overhead caused by the overcomplete MCLT, without the need for iterative algorithms for sparsity reduction. We show that the R/D performance of such MCLT-based encoders can be close to that of MLT-based counterparts, while allowing for reduced warbling artifacts and other advantages of magnitude-phase representations. In Section 2 we review the definition and properties of the MCLT, and in Section 3 we discuss the problem of encoding overcomplete coefficients. In Section 4 we present the proposed structure for the encoder and discuss some practical results. We conclude by noting that efficient audio encoders can be designed with the MCLT as the time-frequency transform.

2. Overcomplete audio representations and the MCLT

The MCLT achieves a nearly shift-invariant representation [3], [4], because it supports a magnitude-phase decomposition that does not suffer from time-domain aliasing [3]. Thus, the MCLT has been successfully applied to problems such as audio noise reduction, acoustic echo cancellation, and audio watermarking. However, the price to be paid is that the MCLT expands the number of samples by a factor of two, because it maps a block with M new real-valued signal samples into M complex-valued transform coefficients. Namely, the MCLT of a block of an audio signal $x(n)$ is given by a block of frequency-domain coefficients $X(k)$, in the form

$$X(k) = X_C(k) + jX_S(k) \tag{1}$$

where k is the frequency index (with $k = 0, 1, \ldots, M-1$), $j \triangleq \sqrt{-1}$ and[1]

$$\begin{aligned} X_C(k) &= \sqrt{\frac{2}{M}} \sum_{n=0}^{2M-1} h(n)x(n)\cos\left[\left(n+\frac{M+1}{2}\right)\left(k+\frac{1}{2}\right)\frac{\pi}{M}\right] \\ X_S(k) &= \sqrt{\frac{2}{M}} \sum_{n=0}^{2M-1} h(n)x(n)\sin\left[\left(n+\frac{M+1}{2}\right)\left(k+\frac{1}{2}\right)\frac{\pi}{M}\right] \end{aligned} \tag{2}$$

We see that the set $\{X_C(k)\}$, the real part of the transform, forms the MLT of the signal. Thus, unlike in Fourier transform, there is a simple reconstruction formula from the real part only, as well as one from the imaginary part only, since each is an orthogonal transform of the signal [3]. The best reconstruction formula is the average of those from the real and imaginary parts. Using both for reconstruction removes time-domain aliasing [3]. Each of the sets $\{X_C(k)\}$ and $\{X_S(k)\}$ forms a complete orthogonal representation of a signal block, and thus the set $\{X(k)\}$ is overcomplete by a factor of two.

[1] Note that the summation extends over $2M$ samples because M samples are new while the other M samples come from overlapping [2].

We can easily convert the real-imaginary representation in (1) to a magnitude-phase representation by

$$X(k) = A(k)e^{j\theta(k)} \tag{3}$$

where $X_C(k) = A(k)\cos\left[\theta(k)\right]$, $X_S(k) = A(k)\sin\left[\theta(k)\right]$, and $A(k)$ and $\theta(k)$ are the magnitude and phase components, respectively.

We can see one of the main advantages of the magnitude-phase MCLT representation in (3): for a constant-amplitude and constant-frequency sinusoid signal, the magnitude coefficients will be constant from block to block. Thus, even under coarse quantization of the magnitude coefficients, a quantized MCLT representation is likely to lead to less warbling artifacts, as we discuss in Section 4. Another advantage is that the magnitude spectrum can be used directly for the computation of auditory models in a perceptual coder [7], without the need of computing an additional Fourier transform, like in MP3 encoders [1], or computing pseudo-spectra for the approximation of the magnitude spectrum from the real-valued (the MLT) coefficients [6].

3. Efficient encoding of MCLT representations

We discussed above that the MCLT has several advantages over the MLT for audio processing. However, for compression applications, an overcomplete representation such as the MCLT creates a data expansion problem: since the best reconstruction formula uses both the real and imaginary components, an encoder has to send both to a decoder. Therefore, we start with twice the data as that in a traditional MLT-based encoder, and we have the problem of how to efficiently encode MCLT coefficients without doubling (or otherwise significantly increasing) the bit rate.

