RD OPTIMAL TEMPORAL NOISE SHAPING FOR TRANSFORM AUDIO CODING

O.A. Niamut and R. Heusdens

Dept. of Mediamatics
Delft University of Technology,
Delft–The Netherlands
E-mail: {O.A.Niamut, R.Heusdens}@ITS.TUDelft.NL

ABSTRACT

In this article we investigate rate-distortion optimal temporal noise shaping for transform audio coding. Temporal noise shaping, or TNS, is a technique for reshaping the quantization noise in the time domain through open-loop linear predictive coding of frequency domain coefficients. Traditionally, a selection mechanism based on prediction gain is employed to determine whether it is advantageous to apply TNS or not. Although this method is effective for reducing coding artifacts in transient and speech signals, critical adjustment of the prediction gain threshold is necessary to avoid excessive bit rate demands. We propose the use of TNS in a rate-distortion optimization framework. Within this framework a jointly optimal selection of the prediction filter order and the quantizer for coding the coefficients can be made, such that the perceptual distortion is minimized for a given target rate. Experimental results for an MDCT-based audio coding system are presented and it is shown that TNS within an RD optimization framework outperforms the existing TNS method.

1. INTRODUCTION

In audio coding applications the goal is to minimize the perceptual distortion introduced by the coding process whilst satisfying a bit rate constraint. This goal can be achieved by the application of an operational rate-distortion (RD) optimization approach [1, 2] that reaches the solution to the audio coding problem for a given coding environment and allows for optimal selection of the involved coding parameters. For example, the work in [3] applies this approach in an MPEG-2/4 AAC audio coder [4, 5].

Most audio coding schemes rely on a time-frequency analysis of the input signal and typically, an MDCT is applied for this purpose. The MDCT has desirable properties, such as good channel separation, strong stopband attenuation, minimum blocking artifacts and the availability of fast algorithms. Additionally, there are various techniques available for efficient resolution switching to further enhance the performance. Temporal noise shaping (TNS) [6, 7] is such a technique that allows for block-continuous adaptation of the time-frequency resolution

In this paper, we propose an operational rate-distortion optimization framework for TNS in an MDCT-based audio coder. The paper is organized as follows. In Section 2 the traditional TNS technique is explained and some of its potential problems are highlighted. Next, we propose an RD optimal temporal noise shaping algorithm in Section 3. Finally, both quantitative and qualitative experimental results for encoding several audio fragments with an MDCT-based audio coding scheme are presented in Section 4.

The research was conducted within the ARDOR project, supported by the E.U. grant no. IST-2001-34095

2. TEMPORAL NOISE SHAPING

TNS [6, 7] is a technique for reshaping and controlling the quantization noise in the time domain through open-loop linear predictive coding (LPC) of frequency domain coefficients. Given a block of signal samples that is transformed to the frequency domain, an LPC filter is applied on a (sub)set of transform coefficients and instead of direct quantization of the frequency domain coefficients, the filtered residual is quantized along with the LPC filter coefficients. Upon reconstruction of the time domain signal, the inverse LPC filter is applied on the quantized residual before the inverse signal transform. This inverse LPC filter acts as a temporal envelope that shapes the quantization noise according to the signal energy distribution over the segment, thereby permitting a coder to exercise control over the temporal structure of quantization noise within a set of frequency coefficients. Thus, most of the quantization noise will reside in signal regions with significant energy in the time domain, thereby avoiding temporal masking problems in coding transient and speech signals, such as pre-echos and reverberation.

TNS is part of the MPEG-2/4 AAC standard [4, 5] where it is applied on coefficients obtained from using a modified discrete cosine transform (MDCT) on the input signal. The MDCT is a so-called 50% overlapped block transform, i.e. a transform where samples from consecutive 50% overlapping segments are windowed and transformed. Given a signal x divided into overlapping segments of length $2M$, a set of M transform coefficients X_i is computed from the ith segment x_i by applying the direct MDCT, which is defined as

$$X_i(k) = \sum_{n=0}^{2M-1} x_i(n)w(n)\cos\Big[\frac{\pi}{4M}(2n+M+1)(2k+1)\Big],$$

where $k = 0, 1, \ldots, M{-}1$ and w an appropriate analysis window.

