PRE-ECHO NOISE REDUCTION IN FREQUENCY-DOMAIN AUDIO CODECS

Jimmy Lapierre and Roch Lefebvre

Speech and Audio Research Group
Faculty of Engineering, Université de Sherbrooke
Sherbrooke (Québec) J1K 2R1 Canada
{jimmy.lapierre, roch.lefebvre} @USherbrooke.ca

ABSTRACT

One of the most common yet detrimental compression artifacts in frequency-domain audio codecs is known as pre-echo, which is perceived as a brief noise preceding transient signals, and is discernable even without direct comparison to the original signal. Because of its substantial negative impact on audio quality, many techniques have been proposed to alleviate it, but not without effect on coding efficiency. This paper presents a novel method to reduce pre-echo noise using only information already available at the decoder, such as scale-factors, that allow an estimation of the quantization noise levels in each frequency band. Doing so allows the proposed method to reduce pre-echo noise based on a precise modeling of the quantization noise spread before the transient signal. This has shown to improve both the subjective and objective quality of the MPEG AAC codec, and requires no modifications to the existent standard-compliant encoders.

Index Terms — Audio coding, Noise reduction, Audio quality enhancement, Transform-domain codecs, Pre-echo.

1. INTRODUCTION

Throughout the past decades, perceptual audio compression technologies have been developed to reduce bandwidth and storage requirements. These codecs generally model the human auditory system by encoding the audio signal in a short-term frequency domain, such as MDCT in the MPEG AAC codec [1, 2]. However, the quantization of transient audio signals in the frequency domain causes the quantization noise to spread before the transient itself, which is known as the pre-echo artifact. It has been shown that this artifact is very harmful to the overall audio quality of a codec [3, 4]. An example of this noise spreading is presented in figure 1.

Given its impact on audio quality, many technologies have been proposed to reduce pre-echo noise. However, existent methods generally require more bits or a reduction in overall coding efficiency. One of the most common methods is short block switching, which limits the spread of quantization noise by detecting transient signals at the encoder and temporarily transitioning to a significantly smaller time frame during those events [5]. However, smaller frames will significantly reduce the coding efficiency of stationary signals, so this technique has a negative impact on audio quality if used in frames with no transient events.

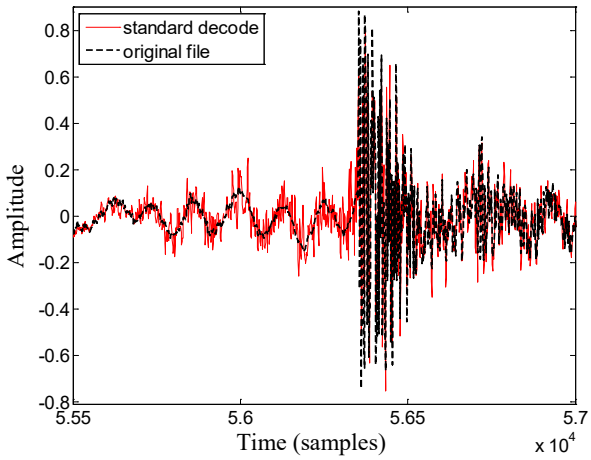

Fig. 1: Pre-echo noise in a transform audio codec

Another existing method to reduce pre-echo noise is known as Temporal Noise Shaping (TNS) [6]. With TNS, a prediction filter is used in frequency domain to shape the quantization noise in the time domain. The total amount of noise is not necessarily reduced, but it becomes less audible, since its energy is concentrated more closely in time to the transient event. However, this technique requires the transmission of TNS filter coefficients in the bitstream, which may reduce the number of bits available for coding the audio signal itself, particularly for constant bitrate applications.

Thus, this paper offers a new method to reduce pre-echo noise without reducing coding efficiency, using information already available at the decoder. Since no modifications are required at the encoder, this method can be applied to existent standard codecs such as MPEG AAC. Therefore, section 2 describes the proposed algorithm in detail, while section 3 presents the corresponding results of subjective and objective tests. Finally, conclusions are drawn in section 4.

