AN MDCT DOMAIN FRAME-LOSS CONCEALMENT TECHNIQUE FOR MPEG ADVANCED AUDIO CODING

Sang-Uk Ryu and Kenneth Rose

Department of Electrical and Computer Engineering
University of California, Santa Barbara, CA 93106-9560, USA
Email: {sang, rose}@ece.ucsb.edu

ABSTRACT

We propose a frame loss concealment technique for decoders compatible with MPEG advanced audio coding (AAC). The spectral information of the lost frame is estimated in the modified discrete cosine transform (MDCT) domain via efficient techniques that are tailored to individual source signal components: In noise-like spectral bins the MDCT coefficients are obtained by shaped-noise insertion, while coefficients in tone-dominant bins are estimated by frame interpolation followed by a refinement procedure so as to optimize the fit of the concealed frames with neighboring frames. Experimental results demonstrate that the proposed technique offers performance superior to techniques adopted in commercial AAC decoders.

Index Terms— Audio Coding, AAC, Frame Loss Concealment, Noise Insertion, Prediction.

1. INTRODUCTION

Frame loss concealment (FLC) can be viewed as the problem of estimating a lost frame, while using all available information, such that the generated output fits, as smoothly as possible, between the neighboring frames. While various approaches for audio FLC have been proposed, they typically fall at either extreme of the tradeoff between post-concealment audio quality and implementation cost. For example, replacing the lost frame with either silence or the previous frame represents one extreme offering low complexity but generally poor performance [1]. Advanced techniques based on source modeling typically fall at the other extreme, as they have been known to produce better quality at high or even prohibitive implementation cost in terms of computational complexity, memory, and delay [2]-[4].

In this paper, we propose an efficient MDCT domain FLC technique, which is tailored to and effectively exploits the characteristics of individual components in the source signal. Specifically, in noise-like MDCT bins the MDCT coefficients are generated by shaped-noise insertion, while in tone-dominant bins a new MDCT estimation technique is employed to adequately handle tonal components. At the spectral positions where MDCT coefficients are completely determined by the characteristics of the underlying sinusoid, they are coarsely estimated by MDCT bin-wise frame interpolation. Refinement is then achieved by means of multiplicative correction. The correction factor is determined by observing the spectral characteristics of the reconstructed frames. Surprisingly, the consideration of sinusoidal energy across the lost and adjacent frames, and imposition of constraints on sinusoidal energy evolution, yield exactly two candidates for the multiplicative correction factor for all coefficients, and the final value selection is made by choosing the value that maximizes the "tone-like" characteristics.

The proposed algorithm is implemented for different decoder setups, mainly in terms of allowed delay, and performance is compared with two techniques that have been adopted in commercial AAC applications – shaped-noise insertion and subband domain prediction. Performance evaluation results, in terms of both subjective quality of post-concealment audio and computational complexity, ascertain that the proposed concealment techniques not only offer substantial quality improvement, especially for tonal signals, but also achieve robust concealment quality for a broad variety of audio signals, while maintaining lower computational complexity than the subband domain prediction approach.

2. PRIOR WORK ON AAC FRAME LOSS CONCEALMENT

For AAC, the FLC problem is posed as that of estimating the MDCT coefficients of a lost frame from the coefficients in the previous and/or future frames. A number of MDCT estimation techniques have been proposed with differing performance in terms of concealment quality and computational complexity. A low-complexity solution with moderate concealment quality, shaped-noise insertion (in MDCT domain), has been adopted in the FLC module of the *aacPlus* decoder for the 3rd generation partnership project (3GPP) [5]. It estimates the MDCT coefficients of the lost frame by fitting a noise model to the signal around the lost frame. While this technique works effectively for noise-like signals at little cost in computation, its performance degrades considerably for audio signals with dominant tonal components.

To enhance concealment quality for general audio, the need for a complementary technique to adequately handle sinusoidal components in the source signal was recognized. Subband domain prediction was adopted in [6][7] as an effective mechanism to estimate the MDCT coefficients for sinusoidal components. The MDCT coefficients of the last two frames are first split into equal width frequency bands and converted to subband samples via an IMDCT of appropriate order. A linear predictor is determined from the statistics of the derived subband samples and predicts subband samples associated with the lost MDCT coefficients. The predicted subband samples are then transformed back to the original MDCT domain. An appropriate control algorithm is incorporated, which is specifically designed to decide whether the given frequency band is tone- or noise-dominant and to switch between the two complementary techniques.

