16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

PRE-ECHO REDUCTION IN THE ITU-T G.729.1 EMBEDDED CODER

Balázs Kövesi[†]*, Stéphane Ragot*[†]*, Martin Gartner*[*]*, Hervé Taddei*[*]

[†]France Telecom R&D, Lannion, France
E-mail: {balazs.kovesi,stephane.ragot}@orange-ftgroup.com

ABSTRACT

Pre-echo is a well-known artefact of transform coding at low bit rates. In this paper we present a new method to address this problem. The input signal is assumed to be coded in two stages: in time domain first, and then in transform domain. This is for instance the case in CELP+transform embedded coding. The first stage reconstructs a signal that is usually free of pre-echo. Therefore transform coding can exploit this reconstructed signal as side information for pre-echo detection and reduction. The proposed method is implemented as an adaptive limiter at the decoder side and does not need transmission of auxiliary data. It is part of the recently standardized ITU-T G.729.1 coder, in which it is used in two separate subbands. Experimental test results show that this method has a significant impact on quality in G.729.1 with very small complexity.

1. INTRODUCTION

Pre-echo is a typical artefact in low-bit-rate transform coding. It is audible especially in regions preceding sharp transients, such as clean speech onsets or percussive sound attacks (e.g. castanets), where pre-masking is not effective [1]. Indeed, pre-echo is coding noise that is injected in transform domain but is spread in time domain over the synthesis window by the transform decoder. For a transient with sharp energy increase, the low-energy region of the input signal preceding the transient is therefore mixed with noise, and the signal to noise ratio is often negative in such low-energy parts. A similar artefact, post-echo, exists after a sudden signal offsets. However post-echo is usually less a problem due to post-masking properties [1]. Also, in real sounds recordings a sudden signal offset is rarely observed due to reverberation. In the rest of this article the term echo is used for both pre-echo and post-echo generated by transform coding.

Many methods have been proposed to tackle the problem of echo in transform audio coding, especially for the case of modified discrete cosine transform (MDCT) coding. The most popular approach is to make the filterbank signal-adaptive, using window switching controlled by transient detection [2, 3, 4] or close-loop decision. Usually window switching implies extra delay and complexity compared with using a non-adaptive filterbank ; furthermore, short windows yield lower transform coding gains than long windows, and side information needs to be sent to the decoder to indicate the switching decision. A similar idea (in frequency domain) is to use adaptive subband decomposition via biorthogonal lapped transform [5]. Another popular approach consists in performing temporal noise shaping (TNS) [6]. Note that TNS requires the transmission of noise shaping filter coefficients [6] or time envelope parameters [7] as side information, which requires extra bitrate. Other methods have been considered, e.g. transient modification prior to transform coding [8] or synthesis window switching controlled by transient detection at the decoder [9].

In this paper we propose a new method to reduce echo artefacts after transform decoding. This method requires the transform coding stage to be conducted in a second stage after a first-stage time-domain coding. This type of constraint is found in embedded audio coders based for instance on code-excited linear predictive (CELP) + transform coding [9, 10]. In this specific case echo reduction can exploit *for free* side information from the first-stage coder. Therefore it does not need to transmit auxiliary information. This principle has been exploited in particular in [10], but not described in details. The echo reduction method described in this paper is part of the ITU-T G.729.1 coder [11].

This paper is organized as follows. Section 2 gives an overview of the G.729.1 codec. In particular, G.729.1 is a split-band embedded coder, in which a lower band is coded by CELP+transform coding, whereas a higher band is coded by time-domain bandwidth extension (TDBWE)+transform coding. Section 3 explains how pre-echo can be reduced in the lower band for the case of CELP+transform coding. Section 4 gives a similar treatment of pre-echo reduction in the higher band, for the case of TDBWE+transform coding. The performance and complexity of the overall echo reduction part of G.729.1 are analyzed in Section 5, before concluding in Section 6.

