ENCODER ASSISTED FRAME LOSS CONCEALMENT FOR MPEG-AAC DECODER

Sang-Uk Ryu[*], *Eddie Choy*[†] *and Kenneth Rose*[*]

[*] ECE Department, University of California, Santa Barbara, CA 93106-9560, USA
[†]QUALCOMM Inc., 5775 Morehouse Drive, San Diego, CA 92121-1714, USA
Email:{sang, rose}@ece.ucsb.edu, echoy@qualcomm.com

ABSTRACT

This paper proposes an efficient method for frame loss concealment within the advanced audio coding (AAC) decoder, which can effectively mitigate the adverse impact of transmission errors on the reconstruction quality. The lost frame information is estimated in the modified discrete cosine transform (MDCT) domain in terms of magnitude and sign of the coefficients. A computationally efficient approach that is capable of providing accurate estimation is employed for the magnitude estimate. Enhanced sign estimation is developed and implemented, which exploits extra information from encoder. Extensive subjective quality evaluation demonstrates that the proposed method achieves substantial quality improvement with minimal encoder assistance.

1. INTRODUCTION

Decoding of a compressed audio bitstream that had been corrupted by transmission errors may result in very annoying artifacts. In order to address such quality impairments, functional blocks for error control are added to the decoder, which consist of error detection and frame loss concealment (FLC). Once the received bitstream is found to be significantly damaged, the corresponding frame is discarded and FLC is invoked to replace the discarded frame with a signal frame that (hopefully) better approximates the original signal and provides a smooth transition to future frames.

FLC can be viewed as the problem of estimating the discarded (lost) frame while using all available relevant information, such that the generated output fits, as smoothly as possible, between the surrounding frames. While various approaches for audio FLC have been proposed, they suffer from the extreme tradeoff between the concealed audio quality and implementation cost in terms of computation and memory. For example, replacing the lost frame with silence or the previous frame represents one extreme tradeoff due to the low complexity but poor performance [1]. Advanced techniques based on source modeling typically fall on the other extreme as they require high or even prohibitive implementation cost to produce better quality [2, 3].

Recently, a computationally efficient approach that is capable of providing a moderate audio concealment quality, has been adopted in the *aacPlus* decoder for the 3rd generation partnership project (3GPP) [4]. The FLC for the core advanced audio coding (AAC) decoder (to be referred to here as "3GPP-AAC-FLC"), estimates the modified discrete cosine transform (MDCT) coefficients of the lost frame using efficient technique that works effectively for noise-like signals. However, its performance with tonal signals degrades significantly and produces audible artifacts or reconstruction seams in the concealed frames.

In this paper, we propose an effective AAC FLC technique, where the signs and magnitudes of the lost frame's MDCT coefficients are separately estimated. The magnitudes are derived using the frame repetition plus energy interpolation method that was adopted in 3GPP-AAC-FLC. The signs of the MDCT coefficients are then obtained via the enhanced sign estimation which takes a different approach to various components of the source signal. Specifically, randomly generated signs are used for MDCT coefficients in noise-like spectral bins. In MDCT bins where the tonal component is dominant, the signs of the original MDCT coefficients are acquired with the encoder's assistance, i.e., via transmission of the sign information to the decoder. A number of subjective listening tests show significant quality improvement over the 3GPP-AAC-FLC method, by minimizing the main potential sources of quality impairment.

2. PERFORMANCE ANALYSIS OF 3GPP-AAC-FLC

AAC FLC may be considered as a problem of estimating the MDCT coefficients of a lost frame from all the available information. In 3GPP-AAC-FLC, an MDCT coefficient of the lost frame $X_m(k)$ is estimated from the corresponding coefficients in the previous and next frames[1], denoted by

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

[1]3GPP's *aacPlus* decoder allows one lookahead frame, and AAC FLC is described with focus on single frame loss which assumes the presence of the surrounding frames during concealment of the lost frame.