2. ALGORITHM DESCRIPTION

The proposed algorithm operates at the decoder using data from the bitstream, as shown in fig. 2. First, the bitstream is decoded in a standard fashion. Then, each frame is tested for the presence of a transient signal that would likely produce a perceivable pre-echo artifact. If such signal is detected, the audio signal is split into two parts, namely the pre-transient (pre-) and post-transient (post-) signals. Finally, both signals, specific transient characteristics and the codec parameters are fed to the noise reduction algorithm, illustrated in fig. 3.

The noise reduction algorithm is described in detail in the following subsections, but can be summarized as follows. First, the amount of quantization noise present in the frame is estimated for each frequency coefficient or frequency band using the scale-factors and coefficient amplitudes from the bitstream. Then, to determine the spectrum of the quantization noise spread from the post- to the pre- signal, that estimate is used to shape a random noise signal that is added to the post- signal in the ODFT or oversampled DFT domain, which is then transformed into the time domain, multiplied by the pre- window and returned to the frequency domain. This provides a suitable noise shape estimate to apply spectral subtraction on the pre- signal without adding any artifacts of its own. To preserve total frame energy, and considering that the quantization noise caused the signal to smear from the post- to the pre- signal, the energy removed from the pre- signal is also added back to the post- signal. Both signals are then added together and transformed to the MDCT domain. The remainder of the decoder may then use the modified MDCT coefficients in replacement of the originals ones.

It is useful to know that added quantization noise or any modifications to MDCT coefficients in a given frame does not have any impact on the coefficients of adjacent frames, even following the overlap-add process. It can be shown that contrary to a FFT-based filterbank with overlap, the perfect reconstruction properties of MDCT-based filterbanks with 50% overlap also hold true for the transformation of MDCT coefficients to the time domain and back [7] (pp. 39-40).

2.1. Transient detection and windowing

By design and for efficiency, the proposed algorithm only processes frames that encompass a transient signal, which can be defined as frames exhibiting a steep increase in energy. Consequently, each frame is split into pre- and post-transient signals by applying two complementary window functions that have a steep cross-over positioned in the frame to maximize the energy ratio between the two resultant signals. Example windows are shown in [7] (pp. 46-48). A frame must show a minimum energy increase from the pre- to the post- signal to be selected for processing by the noise reduction algorithm. If so, the cross-over position and the energy ratio are characteristics that are employed by the noise level estimation step, along with the actual pre- and post- signals.

