SCALABLE AUDIO CODING USING THE NONUNIFORM MODULATED COMPLEX LAPPED TRANSFORM

Anne-Sophie Scheuble and Zixiang Xiong
Dept of Electrical Engineering, Texas A&M University, College Station, TX 77843

ABSTRACT

This paper introduces a scalable audio coder using the nonuniform modulated complex lapped transform (NMCLT) [1], which is a new nonuniform oversampled filter bank with a better combination of time- and frequency-domain localization than previous designs. Masking functions for different critical Bark bands are first calculated directly from the NMCLT coefficients as perceptual weights and arithmetic coding is then used to compress bit planes of the weighted NMCLT coefficients to generate a perceptually scalable audio bitstream. The loss in coding performance due to oversampling is offset by limiting the amount of redundancy in the transform and exploiting the correlations among the NMCLT basis functions. Experiments show that our new coder outperforms a coder with the modulated lapped transform (MLT) [2] both objectively and subjectively.

1. INTRODUCTION

We live in the midst of an Internet revolution that is dramatically changing the way we live, entertain, and communicate. For example, MP3 [3] is taking the music industry by storm. Originally, MP3 is a short-handed name for layer III of the MPEG-1 audio compression standard. Now it encompasses all digital music compressed or stored in MP3 format. Unfortunately, MP3 performs very poorly in terms of compression efficiency because the MPEG-1 coding standard was developed in the early 1990s. With the boom in Internet multimedia, companies such as RealNetworks and Microsoft have successfully developed their own commercial audio coders (e.g., RealAudio and MSAudio). These coders offer better compression than the MP3 coder. The latest MPEG-4 natural audio coder [4] achieves scalability with hierarchical coding, but this scalability is not in the psychoacoustic sense, meaning that the improvement in *perceptual* quality of the decoded audio is not commensurate with the increase in bit rate.

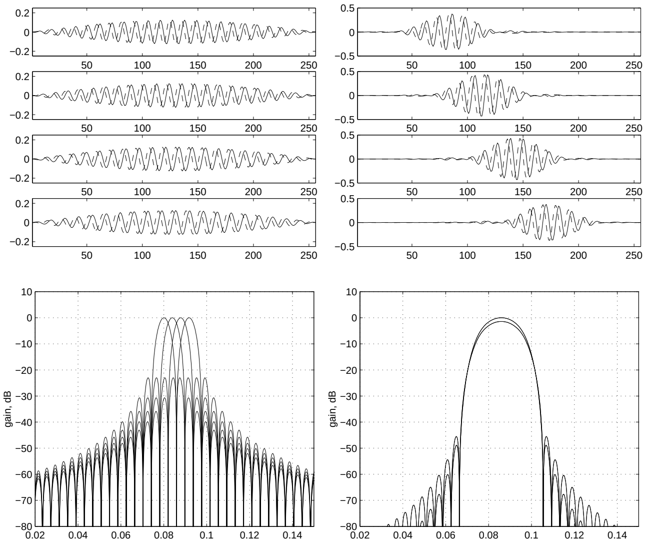
There are two main issues in high performance audio coding: time-frequency (TF) transform and perceptual weighting [5, 6]. As scalability is becoming increasingly important in the emerging world of heterogeneous packet networks and wireless communication, an additional issue in scalable audio coding [7] is bit plane coding of transform coefficients. This paper presents a scalable audio coder using the nonuniform modulated complex lapped transform (NMCLT) and bit plane coding. The NMCLT is a new nonuniform oversampled filter bank with a better TF localization than previous designs. Based on the psychoacoustic principles described in [6], masking functions for different critical Bark bands are first calculated directly from the NMCLT coefficients as perceptual weights and arithmetic coding is then used to compress bit planes of the weighted NMCLT coefficients to generate a perceptually scalable audio bitstream. The loss in coding performance due to oversampling is offset by limiting the amount of redundancy in the transform and exploiting the correlations among the NMCLT basis functions. Experiments show that our new coder outperforms a coder with the modulated lapped transform (MLT) [2] both objectively and subjectively.