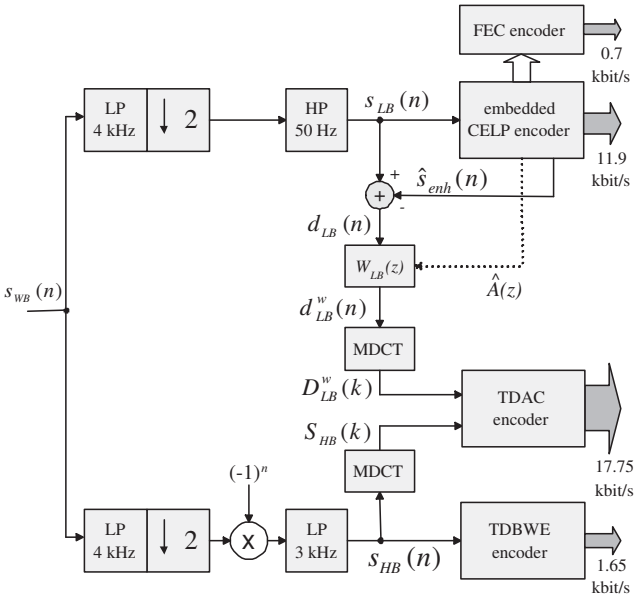
2. OVERVIEW OF ITU-T G.729.1

In May 2006, ITU-T standardized a scalable codec for wideband telephony and voice over IP applications named G.729.1 [11, 12]. It is an 8-32 kbit/s scalable codec producing 12 embedded layers:

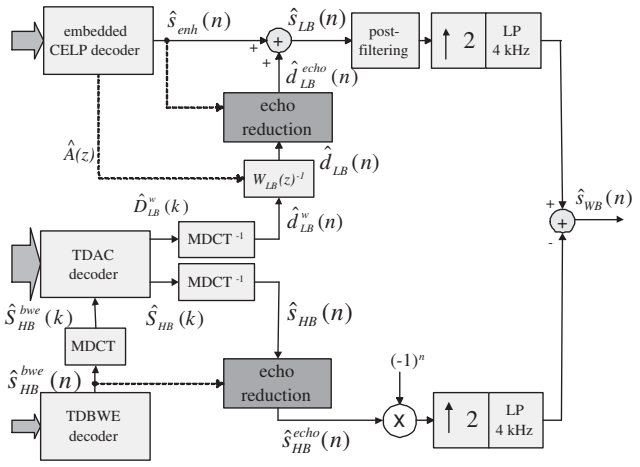
- Two embedded CELP layers, one at 8 kbit/s, called the core layer, compatible with G.729 bitstream [13] and an enhancement layer with a bitrate of 4 kbit/s, that encodes the Lower Band (LB) part of the signal (0.05-4 kHz).
- One bandwidth extension layer coarsely describes the Higher Band (HB) (4-7 kHz) with an additional bitrate of 2 kbit/s.
- Nine 2 kbit/s embedded transform coding based layers refine the whole frequency band. The maximum bitrate of the transform coding layer is 18 kbit/s. At this low bit-rate transform coding artefacts such as pre-echo are clearly present.

The total bitrate is then $8+4+2+9\times 2=32$ kbit/s.

∗ M. Gartner and H. Taddei were with Siemens when this work was done.



(a) encoder



(b) decoder (good frame only)

Figure 1: Block diagrams of the G.729.1 encoder and decoder.

2.1 G.729.1 encoder

The G.729.1 encoder is illustrated in Figure 1 (a). It operates on 20 ms frames. The 16 kHz sampled input signal $s_{WB}(n)$ is decomposed into two frequency bands, a lower band (LB) and a higher band (HB). The LB is pre-processed by a 50 Hz high-pass filter and the resulting signal $s_{LB}(n)$ is encoded by a two stage CELP coder forming the first 2 layers at 8 and 12 kbit/s. The HB is spectrally folded and pre-processed by a 3 kHz low-pass filter. The resulting signal $s_{HB}(n)$ is coded by the TDBWE model using 2 kbit/s. The difference signal $d_{LB}(n)$ between the locally decoded CELP signal at 12 kbit/s and the original signal is computed. The weighted difference signal $d^{w}_{LB}(n)$ is then obtained by applying in time-domain a zero-pole filter $W_{LB}(z)$ derived from decoded linear-predictive parameters $\hat{A}(z)$. Finally the MDCT transform of $d^{w}_{LB}(n)$ and $s_{HB}(n)$ are jointly encoded by a transform coder called Time Domain Aliasing Cancellation (TDAC) that has nine embedded layers with 2 kbit/s granularity. The TDAC coder works in 18 subbands (10 in the LB and 8 in the HB). The first 17 subbands comprise 16 coefficients (400 Hz), the last subband 8 coefficients (200 Hz).