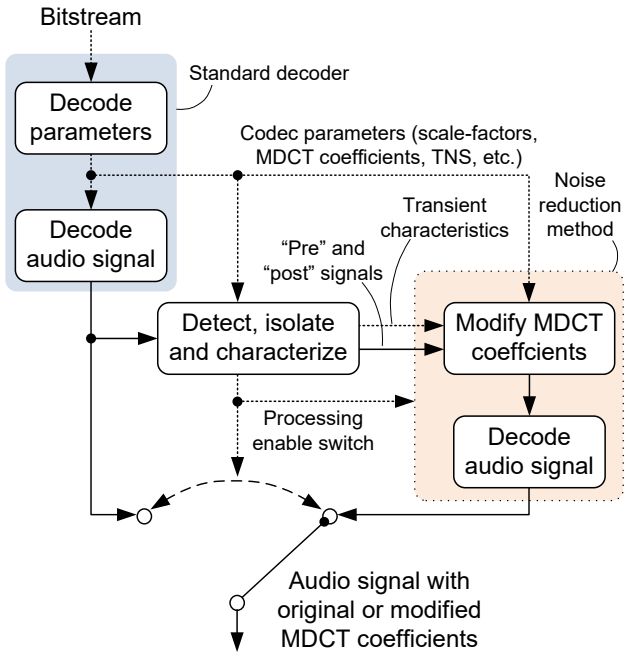

Fig. 2: Algorithm overview

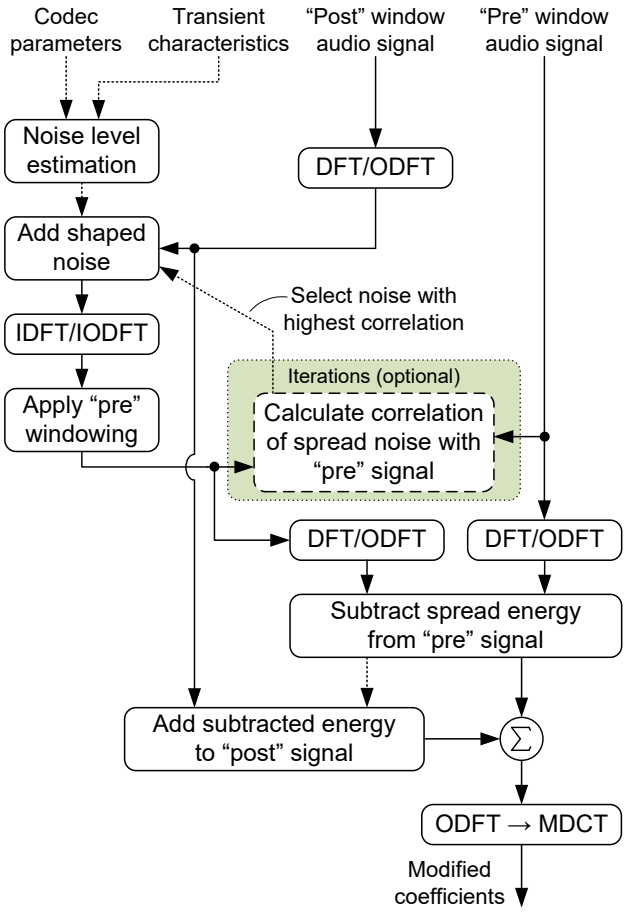

Fig. 3: Noise reduction method detail

2.2. Noise level estimation

The quantization noise levels affecting the post- signal are estimated from the scale-factors, the coefficient amplitudes and the previously calculated energy ratio. Specifically, the k^{th} quantized MDCT coefficient ϑ_k is obtained from the k^{th} MDCT coefficient ξ_k, where each coefficient belongs to a scale-factor band j that has a gain of υ_j. For standard AAC:

$$\vartheta_k = \operatorname{sign}(\xi_k) \cdot \operatorname{round}\left[\left(\frac{|\xi_k|}{2^{1/4} \upsilon_j} \right)^{3/4} + L \right] \quad (1)$$

($L = 0.4054$ or -0.0946, depending on the codec version) [9]. To estimate the noise level, first, the range of the quantization error is obtained by first solving the previous equation for ξ_k:

$$\xi_k = \operatorname{sign}(\vartheta_k)\, 2^{1/4} \upsilon_j \left(|\vartheta_k| - L \right)^{4/3}, \quad (2)$$

then by calculating the adjacent quantizer outputs ξ_k^{+1} and ξ_k^{-1} by substitution of ϑ_k with $\vartheta_k + 1$ and $\vartheta_k - 1$ respectively. Assuming a uniformly distributed quantization noise puts the discrimination levels at the center of the quantization intervals, leading to an interval amplitude of $\left| \xi_k^{+1} - \xi_k^{-1} \right| / 2$. Also, presuming that the quantization noise is spread between the pre- and post- signals proportionally to their relative energy, the factor g_k is also introduced:

$$g_k = \sqrt{\frac{X_O^{post}(k)\, X_O^{post*}(k)}{X_O^{pre}(k)\, X_O^{pre*}(k) + X_O^{post}(k)\, X_O^{post*}(k)}}, \quad (3)$$

where X^* denotes the complex conjugate. Thus, the estimated noise levels for the coefficients are: $n(k) = g_k \left| \xi_k^{+1} - \xi_k^{-1} \right| / 2$.

