CODING OF AUDIO SIGNALS WITH OVERLAPPING BLOCK TRANSFORM AND ADAPTIVE WINDOW FUNCTIONS

Stefan Ringer

ABSTRACT

Modern day audio applications demand for high quality perception while fulfilling data limitation constraints. Audio coding is used for that matter. One quality degrading effect in coded audio is pre-echo distortion that occurs if the signal peaks close to the end of a transform block of otherwise low energy. Such situations occur i.e. in German Male Speech or in percussion. In this paper one method introduced by Edler[1] that relies on adaptively changing the window and transform block length will be reviewed and analyzed with regards to advantages, drawbacks and alternatives.

1. INTRODUCTION

Nearly all audio coders make use of some transformation from time to a domain that makes exploiting masking effects, redundancies and quantizing easier. A common tool for this task are filter banks. They split the signal up into different subbands, which are all handled by individual (often modulated protype lowpass) filters. Since ideal "brickwalling" of the spectrum is not possible (finite frequency responses correspond to infinite thus not easily realizable impulse responses of the filters), overlapping block transforms that fulfill a perfect reconstruction criterion are widely used [2].

A widely applied special case is the MDCT (modified discrete cosine transform) with the condition $L = 2M$ (L being the block transform length and the block advance being M samples). This transform generates M coefficients (despite the transform length $2M$).
It is desirable to maximize coding gain by choosing a big block transform length to exploit stationary parts of the signal in the quantization step. In practice this is not always possible since speech or music tend to only be stationary on very small and varying timescales. Too short transforms will result in degraded coding gain, too long transforms in audible distortions.
One of those are pre-echoes. They occur if a block encompasses a region of low energy followed by a high peak near the end of the block. Since the average spectral estimate (energy) of the block is high due to the peak, the computed masking thresholds are high as well.
When coarse quantization is carried out accordingly and the block is reconstructed, the silent (low energy) region has gained quite a lot of power since the quantization distortion is spread throughout the reconstructed block due to time-frequency uncertainty. Since temporal masking is way stronger forwards than backwards noise (reminding of an echo) is heard before and not after the peak e.g. a percussion hit. Hence the name "pre-echo".
This effect is especially noticeable in overlapping block transforms as the distortion is now spread across $2L$ not L samples compared to non-overlapping block transforms.
Since this is not only a rare condition but i.e. very prominent in German male speech and percussion, this is a very interesting problem to solve.

2. METHODS

Prior to Elder's work the MDCT with $L = 2M$ was already known. Since the transform generates M coefficients from $2M$ samples, obviously aliasing is present in the inverse $M \to 2M$ transform which has to be cancelled out (TDAC) by choosing the window functions of successive blocks according to the Pricen-Bradely conditions [3]:

$$f^2(n) + f^2(N+n) = 1$$
$$f(n) \cdot f(N-1-n) = f(N+n) \cdot f(2N-1-n)$$

Which can be summarized in a simple symmetry condition [1]:

$$f(n) = f(2N-1-n)$$

Since only the window function was constrained for perfect reconstruction, the block (transform) length could be arbitrary – as long it was the same for all blocks.

The innovation introduced by Edler was to demonstrate how one could change the transform block length from block to block when a new constraint was fulfilled for the successive window functions $f(n)$ and $g(n)$:
In an overlapping block transform of length $2N$ the overlapping area is $0 \le n \le N-1$. The input $x(n)$ of length $2N$ is transformed this way:
First multiply it with a window function $f(n)$ (end of first block) respectively $g(n)$ (start of second block). Then apply DCT transform and subsample the result by a factor of 2.

[1] B. Edler, "Codierung von Audiosignalen mit überlappender Transformation und adaptiven Fensterfunktionen," Frequenz, pp. 252-256, 1989.

This obviously leads to time domain aliasing. When inverting the transform the output is different from the input and reads:

$$y(n) = f^2(N+n)\cdot x(n) + f(N+n)\cdot f(2N-1-n)\cdot x(N-1-n) + g^2(n)\cdot x(n) - g(n)\cdot g(N-1-n)\cdot x(N-1-n)$$

Demanding perfect reconstruction ($y(n) = x(n)$) yields the following conditions:

$$f(N+n)\cdot f(2N-1-n) = g(n)\cdot g(N-1-n)$$
$$f^2(N+n) + g^2(n) = 1$$

This allows for the usage of different window forms in successive blocks. Further since only the second half of $f(n)$ (first block) and the first half of $g(n)$ (second block) are important for this reconstruction condition, only the overlapping parts of the window functions are restrained by this property. This allows for asymmetric window functions that on one side can have big overlapping areas and on the other side have small overlapping areas.

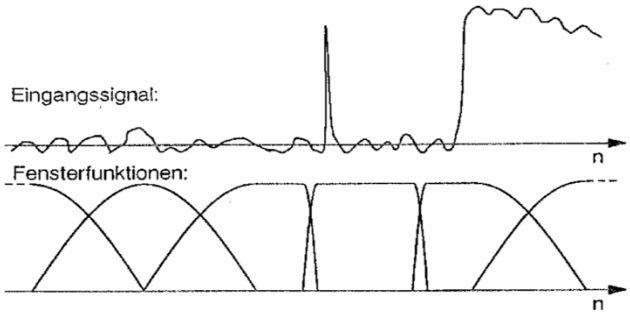

Figure 1: Window functions adapted to signal [1]

This is the key to adaptively changing the block (transform) lengths L – only the overlapping parts – whose lengths can be chosen independently – of the windows are constrained, not the length of the windows. This leads to the final equation [1]:

$$f_L(n)\cdot f_L(L-1-n) = g_L(n)\cdot g_L(L-1-n)$$
$$f_L^2(n) + g_L^2(n) = 1$$

For $0 \le n \le L-1$ where the overlap areas aren't bigger than any of the involved transform lengths L.
Notice how the previously known Princen-Bradely condition from $f(n) = g(n)$ of this more general result.