1.1. NONUNIFORM MODULATED COMPLEX LAPPED TRANSFORM

An ideal TF transformation should compact the energy of the original audio samples into as few TF atoms as possible for the ease of compression. Since such an optimal TF transform is signal dependent, state-of-the-art coders use modulated lapped transform (MLT) [2] or modified discrete cosine transform (MDCT), which is signal independent, for computational reasons. The drawback of the MLT is that it only provides a fixed TF resolution, i.e., all MLT bases have the same TF support (Fig. 1 (a)), making it impossible to arbitrarily adapt to the real TF resolution of the audio signal. This results in annoying pre- or post-echos in low bit rate coded audio signals. One way to alleviate this problem is to switch to a shorter block-size MLT during high-frequency transient sounds, as is done in MP3. Another way is to use hierarchical [9] or tree-structured wavelet packet [7, 8] decompositions with a nonuniform subband structure. This generates highpass basis functions with good time resolution but very poor frequency resolution due to aliasing. An alternative was introduced in [9, 10] to increase the time resolution of MLT basis functions by merging subbands such that basis functions with the same frequency resolution can have different time localizations. However, the number of subbands merged was very small (two or four) and no systematic way of merging subbands was offered in [9, 10].

The modulated complex lapped transform (MCLT) was recently introduced in [11] as a simple extension to the MLT. A key observation made in [11] is that, with a 2× oversampling ratio in the MCLT, time-domain aliasing terms in the real and imaginary reconstruction parts have opposite signs, i.e., they cancel each other automatically. Based on this observation, Xiong and Malvar extended the MCLT to the NMCLT [1] by cascading a MCLT with shorter size MCLTs. Each shorter size MCLT plays the role of merging subbands. Time-aliasing terms in the NMCLT basis functions can be made to cancel each other in the inverse

This work was supported in part by the NSF under the CAREER grant 0096070, the ARO under the YIP grant DAAD19-00-1-0509, and a gift grant from the Microsoft Corporation.



(a) (b)

Figure 1: Time- and frequency-domain responses of different basis functions. (a) The MLT basis functions; (b) The NMCLT basis functions.

NMCLT. This allows one to construct basis functions with desirable frequency resolutions (e.g., the wavelet paket-like decomposition that follows the Bark scale in [8]) and almost ideal time localizations (Fig. 1 (b)). Fast algorithms for computing the modulated lapped sine and cosine transforms can be used to implement the NMCLT.

2. SCALABLE AUDIO CODING USING THE NMCLT

Although the NMCLT basis functions have desirable TF localizations, they have one major drawback for compression: overcompleteness. In motivating the use of the overcomplete NMCLT for audio coding, we note that the reason for introducing redundancy is to cancel time-aliasing terms in the NMCLT basis functions. The condition for aliasing cancellation is equal quantization error for the real and imaginary parts of the NMCLT coefficients [1]. As such, once the real parts of the NMCLT coefficients are quantized, the quantized version of the imaginary parts is also determined. In another words, we only need to focus on quantizing the real parts as the quantizer for the imaginary parts is a "passive one." This is different from quantizing an overcomplete frame representation [12] where the rate-distortion performance will suffer significantly due to oversampling. To boost the performance of our coder, we take advantage of the strong correlations among the NMCLT basis functions by predicting the imaginary parts from the real parts of the NMCLT coefficients. In addition, we limit oversampling (or redundancy in the NMCLT) in Bark bands where quantization noise is most perceptible.