$X_{m-1}(k)$ and $X_{m+1}(k)$, respectively. First, the MDCT coefficient of the previous frame is repeated and its energy is adjusted in scalefactor band resolution as:

$$\hat{X}_m(k) = \alpha(k) X_{m-1}(k), \tag{1}$$

where $\alpha(k)$ is the energy scaling factor and is obtained from

$$\alpha^2(k) = \frac{\sum_{k \in B_b} |X_{m+1}(k)|^2}{\sum_{k \in B_b} |X_{m-1}(k)|^2}, \tag{2}$$

where B_b is the set of MDCT bins in the $b\text{-}th$ scalefactor band. Next, the sign of the estimate $\hat{X}_m(k)$ is randomly determined:

$$\tilde{X}_m(k) = s(k)\hat{X}_m(k) = s(k)\alpha(k)X_{m-1}(k), \tag{3}$$

where $s(k)$ is a random variable with sample space $\{-1, 1\}$.

From an objective evaluation on the performance of the described FLC method, it is found that the method results in considerable estimation error, despite the fact that it has significantly reduced the difference between the absolute values of the original and the estimated MDCT coefficients. Hence, we conclude that (i) the magnitude of the MDCT coefficient can be well estimated by repetition and energy interpolation as above and (ii) the sign mismatch between the original and the estimated MDCT coefficients is a main source of substantial estimation error.

In order to assess the performance of the described FLC method in terms of the subjective quality, the source signal is modeled and analyzed as sinusoids plus noise,

$$x[n] = \sum_{l=1}^{L} A_l \cos(\omega_l n + \phi_l) + v[n], \tag{4}$$

where the tonal component is represented with sum of sinusoids whose parameters (amplitude, frequency and phase) are denoted by A_l, ω_l, and ϕ_l, respectively, and the noise component is denoted as $v[n]$.

For the noise component, the psychoacoustic principle dictates that the spectral energy distribution is perceptually relevant, while its phase properties are not perceptually important [5]. This and the first observation, support the claim that 3GPP-AAC-FLC provides good approximation of the spectral energy distribution and achieves perceptually satisfactory quality for the noise component. Note that the random sign is applied to mimic the synthesis process of the noise component in the Fourier domain, since the random phase used in the Fourier domain synthesis can be realized via random sign change in the MDCT domain. Hence, 3GPP-AAC-FLC in effect estimates MDCT coefficients via noise modeling of the source signal. In other words, 3GPP-AAC-FLC is designed to only work effectively with noise-like signals. Its performance is expected to degrade considerably for signals with dominant tonal components. The performance deterioration for tonal components is mainly attributed to the sign mismatch of the MDCT estimates.

In order to analyze the perceptual impact of the sign mismatch for tonal component, let us consider the case where the tonal component in the source signal is composed of single sinusoid, i.e., $x[n] = A_0 \cos(\omega_0 n + \phi_0)$. In the $m\text{-}th$ frame, let the original samples and the corresponding MDCT coefficients be denoted by $x_m[n]$ and $X_m(k)$, respectively. Now, let us assume that the MDCT estimate $\tilde{X}_m(k)$ has the exact magnitude of $X_m(k)$ but there is a sign mismatch at the most prominent spectral peak, i.e., $\tilde{X}_m(k_0) = -X_m(k_0)$, where $k_0 \equiv nint(f_0), \omega_0 \equiv f_0\pi/M$, M is the number of the MDCT bins, and $nint(\cdot)$ is the nearest integer. Given the sign mismatch at the most prominent peak and assuming that MDCT magnitudes at other bins are relatively negligible, we employ the approximation $\tilde{X}_m(k) \cong -X_m(k)$ for all k for simplicity of analysis.