Assume that X_i has variance $\sigma^2_{X_i}$. A pth order linear prediction $\tilde{X}_i$ of X_i is given by

$$\tilde{X}_i(k) = \sum_{j=1}^{p} a_j X_i(k-j).$$

The prediction error signal $\Delta X_i(k) = X_i(k) - \tilde{X}_i(k)$ with variance $\sigma^2_{\Delta X_i}$ is then computed (in the Z-transform domain) as

$$\Delta X_i(z) = X_i(z)A(z),$$

with $A(z) = 1 - \sum_{j=1}^{p} a_j z^{-j}$. We want to find the filter $A(z)$ that minimizes $\sigma^2_{\Delta X_i}$ and thus maximizes the prediction gain $\sigma^2_{X_i}/\sigma^2_{\Delta X_i}$. This can be done efficiently with the well-known Levinson-Durbin recursion algorithm.

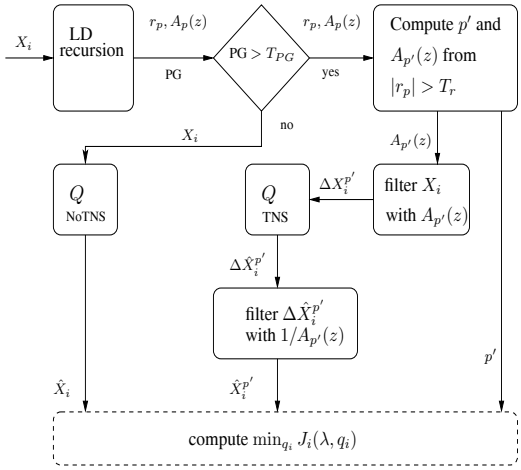

Fig. 1. *Block scheme of a standard TNS implementation based on prediction gain. The RD optimization block is only applied in our experimental tests and is explained in Section 3.*

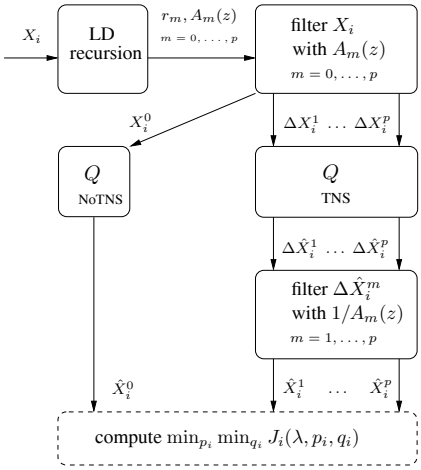

Fig. 2. *Block scheme of the proposed TNS implementation in an RD optimization framework.*

In many implementations of TNS in the AAC standard, the prediction gain is used to determine whether TNS should be applied or not. First, for a block of MDCT coefficients a high order (e.g. 12 or 20) LPC filter is computed with the autocorrelation method using the Levinson-Durbin (LD) algorithm. If the prediction gain is larger than a certain threshold T_{PG}, TNS is applied. The Levinson-Durbin algorithm generates a set of reflection coefficients r. The final LPC filter order p' is determined by subsequently removing reflection coefficients having an absolute value lower than a threshold T_r from the reflection coefficient array. This procedure lowers the side information for sending the filter coefficients. A block scheme of a typical TNS implementation, which we shall refer to as TNS PG (prediction gain), is shown in Fig. 1. We can distinguish two separate quantizer blocks since it is generally efficient to have separate quantizers for unfiltered and filtered MDCT coefficients.

It has been recently recognized that TNS can cause several undesirable coding artifacts [8]. For example, the quantization noise increases with the LPC filter order and is amplified around attacks, which might result in unmasked quantization noise. Moreover, the prediction gain does not always accurately represent the coding performance resulting from TNS usage. If TNS is applied, but the actual coding performance turns out to be insignificant, a large portion of the available bit rate unnecessarily goes to the filter coefficients. Clearly, the choice of the thresholds on prediction gain T_{PG} and reflection coefficients T_r is a delicate one, since it determines both the side information for sending the filter and the occurrence of undesired artifacts. Therefore, the thresholds should be made dependent on the target bit rate, which requires a difficult tuning process. The method in [8] proposes a perceptual entropy measure for determining the usage of TNS. This is basically a one-sided bit rate cost measure, since perceptual entropy can be seen as the minimal bit rate for transparent audio coding. We extend this notion to a two-sided cost measure of both bit rate and perceptual distortion.