Fig. 2 depicts the block diagram of our proposed coder. We denote $\mathbf{T_c}$ as the real part of the NMCLT matrix corresponding to size N MDCT followed by size M $(N > M)$ modified sine transforms (MDSTs); $\mathbf{T_s}$ as the imaginary part of the NMCLT matrix corresponding to size N MDST followed by size M MDCTs[1]. To code the real part of the NMCLT coefficients $\mathbf{X_c}$, we first calculate masking thresholds $\mathbf{w}$ for all Bark bands based on psychoacoustic principles [6], we then uniformly quantize $\mathbf{X_c}/\mathbf{w}$ before applying arithmetic coding on each bit plane of the quantized indices. An embedded bitstream is thus generated for $\mathbf{X_c}$, with the masking thresholds for different Bark bands coded as header information. Note that, because the masking thresholds differ from Bark band to Bark band, we are effectively using nonuniform quantization that exploits the psychoacoustic properties of the human ears.

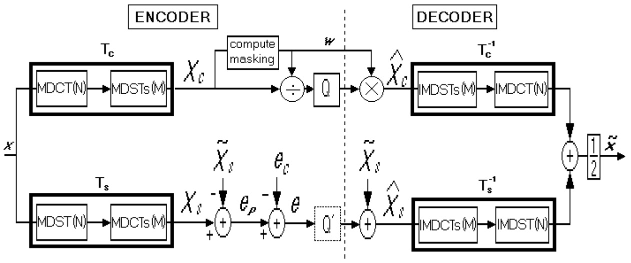


Figure 2: Block diagram of the proposed scalable audio coder.

To efficiently compress the imaginary part of the NMCLT coefficients $\mathbf{X_s}$, we predict $\mathbf{X_s}$ as

$$\mathbf{\tilde{X}_s} = \mathbf{T_s T_c^{-1} \hat{X}_c},$$

where $\mathbf{\hat{X}_c}$ is the quantized version of $\mathbf{X_c}$. The prediction error can be written as

$$\mathbf{e_p} = \mathbf{X_s} - \mathbf{\tilde{X}_s} = \mathbf{T_s T_c^{-1}}(\mathbf{X_c} - \mathbf{\hat{X}_c}) = \mathbf{T_s T_c^{-1} e_c},$$

where $\mathbf{e_c}$ is the quantization error for $\mathbf{X_c}$.

Recall that, to guarantee time-aliasing cancellation in the NMCLT basis functions, the quantization errors for $\mathbf{X_c}$ and $\mathbf{X_s}$ must be equal. This condition is satisfied if we quantize $\mathbf{X_s}$ as

$$\mathbf{\hat{X}_s} = \mathbf{X_s} - \mathbf{e_c} = \mathbf{\tilde{X}_s} + \mathbf{e_p} - \mathbf{e_c} = \mathbf{\tilde{X}_s} + (\mathbf{T_s T_c^{-1}} - \mathbf{I})\mathbf{e_c}.$$

Thus, we only need to code $\mathbf{e} = \mathbf{e_p} - \mathbf{e_c} = (\mathbf{T_s T_c^{-1}} - \mathbf{I})\mathbf{e_c}$ for the imaginary part of the NMCLT coefficients $\mathbf{X_s}$. Ideally, the vector $\mathbf{e}$ should be conveyed with full floating precision, we nevertheless use a very small step size to uniformly quantize it before applying arithmetic coding on bit planes of the resulting quantization indices.

Discussion:

1. In forming the final embedded bitstream, the sub-bitstreams corresponding to $\mathbf{X_c}$ and $\mathbf{X_s}$ (or $\mathbf{e}$) can be either interleaved or concatenated. In the latter case, the sub-bitstream corresponding to $\mathbf{X_s}$ can be considered as an enhancement layer for applications such as audio over IP. In either case, bit allocation between $\mathbf{X_c}$ and $\mathbf{X_s}$ is implicitly accomplished by bit plane coding.

[1] We use fixed size MDCTs or MDSTs in the second stage of the NMCLT in Fig. 2, although transforms of different sizes are used in practice, depending on the desirable TF resolution.