Let a sine window such as $w[n] = \sin(\pi(n+0.5)/2M)$ be employed as analysis/synthesis window, then the closed-form of the time-domain reconstruction via the inverse MDCT (IMDCT) of $\tilde{X}_m(k)$ and overlap/add with that of $X_{m-1}(k)$ can be analytically derived as:

$$\tilde{x}_m[n] \cong \sqrt{2}\cos\Big\{\frac{\pi}{M}(n+0.5) - \frac{\pi}{4}\Big\}x_m[n] \tag{5}$$

$$= \frac{A_0}{\sqrt{2}}\Big\{\cos\Big((\omega_0 - \frac{\pi}{M})n + \phi_0 - \psi\Big) + \cos\Big((\omega_0 + \frac{\pi}{M})n + \phi_0 + \psi\Big)\Big\}, \tag{6}$$

where $\psi = \pi/2M - \pi/4$. From (6), it can be seen that the time-domain reconstruction possesses different spectral characteristics from the original: (i) two strong tones at $\omega_0 \pm \pi/M$, rather than a single tone at ω_0, and (ii) spectral leakage around ω_0 due to the interference of the two nearby spectral peaks. Furthermore, these spectral features are spread to the next frame, since $\tilde{x}_{m+1}[n]$ is reconstructed via the IMDCT of $X_{m+1}(k)$ and overlap/add with that of $\tilde{X}_m(k)$. When $\tilde{x}_m[n]$ and $\tilde{x}_{m+1}[n]$ are injected to the correctly decoded surrounding frames, its spectrum reveals distorted spectral evolution of the sinusoid as well as spectral leakage around ω_0. This spectral dissimilarity with the neighboring frames will produce audible artifact or reconstruction seam that will be very audible when the sinusoid has considerable magnitude.

It was shown in the single sinusoidal model that the sign mismatch at the most prominent spectral peak is a potential source of quality impairment of audio concealment. Now, let us consider the generic case that the tonal component in the source signal is modeled as a sum of multiple sinusoids, wherein the $l\text{-}th$ sinusoid has a prominent spectral peak at the MDCT bin k_l, where $k_l \equiv nint(f_l)$ and $\omega_l \equiv f_l\pi/M$. Let us denote the $m\text{-}th$ frame's index subset of the MDCT bins corresponding to the prominent spectral peaks as I_m, i.e., $I_m \equiv \{k_l,\ 1 \le l \le L\}$. Then, it can be seen that a sign

mismatch at MDCT bin $k \in I_m$ can lead to an artifact in the concealed audio. Hence, in order to minimize the potential for significant quality impairment, the signs of the MDCT estimates at the MDCT bins in I_m should be matched with those of the original MDCT coefficients.

3. PROPOSED FLC TECHNIQUE

In the previous section, we analyzed the performance of 3GPP-AAC-FLC and identified the major sources of quality impairment. Based on the observed features and performance analysis, we propose an enhanced FLC technique, where the signs and magnitudes of the lost frame's MDCT coefficients are separately estimated. The observation that the repetition and energy interpolation method produces good magnitude estimates leads us to adopt this method for the magnitude estimation, i.e.,

$$\hat{X}_m(k) = |\alpha(k) X_{m-1}(k)|. \tag{7}$$

An enhanced sign estimation for the MDCT coefficients, denoted by $s_m^*(k)$ is developed and the proposed MDCT estimation can be reformulated as:

$$\tilde{X}_m^*(k) = s_m^*(k)\hat{X}_m(k) = s_m^*(k)|\alpha(k) X_{m-1}(k)|. \tag{8}$$

Based on the analysis result that sign mismatch at $k \in I_m$ represents a potential artifact, while sign mismatch in the noise-dominant MDCT bins has no serious impact, we propose the enhanced sign estimation:

$$s_m^*(k) = \begin{cases} \mathrm{sgn}\big(X_m(k)\big) & \text{for } k \in I_m \\ s(k) & \text{for } k \notin I_m, \end{cases} \tag{9}$$

where $\mathrm{sgn}(\cdot)$ denotes the sign function and $s(k)$ is a random variable with sample space $\{-1, 1\}$.