This work is supported in part by the University of California MICRO Program, Applied Signal Technology, Inc., Dolby Laboratories, Inc., Mindspeed Technologies, Inc., and Qualcomm, Inc.

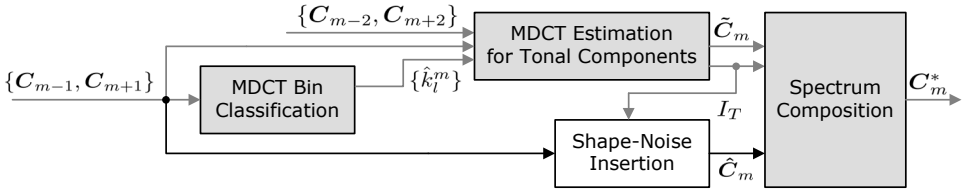

Fig. 1. Block diagram for the proposed overall FLC technique.

3. THE PROPOSED FRAMEWORK AND APPROACH

Although subband domain prediction can adequately handle sinusoidal components of the source signal, it also requires substantial computation for converting MDCT coefficients to subband samples and vice versa. In this section, we propose an efficient MDCT domain FLC technique. Fig.1 shows the framework of the proposed FLC scheme.

3.1. MDCT Bin Classification

As is required by the proposed framework, a functional module to classify MDCT bins into the noise-like and tone-dominant bins is needed prior to MDCT estimation. Suppose that tonal components of the lost frame are modeled with L sinusoids, and the $l\text{-}th$ sinusoid has frequency parameter ω_l. For each spectral peak position $k_l = nint(\omega_l M/\pi)$ where M is the number of MDCT bins, an index subset of the tone-dominant MDCT bins is defined as

$$I_l = \{k \big| |k - k_l| \le \delta_l/2,\ 0 \le k < M\}, \tag{1}$$

where $\delta_l = \min\{W, |k_l - k_{l-1}|, |k_{l+1} - k_l|\}$, and where W is the main-lobe width of the analysis window measured in MDCT bins. The index set of the tone-dominant MDCT bins across all bands, I_T, is defined as the union of the band specific index subsets.

It is clear from the above that MDCT bin classification is straightforward once the exact positions of the spectral peaks corresponding to the underlying sinusoids are known. In practice, approximate peak positions of the sinusoids are typically determined by searching for local maxima in the power spectrum. Unfortunately, the power spectrum of the lost frame is not available during concealment, leading us to adopt the approximation

$$\hat{P}_m(k) = C_{m-1}^2(k) + C_{m+1}^2(k), \tag{2}$$

where the MDCT coefficients of the previous and next frames are denoted as $C_{m-1}(k)$ and $C_{m+1}(k)$, respectively.

Peak detection is implemented by searching for local maxima in the approximated power spectrum, but should be refined by a procedure for screening perceptually irrelevant or spurious peaks. For this purpose, we employ a commonly adopted heuristic, where only local maxima satisfying certain conditions are declared spectral peaks. First, peak detection is applied to a limited spectral range to avoid less relevant sinusoidal components outside the limit. Second, only local maxima that exceed a relative threshold to the absolute maximum of the power spectrum can be considered meaningful spectral peaks. Peaks satisfying the second condition are then sorted in descending order of magnitude, and a pre-specified number of top ranking maxima are classified as tonal peaks.