2. From $\mathbf{e} = (\mathbf{T_s}\mathbf{T_c}^{-1} - \mathbf{I})\mathbf{e_c}$, we see that we are encoding a correction term for the imaginary part of the NMCLT coefficients to cancel time aliasing in the real part. Although refining $\mathbf{e_c}$ will directly reduce the quantization error of the real part, it is not as effective in time-aliasing cancellation as encoding $\mathbf{e}$, which has smaller energy than $\mathbf{e_c}$. Of course, the price paid for aliasing cancellation is extra complexity in computing $\mathbf{e} = (\mathbf{T_s}\mathbf{T_c}^{-1} - \mathbf{I})\mathbf{e_c}$ and $\tilde{\mathbf{X}}_\mathbf{s} = \mathbf{T_s}\mathbf{T_c}^{-1}\hat{\mathbf{X}}_\mathbf{c}$.

In our coder design, we also take advantage of the fact that the amount of redundancy in the NMCLT is controllable. We only apply the second stage MDCTs or MDSTs in the NMCLT on specific Bark bands. For Bark bands that do not go through a second stage transform in the NMCLT, no information from the imaginary part of the transform will be coded. One way is to encode refinement information $\mathbf{e}$ from the imaginary part of the NMCLT in the most important frequency range (500-5000Hz.) From our experiments, this turns out not to be the best option. In our coder, we choose to apply the second stage transforms in two disjoint frequency ranges (520-860Hz and 8300-16500Hz) that span six Bark bands. Fig. 3 shows the masking thresholds for a typical block of NMCLT coefficients with the dark areas indicating these two frequency ranges. The rationale for our choice is as follows: the masking thresholds for Bark bands within 860-8300Hz are relatively low. Time-aliasing terms due to quantization error in these bands will also have low magnitude. They will render the pre-echoes (if any) inaudible. For Bark bands in the low frequency range (<520Hz) or high frequency range (>16500Hz), time-aliasing terms will be masked due to the relative high masking thresholds. In the 500-860Hz and 8300-16500Hz range, the masking thresholds are relatively large for the aliasing terms to be audible without cancellation. We pick these fixed frequency ranges for the sake of simplicity.

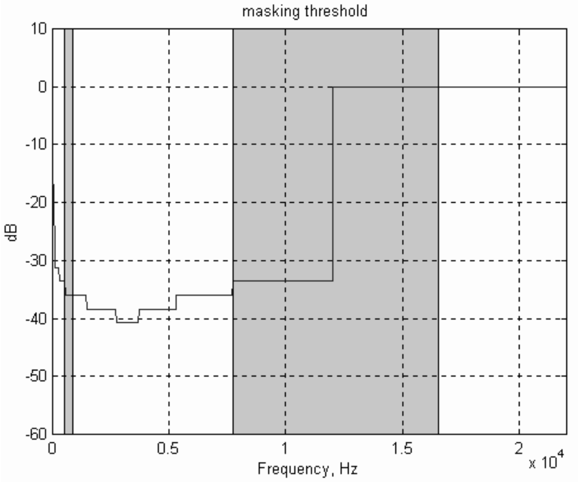


Figure 3: Masking thresholds for a typical block of NMCLT coefficients. Dark areas mark the 520-860Hz and 8300-16500Hz frequency ranges in which the imaginary part of the NMCLT is coded.

3. EXPERIMENTAL RESULTS

Our new coder has been tested with the different types of mono sound clips listed in Table 1. We compare our coder with a benchmark coder that uses MLT as the transform (perceptual masking calculation and arithmetic coding are in principle the same in both coders.) This benchmark coder performs similar to the MSAudio coder, which is better than the MP3 coder. We use the segmental SNR [13] as the objective measure for sound quality. Table 2 gives the segmental SNR values associated with audio signals decoded at various rates for both coders. The rate partition between the real and imaginary parts of NMCLT coefficients in our new coder is also given for each bit rate.

