A MODIFIED DISTORTION METRIC FOR AUDIO CODING

Vinay Melkote and Kenneth Rose

Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106
{melkote, rose}@ece.ucsb.edu

ABSTRACT

Current audio coding standards employ the modified discrete cosine transform (MDCT) where overlapped frames of audio are windowed and transformed to the frequency domain. Encoding parameters are chosen so as to minimize a distortion measure subject to a rate constraint. At the decoder, inverse transformation involves additional windowing and overlap-add of frames. An analysis of the time domain error in the reconstructed frame reveals that distortion metrics based solely on the MDCT domain error are in fact unable to capture the effects of windowing and overlap-add at the decoder. The main contribution of this paper is a modified distortion metric that does capture these effects via modified discrete sine transform analysis. When incorporated into an Advanced Audio Coder the proposed distortion metric significantly improves subjective quality of reconstructed audio.

Index Terms— audio coding, perceptual distortion, lapped transform, modified discrete sine transform

1. INTRODUCTION

Audio coding methods such as Advanced Audio Coding (AAC) [1] convert overlapped frames of audio to the frequency domain using a suitable transform which in many cases (including AAC) is the modified discrete cosine transform (MDCT) [2]-[4]. The transform coefficients are grouped into psychoacoustically relevant partitions, quantized and entropy coded. The quantization and coding parameters are chosen so that a distortion measure such as the noise-to-mask ratio (NMR) based on quantization error and masking thresholds (provided by a psychoacoustic model) is minimized subject to a bit-rate constraint. At the decoder the frame's quantized coefficients are inverse transformed and overlap-added with neighboring frames to reconstruct the time domain audio signal. This is illustrated in Fig. 1. Each vector $\underline{x}_k$ denotes a 'frame shift' of audio samples. Frame k, composed of $\underline{x}_k$ and $\underline{x}_{k+1}$, is used to obtain the vector of transform coefficients $\underline{X}_k$. This when quantized yields $\hat{\underline{X}}_k$ which is entropy coded losslessly and hence received intact at the decoder. The reconstruction $\hat{\underline{x}}_k$ is obtained by the overlap-add of the inverse transforms $\underline{z}_{k-1}$ and $\underline{z}_k$ of $\hat{\underline{X}}_{k-1}$ and $\hat{\underline{X}}_k$, respectively. Prior to the transformation at the encoder and post inverse transformation at the decoder, the frames are multiplied by a suitable window choice to avoid blocking effects. This operation can in fact be embedded in the transform (and its inverse) as is the case with MDCT (see Sec. 2) and is implicit in the corresponding stages of Fig. 1.

Note that the reconstructed frame k comprising of $\hat{\underline{x}}_k$ and $\hat{\underline{x}}_{k+1}$ has error contributions due to quantization of not just $\underline{X}_k$ but also $\underline{X}_{k-1}$ and $\underline{X}_{k+1}$. But current encoders such as the publicly available MPEG Verification Model (VM) [5] calculate distortion for each frame individually, i.e., using a metric of the form $D(\underline{X}_k, \hat{\underline{X}}_k)$ which ignores the effect of any decoder based operation such as overlap-add. Thus it is instructive to see if analysis (in the frequency domain) of the decoded time domain signal can capture these effects. To this end, consider applying the same transform and framing as in the encoder to the reconstructed time domain signal. The resulting transform coefficients are shown as $\hat{\underline{X}}_k^D$ in Fig. 1. The same metric as before could be used to define the "end-to-end" distortion $D(\underline{X}_k, \hat{\underline{X}}_k^D)$. It is observed that in the case of lapped orthogonal transforms (LOTs) [2], [6], to which class the MDCT belongs, $\hat{\underline{X}}_k^D = \hat{\underline{X}}_k$ and hence $D(\underline{X}_k, \hat{\underline{X}}_k) = D(\underline{X}_k, \hat{\underline{X}}_k^D)$. This is not true for other well known transforms including the discrete Fourier transform (DFT). The latter fact is rightly demonstrated in [7], where the authors using an audio encoder based on DFT of 50% overlapped frames show that $D(\underline{X}_k, \hat{\underline{X}}_k) \neq D(\underline{X}_k, \hat{\underline{X}}_k^D)$.