3.2. MDCT Estimation for Sinusoidal Components

For an explicit relationship between parametric representation of sinusoidal components and MDCT coefficients, let us first consider a single stationary sinusoid, i.e.,

$$x_m(n) = A_l^m \cos(\omega_l^m n + \phi_l^m), \tag{3}$$

where $A_l^m = A_l$, $\omega_l^m = \omega_l$, and $\phi_l^m = \omega_l M + \phi_l^{m-1}$. Employing the sine analysis window, the $m\text{-}th$ frame's MDCT coefficients on index set I_l can be explicitly derived as

$$C_m(k) \cong G_l(k) \sin\Big(\frac{\pi}{2}k + \psi_l^m\Big),\ \ k \in I_l, \tag{4}$$

$$G_l(k) = -\frac{A_l\sqrt{N}}{4\pi} \frac{\sin\{\pi(f_l - k)\}}{(f_l - k)(f_l - k - 1)}, \tag{5}$$

$$\psi_l^m = \phi_l^m - \frac{\pi}{4} + \frac{N-1}{N}\pi f_l, \tag{6}$$

where $f_l = \omega_l M/\pi$ and $N = 2M$. Let $\bar{C}_m(k)$ be the MDCT estimate obtained by (MDCT bin-wise) frame interpolation, i.e.,

$$\bar{C}_m(k) = \frac{1}{2}\big\{C_{m-1}(k) + C_{m+1}(k)\big\}. \tag{7}$$

Substituting the expressions for $C_{m-1}(k)$ and $C_{m+1}(k)$ from (4) into (7), we obtain

$$\bar{C}_m(k) \cong \cos(\pi f_l) C_m(k),\ \ k \in I_l. \tag{8}$$

From (8), we observe that MDCT estimates by frame interpolation are a down-scaled version of the original coefficients by the *constant* multiplicative factor $\cos(\pi f_l)$. Clearly, the interpolated MDCT coefficients can be further refined by applying a *constant* multiplicative correction for all coefficients in I_l. Let α_l be the multiplicative correction factor for the interpolated MDCT coefficients. Then, the refined MDCT estimate can be expressed as

$$\tilde{C}_m(k) = \alpha_l \bar{C}_m(k),\ \ k \in I_l. \tag{9}$$

It can be seen from (9) that MDCT coefficients dominated by a single sinusoid are effectively estimated by frame interpolation followed by multiplicative correction. In the general case that tonal components in the source signal are modeled with multiple sinusoids, this technique is performed per index subset, and in this manner all MDCT coefficients for sinusoidal components can be effectively estimated.

A crucial component of the proposed technique is to determine the multiplicative correction factor of the interpolated MDCT estimates within an index subset. It will be determined by observing the spectral characteristics of the reconstructed frames in terms of the sinusoidal energy as well as spectral energy distribution within the index subset. To facilitate it, the power spectrum of the reconstructed frames needs to be computed locally for the MDCT bins in the index

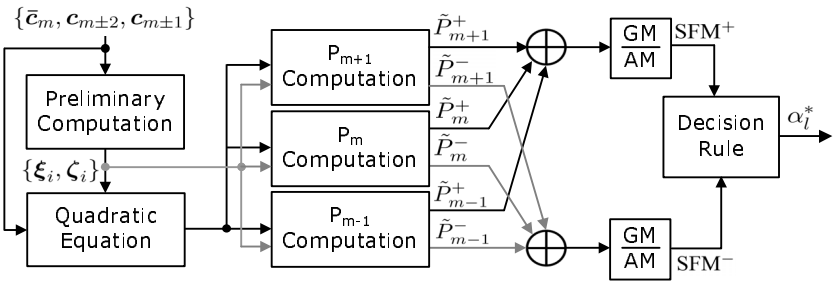

Fig. 2. Block diagram for the correction factor computation.

subset. As shown in [8][9], the power spectrum can be equivalently derived from the squared sum of the MDCT and MDST (modified discrete sine transform) coefficients. Also, the MDST coefficients can be computed from a set of the MDCT coefficients, i.e.,

$$\boldsymbol{S}_m = \mathbf{A}_1\boldsymbol{C}_{m-1} + \mathbf{A}_2\boldsymbol{C}_m + \mathbf{A}_3\boldsymbol{C}_{m+1}, \tag{10}$$

where $\boldsymbol{C}_r$ and $\boldsymbol{S}_r$ are the respective MDCT and MDST coefficient vectors for frame $r \in \{m-1, m, m+1\}$, and $\mathbf{A}_i$ is constant matrix specified in [9] for $i = 1, 2, 3$. The above MDST computation can be simplified by exploiting the stationary sinusoidal model and the observation that MDCT magnitudes at all other bins are relatively negligible. Employing $C_r(k) \cong 0$ for $k \notin I_l$ and simple vector notation for MDCT coefficients within an index set (e.g., $\boldsymbol{c}^l_r$ is the subvector of $\boldsymbol{C}_r$ that falls in I_l), we obtain