To evaluate how $n(k)$ is spread between the real ($\Re e$) and imaginary ($\Im m$) parts of ODFT coefficients $X_O(k)$, the relation between MDCT coefficients $X_M(k)$ and ODFT coefficients is used [8], which is, for $k = [0, N-1]$:

$$\begin{aligned} X_M(k) = {} & \Re e\{X_O(k)\} \cos[\theta(k)] + \\ & \Im m\{X_O(k)\} \sin[\theta(k)], \end{aligned} \quad (4)$$

where: $$\theta(k) = \frac{\pi}{N}\left(k + \frac{1}{2}\right)\left(1 + \frac{N}{2}\right). \quad (5)$$

Then, knowing that the quantization noise $q(k)$ of the MDCT coefficients $X_M(k)$ is spread between the real and imaginary parts of the ODFT coefficients $X_O(k)$:

$$\begin{aligned} X_M(k) + q(k) = {} & \Re e\{X_O(k) + q_{\Re}(k)\} \cos[\theta(k)] + \\ & \Im m\{X_O(k) + j \cdot q_{\Im}(k)\} \sin[\theta(k)]. \end{aligned} \quad (6)$$

Assuming that $q(k)$ affects the real and imaginary parts of the ODFT coefficients proportionally to their contribution to their corresponding MDCT coefficients, the real and imaginary parts of the quantization noises $q_{\Re}(k)$ and $q_{\Im}(k)$ affecting the ODFT coefficients are defined as:

$$q(k) = q_{\Re}(k) \cos[\theta(k)] + q_{\Im}(k) \sin[\theta(k)]. \quad (7)$$

Consequently, the simulated shaped quantization noise generated for the set of ODFT coefficients is:

$$b(k) = n(k)\left(\cos[\theta(k)] + j \sin[\theta(k)]\right)\varepsilon(k), \quad (8)$$

where $\varepsilon(k)$ is a pseudo-random noise with a uniform distribution between -½ and ½. Alternatively, independant pseudo-random functions $\varepsilon_1(k)$ and $\varepsilon_2(k)$ can be used to model the quantization noise for the real and imaginary parts :

$$b(k) = n(k)\left(\cos[\theta(k)]\varepsilon_1(k) + j \sin[\theta(k)]\varepsilon_2(k)\right). \quad (9)$$

Optionally, an additional factor could be applied to $b(k)$ to modify the overall pre-echo noise reduction strength. That factor could be a constant value or a function of k.

To increase the SNR between the original signal and the signal produced by this algorithm, the correlation of the actual pre- signal with the ones produced from a multitude of pseudo-random signals $\varepsilon(k)$ can be compared, where the $\varepsilon(k)$ producing the highest correlation is selected for the remainder of the algorithm. However, such additional iterations increase algorithmic complexity for a benefit to subjective quality that is mostly negligible.

Next, the pre-transient frequency-domain noise $Y_n^{pre}(k)$ is estimated using the spread of the frequency-domain post-transient signal $Y_O^{post}(k)$ affected by the previously estimated quantization noise. As shown in fig. 3, that is:

$$Y_n^{pre}(k) = \text{ODFT}\left\{ \left[\text{IODFT}\left\{ Y_O^{post}(k) + b(k) \right\} \right] w_1(n) \right\}, \quad (10)$$

where $w_1(n)$ is the previously employed pre- window.

2.3. Spectral subtraction and energy compensation

With the frequency-domain noise $Y_n^{pre}(k)$ calculated, the next step is to subtract that energy from the pre-transient signal and to add it back to the post-transient signal. This is achieved by modifying the amplitude of the pre-transient signal $Y_O^{pre}(k)$ while preserving its phase, so:

$$Y_O^{pre+n}(k) = \left(\left| Y_O^{pre}(k) \right| - \left| Y_n^{pre}(k) \right| \right) \cdot e^{\, j \angle Y_O^{pre}(k)} \quad (11)$$

for $\left| Y_O^{pre}(k) \right| > \left| Y_n^{pre}(k) \right|$ and 0 otherwise. Then, the energy subtracted by equation (11) is added to the post-transient signal $Y_O^{post}(k)$, again while preserving its phase:

