Robust Block Switching Decision for Transform-based Audio Coder

Yan-Chen Lu, Cheng-Ching Huang and Wan-Kuei Lin
Vivotek Inc., Taipei County, Taiwan, R.O.C.
yanchen@vivotek.com

ABSTRACT

In this paper, we propose a robust block switching mechanism to improve the adaptive window switching algorithm for reducing smearing effect in transform-based perceptual audio coder. In addition to monitor abrupt energy change, we propose to use spectral consistency as better characteristic mark for block switching. Specifically, we check the signal uniformity after the removal of interference to identify the proper block type. Besides, we use an adaptability control to adjust the analysis process to address the non-stationary characteristic of audio signal. Particularly, a practical implementation which integrates with MPEG-2/4 AAC demonstrates our algorithm's efficiency both in respects of complexity and quality enhancement. Experimental results show that our scheme can achieve better encoding quality and spend about one-fourth computation power as compared to the perceptual entropy based MPEG method.

1. INTRODUCTION

Wide-band audio compressor suffers from temporal mismatch between quantization noise and perceptual masking due to blockwise signal representation in frequency domain. To suppress this artifact, adaptive window switching [1] was proposed to adjust the filter bank windows for matching the time-frequency resolution of input signal. Although the adaptive window switching introduces additional complexity, many modern audio compressors including MPEG-4 extension [2] still embed such a technique because of its superior performance on suppressing noise-spreading.

For preserving the best time-frequency localization, a block switching scheme is used to adaptively switch the window (block) type between "long" and "short". Particularly, to generate the block switching decision, MPEG VM [3] monitors the variation of perceptual entropies from psychoacoustic analysis about consecutive long blocks. When a significant change is detected, short window type is employed for better coding gain. An inevitable computation of long block masking threshold and the tweaking effort of psychoacoustic model make this approach a cumbersome choice for audio coder.

To obtain a block type decision earlier and avoid a long block psychoacoustic analysis, a temporal transient detection mechanism is utilized to assess the nature of input signal at Dolby AC-3 [4]. A comparison of peak values corresponding to segmented high-passed samples signifies the attack in signal energy. Similar approach was proposed by Smithers in [5]. However, using a pre-defined threshold to locate transient nature and screen peak values is not a robust paradigm for non-stationary audio signal.

In this paper, we offer a different perspective of view for interpreting the pre-echo artifact and propose a new strategy to perform block switching. Specifically, our analysis shows that using an adaptive mechanism to generate process parameters and a better cue-search for the possibility of artifact's existence can deliver a superior solution for adaptive window switching. Performance evaluation results with MPEG-2/4 AAC illustrate that our approach can provide the best encoding quality while maintaining lower and similar complexity as compared to other state-of-the-art methods [3][4][5].

This paper is organized as the following. Sec. 2 provides a novel explanation about temporal masking mismatch. Sec. 3 introduces our proposed block switching scheme for solving inadequate-masking problems. Sec. 4 presents a practical implementation of our scheme with MPEG-2/4 AAC. Sec. 5 demonstrates the effectiveness of proposed algorithm. Lastly, Sec. 6 gives our conclusions of this work.

2. PROBLEM FORMULATION

Regarding to the quantization process of transform-based audio coder, the spectral coefficients are classified into different subbands before the real bit-rate reduction is achieved. The quantization noise power of *k*-th subband is described as follows:

$$s_Q^2(k) = s_X^2(k)\, g\big(b(k), k\big),$$

where $s_X^2(k)$ is the estimation of signal power, *b(k)* is the number of bits allocated to the *k*-th quantizer, and *g* is characteristic of the quantizer. Since the quantizer has a descending R-D curve, we can obtain a smaller *g* by increasing *b(k)* which is originally proportional to *Sx(j)* / *Sm(j)*, where *Sx(j)* and *Sm(j)* are the power spectral density estimation of the signal and masked threshold in the scale of psychoacoustic analysis partition.