$$\boldsymbol{s}^l_m \cong \mathbf{A}^l_1\boldsymbol{c}^l_{m-1} + \mathbf{A}^l_2\boldsymbol{c}^l_m + \mathbf{A}^l_3\boldsymbol{c}^l_{m+1}, \tag{11}$$

where we denote the $(\delta_l + 1)$ dimensional sub-matrix sampled from $\mathbf{A}_i$ on the index grid $I_l \times I_l$ as $\mathbf{A}^l_i$ for $i = 1, 2, 3$.

Armed with the above MDST derivation and the power spectrum approximation, we now describe the correction factor computation procedure. The correction factor is determined based on the primary assumption that sinusoidal energy evolves smoothly over frames. Given the MDCT estimate $\bar{\boldsymbol{c}}^l_m$ and a correction factor α_l, the MDST coefficients on I_l can be approximated as

$$\tilde{\boldsymbol{s}}^l_m \cong \mathbf{A}^l_1\boldsymbol{c}^l_{m-1} + \mathbf{A}^l_3\boldsymbol{c}^l_{m+1} + \alpha_l\mathbf{A}^l_2\bar{\boldsymbol{c}}^l_m = \boldsymbol{\xi}^l_2 + \alpha_l\boldsymbol{\zeta}^l_2, \tag{12}$$

where $\boldsymbol{\zeta}^l_2 = \mathbf{A}^l_2\bar{\boldsymbol{c}}^l_m$ and the sum of the first and second terms in (12) is compactly denoted by $\boldsymbol{\xi}^l_2$. Hereafter, we will omit the super/subscript l employed to distinguish the index subset. The m-th frame's sinusoidal energy measured on I is hence approximately given as functions of α

$$\tilde{E}_m(\alpha) \cong \alpha^2|\bar{\boldsymbol{c}}_m|^2 + |\tilde{\boldsymbol{s}}_m|^2 = \alpha^2|\bar{\boldsymbol{c}}_m|^2 + |\boldsymbol{\xi}_2 + \alpha\boldsymbol{\zeta}_2|^2. \tag{13}$$

Adopting a similar procedure and notation for the MDST coefficients of adjacent frames, we obtain:

$$\tilde{\boldsymbol{s}}_{m-1} \cong \mathbf{A}_1\boldsymbol{c}_{m-2} + \mathbf{A}_2\boldsymbol{c}_{m-1} + \alpha\mathbf{A}_3\bar{\boldsymbol{c}}_m = \boldsymbol{\xi}_1 + \alpha\boldsymbol{\zeta}_1, \tag{14}$$

$$\tilde{\boldsymbol{s}}_{m+1} \cong \mathbf{A}_2\boldsymbol{c}_{m+1} + \mathbf{A}_3\boldsymbol{c}_{m+2} + \alpha\mathbf{A}_1\bar{\boldsymbol{c}}_m = \boldsymbol{\xi}_3 + \alpha\boldsymbol{\zeta}_3, \tag{15}$$

and the corresponding sinusoidal energies are

$$\tilde{E}_{m-1}(\alpha) \cong |\boldsymbol{c}_{m-1}|^2 + |\boldsymbol{\xi}_1 + \alpha\boldsymbol{\zeta}_1|^2, \tag{16}$$

$$\tilde{E}_{m+1}(\alpha) \cong |\boldsymbol{c}_{m+1}|^2 + |\boldsymbol{\xi}_3 + \alpha\boldsymbol{\zeta}_3|^2. \tag{17}$$

It is clear from (13), (16), and (17) that the correction factor impacts the sinusoidal energies of three consecutive frames, and the energy distribution can be controlled by adjusting this correction factor. Therefore, the correction factor can be determined to ensure smooth energy evolution,

$$\tilde{E}_m(\alpha) = \frac{1}{2}\big\{\tilde{E}_{m-1}(\alpha) + \tilde{E}_{m+1}(\alpha)\big\}. \tag{18}$$