Assuming a given quantization threshold, one approach to reduce redundancy in having both real and imaginary MCLT coefficients is to try to shrink the number of nonzero coefficients via iterative thresholding methods [4], [8]. For image coding, such methods are capable of essentially eliminating redundancy in terms of R/D performance, when using the also overcomplete dual-tree complex wavelet [9]. There are two main disadvantages of those methods, though. First, convergence is slow, so the dozens of required iterations are likely to increase encoding time considerably. Second, and most important for audio, the method does not guarantee that if $X_C(k)$ is nonzero at a particular frequency k, then $X_S(k)$ will also be nonzero, or vice-versa. Thus, we lose the magnitude and phase information and introduce time-domain aliasing artifacts at that frequency.

Another approach is to predict the imaginary coefficients from the real ones. For a given block, if both the previous and next block were available, then the time-domain waveform could be reconstructed, and from it $X_S(k)$ could be computed exactly. However, that would introduce an extra block delay, which is undesirable in many applications. Using only the current and previous block, it is possible to approximately predict $X_S(k)$ from $X_C(k)$ [10]. Then the prediction error from the actual values of $X_S(k)$ can be encoded and transmitted. We can first encode $X_C(k)$, and predict $X_S(k)$ for the

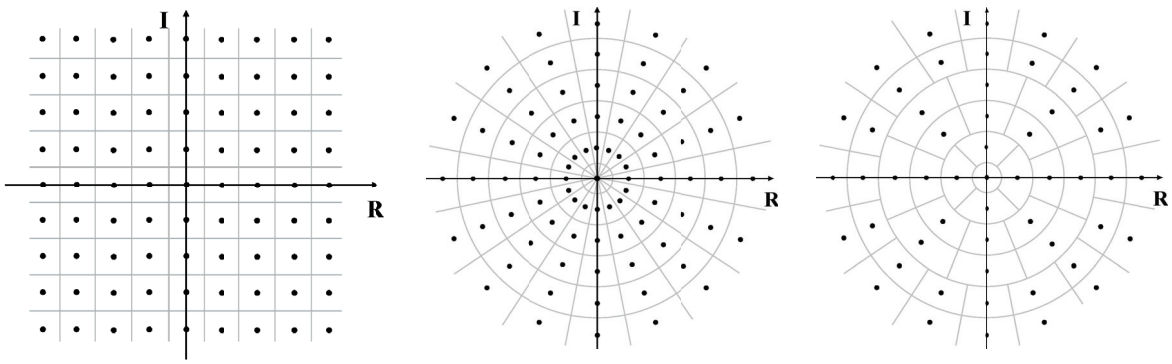

Figure 1. Different approaches to quantization of the real and imaginary parts of each MCLT coefficient. Left: independent scalar quantization of real and imaginary parts (rectangular quantization); middle: independent scalar quantization of magnitude and phase (uniform polar quantization); right: unrestricted polar quantization.

frequencies k for which $X_C(k)$ is nonzero. That way, for every frequency k for which data is transmitted, both the real and imaginary coefficients are transmitted. However, that approach still leads to a significant rate overhead, mainly because the prediction of the imaginary part from the real part without using future data is not very efficient.

3.1. Magnitude-phase quantization

As we discussed in the previous version, to attenuate the warbling artifacts, we need an explicit magnitude-phase representation. Therefore, we should consider quantizing the magnitude and phase coefficients $A(k)$ and $\theta(k)$ (polar quantization), instead of the real and imaginary coefficients $X_C(k)$ and $X_S(k)$ (rectangular quantization). It is well known that polar quantization can lead to essentially the same rate-distortion performance of rectangular quantization, as long as the phase quantization is made coarser for smaller magnitude values [11], as shown in Figure 1 (right); that is usually referred to as unrestricted polar quantization (UPQ) [12]. That is an intuitive result, because if the number of phase quantization levels were to be set independent of magnitude, then the quantization bins near the origin would have much smaller areas, thus leading to an increase in entropy.