By default, the G.729.1 encoder operates at the maximum bit-rate of 32 kbit/s. However, this bitrate can be lowered at any moment before the decoding by shortening the bitsream (cutting off the latest layers), as long as the core bitstream is preserved.

2.2 G.729.1 decoder

The decoding process depends on the number of received layers. Only the LB is decoded at 8 and 12 kbit/s. At 14 kbit/s, the TDBWE decoder extends the output bandwidth up to 7 kHz. From 14 to 32 kbit/s, any additional TDAC layer refines this signal in a different manner in the 2 frequency bands:

- The decoded MDCT coefficients $\hat{D}^{w}_{LB}(k)$ of the first 10 sub-bands represent the perceptually weighted difference signal. The MDCT coefficients in the non received subbands are set to 0. After inverse MDCT transform, inverse weighting filtering and echo reduction the resulting signal $\hat{d}_{LB}(n)$ is added to the CELP output $\hat{s}_{enh}(n)$ to form the LB part of the output signal $\hat{s}_{LB}(n)$.
- The decoded MDCT coefficients $\hat{S}_{HB}(k)$ of the last 8 subbands represent the quantized HB part of the original signal. When they are received, they replace the MDCT coefficients $\hat{S}^{tdbwe}_{HB}(k)$ obtained from the TDBWE decoder. This combined TDBWE/MDCT spectrum $\hat{S}_{HB}(k)$ is then transformed to the time domain by inverse MDCT transform. After echo reduction the HB band part of the output $\hat{s}^{echo}_{HB}(n)$ is obtained.

The block diagram of the decoder including the echo reduction modules is illustrated in Figure 1 (b).

3. ECHO REDUCTION IN LOWER BAND: CASE OF CELP+TRANSFORM DECODING

In G.729.1 the lower band (LB) is encoded using a combination of CELP and TDAC techniques. Echo introduced by the TDAC transform coder can therefore be reduced at the decoder side by exploiting information from the time-domain decoded CELP layers. This idea is illustrated in Figure 2.

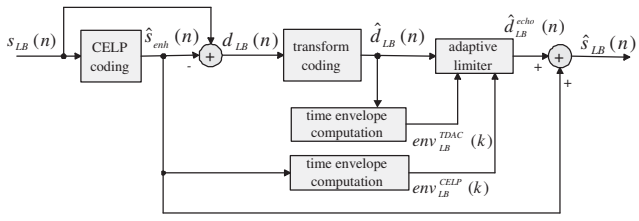


Figure 2: Echo reduction in CELP+transform coding.

3.1 Preliminary echo reduction algorithm: adaptive limiter based on energy envelopes

First the time characteristics of the CELP output and of the first 10 subbands of the TDAC decoder are analysed. We noticed that better results are obtained when using signals

filtered by $H_{hp}(z) = 1 - z^{-1}$. The resulting signals are divided into 4 segments of 40 samples each (5 ms). For each segment, the time envelope is computed. It corresponds to the square root of the energy of the signal. Then, the time envelope of the k^{th} segment of the TDAC layer $env_{LB}^{TDAC}(k)$ is compared to the corresponding time envelope of the CELP layers $env_{LB}^{CELP}(k)$. When the energy of the TDAC difference signal is higher than the energy of the CELP layers output ($env_{LB}^{TDAC}(k) > env_{LB}^{CELP}(k)$), we consider that echo is present. This is typically the case when silence is preceding an onset. For the segment before the onset, the envelope of the TDAC layer $env_{LB}^{TDAC}(k)$ would then be much higher than the envelope of the CELP layers $env_{LB}^{CELP}(k)$, as the CELP layer does not produce any echo and will have no energy. For these cases, a scaling factor $g_L(n)$ is computed for each sample of the given segment as the inverse ratio of the time envelopes

$$g_L(n) = \begin{cases} \frac{env_{LB}^{CELP}(k)}{env_{LB}^{TDAC}(k)} & if\ env_{LB}^{TDAC}(k) \geq Env_{LB}^{CELP}(k) \\ 1 & otherwise \end{cases}$$
$$\text{for} \quad n \in [40k, 40(k+1)-1] \tag{1}$$

The case where the envelope of the CELP layers is higher than the one of the TDAC layer corresponds to the normal case (energy of the difference signal lower than the energy of the signal encoded by the CELP layers). Nothing needs to be done, $g_L(n)$ is set to 1.