3. RD OPTIMAL TNS

In an operational RD optimization framework [1, 2], the distortion D is minimized subject to a bit rate constraint R_T for a given coding environment. In the TNS case the coding environment consists of an MDCT-based audio coding system for coding a signal consisting of, say N, overlapping segments. Given a set of LPC filter orders $\mathcal{P}$ and a set of quantizer stepsizes $\mathcal{Q}$ for coding the MDCT coefficients, either directly or after LPC filtering, we want to select for the ith segment x_i the LPC filter order $p_i \in \mathcal{P}$ and the quantizer stepsize $q_i \in \mathcal{Q}$ that jointly minimize the perceptual distortion, denoted by D_π for a given target rate R_T. Note that a zero-order LPC filter, i.e. $p_i = 0$, indicates direct coding of the MDCT coefficients. Fig. 2 shows a block scheme of the RD-based TNS algorithm, referred to as TNS RD.

Let $\mathbf{p}$ be the N-dimensional vector consisting of the selected filter orders for all N segments and let $\mathbf{q}$ denote the vector containing the N quantizer stepsizes. The rate-constrained problem of minimizing D_π subject to R_T can then be written as

$$\min_{\mathbf{p}} \min_{\mathbf{q}} D_\pi(\mathbf{p}, \mathbf{q}) \quad \text{s.t.} \quad R(\mathbf{p}, \mathbf{q}) \leq R_T,$$

which can be converted into the unconstrained problem

$$\min_{\mathbf{p}} \min_{\mathbf{q}} J(\lambda, \mathbf{p}, \mathbf{q}), \tag{1}$$

where $J(\lambda, \mathbf{p}, \mathbf{q}) = D_\pi(\mathbf{p}, \mathbf{q}) + \lambda R(\mathbf{p}, \mathbf{q})$ is a Lagrangian cost function of perceptual distortion D_π and bit rate R for coding the complete signal. The parameter $\lambda \geq 0$ has to be chosen such that the target rate is met, i.e. $R(\mathbf{p}, \mathbf{q}) = R_T$. The bit rate can be split up into contributions from the coded -and possibly filtered- MDCT coefficients and side information, such as the LPC filter order, the quantizer stepsize and the coded LPC filter coefficients. Under the assumptions of additivity of the cost measure and independent coding over segments (or, equivalently, independence of the cost measure over segments), the problem from Eq. (1) can be formulated as

$$\min_{\mathbf{p}} \min_{\mathbf{q}} \sum_{i=1}^{N} J_i(\lambda, p_i, q_i) = \sum_{i=1}^{N} \min_{p_i \in \mathcal{P}} \min_{q_i \in \mathcal{Q}} J_i(\lambda, p_i, q_i), \tag{2}$$

where $J_i(\lambda, p_i, q_i) = D_{\pi,i}(p_i, q_i) + \lambda R_i(p_i, q_i)$ denotes the cost for segment x_i. As previously mentioned, λ has to be chosen such that the target rate is met. Hence, if $R(\mathbf{p}, \mathbf{q}) \neq R_T$, we have to modify λ and solve Eq. (2) for the new λ. This is a convex problem

in λ and fast algorithms exist to solve the problem, e.g. the bisection algorithm in [1].

Given that the MDCT coefficients obtained for segment x_i are filtered with an LPC filter of order p and quantized with stepsize q, the perceptual distortion $D_{\pi,i}(p,q)$ is derived as a perceptually weighted squared sum of the difference between the original MDCT coefficients X_i before TNS and the quantized coefficients $\hat{X}_i^p$ after the inverse TNS operation, that is,

$$D_{\pi,i}(p,q) = \sum_{k=0}^{M-1} \alpha_i(k)\big[X_i(k) - \hat{X}_i^p(k)\big]^2. \qquad (3)$$

The set of M perceptual weighting coefficients α_i is taken as the inverse masking curve for segment x_i evaluated at the MDCT center frequencies, i.e. $\alpha_i(k) = msk_i^{-1}(\frac{\pi}{M}(k+\frac{1}{2}))$, $k = 0, 1, \ldots, M$–1. This has the desired effect that spectral distortions in frequency regions with strong masking power are de-emphasized. Moreover, if the TNS and subsequent quantization operations are performed on the weighted coefficients $X_i' = \alpha_i X_i$, computing the ℓ_2 distortion values for the weighted coefficients is equivalent to computing the perceptual distortions with Eq. (3), under high-resolution assumptions.