We note that the near-optimal properties of UPQ apply for quantization of uncorrelated complex-valued Gaussian random variables [12]. However, for our application two unrelated properties make it difficult to directly apply such results. First, for many short-time music segment, amplitudes of tones tend to vary slowly from block to block, thus the values of a particular MCLT magnitude coefficient $A(k)$ are correlated from block to block. Second, the human ear is relatively insensitive to phase [13]; thus, phase quantization errors may lead to increases in root-mean-square (RMS) error that may not lead to proportional decreases in perceived quality. Therefore, straight R/D results may not apply, and some trial-and-error is needed for the proper adjustment of the quantization bins in the UPQ in Figure 1 (right).

range for magnitude X_M	0–0.5	0.5–1.5	1.5–2.5	2.5–3.5	3.5–4.5	> 4.5
# of bits for phase ϕ	0	2	3	3	4	4

Table 1. Practical parameter values for UPQ quantization.

In our experiments, we noticed that for most content random phase errors in MCLT coefficients of up to $\pi/8$ are nearly imperceptible, even when listening with high-quality headphones. Coarser quantization may bring warbling and echo artifacts. Thus, we may not need more than 4 bits to quantize the phase of high-magnitude coefficients, and fewer bits for lower-magnitude ones. We found that the UPQ shown in Figure 1 led to best results. If the magnitude is quantized to zero, then of course no phase information is needed; for nonzero magnitude values, the number of bits for phase is assigned as indicated in Table 1, which corresponds to the UPQ plot in Figure 1.

With the UPQ as defined above, the rate-distortion performance is then controlled by a single parameter: the scaling factor applied to the MCLT coefficients prior to magnitude-phase quantization; the higher the scaling factor, the higher the bit rate and fidelity.

Even with the relatively coarse phase quantization as proposed above, warbling artifacts are reduced, when compared to quantization of MLT coefficients. That can be verified in Figure 2, where we plot the spectrogram of quantized MLT coefficients versus that of quantized magnitude MCLT coefficients, for a piano test signal sampled at 16 kHz. The quantization step size for the MLT and the scaling factor for the MCLT with

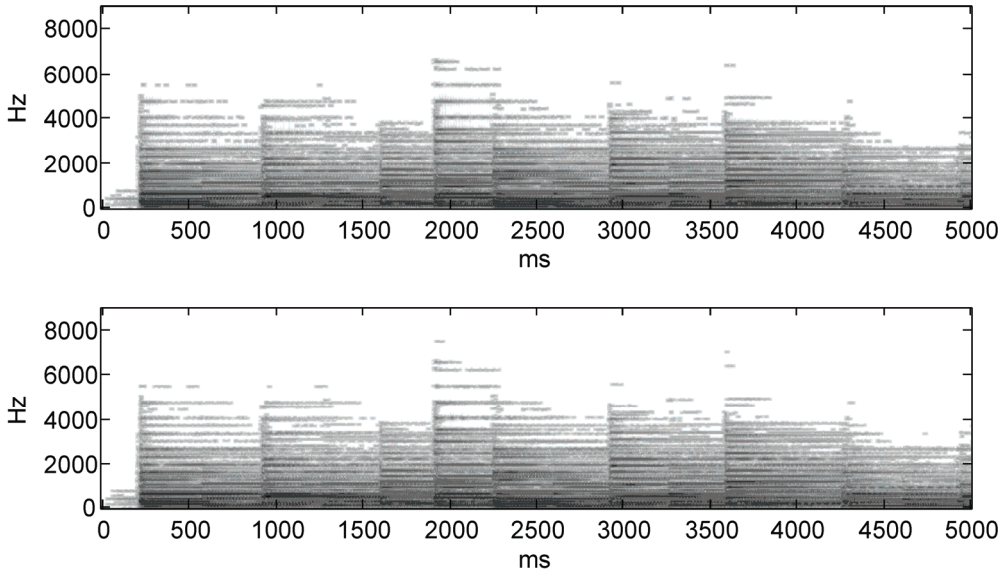

Figure 2. Spectrogram of quantized coefficients; darker regions indicate higher energy. Top: using the MLT and scalar quantization; bottom: using the MCLT and UPQ quantization.