Substituting (13), (16), and (17) into (18), the above interpolation yields an expression that is quadratic in α. Hence, for the given MDCT estimate $\bar{\boldsymbol{c}}_m$, there exist two candidates (with opposite signs as verified by simple algebra) for the multiplicative correction factor. One of them will yield MDCT estimates that are *sign-matched* with the original, so that the reconstructed frames will possess spectral characteristics similar with that of the correctly decoded frames, i.e., a strong spectral peak. However, the other candidate could produce *sign-mismatched* MDCT estimates, which will result in significant spectral leakage around the peak position. These spectral features provide a means to distinguish between the sign-matched and mismatched cases and thereby choose the proper correction factor. To concretize this idea, we select as quantitative measure the spectral flatness measure (SFM), which is defined here as the ratio of the geometric mean to the arithmetic mean of power spectrum coefficients in I_l, i.e., $\text{SFM}_l = \text{GM}_l/\text{AM}_l$, where the geometric and arithmetic means are computed as

$$\text{GM}_l = \Big\{\prod_{k\in I_l}\tilde{P}_m(k)\Big\}^{1/|I_l|} \quad \text{and} \quad \text{AM}_l = \frac{1}{|I_l|}\sum_{k\in I_l}\tilde{P}_m(k). \tag{19}$$

Given the observation that the sign-mismatched case produces considerably flatter spectrum, we select as correction factor the candidate that minimizes the spectral flatness. Denoting the two candidates as α^+_l and α^-_l, the decision rule is described as

$$\alpha^*_l = \begin{cases} \alpha^+_l & \text{if } \text{SFM}^+_l < \text{SFM}^-_l \\ \alpha^-_l & \text{otherwise,} \end{cases} \tag{20}$$

where the SFMs associated with α^+_l and α^-_l are denoted by SFM^+_l and SFM^-_l, respectively. We note further that the spectral flatness considerations are extendible to the neighboring frames. When the sign-mismatched MDCT estimates are used in the power spectrum approximation of the previous and next frames, similar spectral flatness features will be observed there. Therefore, more robust decision can be made by taking into account the power spectra of the adjacent frames. Fig.2 depicts the overall correction factor computation procedure involving power spectra of three consecutive frames.

4. EXPERIMENTS AND OBSERVATIONS

In order to evaluate the performance of the proposed concealment technique, we implemented it within different decoder setups. First,

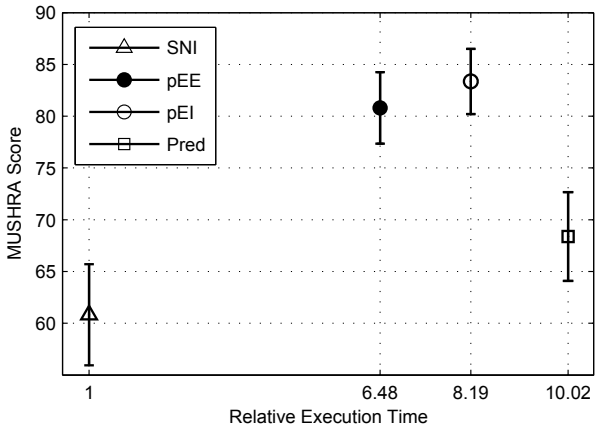

Fig. 3. Average MUSHRA scores with 95 % confidence intervals versus average execution time for concealment methods: shaped-noise insertion (SNI); SNI complemented with subband domain prediction (Pred); two variants of the proposed technique - with energy interpolation (pEI) and energy extrapolation (pEE). Averaged MUSHRA scores for the hidden reference and two anchors are 97, 39, and 64, respectively.

we considered the case where the AAC decoder allows two look-ahead frames so that the correction factor can be determined using energy interpolation of (18). The second implementation accounted for decoder configuration of one look-ahead frame buffer as in the 3GPP's AAC-FLC module and employed energy extrapolation $\tilde{E}_m(\alpha) = \tilde{E}_{m-1}(\alpha)$. As an important parameter involved in the tradeoff between the concealment quality and the required computation, the peak detection in the MDCT bin classification was restricted to the spectral range of 10kHz. We selected the local maxima within 60dB of the absolute maximum of the power spectrum, and 20 peaks having the highest magnitude were classified as meaningful peaks.