An incorrect *Sm(j)* may cause inadequate *b(k)* that creates perceptible $s_Q^2(k)$. Therefore, a transient event requires a distinguished *Sm(j)* other than its nearby signal to obtain a transparent coding. The inconsistent requirement for *Sm(j)* introduces the necessity of short block to localize the masking ability. On the other hand, repeated transients in a frame with similar spectral characteristic may obtain a consistent *Sm(j)* requirement which is sustainable by long block.

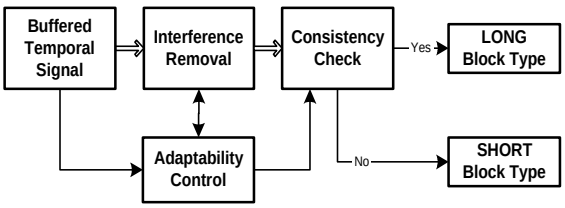

Figure 1. Block diagram of proposed block switching.

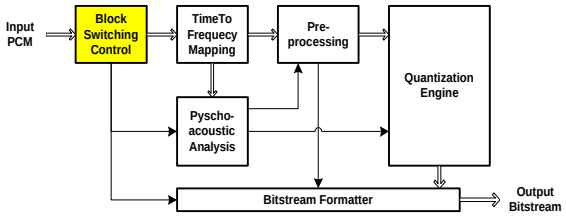

Figure 2. AAC with proposed block switching control.

Audio coder possesses an inherent coding limitation for "pitch-based" signals such as speech. The pseudo-stationary series of impulse-like signal requires a large amount of bits to completely symbolize its structure. Instead of switching to short block for transient nature, a better strategy for dealing with pitch signal is remaining long block without extra side information. Besides, similar *Sm(j)* for each pitch transient may not bring an inconsistent masking requirement when staying in long block which is also preferable for subsequent tool like TNS [6].

Transient-detection oriented algorithm narrows the scope of reason for generating smearing effect. From a perspective on spectral-consistency, we develop a more robust tool to adapt the block type. It does not lose the perceptibility of Meixner / step transient but also fulfills any case individual *Sm(j)* is required.

3. ROBUST BLOCK SWITCHING DECISION

The clue which signifies the violation of consistency is often derived from the differential part of signal, e.g. the high frequency component (HF). The low frequency component (LF) is unlikely changing a lot during a frame period, and its smear is also hardly detected by human perception. The existence of ambient noise and LF conceal the cue of non-stationary. We should remove their disturbance to enhance the detection of consistency. Since it is possible to isolate particular subband signal by using temporal techniques, we can realize the whole operation without the penalty of time-to-frequency mapping.

Fig. 1 depicts our proposed block switching decision. As shown, the processed audio signal is first feed to an interference removal block. Any signal that is not helpful to identify the consistency would be deemed as interference and eliminated purposely. After the interference removal, the residual output is then used for consistency checking. The consistency checking is to define the consistent distribution of current audio spectra and identify this status from the interference-free signal. To enhance the manipulation for non-stationary nature of audio, we further use an adaptability control unit to monitor the characteristics of input signal and adjust the parameters for the above two blocks. At the end, a preferable block type is output to direct the subsequent windowing and transforming.

The following section shows a practical implementation of proposed method in perceptual audio coder to demonstrate the design of each major block more specifically.

4. APPLICATION TO MPEG-2/4 AAC

AAC [3] is the successor of MP3, and is a standard perceptual transform-based (MDCT) audio coder. It also employs the long/short block mechanism to allow adaptive window switching. The adaptation of the time-frequency resolution is done by shifting between transforms with input lengths to be either 2048 or 256 samples. The following illustrates how we embed the proposed method inside the AAC and tweak the system to achieve a best performance.

4.1. Integration with Encoder

Fig. 2 depicts the block diagram of AAC encoder where our block switching control unit is placed in the very first of the entire encoding process. There is no additional overhead for the analysis of long block once the short block is adopted. However, a long block may be forced into short block, if a "short/long-stop/short" sequence is generated. The psycho-acoustic-based method has to buffer previous data frame to execute such turn-round calculation which is not necessary for temporal-based method. In this paper, we design an efficient mapping of interference removal and consistency-checking to obtain a merit in speed without sacrificing the correctness of block switching decision. Fig. 3 shows the details of our block switching control. The necessity of each component will be discussed subsequently.