UPQ quantization were chosen such that the mean-square errors in both cases were comparable, and the distortions were noticeable but barely so. We see from Figure 2 that, as the piano notes decay, the MCLT coefficients are less likely to be quantized to zero in a particular block and then to a nonzero value in the next block. In other words, there are fewer discontinuities within each of the horizontal lines in Fig. 2, which correspond to harmonics of each piano note. That leads to a reduction in warbling artifacts.

3.2. Magnitude and phase prediction

Figure 3 shows plots of the real part $X_C(k)$ and the magnitude $A(k)$ of the MCLT of a piano test signal sampled at 16 kHz, for subband $k = 5$, in a representation with $M = 512$ subbands. We see that there is significant correlation between consecutive samples $A(k,m-1)$ and $A(k,m)$, where m is the block index. Thus, instead of encoding $A(k,m)$, we can encode the residual from a linear prediction based on previously-transmitted samples

$$E(k,m) \triangleq A(k,m) - \sum_{r=1}^{L} b_r \, A(k,m-r) \tag{4}$$

where L is the predictor order and $\{b_r\}$ is the set of predictor coefficients, which can be computed via a traditional autocorrelation analysis [14]. For most blocks the optimal predictor order L can be very low, from $L = 1$ to $L = 3$. The values of L and $\{b_r\}$ can be encoded in the header for each block.

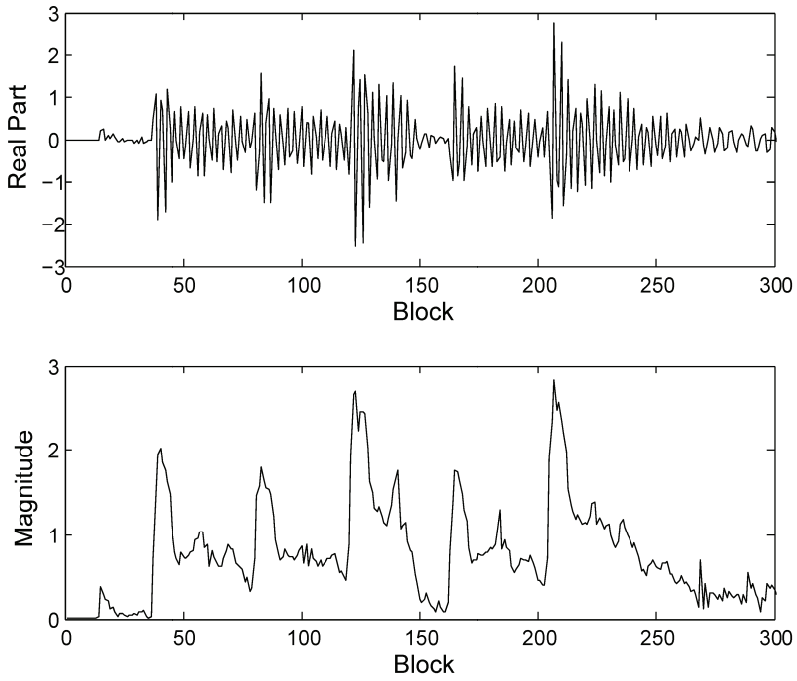

Figure 3. Plot of MCLT coefficients for a particular frequency, for a piano signal. We see that magnitude values are strongly correlated from block to block.

We can also predict the phase of MCLT coefficients. From the analyses presented in [6] for the computation of MLT coefficients for sinusoidal inputs, we conclude that if the input signal is a sinusoid at the center frequency of the kth subband, then the phase of two consecutive blocks satisfies $\theta(k,m) = \theta(k,m-1) + (k+\frac{1}{2})\pi$. Therefore, we can encode just the phase difference between $\theta(k)$ and the value predicted by that formula, namely

$$p(k,m) \triangleq \theta(k,m) - \theta(k,m-1) - \left(k + \frac{1}{2}\right)\pi \tag{5}$$

For most audio signals, components are not exactly sinusoidal, and their frequencies are not at the center of the subbands. Thus, prediction efficiency varies from block to block.