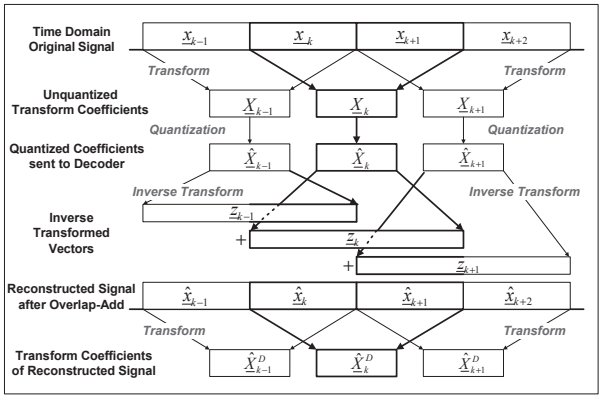

Fig. 1. Signal analysis in audio coding. The frequency domain reconstructed signal $\hat{\underline{X}}_k^D$ is added here to illustrate the discussion.

The overlap error components from neighboring frames are orthogonal to the MDCT basis vectors of the current frame. Thus distortion metrics based solely on MDCT domain error do not capture overlap-add effects. The error orthogonal to the MDCT bases can be analyzed using the modified discrete sine transform (MDST). Such analysis reveals that in addition to the overlap contributions, the orthogonal error has a component from quantization in the current frame itself due to the non-rectangular window used. In other words, the decoder based windowing leads to a spreading of quantization noise from the MDCT domain to the MDST domain. Since the human ear is sensitive to the magnitude of noise at any frequency rather than its projections only on cosine or sine bases, a modified distortion measure is proposed that accounts for the MDST domain

This work was supported in part by the University of California MICRO program, Applied Signal Technology Inc., Cisco Systems Inc., Dolby Laboratories Inc., Qualcomm Inc., and Sony Ericsson, Inc.

error. The fact that the windows used in these transforms are heavy centered and taper at the ends leads to the MDST domain error being dominated by the effect of decoder based windowing rather than overlap-add. Thus a simplified version of the distortion metric which accounts only for the window effect is implemented in the Two Loop Search for quantization and coding parameters of the MPEG VM AAC encoder [5]. Subjective tests indicate a preference for audio encoded in light of this modification rather than the usual NMR metric. Experiments are performed using both window choices, sine and Kaiser Bessel derived (KBD), available in the AAC standard. The advantages of one over the other with respect to the new metric are discussed.

2. BACKGROUND

We introduce here notation as well as relevant background information on MDCT with reference to Fig. 1. Segment $\underline{x}_k$ of the original signal and corresponding reconstruction $\underline{\hat{x}}_k$ are column vectors of M audio samples. The k^{th} original and reconstructed frames of length $2M$ are, respectively,

$$\mathbf{x}_k = \begin{bmatrix} \underline{x}_k \\ \underline{x}_{k+1} \end{bmatrix} \text{ and } \hat{\mathbf{x}}_k = \begin{bmatrix} \underline{\hat{x}}_k \\ \underline{\hat{x}}_{k+1} \end{bmatrix} \tag{1}$$

Thus frames are 50% overlapped. MDCT of $2M$ audio samples yields M coefficients and the $M \times 2M$ forward MDCT matrix is,

$$P = CH \tag{2}$$

$$\text{with} \qquad H = \begin{bmatrix} h(0) & 0 & \cdots & 0 \\ 0 & h(1) & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & h(2M-1) \end{bmatrix}_{2M \times 2M} \tag{3}$$

$$\text{and} \quad C = \left[\sqrt{\frac{2}{M}} cos\Big[\frac{\pi}{M} \Big(m + \frac{1}{2}\Big) \Big(n + \frac{M+1}{2}\Big) \Big] \right]_{\substack{M \times 2M \\ 0 \le m \le M-1,\ 0 \le n \le 2M-1}} \tag{4}$$