To avoid discontinuities, the reduction gain $g_L(n)$ is smoothed:

$$g'_L(n) = 0.85 g'_L(n-1) + 0.15 g_L(n) \tag{2}$$

$g'_L(n)$ is then applied to the TDAC output signal and the obtained signal is added to the CELP layers output to produce the output of the LB layers:

$$\hat{s}_{LB}(n) = \hat{s}_{enh}(n) + g'_L(n)\hat{d}_{LB}(n) \tag{3}$$

3.2 Improved echo reduction algorithm using non echo zone detection

Applied on its own, the echo reduction technique described above did not give satisfactory results because the time envelopes of the time domain layers are not always perfect. Fig. 3 gives an example where the CELP decoder does not manage to follow a sharp onset. The pre-echo of the first two segments is correctly reduced. However, as the time envelope of the TDAC layer is limited to the level of the CELP time envelope, the onset of the third segment in the final output is destroyed. Note that in Fig. 3 the gain smoothing applied on $g_L(n)$ is clearly noticeable.

These problems show that the application of the gain computed as in Eq. (1) has to distinguish between non echo and echo zones. In non echo zones, the previously described algorithm is not applied. The echo reduction is only necessary in zones where the echo can be present, which are the following zones:

- Zones where there is a significant energy increase in some part of the synthesis window. In that case, the reduction has to be limited to the low energy part of the synthesis window;
- Zones where the energy of the previous decoded frame was significantly higher than that of the present frame (post-echo situation).

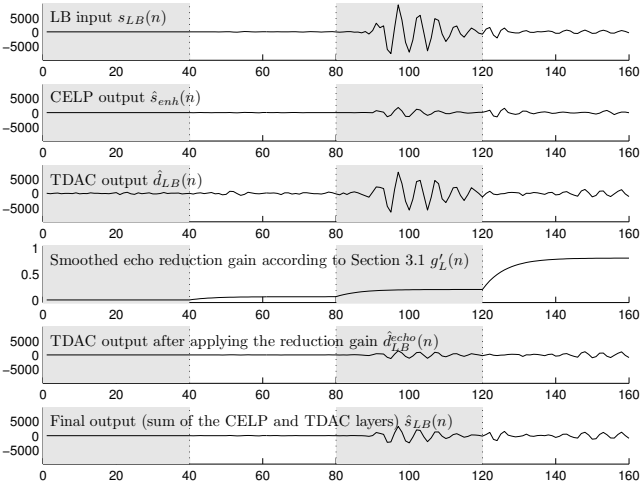


Figure 3: An illustration of the limits of the initial algorithm.

Note that the below described non echo zone detection is based only on the overlap-add operation of the inverse MDCT transform and is independent from the time domain layers and so from G.729.1 scalable structure.

The reconstructed output signal $s_{rec}(n)$ of the MDCT transform coding is obtained by the following overlap-add operation:

$$s_{rec}(n) = h(n+L).s_{prev}(n+L) + h(n).s_{cur}(n)$$
$$\text{for} \quad n \in [0, L-1] \tag{4}$$

where L is the frame length (20 ms, 160 samples in G.729.1), $h(n)$ is the MDCT window of length $2L$, $s_{prev}(n)$ is a component of length $2L$ obtained from the previous inverse MDCT transform, $s_{cur}(n)$ is a component of length $2L$ obtained from the current inverse MDCT transform. The second half of the component $s_{cur}(n)$ is used in the overlap-add operation of the next frame. Note that this signal is symmetric.

Signals $s_{rec}(n)$ and $h(L)s_{cur}(L+n)$ are concatenated to form an auxiliary signal $x_{conc}(n)$ of length $2L$ corresponding to the length of the current MDCT synthesis window. Then $x_{conc}(n)$ is divided into 8 segments of $N = 2L/8$ samples whose energy is computed:

$$E_n(k) = \sum_{l=kN}^{(k+1)N-1} x_{conc}^2(i); \quad k \in [0,5] \tag{5}$$

Note that only 6 energies are different due to the symmetry in $s_{cur}(n)$.