An additional prediction step can be applied to the phase. From (3) we see that if we know just $|\theta(k)|$ we can reconstruct the real part $X_C(k)$, because $\cos[\theta(k)] = \cos[-\theta(k)]$. We only need to know the sign of $\theta(k)$ to reconstruct $X_S(k)$. We mentioned before that predicting $X_S(k)$ from $X_C(k)$ may not be very precise [10], but if the precision is good enough to at least get the sign of $X_S(k)$ correctly, then we know the sign of $\theta(k)$ and thus we don't need to encode it. Therefore, we can aggregate the signs of all encoded phase coefficients into a vector and replace them by predicted signs computed from the real-to-imaginary component prediction. Without prediction, the phase signs would have roughly an entropy of one bit per encoded value (because signs are equally likely to be positive or negative), but after prediction the entropy can be reduced, as discussed next.

4. Proposed audio encoder structure

Based on the results of the previous section, we propose the audio encoding and decoding structure shown Figure 4. For each block of the input signal $x(n)$, we first compute its MCLT coefficients $X_C(k, m)$ and $X_s(k, m)$, and from them we compute the corresponding magnitude and phase coefficients $A(k, m)$ and $\theta(k, m)$, where m denotes the block index. For audio signals sampled at 16 kHz, a block length of $M = 512$ samples leads to best results, whereas for CD-quality audio sampled at 44.1 or 48 kHz, a block size of $M =$ 2,048 samples works best[2]. We then quantize the magnitude and phase coefficients using the UPQ polar quantizer in Figure 1 (right), producing the corresponding quantized values $A_Q(k, m)$ and $\theta_Q(k, m)$. Not explicitly indicated in Figure 4 is the scaling factor α, by which the MCLT coefficients are multiplied prior to the polar conversion. That parameter controls rate/distortion; the higher its value, the higher the fidelity and the bit rate. At the decoder, the coefficients are multiplied by $1/\alpha$ prior to the inverse MCLT.

The quantized magnitude and phase coefficients then go through the prediction steps described in the previous section. Note that in computing the predictors in (4) and (5) we should use the quantized values $A_Q(k, m)$ and $\theta_Q(k, m)$, so the decoder can recompute the

[2] For CD-quality audio, usually a fixed time-frequency resolution does not produce good reproduction of transient sounds. Thus, usually a block-size switching technique is employed, e.g. using $M = 2{,}048$ for blocks with mostly tonal components, and $M = 256$ for blocks with mostly transient components. These techniques [1] can be directly applied to the proposed encoder, except that we cannot predict the quantized coefficients for the first block after size switching.

predictors. Note that in (5) the phase prediction is indicated in the original continuous-valued domain. To map it to a prediction in the UPQ-quantized domain, we observe that for every cell in the UPQ diagram in Figure 1, a cell with the same magnitude but with a phase equal to the original phase plus an integer multiple of $\pi/2$ is also in the diagram.

The final step is to entropy encode the quantized prediction residuals. In our experiments we estimated entropies via data statistics, and in practice these estimates should predict reasonably well the actual bit rates if we use adaptive arithmetic encoders or adaptive run-length Golomb-Rice (RLGR) encoders [15].