For useful implementation a control algorithm is needed that decides on the length of the used transform. The algorithm proposed by Edler employs two coupled criteria for the impulse detection, one being the relative energies of two successive blocks and the other being the max-average ratio of the samples within a block. When a peak is detected, the coder switches to a shorter transform length and thus reduces pre-echoes while still exploiting stationary signal parts by not generally employing short transform lengths.

Comparison of the proposed algorithm to the naïve approach without length adaption gave improved perceived quality scores [1].

3. ALTERNATIVES IN THE LITERATURE

To this day similar strategies are employed in widespread codecs like the famous MP3 [4], partly because it is compatible with MDCT-based algorithms for which fast FFT based implementations are available [5]. Still this method has significant drawbacks as pointed out by [2]:

- Perceptual model and lossless coding algorithms need to support multiple time resolutions
- To meet window constraints, time and frequency localization performance must be sacrificed which in turn leads to degraded coding gain [6].
- High coder delay
- Overusage of short windows for "pitched" signals

There are also alternative strategies to combat pre-echoes:

Bit Reservoir [7]: Since transients need more bits to be properly encoded and other parts need less, the idea is to store bits from intensive parts in less resource intensive parts. A problem is though that in signals with lots of transients unfeasibly large bit reservoirs are needed.

Hybrid and Switched Filter Banks: In Window Switching (as discussed above) only the transform and window length are adapted. Contrary in Switched Filter Banks [8] the filter banks themselves are exchanged depending on the current signal statistics. In Hybrid Filter Banks [9] they are cascaded to achieve optimal time-frequency tiling.

Gain Modification [9]: Before coding, transients are smoothed and then processed like stationary blocks. Since the block is smoothed the temporal spread of quantization noise is not as significant and thus results in a reduced pre-echo. Though this approach can lead to issues of broadening the filter bank responses at low frequencies beyond critical bandwidth while measures to tackle this problem struggle with coder complexity.

4. CONCLUSIONS

Pre-echoes are significant distortions caused by quantization in audio coding that happen in occurrence of transients in block-based transform coding.

Adaptive window and transform length switching showed a novel way to reduce these issues while retaining advantages of overlapping block transforms such as easy and fast FFT based implementation and high coding gains. Its significance is underlined by its widespread usage.

Yet with this approach new issues arise which might be better addressed with alternative approaches such as the above-mentioned alternatives.

5. REFERENCES

[1] B. Edler, “Codierung von Audiosignalen mit überlappender Transformation und adaptiven Fensterfunktionen,” Frequenz, pp. 252-256, 1989.

[2] Painter, T & Spanias, A 2000, 'Perceptual coding of digital audio', Proceedings of the IEEE, vol. 88, no. 4, pp. 451-512.

[3] Princen, John P.; Johnson, A.W.; Bradley, Alan B. (1987). "Subband/Transform coding using filter bank designs based on time domain aliasing cancellation". ICASSP '87. IEEE International Conference on Acoustics, Speech, and Signal Processing. 12: 2161–2164.

[4] ISO/IEC JTC1/SC29/WG11 MPEG, IS11172-3 “Information Technology - Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to About 1.5 Mbit/s, Part 3: Audio” 1992. (“MPEG-1”)

[5] P. Duhamel, et al., “A Fast Algorithm for the Implementation of Filter Banks Based on Time Domain Aliasing Cancellation,” in Proc. Int. Conf. Acous., Speech, and Sig. Process. (ICASSP-91), pp. 2209-2212, May 1991.

[6] S. Shlien, “The Modulated Lapped Transform, Its Time-Varying Forms, and Its Applications to Audio Coding Standards,” IEEE Trans. on Spch. and Aud. Proc., v. 5, n. 4, pp. 359-366, Jul. 1997.

[7] J. Johnston, et al., “AT&T Perceptual Audio Coding (PAC),” in Collected Papers on Digital Audio Bit-Rate Reduction, N. Gilchrist and C. Grewin, Eds., Aud. Eng. Soc., pp. 73-81, 1996

[8] D. Sinha and J. Johnston, “Audio Compression at Low Bit Rates Using a Signal Adaptive Switched Filterbank,” in Proc. Int. Conf. Acous., Speech, and Sig. Proc. (ICASSP-96), pp. 1053-1056, May 1996.

[9] J. Princen and J. Johnston, “Audio Coding with Signal Adaptive Filterbanks,” in Proc. Int. Conf. Acous., Speech, and Sig. Proc. (ICASSP-95), pp. 3071-3074, May 1995.

[10] T. Vaupel, “Ein Beitrag zur Transformationscodierung von Audiosignalen unter Verwendung der Methode der ‘Time Domain Aliasing Cancellation (TDAC)’ und einer Signalkompandierung in Zeitbereich,” Ph.D. Thesis, 1991.