JOINT PRE-ECHO CONTROL AND FRAME ERASURE CONCEALMENT FOR VOIP AUDIO CODECS

Bernd Geiser and Peter Vary

Institute of Communication Systems and Data Processing (ind)
RWTH Aachen University, Germany
{geiser|vary}@ind.rwth-aachen.de

ABSTRACT

Pre-echo artifacts are a common problem of transform audio codecs owing to their comparatively large block lengths. If such codecs are used in Voice-over-IP (VoIP) applications, the concealment of lost audio packets poses another major issue. These two problems are usually solved by two separate algorithmic components: A mechanism for pre-echo control (PEC) and a frame erasure concealment (FEC) algorithm. In this contribution we propose a combined PEC/FEC approach which elegantly solves both problems based on a single side information stream. In particular, our proposal is suitable for (subband) transform audio codecs that are based on a frequency transform with half-overlapped windowing (such as the popular MDCT).

The described method is part of the Huawei/ETRI candidate for the future ITU-T super-wideband extensions of Rec. G.729.1 and G.718 where it has been successfully applied in the context of high frequency (8–14 kHz) encoding.

1. INTRODUCTION

The current technology shift from circuit- towards packet-switched voice communication has led to a renewed interest in high-quality speech and audio codecs for real-time applications. Typical use cases for such codecs are high-quality conferencing scenarios and Voice-over-IP (VoIP) telephony but they also include other audio services such as streaming, e-learning or remote monitoring.

High audio quality is known to be achievable with so-called perceptual transform codecs [1] where, typically, a comparatively high algorithmic delay (up to several 100 ms) is tolerated in favor of a high coding efficiency. For conversational applications like VoIP, a significantly lower algorithmic delay (e.g. a maximum of 50 ms) is desirable. As an example, the wideband (50–7000 Hz) VoIP codec ITU-T G.729.1 [2, 3], which is a hybrid time/transform-domain coder, offers a total delay of 48.9375 ms whereby the transform length is 40 ms. Still, despite the reduced delay, the well-known problem of *pre-echo artifacts* [1] persists, in particular at low bit rates.

Typically, pre-echo problems arise in lossy block-based audio codecs, in particular transform codecs, when the introduced quantization noise is "smeared" over a complete audio block instead of being temporally shaped such that it remains below the temporal masking threshold of the human auditory system. Consequently, pre-echo artifacts are predominant for heavily transient signals that are encoded with a low bit rate. Then a dedicated mechanisms for pre-echo control (PEC) must be used, e.g. [4]. The problem is even more relevant for *parametric coding schemes* where the transmission of explicit (and fine-grained) temporal information is advisable, e.g. [5].

If a "conversational audio codec" is used within a *packet switched* network environment (VoIP), also the problem of *lost data packets* has to be considered and adequate means to *conceal* possibly missing speech or audio frames have to be provided. Usually, this is accomplished by an algorithm for *frame erasure concealment* (FEC). Clearly, both PEC and FEC can benefit from the transmission of additional side information which can be determined at the encoder side. Moreover, for the case of PEC side information, it is appropriate to distinguish between transient and stationary signal segments (or frames) because transient sounds are particularly susceptible to pre-echo artifacts.

In this contribution we propose to integrate the thus far separate PEC and FEC modules, resulting in a joint PEC/FEC approach based on a single side information stream which describes the temporal energy envelope of subband signals. For transient sounds, the transmitted side information is usable for PEC and, in case the transient audio frame is lost, also for FEC. During stationary sounds, the problem of pre-echo artifacts is not as relevant. Here, a few FEC bits have to be added as redundant information.

The proposed PEC/FEC method has been successfully implemented in the Huawei/ETRI candidate codec [6] for the super-wideband extensions of ITU-T Rec. G.729.1 [2, 3] and G.718 [7, 8], where it has been applied in the context of high frequency (8–14 kHz) encoding. Both a clear reduction of pre-echos (especially for low bit rates and for the parametric coding modes) and good performance under frame erasure conditions could be observed.

The present paper is structured as follows: First, we review state-of-the-art techniques for frame erasure concealment (Sec. 2) and pre-echo control (Sec. 3). Then, we explain our proposal for joint PEC/FEC in detail (Sec. 4) and present an evaluation of the scheme in the framework of the Huawei/ETRI G.729.1-SWB candidate codec (Sec. 5).

2. FRAME ERASURE CONCEALMENT (FEC)

FEC is an essential component of audio codecs that are being used in packet switched network environments. Therefore, FEC modules have recently been added to several standardized codecs such as [9]. New codecs, e.g. [2, 7], are directly designed under such constraints.

To support the FEC algorithm, the encoder may add a certain amount of additional information to the bitstream of the

This work has been conducted in cooperation with Huawei Technologies Co., Ltd., Core Network Research Dept., Beijing, P.R. of China

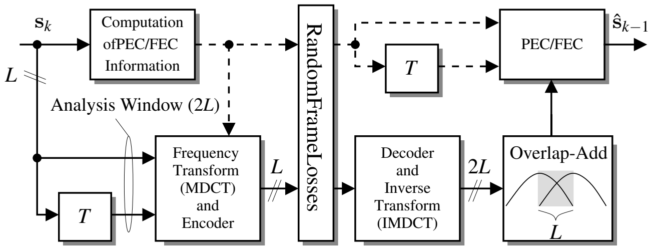

Figure 1: Joint Pre-Echo Control (PEC) and Frame Erasure Concealment (FEC) in a transform codec with 50% overlapping analysis windows of length $2L$. The PEC/FEC information is extracted from the time domain input vectors $\mathbf{s}_k$. T is a one-frame delay, no additional delay is needed for FEC.

codec which leads, when compared with blind (extrapolation based) FEC approaches, to a substantially improved quality of the FEC algorithm [2, 10]. For example, the transmitted FEC side information might comprise a coarse but relevant description of *past* signal frames, e.g. their energy envelope and/or rough phase information (for strongly periodic signals). Therefore, if a frame has been lost during transmission, an approximate version can still be reproduced based on the decoder's memory and on the available FEC information.