m and n in C are row and column indices, respectively. $h(n)$, a window of length $2M$, satisfies the constraints

$$h(2M-1-n) = h(n) \text{ and } h^2(n) + h^2(n+M) = 1 \tag{5}$$

The inverse MDCT (IMDCT) matrix is P^T and obtained by transposition. Information about window prototypes and the use of MDCT in audio coding can be found in [4]. We alternatively write P as,

$$P = [P_A \; P_B] \tag{6}$$

where P_A and P_B are $M \times M$ sub-matrices. Applying MDCT to the original signal one obtains

$$\underline{X}_k = P\mathbf{x}_k = P_A \underline{x}_k + P_B \underline{x}_{k+1} \tag{7}$$

We will also consider MDCT of the reconstructed signal:

$$\underline{\hat{X}}_k^D = P\hat{\mathbf{x}}_k = P_A \underline{\hat{x}}_k + P_B \underline{\hat{x}}_{k+1} \tag{8}$$

The vector $\underline{X}_k$ is quantized to $\underline{\hat{X}}_k$ and the quantization error is,

$$\underline{E}_k = \underline{X}_k - \underline{\hat{X}}_k \tag{9}$$

The vectors $\underline{z}_k$ in Fig. 1 are obtained by IMDCT,

$$\underline{z}_k = P^T \underline{\hat{X}}_k = \begin{bmatrix} P_A^T \\ P_B^T \end{bmatrix} \underline{\hat{X}}_k \tag{10}$$

Since the MDCT belongs to the class of LOTs it satisfies the following conditions [2],

$$PP^T = P_A P_A^T + P_B P_B^T = \mathbf{I} \tag{11}$$

$$\text{and} \qquad P \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} P^T = \mathbf{0} \tag{12}$$

$$\Rightarrow \; P_A P_B^T = \mathbf{0} = P_B P_A^T \tag{13}$$

where $\mathbf{0}$ and $\mathbf{I}$ are each $M \times M$ in dimension. The above conditions enable perfect reconstruction and time domain aliasing cancellation properties that are characteristic of LOTs.

The reconstruction segments $\underline{\hat{x}}_k$ and $\underline{\hat{x}}_{k-1}$ are formed by overlap-add of corresponding IMDCT vectors:

$$\underline{\hat{x}}_k = \begin{bmatrix} \mathbf{0} & \mathbf{I} \end{bmatrix} \underline{z}_{k-1} + \begin{bmatrix} \mathbf{I} & \mathbf{0} \end{bmatrix} \underline{z}_k = P_B^T \underline{\hat{X}}_{k-1} + P_A^T \underline{\hat{X}}_k \tag{14}$$

$$\underline{\hat{x}}_{k+1} = \begin{bmatrix} \mathbf{0} & \mathbf{I} \end{bmatrix} \underline{z}_k + \begin{bmatrix} \mathbf{I} & \mathbf{0} \end{bmatrix} \underline{z}_{k+1} = P_B^T \underline{\hat{X}}_k + P_A^T \underline{\hat{X}}_{k+1} \tag{15}$$

where $\mathbf{0}$ and $\mathbf{I}$ are of dimensions $M \times M$. Substituting into (8) we obtain

$$\underline{\hat{X}}_k^D = P_A P_B^T \underline{\hat{X}}_{k-1} + (P_A P_A^T + P_B P_B^T) \underline{\hat{X}}_k + P_B P_A^T \underline{\hat{X}}_{k+1} \tag{16}$$

$$\text{and by (11), (13)} \qquad \underline{\hat{X}}_k^D = \underline{\hat{X}}_k \tag{17}$$

which subsequently leads to,

$$D(\underline{X}, \underline{\hat{X}}_k) = D(\underline{X}, \underline{\hat{X}}_k^D) \tag{18}$$

Thus a metric such as NMR defined as quantization noise in the MDCT coefficients divided by the masking thresholds, is not altered by decoder based operations such as overlap-add and hence is deficient in its ability to capture corresponding psychoacoustic effects. The derivation of (18) has not explicitly used the MDCT kernel but the more general LOT properties (11) and (12). Hence (18) holds true for other LOTs also. Note that, as evidenced by the system of [7], (18) is not valid for all perfect reconstruction systems employing overlapped transforms.