Clip	Type	Sampling rate	Length
Castanets	Percussive sound	44.1kHz	7.95s
Vega	Voice & music	16kHz	9.76s
Seal	Pop music	44.1kHz	12.94s
Vollenweider	Cappella song	44.1kHz	12.94s
Copland	Symphony	44.1kHz	12.94s

Table 1: Characteristics of test sound clips.

Table 2 shows that the new coder leads to mostly better results than the benchmark coder. In fact, for sounds with a presence of percussions, like the Castanets clip and even the Seal clip, the results are as good as expected. Likewise, the encoding result of the Vega clip (voice and very light musical background) is improved compared to the result of the benchmark coder. In contrast, sound clips with sophisticated musical contents like symphonic music do not really benefit from the new coder.

Clip	Benchmark coder	New coder
Castanets	22.94 (50.40) 18.57 (35.68) 14.44 (24.70)	25.08 (45.44+4.96) 20.98 (31.84+3.84) 16.38 (19.86+4.84)
Vega	36.76 (25.48) 33.41 (17.10) 27.00 (8.05)	39.46 (22.75+2.73) 35.11 (15.33+1.77) 27.50 (7.01+1.04)
Seal	28.33 (48.69) 26.86 (32.58) 25.37 (23.92)	33.30 (43.56+5.13) 30.76 (25.75+6.83) 28.61 (19.95+3.97)
Vollenweider	47.36 (43.47) 41.74 (33.06) 33.80 (20.17)	49.16 (39.69+3.78) 43.32 (29.40+3.66) 33.57 (17.89+2.28)
Copland	45.20 (38.19) 42.25 (30.26) 36.95 (22.77)	45.53 (34.30+3.89) 41.96 (25.91+4.35) 35.66 (19.18+3.59)

Table 2: Segmental SNR comparisons (in dB) for test sound clips. Numbers in parentheses are bit rates (in kbps). For the new coder, we show the partition of the total bit rate in coding the real and imaginary part of the NMCLT coefficients.

To give an indication of the subjective improvement in sound quality using our new coder, we plot in Figs. 4 and 5 the reconstruction errors for Castanets and Vega resulting from the two encoders. The first plot in each of these figures show the reconstruction error when both the real and imaginary parts of the NMCLT are coded, the second showing the error when only the real part of the NMCLT is coded, and the third depicting the error using the benchmark coder. In all three cases, the rate is the same for each sound clip. It is obvious that the new coder gives the smallest reconstruction error when both the real and imaginary parts of the NMCLT coefficients are coded. This confirms the advantage of using the NMCLT for audio coding.

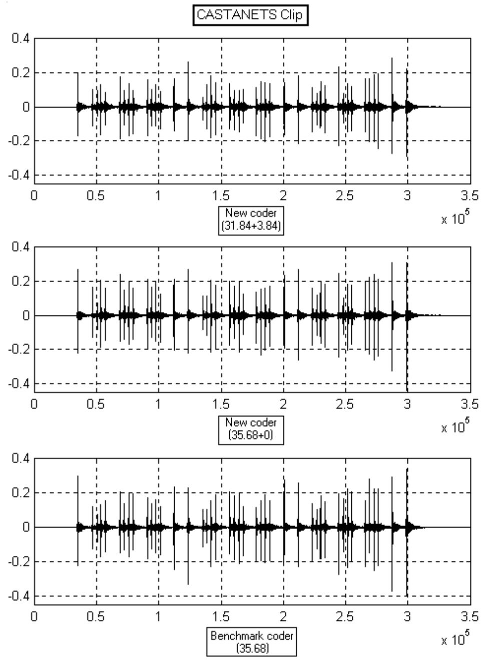


Figure 4: Errors between the original Castanets clip and different decoded versions at 35.68 kbps.