The following maxima and minimum are also computed:

$$\begin{aligned} max_{en} &= max(E_n(k)); \quad k \in [0,5]; \\ max_{rec} &= max(E_n(k)); \quad k \in [0,3] \\ min_{rec} &= min(E_n(k)); \quad k \in [0,3] \end{aligned} \tag{6}$$

The value of max_{rec} is memorized from frame to frame, its value for the previous frames is called max_{prev}. In the following, the indices ind_{LB1} and ind_{LB2} delimit the non echo zone of the current frame in the lower band; $ind_{LB1} > ind_{LB2}$ indicates that the whole frame is considered as a echo zone. The determination procedure of a non echo is described below:

- if $max_{prev} > 32.max_{rec}$: then we are in a post-echo situation. The previous frame has a much higher energy than then current one, the whole current frame is considered as an echo zone ($ind_{LB1} = L$, $ind_{LB2} = L-1$).
- else if $max_{en} < 16.min_{rec}$: then there is no significant energy increase in the synthesis window; the whole frame is considered as non echo zone ($ind_{LB1} = 0$, $ind_{LB2} = L-1$)
- otherwise, ind_{LB1} is set at the beginning of the subframe with maximal energy; $ind_{LB2} = min(ind_{LB1} + 2N, L-1)$ (Notice that $ind_{LB1} > L-1$ indicates that the maximum is in the next frame and thus the whole current frame is considered as an echo zone.)

The constants of the above equations (32 and 16) were found in an experimental way.

3.3 Echo reduction including false alarm detection

The echo reduction in G.729.1 decoder takes into account both the initial scaling factors computed (described in Section 3.1) and the detected non echo zones. The initial scaling factors are set to 1 in the non echo zone:

$$g_L(n) = 1;\ n \in [ind_{LB1}, ind_{LB2}] \tag{7}$$

Fig. 4 gives an example of the non echo zone detection. The top figure represents the signal x_{conc}, when the current frame contains pre-echo and the onset is in the next frame. The maximum energy is found in the 6^{th} segment, the whole current frame is considered as echo zone.

The bottom figure represents the signal x_{conc} one frame later. Now the same onset is in the current frame. As the maximum energy is in the 3^{rd} segment, the 3^{rd} and 4^{th} segments are detected as non echo zone.

Note that the symmetry in the second half of x_{conc} can be clearly observed on thess figures.

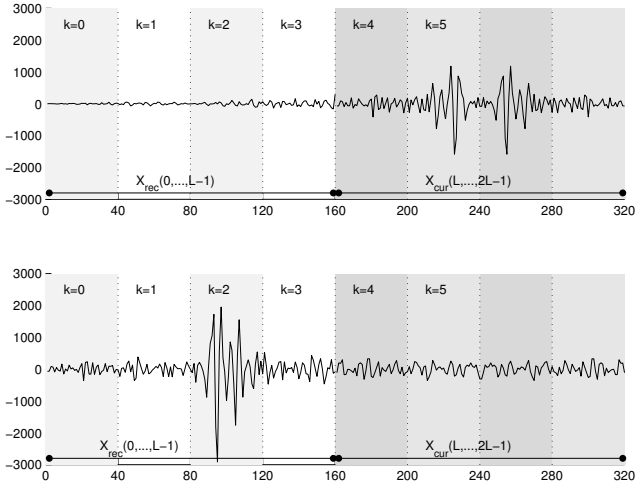


Figure 4: Examples of the concatenated x_{conc} signal.

4. ECHO REDUCTION IN HIGHER BAND: CASE OF BANDWIDTH EXTENSION+TRANSFORM DECODING

The scale factor computation in the HB part is done in a similar way as for the NB part, but without the high-pass filtering.

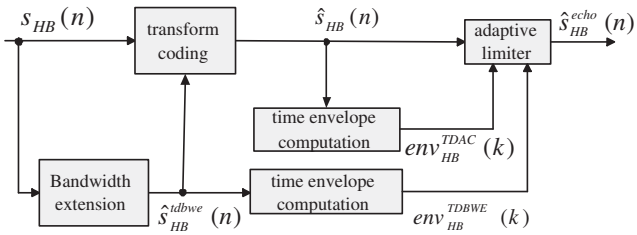


Figure 5: Echo reduction in bandwidth extension+transform coding.