4. EXPERIMENTAL RESULTS

In our experiments, a simple implementation of TNS in an MDCT-based audio coding system was considered. The LPC filters were applied on the complete set of MDCT coefficients. Although the quantization of the LPC filter coefficients can be taken into account, errors due to imperfect modelling of the LPC filter were neglected in this study. As discussed in the previous section, for every segment of the input signal both the LPC filter order and the coding template were selected that minimized the perceptual distortion over all segments, subject to a target rate constraint for coding the complete signal. We compared this implementation with both a system not employing TNS and the existing TNS method, i.e. based on prediction gains. In all three systems the selection of quantizers was performed in an RD optimal manner, as outlined in the previous section.

A 1024-channel MDCT was used, similar to the AAC long block operation, along with a 2048-sample Kaiser-Bessel derived window. Masking curves for each segment were computed according to the perceptual model in [9]. Based on settings in [5], the maximum LPC filter order was set to 20 and the prediction gain threshold T_{PG} for selection of TNS was set to 1.4 dB. The threshold for discarding high-order reflection coefficients was set to $T_r = 0.1$. For both direct and TNS filtered coefficients, eight quantizer stepsizes were available, including a stepsize for quantizing all coefficients to zero, and corresponding Huffman codebooks were designed. Additionally, efficient coding of long zero regions was applied. As a distortion measure, the perceptual distortion measure D_π from Eq. (3) was taken and the Huffman codewords were used to determine the bit rate. The side information consisted of the LPC filter order (5 bits), the filter coefficients (4 bits per filter coefficient) and the quantizer stepsizes (3 bits). The perceptual weighting coefficients, derived from the masking curve were assumed to be coded at 4 kbps, in line with results from [10]. A set of four audio fragments (48 kHz, 16-bit, mono) was used for evaluation of the three algorithms, consisting of the castanet, German male speech, bass guitar and English female speech signals.