3. DISTORTION IN THE MDCT AND MDST DOMAINS

We now analyze the time domain error in a reconstructed frame. From (2), taking the MDCT of frame $\mathbf{x}_k$ implies applying the cosine based transform C to the 'windowed' frame $H\mathbf{x}_k$. The time domain reconstruction error in the k^{th} frame is $\mathbf{x}_k - \hat{\mathbf{x}}_k$. The 'windowed' error is

$$\mathbf{e}_k = H[\mathbf{x}_k - \hat{\mathbf{x}}_k] = H \begin{bmatrix} \underline{x}_k - \underline{\hat{x}}_k \\ \underline{x}_{k+1} - \underline{\hat{x}}_{k+1} \end{bmatrix} \tag{19}$$

By the perfect reconstruction property, absent quantization, IMDCT followed by overlap-add yields back the original samples:

$$\underline{x}_k = P_B^T \underline{X}_{k-1} + P_A^T \underline{X}_k \tag{20}$$

$$\underline{x}_{k+1} = P_B^T \underline{X}_k + P_A^T \underline{X}_{k+1} \tag{21}$$

Substituting (14), (15) and the above in (19) and using (9) we have,

$$\mathbf{e}_k = H \begin{bmatrix} P_B^T \underline{E}_{k-1} + P_A^T \underline{E}_k \\ P_B^T \underline{E}_k + P_A^T \underline{E}_{k+1} \end{bmatrix} \tag{22}$$

$$(2) \Rightarrow \; C\mathbf{e}_k = P \begin{bmatrix} P_B^T \underline{E}_{k-1} + P_A^T \underline{E}_k \\ P_B^T \underline{E}_k + P_A^T \underline{E}_{k+1} \end{bmatrix} \tag{23}$$

$$(6), (11), (13) \Rightarrow \; C\mathbf{e}_k = \mathbf{0}\underline{E}_{k-1} + \mathbf{I}\underline{E}_k + \mathbf{0}\underline{E}_{k+1} \tag{24}$$

This indicates that the cosine basis vectors (rows of C) are orthogonal to error components in $\mathbf{e}_k$ that result from the overlap of $\hat{\mathbf{x}}_k$ with neighboring frames. On the other hand these components can be captured using a basis set that is orthogonal to the row space of C. The sine transform S given by

$$S = \left[\sqrt{\frac{2}{M}} sin\Big[\frac{\pi}{M}\Big(m+\frac{1}{2}\Big)\Big(n+\frac{M+1}{2}\Big)\Big]\right]_{M\times 2M} \quad (25)$$

is one possible orthogonal basis set, i.e., $SC^T = \mathbf{0}$. Note that both C and S are of rank M and together form a 'complete basis' for the $2M$ dimensional space. By straightforward manipulations, it can be shown that

$$C^TC + S^TS = 2\mathbf{I} \quad (26)$$

$$\Rightarrow \mathbf{e}_k^T\mathbf{e}_k = \frac{1}{2}\Big[(C\mathbf{e}_k)^T(C\mathbf{e}_k) + (S\mathbf{e}_k)^T(S\mathbf{e}_k)\Big] \quad (27)$$