$$\begin{aligned} & Y_O^{post+n}(k) = \\ & \left(\sqrt{Y_O^{post}(k) \cdot Y_O^{post*}(k) + Y_O^{n}(k) \cdot Y_O^{n*}(k)} \right) \cdot e^{\, j \angle Y_O^{post}(k)}, \end{aligned} \quad (12)$$

where: $$Y_O^{n}(k) = Y_O^{pre}(k) - Y_O^{pre+n}(k). \quad (13)$$

Finally, the ODFT coefficients (or the coefficients of the oversampled DFT that correspond to them) $Y_O^{pre+n}(k)$ and $Y_O^{post+n}(k)$ are added together and converted to MDCT coefficients with equation (4). These new coefficients replace those that were previously decoded for the transient frame, and they can be used by the decoder as if they were the actual coefficients that were received in the first place.

3. EVALUATION RESULTS

First, the performance of the proposed algorithm was assessed with a MUSHRA blind subjective listening test [10]. In this test, listeners choose to hear the reference (original) signal or any version from A through F, consisting of the hidden reference, a 3.5kHz low-pass filtered reference, the standard decoder and the post-processed signals at both 24 and 28 kbps, in random order. Each item was voted on a scale of 1 to 100. Overall, there were 12 different sample sets evaluated by 10 expert listeners. The results were 54.6 for the processed signal compared to 52.6 for the standard decoder at 24 kbps, and 71.1 for the processed signal against 68.8 for the standard decoder at 28 kbps. However, as the range of values used to score both 24 kbps and 28 kbps versions vary significantly between listeners, the confidence intervals were slightly too wide to observe a statistically significant gain. Therefore, the results were examined with a differential analysis, compiling only the differences between the scores of processed and standard versions for each set at each bitrate. These results are shown for each sample set (S1-S12) and for all sets combined (ALL) in fig. 5, along with 95% confidence interval bars. Although not every set shows a statistically significant improvement, combining the results of all 12 sets produces a smaller confidence interval caused by a 12-times larger sample size. Consequently, this shows that the average gain for all samples is approximately 2 MUSHRA points, which is substantial when considering that it cost no extra bits.

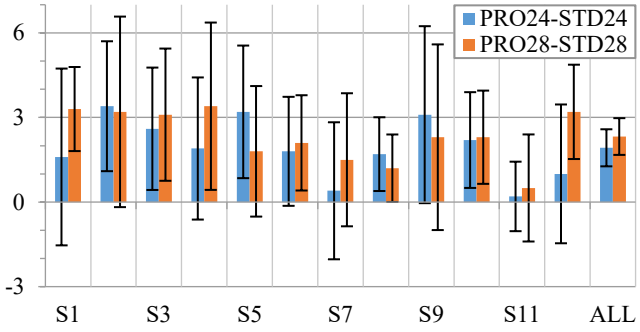

Fig. 4: Differential MUSHRA subjective test results

Next, table 1 presents the SNRs of processed frames for the standard decoder and the post-processed version, using a total of 500 various files that totalized 173810 frames. The percentage of frames that were processed for each bitrate is included. The fact that an improvement in achieved without adding new data from the encoder demonstrates that the algorithm effectively makes use of underexploited information. Also, gains of up to 6 dB were observed for some frames. Note that the optional noise generator/correlation iterations (in fig. 3) were active for compiling the results in table 1.

Finally, fig. 6 shows an example of the output of the post-processing algorithm for the same "standard decode" signal that was presented in fig. 1.