In order to generate $s_m^*(k)$ at the decoder's FLC block, the position information I_m as well as the corresponding signs of the original MDCT coefficients must be made available to the decoder. One simple way would be to transmit this information to the decoder. Rather than transmit the information on I_m to the decoder, it can be derived at the decoder side (as well as at the encoder side) in closed-loop manner (in analogy to predictive coding). Recall that I_m is defined as the index subset of the MDCT bins of prominent spectral peaks. However, the detection of sinusoidal components requires extensive computation (e.g., peak picking and tracking in the Fourier domain). Instead, we employ an efficient approximation:

$$I_m \cong \big\{k \big| |X_m(k)| > Thr,\ 0 \le k < M\big\}, \tag{10}$$

where Thr is determined such that $|I_m| = B_m$ and B_m is the number of signs to be transmitted. The MDCT bins having the most prominent magnitudes are chosen from the MDCT coefficients of the m-th frame, which are obviously not available during concealment at the decoder. Hence, another approximation is employed for the magnitude of the MDCT coefficient: $|X_m(k)| \cong \hat{X}_m(k)$, which is already computed at the magnitude estimation in (7). The approximated index subset for the frame m is finally given by:

$$\hat{I}_m \equiv \big\{k \big| |\hat{X}_m(k)| > Thr,\ 0 \le k < M\big\}. \tag{11}$$

Then, the proposed FLC employing the enhanced sign estimation $s^*(k)$ can be realized with the encoder's assistance that the signs of the MDCT coefficients corresponding to the derived index subset are transmitted to the decoder. For this, AAC encoder side is necessarily modified as in Fig. 1. Note the closed loop implementation. When frame $m+1$ is encoded, a subset of signs of the MDCT coefficients for the frame m is extracted and attached to the AAC bitstream for the frame $m+1$.

In order for the decoder's FLC to exploit the transmitted sign information, three additional functional blocks are incorporated within the AAC decoder as described in Fig. 2: (i) magnitude estimation (ii) component selection and (iii) sign extraction. In the magnitude estimation block, the magnitude of the MDCT coefficients are estimated by the repetition and energy scaling method. Then, from the estimated magnitudes, the index subset for the prominent tonal components are determined in the component selection block using (11). The signs for the selected tonal components

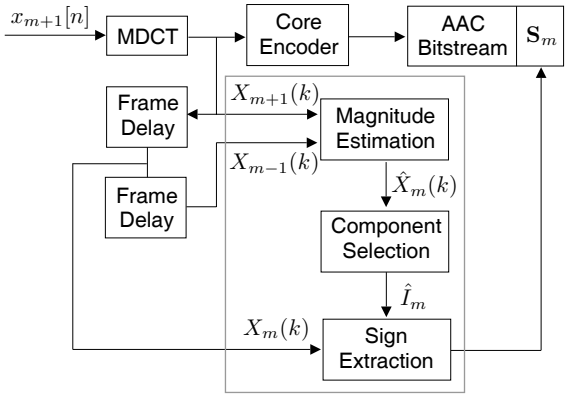

Fig. 1. Block diagram of the modified encoder for extraction of the extra sign information.

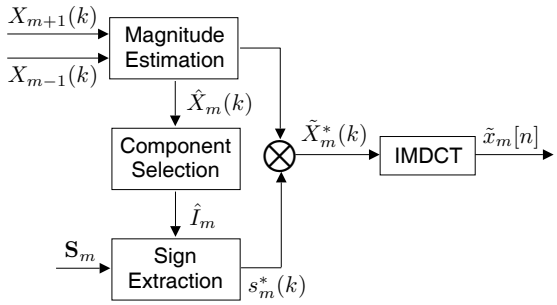

Fig. 2. Block diagram for the proposed FLC at the decoder side.

are obtained from the derived index subset and the transmitted sign bits. Then, the enhanced sign estimation parameter $s_m^*(k)$ is determined as in (9) and the ultimate MDCT coefficient estimates are achieved by multiplication with the magnitude estimates. In order to guarantee that the encoder chooses the same MDCT bins as will be selected by the decoder, the encoder has to mimic the decoder's status. Hence, the magnitude estimation and component selection blocks used by the decoder are also employed at the encoder side. Then, the signs of the selected components are extracted at the sign extraction block of the encoder and the signs are attached to the encoded bitstream as side information.