Thus the time domain error in a windowed frame can be completely analyzed using both cosine and sine transforms. Define $\underline{\mathcal{E}}_k = S\mathbf{e}_k$. By (22) ,

$$\underline{\mathcal{E}}_k = SH\begin{bmatrix}P_B^T\\ \mathbf{0}\end{bmatrix}\underline{E}_{k-1} + SH\begin{bmatrix}P_A^T\\ P_B^T\end{bmatrix}\underline{E}_k + SH\begin{bmatrix}\mathbf{0}\\ P_A^T\end{bmatrix}\underline{E}_{k+1} \quad (28)$$

$$= P_S\begin{bmatrix}P_B^T\\ \mathbf{0}\end{bmatrix}\underline{E}_{k-1} + P_SP^T\underline{E}_k + P_S\begin{bmatrix}\mathbf{0}\\ P_A^T\end{bmatrix}\underline{E}_{k+1} \quad (29)$$

where paralleling the treatment of MDCT we define the MDST matrix as

$$P_S = SH \quad (30)$$

The error $\underline{\mathcal{E}}_k$ will be referred to as the MDST domain error, as it is the MDST of the actual (not windowed) time domain error $\mathbf{x}_k - \hat{\mathbf{x}}_k$. Note that despite $SC^T = \mathbf{0}$,

$$P_SP^T = SH^2C^T \neq \mathbf{0} \quad (31)$$

for windows not satisfying $H^2 = \mathbf{I}$. A rigorous proof of the prior statement is left out for conciseness. It can specifically be verified for the sine and KBD windows specified by the AAC standard [1]. Thus, by (29), in addition to quantization error contributions from neighboring frames, part of the MDST domain error for a frame, i.e., $P_SP^T\underline{E}_k$ results from quantizing the MDCT coefficients of the concerned frame itself. In other words, the non-rectangular window used in these transforms results in 'spreading' the MDCT quantization error into the MDST domain.

In the AAC framework, prior to quantization the M MDCT coefficients of a frame are divided into partitions called scale factor bands (SFBs) each of which is associated with a scale factor (SF) and Huffman code book (HCB). The SF and HCB for each SFB is selected from a finite set specified by the standard. The SF choice decides the quantization granularity for MDCT coefficients in the SFB. HCB choices determine the number of coding bits. The AAC encoder uses an iterative search to find the choice of parameters that minimize a distortion metric subject to the prescribed rate. The distortion metric used needs to properly account for quantization effects (i.e., choice of SF) in various SFBs. The common metric of choice is the NMR, which is defined for SFB i of frame k as

$$NMR_{k,i} = \frac{\sum_{j\in SFB\ i}\underline{E}_k^2(j)}{T_i} \quad (32)$$

Here $\underline{E}_k(j)$ is the j^{th} element of $\underline{E}_k$ and the masking threshold T_i for each SFB is provided by a psychoacoustic model. It is well known that the human ear is sensitive to the spectral magnitude rather than any one individual orthogonal component (sine or cosine). Thus a distortion metric that accounts for the magnitude of error in different frequency bins, rather than its projection only in the MDCT domain, yields a better comparison of the effects of quantization in different coding bands. Therefore we propose an enhanced distortion measure, NMR$^+$, which, in addition to the MDCT error, accounts for the error $\underline{\mathcal{E}}_k$ (29) present in the MDST domain. Specifically,

$$NMR^+_{k,i} = \frac{\sum_{j\in SFB\ i}[\underline{E}_k^2(j) + \underline{\mathcal{E}}_k^2(j)]}{T'_i} \quad (33)$$

It follows from (29) that NMR$^+$ depends on the MDCT errors of neighboring frames and hence cannot be incorporated into an encoder that analyzes each frame separately, e.g., the MPEG VM [5]. Note that the masking thresholds in (32) and (33) are not the same. Usually the psychoacoustic model performs an FFT of the windowed frame and finds thresholds in different bands. The FFT thresholds are eventually scaled to reflect the energy in the MDCT domain. In the case of (33) the threshold T'_i should additionally account for MDST domain energy.