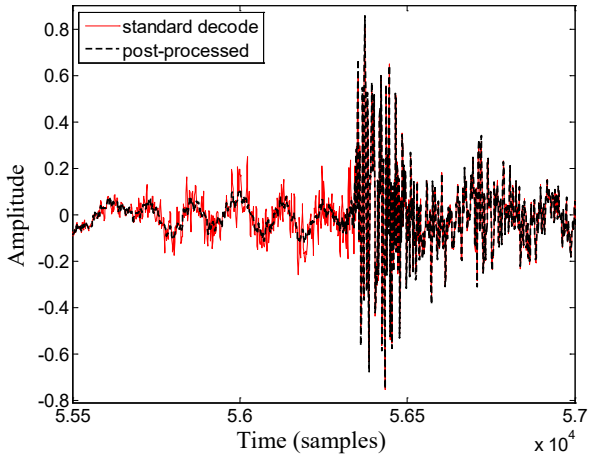

Fig. 5: Result of pre-echo post-processing

Bitrate (kbps)	Processing ratio (%)	Standard (SNR)	Processed (SNR)	Processing gain (SNR)
12	5.81	-0.48	-0.26	0.223
16	5.54	3.80	4.19	0.393
20	5.31	5.70	6.15	0.444
24	5.47	9.17	9.47	0.302
28	5.43	9.08	9.40	0.319
32	5.52	11.25	11.50	0.253
48	5.64	14.89	15.05	0.159

Table 1: Objective results for transient frames at 7 bitrates

4. CONCLUSION

A novel algorithm has been proposed to reduce pre-echo by better exploiting information that is already available to the decoder. Since no modification is required to the encoder, the algorithm can be used in conjunction with existing encoders. Also, the fact that the encoder can remain as-is demonstrates a clear advantage over previous pre-echo reduction methods: there is no reduction of coding efficiency, no extra bits to transmit and no inefficient window shapes or frame lengths.

The changes made to the MDCT coefficients by the proposed algorithm are essentially limited to the quantization intervals of the received coefficients, so if the processed signal was to be re-quantized with the same scale-factors, the result would be identical to the unprocessed signal. This substantiates the claim that the chances of over-processing the signal and introducing new artifacts are very low. Also, it was shown that objective quality (SNR) is also improved by the algorithm's use of already available information.

Finally, the algorithm demonstrates a clear gain in subjective audio quality, with no drawback other than an increase in computations at the decoder. Therefore, since no modifications are required to the encoder nor the bitstream, the proposed algorithm can be readily implemented in AAC or comparable decoders in use in present-day systems.

5. REFERENCES

[1] ISO/IEC 14496-3:2001: “Information technology - Coding of audio-visual objects - Part 3: Audio.”

[2] 3GPP TS 26.403, V12.0.0, “General audio codec audio processing functions; Enhanced aacPlus general audio codec; Encoder specification; Advanced Audio Coding (AAC) part.”

[3] P. Marins, F. Rumsey, and S. K. Zielinski, “The Relationship between Selected Artifacts and Basic Audio Quality in Perceptual Audio Codecs,” *120th AES Conv.*, Paris, France, 2006.

[4] P. Marins, F. Rumsey, and S. K. Zielinski, “The Relationship between Basic Audio Quality Selected Artefacts and in Perceptual Audio Codecs – Part II: Validation Experiment,” *122nd AES Conv.*, Vienna, Austria, 2007.

[5] B. Karlheinz and G. Stoll, “The ISO/MPEG-Audio Codec: A Generic Standard for Coding of High Quality Digital Audio,” *92nd AES Conv.*, Vienna, Austria, 1992.

[6] J. Herre and J. D. Johnston, “Enhancing the performance of perceptual audio coders by using temporal noise shaping (TNS),” *101st AES Conv.*, Los Angeles, California, USA, 1996.

[7] J. Lapierre, “Amélioration de codecs audio standardisés avec maintien de l’interopérabilité,” Ph.D. thesis, Université de Sherbrooke, Sherbrooke, Québec, Canada, 2016.

[8] A. Ferreira, “Perceptual coding using sinusoidal modeling in the MDCT domain,” *112th AES Conv.*, Munich, Germany, 2002.

[9] D. Salomon and G. Motta, “Handbook of Data Compression,” Springer-Verlag, London, 2010.

[10] ITU-R Recommendation BS.1543-1: “Method for the Subjective Assessment of Intermediate Sound Quality (MUSHRA),” *ITU-T*, Genève, 2001.