4. CONCLUSIONS

This paper presents a scalable audio coder based on the NMCLT, whose basis functions have desirable TF resolutions that allow us to reduce pre-echoes in the coded sound. Two strategies (prediction and control of redundancy) are used to overcome the major drawback of overcompleteness in the transform. Experimental results show that our new coder significantly reduces pre-echoes and improves the sound quality of music clips with transient sounds. Although our coder does not perform as well for symphonic sound clips, it is still competitive with the benchmark coder. Current research is focusing on adaptive choices of Bark bands for which the imaginary parts of the NMCLT are coded.

5. ACKNOWLEDGEMENT

Z. Xiong would like to thank Rico Malvar at Microsoft Research for teaching him everything about audio coding and Phil Chou for fruitful discussions.

6. REFERENCES

[1] Z. Xiong and H. S. Malvar, "A nonuniform modulated complex lapped transform," submitted to *IEEE Signal Processing Letters*, October 2000.

[2] 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.

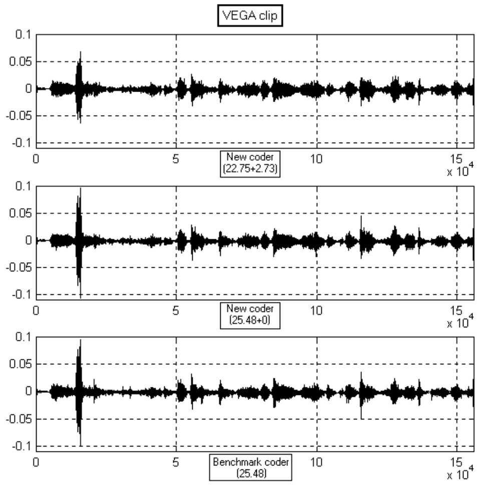


Figure 5: Errors between the original Vega clip and different decoded versions at 25.48 kbps.

[3] ISO/IEC DIS 11172-3, *Coding of Moving Pictures and Associated Audio for Digital Storage Media up to about 1.5 Mb/s*, Part 3: Audio, 1993.

[4] K. Brandenburg, O. Kunz, and A. Sugiyama, "MPEG-4 natural audio coding," *Signal Processing: Image communication*, vol. 15, January 2000.

[5] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE Journal on Selected Areas in Communications*, vol.6, pp. 314-323, February 1988.

[6] T. Painter and A. S. Spanias, "Perceptual coding of digital audio," *Proc. of the IEEE*, vol.88, pp. 451-513, April 2000.

[7] Z. Lu and W. A. Pearlman, "An efficient, low-complexity audio coder delivering multiple levels of quality for interactive application," *Proc. Workshop on Multimedia Signal Processing*, pp. 529-534, Redondo Beach, CA, December 1998.

[8] D. Sinha and A. Tewfik, "Low bit rate transparent audio compression using adapted wavelets," *IEEE Trans. Signal Processing,* vol. 41, pp. 3463 -3479, December 1993.

[9] H. S. Malvar, "Biorthogonal and nonuniform lapped transforms for transform coding with reduced blocking and ringing artifacts," *IEEE Trans. Signal Processing,* vol. 46, pp. 1043-1053, April 1998.

[10] H. S. Malvar, "Enhancing the performance of subband audio coders for speech signals," *Proc. ISCAS'98*, pp. 98-101, Monterey, CA, June 1998.

[11] H. S. Malvar, "A modulated complex lapped transform and its applications to audio processing," *Proc. ICASSP'99*, pp. 1421 -1424, Phoenix, AZ, March 1999.

[12] V. Goyal, M. Vetterli, and N. Thao, "Quantized overcomplete expansions in R^N: analysis, synthesis, and algorithms," *IEEE Trans. on Information Theory*, vol. 44, pp. 16-31, January 1988.

[13] A. S. Spanias, "Speech coding: a tutorial review," *Proc. of the IEEE*, vol. 82, pp. 1541-1582, October 1994.