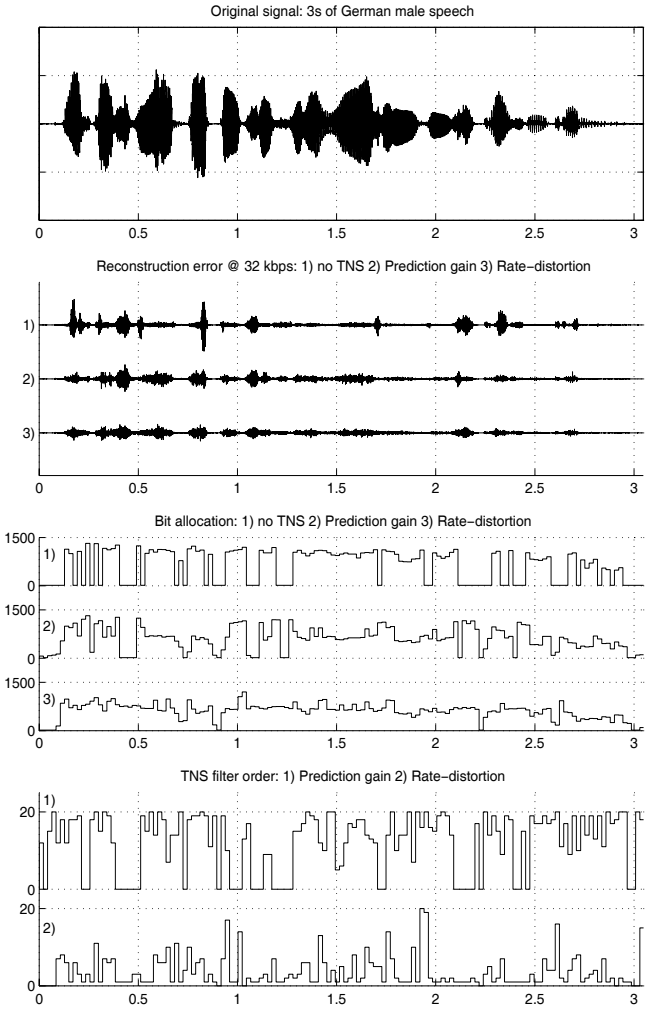

Fig. 3. *Upper plot: 3s of German male speech signal. 2nd plot: Reconstruction error signals for the 3 methods. 3rd plot: Bit allocation over segments for the 3 methods. Lower plot: LPC filter order for PG-based method and RD-based method.*

4.1. Results for single fragment

Fig. 3 presents coding results for 3 seconds of the German male speech fragment (upper plot), coded at 32 kbps. In the 2nd plot, the reconstruction error signals are displayed for the three algorithms and we see that the RD-based algorithm leads to reconstruction noise that is shaped similarly to the PG-based method. The 3rd plot shows the bit allocation over segments, where the number of bits for both transform coefficients and side information per segment is shown. The standard TNS method lowers the peak bit rate demand compared to the system without TNS, however, bit shortages still occur at several positions in the signal, mainly during segments with low energy signal content. This can be explained from the fact that at low bit rates, the existing schemes frequently run out of bits in various segments. Since selection of the stepsize that quantizes all coefficients to zero requires very few bits, this remains the only choice at these segments, thereby creating gaps in the reconstructed signal. Although this gap artifact can not be attributed to the TNS algorithm directly, it is a clear example of inefficient use of the available bits when the TNS algorithm operates outside the rate-distortion con-

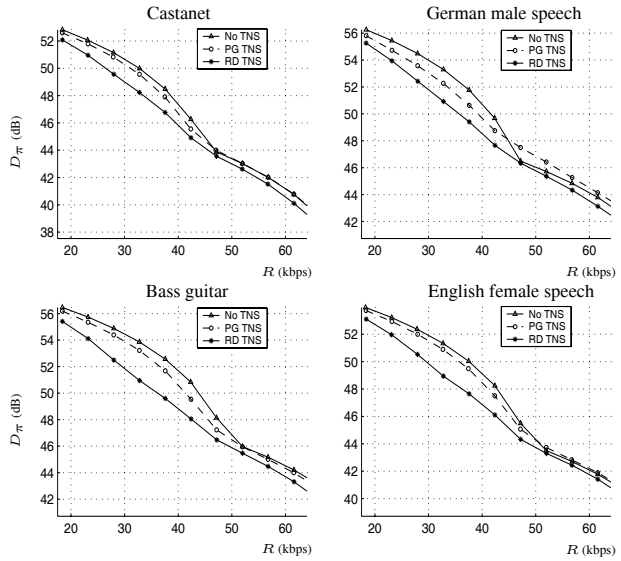

Fig. 4. *Rate vs. perceptual distortion plots for the set of four audio fragments.*

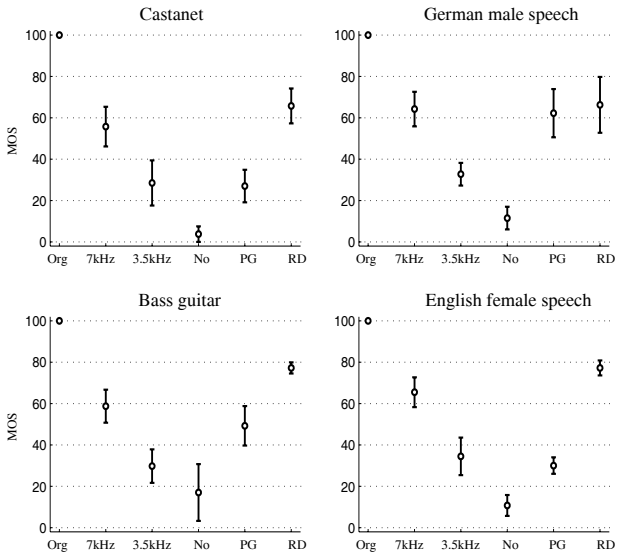

Fig. 5. *MUSHRA test results for the original, the anchors and the coded versions without TNS, PG-based TNS and RD-based TNS.*

trol in this coding framework, or when not using TNS. In contrast, the RD-based method reduces the bit demand even further and facilitates a more continuous bit rate distribution over various signal parts. This allows for a more continuous signal modelling. The bit savings can be explained from the lower plot, where it is seen that the RD-based algorithm selects high order LPC filters only at critical signal parts. In contrast, the standard method uses a high order LPC filter almost exclusively, which in some cases requires an unnecessary high amount of bits.