4.2. Interference removal

The HF signal may cause a spectral variation within a small time interval. To extract this significant portion, we perform high pass filtering and center-clipping for the signal in interest. Since the frame length is fixed (temporal resolution is varied) at different sample rate in AAC, we use Kaiser Window method to design a fixed 7-tap non-casual type-I FIR half-band filter. After quantization, only three filter coefficients are non-zero. Thus, we offer a low-complexity filter.

To further enhance the filtering of HF component, we place a center-clipping block right after high pass filter for flattening the spectrum nonlinearly. In addition, the center-clipping removes other unwanted signals like small energy fluctuation and the big DC spike. To compensate the rigid half-band high pass filter, we adaptively set the clipping threshold according to the energy difference between original signal and high pass residual. Specifically, Eq. (1) shows our formulation for the clipping-threshold (IRT) where "*offset_const*" and "*weight_const*" are parameters from experiential data to determine the absolute value of "*IRT*". An in-

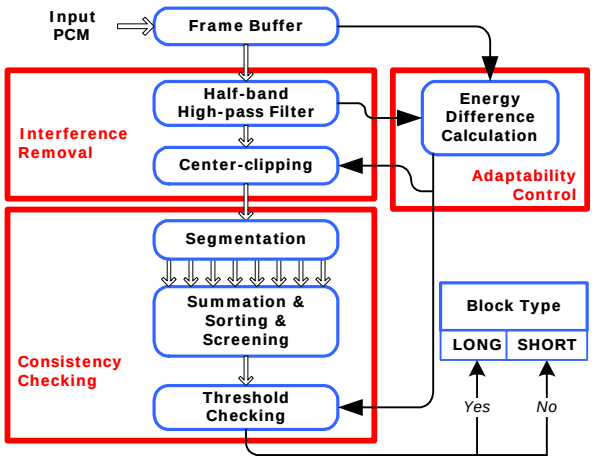

Figure 3. Detailed block diagram of proposed block switching control.

adequate training parameter might remove useful data and mislead the analysis. As shown, IRT is proportional to the logarithmic scale of HF energy. If the energy difference monitored by adaptability control unit is low which means current signal is HF dominant, we only have to additionally boost *IRT* to improve the signature of inconsistency.

$$IRT = offset\_const + log(HF) \times weight\_const \quad (1)$$

4.3. Consistency-checking

The signal (shaping residual) output by interference removal is screened by consistency-checking block to generate the decision for block switching. At first, the shaping residual is separated into eight groups. In each group, the samples are averaged to generate an energy coefficient. According to these energy coefficients of different groups, we calculate a ratio by dividing the maximal energy coefficient with the average of least non-zero three energy coefficients. Furthermore, we exploit such ratio to compare with an adaptive threshold "*CCT*" for determining the block type. Specifically, the calculation of "*CCT*" is formulated in Eq. (2):

$$CCT = (offset\_const - log(difference)) \times weight\_const \quad (2)$$

As shown in Eq. (2), another set of constants are employed but the energy difference is included in logarithmic scale. If the difference is high, "*CCT*" is supposed low to scrutinize the weaker HF component. Audio frame should be processed with short block type if the threshold is surpassed by the derived ratio in consistency-checking block.

Noticeably, beside the embodiment of adaptability, our energy segmentation is different from [3][4] as well. Because of the emphasis on consistency, the resolution of segmentation can be coarser. The energy coefficient is calculated by averaging all the samples but just choosing the largest one. We abandon the temporal relation and diminish it by sorting. Thus, subsequent threshold checking is no longer limited to adjacent energy coefficients.