4. EXPERIMENTAL RESULTS

In order to evaluate the performance of the proposed FLC, we carried out a number of quality evaluations of the concealed audio under various frame loss rates (FLR). The AAC encoder and decoder were modified from the 3GPP's *aacPlus* implementation [6] and were employed as the target audio codec. A number of MONO audio sequences sampled from the commercial CDs were encoded at the bitrate of 48 kbps, and the encoded frames were randomly dropped at the specified rates with restriction to single frame loss.

For the proposed FLC system, all supplementary functional blocks for extracting the signs of the prominent tonal components were incorporated within the modified AAC encoder. At the decoder side, all necessary blocks for the proposed sign estimation were implemented within the modified AAC decoder. The number of signs that are transmitted as side information was fixed for all frames and restricted to 10 bits/frame, which is equivalent to the bitrate of 0.43 kbps (less than 1% of the bitrate for AAC bitstream). Two different bitstreams were generated: (i) 48 kbps AAC bitstream for the competing 3GPP-AAC-FLC and (ii) 47.57 kbps AAC bitstream complemented with sign information at the bitrate of 0.43 kbps for the proposed FLC.

For subjective evaluation of the concealed audio quality, various genres of polyphonic audio sequences with 44.1 kHz sampling rate were selected, and the decoder reconstructions by both methods under various FLRs were compared. We employed the multi-stimulus test with a hidden reference (MUSHRA) setup [7] and it was performed by eleven listeners. Fig. 3 presents the overall subjective quality comparison of the reconstructions by the two concealment methods under FLR of 0, 5, 10, 15, 20%. From the figure, it can be seen that the proposed concealment algorithm significantly improves the decoder reconstruction quality at all FLRs. For example, the proposed method maintains reconstruction quality that is better than 80 point MUSHRA score at moderate (5 and 10%) FLR. Furthermore, the reconstruction quality of the proposed FLC method at 15% FLR is statistically equivalent to that of the existing method at 5% FLR, demonstrating the enhanced error-resilience offered by the proposed technique.

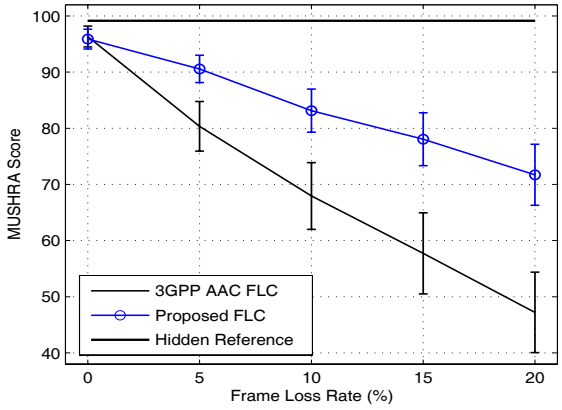

Fig. 3. Results of the MUSHRA test with 95% confidence intervals for the 3GPP AAC FLC and the proposed FLC under 0, 5, 10, 15, 20% frame loss rates.

5. CONCLUDING REMARKS

We proposed an efficient frame loss concealment approach for AAC with minimal assistance by the encoder. The approach differentiates between tonal and noise components of the source signal (at the MDCT bin level) in order to minimize the potential for quality impairment of the audio generated for concealment. Subjective quality evaluation demonstrates that the proposed method achieves substantial quality improvement and significantly enhances error-resilience. Further research on the optimal bit allocation of the side information, more accurate magnitude estimation, and extension to multiple frame losses should provide additional performance improvement.

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] 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.

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

[4] 3GPP TS 26.404: "Enhanced aacPlus encoder SBR part," June 2004.

[5] M. Goodwin, "Adaptive singal models: theory, algorithm, audio applications," PhD thesis, University of California, Berkeley, 1997.

[6] 3GPP TS 26.410: "General audio codec audio processing functions; Enhanced aacPlus general audio codec; Floating-point ANSI-C code," Jan. 2005.

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