The time envelope of the k^{th} segment issued from the output of the inverse MDCT transform of the combined spectrum $Env_{HB}^{TDAC}(k)$ is compared to the corresponding time envelope issued from the TDBWE layer $Env_{HB}^{TDBWE}(k)$ multiplied by a factor $\frac{10}{9}$. The factor $\frac{10}{9}$ was found in an experimental way. If $Env_{HB}^{TDAC}(k)$ is found bigger than this value, then a scaling factor $g_H(n)$ is computed for each sample of the given segment as the inverse ratio of the time envelopes, otherwise $g_H(n)$ is set to 1:

$$g_H(n) = \begin{cases} \frac{Env_{HB}^{TDBWE}(k)}{Env_{HB}^{TDAC}(k)} & if\ Env_{HB}^{TDAC}(k) \geq \frac{10}{9} Env_{HB}^{TDBWE}(k) \\ 1 & otherwise \end{cases}$$
$$\text{for} \quad n \in [40k, 40(k+1)-1] \tag{8}$$

Finally $g_H(n)$ is smoothed in a same way as $g_L(n)$ in the lower band:

$$g'_H(n) = 0.85 g'_H(n-1) + 0.15 g_H(n) \tag{9}$$

The HB output is the TDAC higher layer output signal weighted by $g'_H(n)$:

$$\hat{s}_{HB}^{echo}(n) = g'_H(n) \hat{s}_{HB}(n) \tag{10}$$

4.1 Limits of the initial algorithm in the HB part

Fig. 6 shows the TDBWE decoder output when the input signal is a pure sinusoid. The bandwidth extension module cannot correctly reproduce this signal. It is evident that the time envelope of such output should not be used to limit the time envelope of the TDAC layer. Such problems were correctly suppressed when using the echo zone detection as described in Section 3.2. In the example illustrated on Fig. 6 there is no significant energy increase in the synthesis window and so the whole frame is considered as non echo zone.

5. PERFORMANCE AND COMPLEXITY

5.1 Subjective quality

The proposed echo reduction technique has been intensively tested during the development of the G.729.1 standard. An informal AB test has been performed to compare G.729.1 with or without pre-echo reduction. Very clear preference was observed for the version with echo reduction, especially in case of clean speech or music samples with percussive sound attacks.

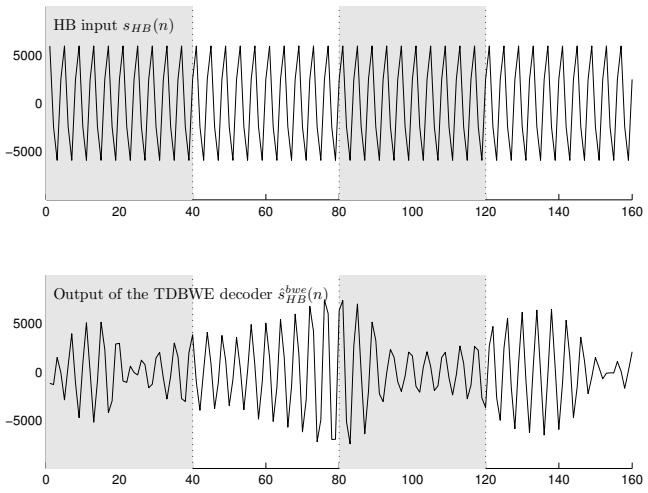


Figure 6: Sinusoidal input for the TDBWE codec.

5.2 Objective quality

Table 1 presents objective test results in terms of WB-PESQ scores [14]. These results were obtained with 24 clean speech samples in French language: 6 talkers (3 females and 3 males) and 4 double sentences per talker. Two bitrates (24 kbit/s and 32 kbit/s) and 3 levels (-26 dBov, -16 dBov and -36 dBov) were considered. Note that WB-PESQ scores have been compared with subjective scores during G.729.1 standardization, and the use of WB-PESQ was validated in this framework.

level dBov	24 kbit/s without ER	24 kbit/s with ER	32 kbit/s without ER	32 kbit/s with ER
-26	4.05	4.08	4.10	4.19
-16	4.00	4.03	4.11	4.20
-36	3.68	3.69	3.72	3.76

Table 1: WB-PESQ MOS-LQ0 results with or without Echo Reduction (ER)

On another database of 10 speakers (4 females, 4 males, and 2 children) at -22 dBov, with 10 double sentences per talker the mean improvement was more significant: 0.07 MOS-LQ0 at 24 kbit/s and 0.14 MOS-LQ0 at 32 kbit/s. Note that more improvement can be observed for higher bitrates.