As suggested by (31), the error $\underline{E}_k$ propogates to the MDST domain through H^2 (or $h^2(n)$) which is plotted in Fig. 2 for sine and KBD windows. The KBD window provides reduced overlap. Under the assumption that all M elements of $\underline{E}_{k-1}$, $\underline{E}_k$ and $\underline{E}_{k+1}$ are independent random variables with equal variance, (29) can be used to calculate the variance of elements in $\underline{\mathcal{E}}_k$ (the MDST domain error) for any specific window choice. The MDST domain error turns out to have the same variance as $\underline{E}_k$ suggesting that the orthogonal domain error is as important to account for as the MDCT domain error. In case of the sine window the errors $\underline{E}_{k-1}$ and $\underline{E}_{k+1}$ can be shown to contribute 25% each to the MDST domain error of the k^{th} frame while the remaining 50% is due to $\underline{E}_k$. For the KBD window only 15% of the MDST domain error is due to each of the neighboring frames and 70% due to MDCT quantization in the current frame. Therefore we approximate $\underline{\mathcal{E}}_k$ by $\hat{\underline{\mathcal{E}}}_k = P_SP^T\underline{E}_k$ and NMR$^+$ by,

$$NMR^+_{k,i} \approx \frac{\sum_{j\in SFB\ i}[\underline{E}_k^2(j) + \hat{\underline{\mathcal{E}}}_k^2(j)]}{T'_i} \quad (34)$$

This simplified NMR$^+$ accounts for most of $\mathbf{e}_k^T\mathbf{e}_k$ in (27), especially in the case of the KBD window.

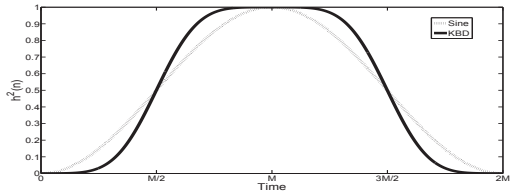

Fig. 2. Comparison of the squares of sine and KBD windows. The KBD window results in reduced overlap error due to faster tapering.

Since this approximate NMR$^+$ depends only on the MDCT error in the current frame itself, it can be incorporated in an encoder like the MPEG VM by simple substitution of the usual NMR. Whenever the SF (and hence the MDCT error $\underline{E}_k$) for an SFB is altered, $\hat{\underline{\mathcal{E}}}_k = P_SP^T\underline{E}_k$ is re-computed and the NMR$^+$ value updated.

Multiplication with the $M \times M$ matrix P_SP^T is performed efficiently by recognizing the fact that, for good window choices such

as sine and KBD, this matrix has its most dominant elements close to the principal diagonal. This band-like structure of $P_S P^T$ is the result of critically located spectral zeroes in the case of the sine window and very good anti-aliasing (side lobe reduction) properties in the case of KBD. Therefore for any j, $\hat{\underline{\mathcal{E}}}_k(j)$ is constructed from elements of $\underline{E}_k$ with indices in a very small neighborhood of j. When the sine window is used it can be shown that $\hat{\underline{\mathcal{E}}}_k(j)$ depends exactly on $\underline{E}_k(j+1)$ and $\underline{E}_k(j-1)$. In case of the KBD window 4 to 6 $\underline{E}_k$ coefficients are sufficient to calculate each $\hat{\underline{\mathcal{E}}}_k(j)$. Thus the M multiplications (and additions) to calculate each $\hat{\underline{\mathcal{E}}}_k(j)$ can be reduced to a modest number. Efficient computation of MDST coefficients from MDCT coefficients has been used previously, for example in [8] to estimate the power spectrum of the frame. Since the sine and cosine bases in P_S and P are uniformly spaced in frequency, most of the rows of $P_S P^T$ (except a few at the top and bottom ends) are shifted repetitions of each other enabling efficient storage of the matrix.