4.2. Results for multiple fragments

For the four signals, objective perceptual distortions for different target bit rates ranging from 18 to 64 kbps are presented in Fig. 4. It can be seen that the RD-based method results in the lowest perceptual distortion at every bit rate for all fragments. In order to subjectively evaluate our scheme, a MUSHRA [11] listening test was performed for evaluating the four fragments at a bit rate of 36 kbps. At this bit rate, no obvious time gap artifacts where present in the coded fragments where TNS was applied. A total of ten listeners participated (authors not included). The results averaged over all listeners are displayed per fragment in Fig. 5. For three fragments, the RD-based method significantly outperforms the PG-based method. Only for the German male speech fragment no significant improvement is observed. This can be explained as follows. Since a perceptual model is used that accounted for simultaneous masking only, the RD-based method does not concentrate specifically on time domain coding artifacts such as speech reverberation. The PG-based method determines TNS usage outside the perceptual model, hence a larger reduction of these artifacts is obtained than with the RD-based method. Therefore, we expect improved performance when a perceptual model that incorporates temporal masking is applied. We conclude that the main contribution of the proposed algorithm lies in increased bit rate reduction compared to the existing method. This performance gain is obtained at a higher complexity, mostly determined by the repeated inverse IIR filtering. Therefore, complexity reductions will focus at estimating the perceptual distortion in the LPC filtered domain.

5. REFERENCES

[1] K. Ramchandran and M. Vetterli, "Best wavelet packet bases in a rate-distortion sense," *IEEE Trans. Image Processing*, vol. 2, no. 2, pp. 160–175, Apr. 1993.

[2] P. Prandoni and M. Vetterli, "R/D optimal linear prediction," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 6, pp. 646–655, Nov. 2000.

[3] A. Aggarwal, S. Regunathan, and K. Rose, "Trellis-based optimization of MPEG-4 advanced audio coding," in *Proc. 2000 IEEE Workshop on Speech Coding*, Wisconsin, USA, Sept. 2000, pp. 142–144.

[4] M. Bosi and et al, "ISO/IEC MPEG-2 advanced audio coding," *Journal AES*, vol. 45, pp. 789–812, Oct. 1997.

[5] 3GPP, "TS 26.403 V6.0.0: Encoder specification AAC part (release 6)," ftp://ftp.3gpp.org/specs/2004-12/Rel-6/26_series/26403-600.zip, Nov. 2004.

[6] J. Herre and J.D. Johnston, "Enhancing the performance of perceptual audio coders by using temporal noise shaping," in *Proc. 101st Aes Conv.*, Los Angeles, USA, Nov. 1996.

[7] J. Herre and J.D. Johnston, "Continuously signal-adaptive filterbank for high-quality perceptual audio coding," in *Proc. Workshop Appl. Signal Proc. to Audio and Acoustics*, New Paltz, USA, Oct. 1997.

[8] C.M. Liu, W.C. Lee, and T.W. Chang, "The efficient temporal noise shaping method," in *Proc. 116th Aes Conv.*, Berlin, Germany, May 2004.

[9] S. van de Par, A. Kohlrausch, G. Charestan, and R. Heusdens, "A new psychoacoustical masking model for audio coding applications," in *Proc. Int. Conf. Acoustics, Speech and Signal Proc.*, Orlando, USA, May 2002, pp. 1805–1808.

[10] O. Niemeyer and B. Edler, "Efficient coding of excitation patterns combined with a transform audio coder," in *Proc. 118th Aes Conv.*, Barcelona, Spain, May 2005.

[11] ITU-R BS.1534-1, "Method for the subjective assessment of intermediate quality level of coding systems," 2003.