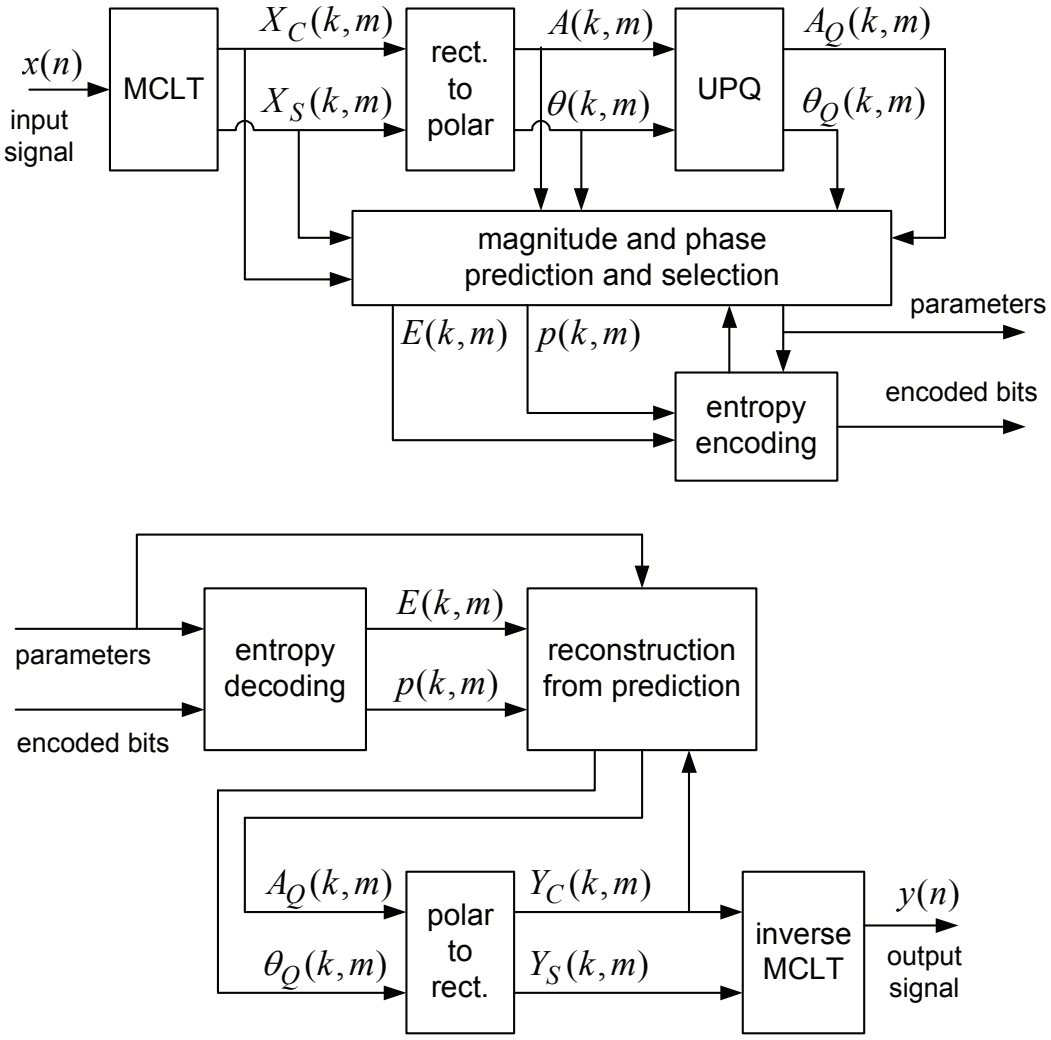

Figure 4. Proposed structure for audio encoding with an MCLT-based magnitude-phase representation. Top: encoder; bottom: decoder.

Besides the encoded bits corresponding to the processed MCLT coefficients, additional parameters should be encoded and added to the bitstream. Those include the scaling factor α, the number of subbands M, the predictor order L, the prediction coefficients $\{b_r\}$, and maybe additional parameters that control the entropy coders. Unless compression ratios are high enough for artifacts to be very strong, usually the bit rate used by the parameters is less than 5% of that used for the processed MCLT coefficients.

We ran several sets of experiments with audio sampled at 16 kHz, at several rate-distortion points. Our main goal was to determine if the use of MCLT instead of the MLT would lead to significant increase in bit rate. Thus, we compared our proposed encoder in Figure 4 with encoding only the real coefficients (recall that the MLT is the real part of the MCLT), without prediction. We did not make comparisons with full-blown encoders such as WMA or AAC [1] because our proposed encoder would have to be enhanced and fine-tuned to achieve performance comparable to those.