4.4. Parameter Calibration

The signal quality of a codec not only depends on the choice of block type but also on the coding process itself. For the parameter adjustment in block switching control, one would wish both influences to be decoupled. A higher bit-rate may compensate the incorrect block choice where lower bit-rate may complicate the reason of quality degradation. Given a reasonable bit-rate setting that sustains the quantization noise just below the masking curve such as 128 kbps for 44.1 kHz sampled signal, we pre-set the block type by mandatory. A wrong block choice will generate the unbearable smearing artifact, and every frame of the audio signal should be experimented to secure its correct block type. Mandatory results are the calibrating targets we want to approach and the referencing patterns when we try to examine the efficiency of particular algorithm.

Aforementioned four constant parameters may vary with different sample-rate of audio signal. Several test sequences notorious for their hard-coding nature are collected in [7] as the training data to generate the constant parameters. The construction of database contains correct block type for different test sequences and sample-rate rely greatly on the effort of subjective listening test. Not every frame belong to long block is also suitable for short block. The identification of block-insensitive frame can add some flexibility during calibration process. We can successfully converge the calibration toward a set of integer parameters in the range of 3~20. It proves the feasibility of our algorithm.

5. PERFORMANCE EVALUATION

Fig. 4 demonstrates the process of block switching method. Fig. 4 (a) shows the target audio signal and its shaping residual. The time line is separated into eight sections and each section confines the data as an individual analysis group. In Fig. 4 (b), eight diagrams show the spectral distributions of eight different groups in original signal (dark color) and shaping residual (light color). Obviously, we can discover that three transients are respectively at group 1, 4 and 6. Besides, these three groups also possess different spectral characteristics from other five groups. Fig. 4 (c) illustrates the shaping residual averages of eight groups. We can easily recognize the inconsistency and conclude that the current frame is better coded with short block type by observing three conspicuous pillars.

Fig. 5 expresses the comparison between other methods and ours. In Fig. 5 (a), we use the results obtained in the parameter calibration phase to evaluate the correctness of block switching in terms of accuracy. Evaluation results of eight test sequences [7] are listed for assessment. Among these sequences, "Fatboy" sounds like robot speaking. It sustains more energy at HF and slower rate when comparing with human pitch. Its obvious energy abrupt makes the job of block switching effortless. Similar characteristic also exists in famous "Castanets" and others. After the algorithms

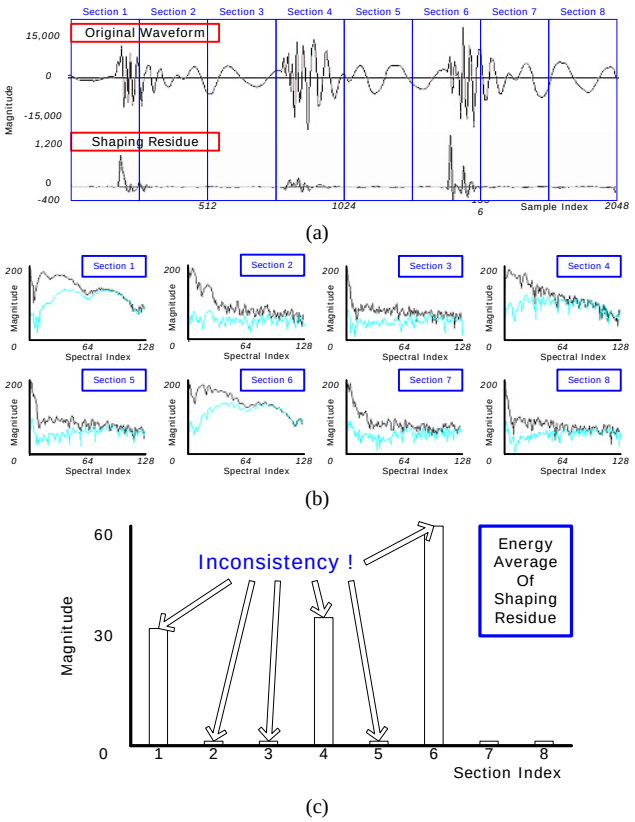

Figure 4. The process of proposed method: (a) target audio signal and its shaping residual, (b) spectral distribution of original waveform and shaping residual (dark/light color) in different groups and (c) shaping residual average.

are well-tuned for identifying the transient event, transient-locating methods, [4][5], inevitably meet obstacles on processing "German Speech". However, by monitoring the consistency, our approach can successfully handle all sequences. In addition, our adaptability control can compensate the over-switching caused by abrupt-sensitive behavior. The ODG (object difference grade) scores of our method in Fig. 5 (a) generated by PEAQ (perceptual evaluation of audio quality) depicts the high accuracy is also conform to a high encoding quality.