Various types of MONO audio sequences were selected from EBU SQAM [10] and the test items used in the MPEG-4 HE-AAC verification test [11]. The sampled sequences consist of three pieces of single instrumental sound, three pieces of pop music, and four singing vocal and speech pieces. The audio was AAC-encoded at the bit-rate of 64 kbps, and the encoded frames were randomly dropped at 10% frame loss rate. The dropped frames were concealed by the above two implementations of the proposed AAC-FLC scheme as well as shaped-noise insertion with/without the complement of subband domain prediction. To evaluate the subjective quality of post-concealment audio, we carried out the multi-stimulus test with hidden reference and anchors (MUSHRA), including low-pass filtered anchors with 3.5 and 7 kHz bandwidth [12]. The listening tests were performed by ten listeners. As another important measure of FLC system performance, we evaluated computational complexity of all competing techniques in terms of the execution time spent on the concealment process. Since shaped-noise insertion is the simplest, its execution time was used as complexity benchmark. In other words, we normalized the execution time of each concealment algorithm by that of shaped-noise insertion.

The MUSHRA scores and the execution times for concealment by each method were averaged over all test items, and the mean values are marked in Fig.3 to indicate the performances. We observed that shaped-noise insertion offers unsatisfactory concealment quality for most types of audio (except for complex pop music). Complementing it with subband domain prediction improves concealment quality with modest gains (about 9 points of MUSHRA score), while consuming a significant amount of additional computation (about 10 times that of noise insertion only). The proposed FLC techniques outperform subband domain prediction in terms of both quality and complexity and could be an efficient solution for practical applications. Moreover, the energy extrapolation variant of the proposed approach may be the practical favorite among the two, as it consumes a moderate amount of computation and involves less delay, while achieving concealment quality that slightly trails that of the energy interpolation variant.

5. CONCLUSION

We proposed an advanced frame loss concealment technique for the AAC decoder. An efficient technique for estimating the MDCT coefficients of sinusoidal components was developed in terms of frame interpolation and multiplicative refinement. The correction factor was determined by observing the spectral characteristics of the reconstructed frames and ensuring smooth inter-frame energy evolution and tonal characteristics of the reconstructed frames. Performance evaluation demonstrated that the proposed concealment techniques offer better concealment quality than existing techniques, while maintaining moderate computational complexity.

6. REFERENCES

[1] C. Perkins *et al.*, "A survey of packet loss recovery techniques for streaming audio," *IEEE Network,* vol. 12, issue 5, pp. 40-48, 1998.

[2] R.C. Maher, "A method for extrapolation of missing digital audio data," *95th AES Convention*, Preprint 3715, Oct. 1993.

[3] V.N. Parikh *et al.*, "Frame erasure concealment using sinusoidal analysis-synthesis and its application to MDCT-based codecs," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing,* vol. 3, pp. 1483-1486, June 2000.

[4] S.-U. Ryu and K. Rose, "Advances in sinusoidal analysis synthesis based error concealment in audio networking," *116th AES Convention*, Preprint 5997, May. 2004.

[5] 3GPP TS 26.404: "Enhanced aacPlus encoder SBR part," June 2004.

[6] J. Herre, "Evaluation of concealment techniques for compressed digital audio," *94th AES Convention*, Preprint 3460, Feb. 1993.

[7] R. Sperschneider and P. Lauber, "Error concealment for compressed digital audio," *111th AES Convention*, Preprint 5460, Nov. 2003.

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

[9] S.-U. Ryu and K. Rose, "A frame loss concealment technique for MPEG-AAC," *120th AES Convention*, Preprint 6662, May. 2006.

[10] "Sound quality assessment material of the european broadcasting union," *http://www.ebu.ch/en/technical/publications/tech 3000_series/tech3253/index.php*

[11] ISO/IEC JTC1/SC29 WG11 MPEG, "Report on the verification tests of MPEG-4 high efficiency AAC," N6009, Oct. 2003.

[12] "Method for the subjective assessment of intermediate quality levels of coding system," ITU-R BS.1534-1, Jan. 2003.