4. EXPERIMENTS

The MPEG VM [5] implementation of the AAC encoder uses a 'Two Loop Search' where an inner loop finds the SFs that achieve a target NMR while an outer loop monitors the rate. The (approximate) modified metric NMR$^+$ can be used in lieu of the NMR in the inner loop. The two implementations are respectively termed VM-NMR and VM-NMR$^+$. The encoders were constrained to work only in the 'LONG' window mode of AAC (i.e., M was fixed at 1024). 5 audio files each at sampling rate 44.1kHz were encoded at a bit-rate of 48kbps by both methods and with both window choices, sine and KBD. Blind listening tests in the A-B style were conducted with 15 subjects, with access to the original audio file and randomly ordered samples encoded by the two methods when using the same window. They could switch near instantaneously between any of these 3 files. Since the choice of bit-rate is relatively high, the original helps listeners to identify artifacts in either coded sample. They could pick one as preferred or state that they were unable to decide. The results of the tests are given in Table 1. M1, M2 and M3 indicate instrumental music samples harpsichord, organ and accordion, respectively. S1 corresponds to male german speech and S2 is female english speech. Considerable subjective gains of using the new measure are seen with either window choice. Only in the case of the accordion piece there was no clear preference.

Audio Sample	sine			KBD		
	VM-NMR$^+$	VM-NMR	No Pref	VM-NMR$^+$	VM-NMR	No Pref
M1	58.33	0	41.66	75	0	25
M2	50	0	50	66.67	0	33.33
M3	25	25	50	41.67	33.33	25
S1	91.67	8.33	0	100	0	0
S2	91.67	8.33	0	91.67	8.33	0

Table 1. Subjective comparison tests of VM-NMR and VM-NMR$^+$ with both sine and KBD windows: figures indicate the percentage of listeners who preferred audio encoded using corresponding method.

5. GENERALIZATION TO OTHER LOT BASED CODECS

We consider here audio coding with generic LOT matrices of dimensions $M \times 2M$. Additionally, let us suppose the forward LOT matrix P of dimensions $M \times 2M$ is decomposable into the form $C'H$ as in (2), with the rows of C' being orthogonal basis vectors spanning an M dimensional sub-space of the $2M$ dimensional space. Using a matrix S' with rows as orthogonal basis vectors of the complementary M dimensional sub-space, similar to the definition of the MDST, we could now define corresponding P_S and hence proceed to a time domain error analysis similar to (29). Thus the use of a distortion measure similar to NMR$^+$ is conceivable even in such generic encoders, although perceptual considerations may need to be revisited in light of the actual choice of transform.

6. CONCLUSIONS

Distortion metrics for audio coding based solely in the MDCT domain of a frame are invariant to necessary windowing and overlap-add operations at the decoder. An analysis of the time domain error of a frame reveals that the corresponding error components are orthogonal to the MDCT basis vectors. An enhanced distortion measure is suggested that incorporates these components via MDST domain analysis. Subjective tests, using a simplified version of this metric accounting only for the windowing effects, evidence a preference for audio encoded by employing this modification. The improved metric captures the magnitude of the frequency domain error rather than its projection onto the cosine basis vectors of MDCT. Future improvements involve implementing the proposed metric without simplifying approximations, in an audio encoder that analyzes multiple frames at a time.

7. REFERENCES

[1] ISO/IEC std, "Information technology - generic coding of moving pictures and associated audio," *ISO/IEC JTC1/SC29 13818-7:1997(E)*, 1997.

[2] H. S. Malvar, "Lapped transforms for efficient transform/subband coding," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 38, pp. 969–978, Jun 1990.

[3] J. P. Princen, A. W Johnson, and A. B. Bradley, "Subband/transform coding using filter bank designs based on time domain aliasing cancellation," in *Proc. IEEE ICASSP*, Apr 1987, pp. 2161–2164.

[4] 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, Jul 1997.

[5] *MPEG Verification Model*, `http://standards.iso.org/ittf/PubliclyAvailableStandards/ISO_IEC_14496-5_2001_Software_Reference`.

[6] H. S. Malvar and D. H. Staelin, "The LOT: transform coding without blocking effects," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 37, pp. 553–559, Apr 1989.

[7] R. Der, P. Kabal, and W-Y. Chan, "Rate-distortion allocation for time-frequency dependent audio coding," in *Proc. IEEE ICASSP*, 2005, pp. 197–200.

[8] C. Cheng, "Method for estimating magnitude and phase in the MDCT domain," in *Proc. 116th AES convention*, 2004, preprint 6091.