Regarding to the computation power, we assume the psychoacoustic analysis of each method is realized at FFT-domain. The difference in run-time between spectral method and temporal one will increase as the growing of short blocks. Fig. 5 (b) illustrates the case of Fatboy with maximal number of short block. Particularly, we use MPEG VM for baseline and normalize its computation power as 1. As shown in Fig. 5 (b), we spend a comparable or lower effort with other temporal methods but acquire a tremendous advantage in correctness.

6. CONCLUSION

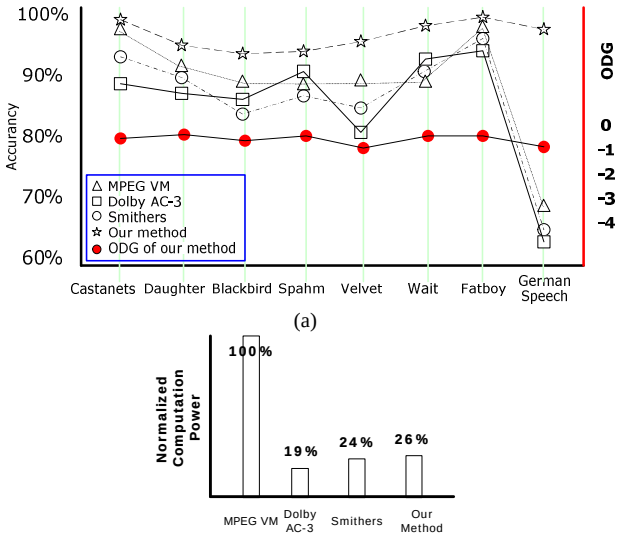

Figure 5. Comparisons between our method and other three approaches in (a) accuracy, quality (ODG) and (b) computation power.

In this paper, we propose a temporal method to perform block switching for transform-based perceptual audio coder. Specifically, we scrutinize the signal energy to examine the consistency of spectral characteristics within a short time interval. Furthermore, we use such consistency information as a criterion for block type decision. Besides, we propose an adaptability control to monitor the variation of energy for addressing the non-stationary characteristic of audio signal. Performance evaluation results with MPEG-2/4 AAC show that our approach can provide the best encoding quality while maintaining lower and similar complexity as compared to other state-of-the-art methods. This work shows that consistency check is a better criterion than abrupt energy change for block type decision. Also, the adaptability control can effectively secure the performance in various audio signals. With proper modification, the proposed scheme can be applied in other perceptual audio coders.

7. REFERENCE

[1] B. Edler, "Coding of Audio Signals with Overlapping Block Transform and Adaptive Window Functions," (in German), *Frequenz*, vol. 43, pp. 252-256, 1989.

[2] T. Liebchen, "MPEG-4 Lossless Coding for High-Definition Audio," *115th AES convention*, New York 2003.

[3] ISO/IEC JTC1/SC29/WG11 (MPEG), International Standard ISO/IEC 14496-3: "Coding of Audio Visual Objects, Part 3: Audio", 1999 - 2001.

[4] United States Advanced Television Systems Committee Digital Audio Compression (AC-3) Standard, Doc. A/52/10, Dec. 1995.

[5] M. J. Smithers, M. C. Fellers, "Increased efficiency MPEG-2 AAC Encoding," *111th AES convention*, New York 2001.

[6] J. Herre, J. D. Johnson, "Enhancing the Performance of Perceptual Audio Coders by Using Temporal Noise Shaping (TNS)," *101st AES convention*, Los Angeles 1996.

[7] http://www.ff123.net