In our experiments, we started by turning off the predictors and measuring the entropy, for several values of the scale factor α and various test signals. When we turned on magnitude prediction only, entropy was reduced between 0.2 to 0.6 bits/sample. As expected, the gains were stronger for music signals than for speech, because music usually has more strongly tonal components. Then we turned on phase prediction, and that led to an additional entropy reduction of 0.1 to 0.3 bits/sample. The overall reduction in entropy is on the order of 0.4 to 0.8 bits/sample. That is enough to bring the overall entropy quite close to that of encoding the MLT coefficients. At bit rates where small distortions are barely audible (around 1.5 to 2.5 bits/sample), the overall quality of the MCLT-reconstructed audio was similar to that of the MLT-reconstructed audio; however the MCLT-reconstructed signal had significantly less warbling artifacts.

5. Conclusion

We presented an approach for efficient audio compression using overcomplete signal expansions, in particular the MCLT. Our proposed basic encoding structure uses unrestricted polar quantization of the MCLT magnitude and phase. Additionally, we use prediction of the quantized magnitude and phase coefficients, based on properties of the MCLT and properties of audio signals. Preliminary results show that these predictors can reduce the bit rate overhead in encoding an overcomplete representation (the MCLT) to the point where the overall rate is comparable to that of encoding an orthogonal representation (the MLT).

There are several advantages of using the MCLT instead of the MLT. First, we can achieve better continuity of the magnitude of spectral components across blocks, thus reducing warbling artifacts. Second, we can run more precise auditory models directly from the MCLT coefficients, without having to compute additional Fourier transforms, as in MP3 encoders. The models can also be more precise because for a given block length the MCLT has twice the frequency resolution of the discrete Fourier transform [3].

References

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

[2] H. S. Malvar, *Signal Processing with Lapped Transforms*. Boston, MA: Artech House, 1992.

[3] H. S. Malvar, "A modulated complex lapped transform and its applications to audio processing,", *Proc. Int. Conf. Acoustics, Speech, Signal Processing,* Phoenix, AZ, pp. 1421–1424, Mar. 1999.

[4] M. E. Davies and L. Daudet, "Sparse audio representations using the MCLT," *Signal Processing*, vol. 86, pp. 457–470, Mar. 2006.

[5] N. G. Kingsbury, "Complex wavelets for shift invariant analysis and filtering of signals," *Applied and Computational Harmonic Analysis*, vol. 10, pp. 234–253, May 2001.

[6] L. Daudet and M. Sandler, "MDCT analysis of sinusoids: exact results and applications to coding artifacts reduction," *IEEE Trans. Speech and Audio Processing*, vol. 12, pp. 302–312, May 2004.

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

[8] M. Yaghoobi, T. Blumensath, and M. Davies, "Quantized sparse approximation with iterative thresholding for audio coding," *Proc. Int. Conf. Acoustics, Speech, Signal Processing,* Honolulu, HI, pp. I-257–I-260, Apr. 2007.

[9] N. G. Kingsbury and T. H. Reeves, "Iterative image coding with overcomplete complex wavelet transforms," *Proc. Conf. Visual Comm. and Image Processing*, Lugano, Switzerland, pp. 1253–1264, July 2003.

[10] S. Cheng and Z. Xiong, "Audio coding and image denoising based on the nonuniform modulated complex lapped transform," *IEEE Trans. Multimedia*, vol. 7, pp. 817–827, Oct. 2005.

[11] S. Wilson, "Magnitude/phase quantization of independent Gaussian variates," *IEEE Trans. Communications*, vol. COM-28, pp. 1924–1929, Nov. 1980.

[12] R. Vafin and W. B. Kleijn, "Entropy-constrained polar quantization and its applications to audio coding," *IEEE Trans. Speech and Audio Processing*, vol. 13, pp. 220–232, Mar. 2005.

[13] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 5th ed. San Diego, CA: Academic Press, 2003.

[14] A. M. Kondoz, *Digital Speech*. New York: Wiley, 1994; Chapter 3.

[15] H. S. Malvar, "Adaptive run-length/Golomb-Rice encoding of quantized generalized Gaussian sources with unknown statistics," *Proc. Data Compression Conf.*, Snowbird, UT, pp. 23–32, Mar. 2006.