5.3 Complexity

The computational complexity was also evaluated. The integration of echo reduction in the decoder (including the non-echo zone detection) increases both the mean and the worst-case observed complexity by about 0.3 WMOPS (Weighted Million Operations Per Second). This complexity is negligible when compared with the worst-case complexity of the G.729.1 decoder at 32 kbit/s, which is around 14.13 WMOPS.

6. CONCLUSION

In this paper, we described how pre-echo artefacts were addressed in the design of the G.729.1 coder. Pre-echo reduction in G.729.1 is used at the decoder above 14 kbit/s. It is implemented as a split-band adaptive limiter, in which gains are applied to the transform decoder synthesis. The applied gains are first computed in function of the time envelopes of the lower time-domain decoded layers and that of the transform layers, but this pre-echo reduction is inhibited in zones identified by a non pre-echo zone detection algorithm. Contrary to other echo reduction methods, the proposed method does not need to transmit auxiliary information.

REFERENCES

[1] B. Moore, "Characterisation of simultaneous, forward and backward masking," in *12th Int. AES Conf. on the Perception of Reproduced Sound*, 1993, pp. 22–33.

[2] B. Edler, "Method for transmitting a signal," PCT patent WO90/09063, Jan. 1990.

[3] A. Sugiyama, F. Hazu, M. Iwadare, and T. Nishitani, "Adaptive transform coding with an adaptive block size (ATC-ABS)," in *Proc. ICASSP*, Apr. 1990, vol. 5, pp. 1093 – 1096.

[4] S. Shlien, "The modulated lapped transform, its timevarying forms, and its applications to audio coding standards," *IEEE Trans. on Speech and Audio Processing*, vol. 5, pp. 359–366, July 1997.

[5] H.S. Malvar, "Enhancing the performance of subband audio coders for speech signals," in *Proc. IEEE Int. Symp. on Circuits and Systems*, June 1998.

[6] J. Herre and J.D. Johnston, "Enhancing the Performance of Perceptual Audio Coders by Using Temporal Noise Shaping (TNS)," 101st AES convention, 1996, Preprint 4384.

[7] 3GPP TS 26.290, "Audio codec processing functions; Extended AMR wideband codec; Transcoding functions," .

[8] R. Vafin, R. Heusden, and W.B. Kleijn, "Modifying transients for efficient coding of audio," in *Proc. ICASSP*, May 2001, vol. 5, pp. 3285 – 3288.

[9] S. Ragot, B. Kövesi, D. Virette, R. Trilling, and D. Massaloux, "A 8-32 kbit/s scalable wideband speech and audio coding candidate for ITU-T G.729EV standardization," in *Proc. ICASSP*, May 2006.

[10] B. Geiser, P. Jax, P. Vary, H. Taddei, M. Gartner, and S. Schandl, "A Qualified ITU-T G.729EV Codec Candidate for Hierarchical Speech and Audio Coding," in *Proc. IEEE 8th Workshop on Multimedia Signal Processing*, Oct. 2006, pp. 114 – 118.

[11] ITU-T Rec. G.729.1, "An 8-32 kbit/s scalable wideband coder bitstream interoperable with G.729," May 2006.

[12] S. Ragot and al., "ITU-T G.729.1: An 8-32 kbit/s Scalable Coder Interoperable with G.729 for Wideband Telephony and Voice over IP," in *Proc. ICASSP, Honolulu, HI, USA*, Apr. 2007, pp. 114 – 118.

[13] ITU-T Rec. G.729, "Coding of speech at 8 kbit/s using conjugate structure algebraic-code-excited linear-prediction (CS-ACELP)," November 1996.

[14] ITU-T Rec. P.862.2, "Wideband extension to recommendation p.862 for the assessment of wideband telephone networks and speech codecs," November 2005.