The distribution of information over successive frames can be interpreted as a parallel transmission over more than one transmission path (diversity) and it is clear that *multiple description coding* schemes [11] can be employed to achieve robustness against packet losses. Yet, temporally distributed FEC information requires at least a *one-frame delay* at the decoder side in order to handle (single) frame losses. However, as illustrated in Fig. 1, audio codecs based on frequency transforms using half-overlapped windowing (lapped transforms, e.g. the popular MDCT) inherently incur a one-frame delay within the decoder because of the overlap-add operation. This property can be elegantly exploited to transmit FEC information for single frame losses *without any additional delay-penalty* provided that the computation of this information does not use additional look-ahead samples. An example realization which extracts the FEC side information directly from the time domain signal is used in [2].

3. PRE-ECHO CONTROL (PEC)

In the audio coding literature and in standardized codecs several methods for PEC can be identified which will be briefly reviewed in the following, cf. [1]. We will also discuss the applicability of the PEC methods to the FEC problem.

3.1 Window Switching

A popular solution for PEC is the so called "window switching" mechanism where the temporal spread of the quantization noise is confined by using shorter analysis windows (e.g. 4 ms) for transient signal segments. Naturally, the frequency selectivity is decreased in that case and some signaling overhead is required. To ensure the perfect reconstruction property of the frequency transform during stationary/transient transitions, special transitional "start/stop" analysis windows must be used resulting in a time-varying transform and, hence, in increased design complexity. Moreover, these start/stop windows make it difficult to achieve an instantaneous reaction to a sharp transient.

Since the window switching method modifies the frequency transform itself, the overlap-add delay can not be exploited, hence, there is no direct way to integrate this scheme with FEC unless a significant amount of additional algorithmic delay is spent.

3.2 Spectral Prediction

Another possibility to achieve the desired temporal shaping of the quantization noise is to apply linear prediction over frequency (transform coefficients). The dual operation of frequency domain linear predictive filtering is multiplication in the time domain. Therefore, spectral prediction is an efficient tool to cope with pre-echo artifacts. With this method, even a *frequency selective* temporal shaping can be easily achieved by applying different predictors towards different subsets (i.e. subbands) of the transform coefficients. Naturally, the prediction coefficients need to be quantized and transmitted.

However, similar to window switching, the transmitted temporal information can not be reused for FEC purposes without accepting additional algorithmic delay since the prediction coefficients are extracted from the frequency domain.

3.3 Time Domain Gain Manipulation

A very intuitive PEC approach is to apply gain manipulation directly in the time domain (as opposed to spectral prediction). This approach is for example realized in the PEC module of G.729.1 [4]. Naturally, the elegant ability of spectral prediction to apply frequency selective temporal shaping is lost, i.e., a dedicated filterbank is required to obtain such frequency selectivity. However, the related side information can be directly extracted from the time domain signal. Therefore, this information can potentially be useful for FEC purposes without spending additional algorithmic delay. In the following section, we describe our proposal which employs a time domain gain manipulation scheme for *both* PEC and FEC.

4. JOINT PEC AND FEC

The principle structure of our proposal for joint PEC and FEC is illustrated in Fig. 1. The respective encoder- and decoder-side operations are detailed in the following.

4.1 Encoder

At the encoder side, the PEC/FEC side information is computed based on the k-th input frame of the input signal $s(n)$ (frame length: L). Concretely, we use the overall gain

$$g(k) = \frac{1}{2} \log_2 \frac{\sum\limits_{n=0}^{L-1} s^2(kL+n)}{L}, \quad k \in \mathbb{N}_0, \tag{1}$$

and N_{SF} subframe gains (subframe length $L_{\mathrm{SF}} = L/N_{\mathrm{SF}}$)

$$g_{\mathrm{SF}}(k, k_{\mathrm{SF}}) = \frac{1}{2} \log_2 \frac{\sum\limits_{n=0}^{L_{\mathrm{SF}}-1} s^2(kL + k_{\mathrm{SF}} L_{\mathrm{SF}} + n)}{L_{\mathrm{SF}}}, \tag{2}$$

where $k_{\mathrm{SF}} \in \{0, \ldots, N_{\mathrm{SF}} - 1\}$. Note that the input signal $s(n)$ may also be a subband signal, e.g., obtained from a preceding QMF bank stage.

Table 1: Exemplary PEC/FEC bitstream for a stationary/stationary/transient/stationary frame sequence.

Frame index k	1	2	3
$t(k)$	0	1	0
$t(k-1)$	0	0	1
$\hat{g}(k)$	$\hat{g}(1)$	–	$\hat{g}(3)$
$\hat{g}(k-1)$	$\hat{g}(0)$	$\hat{g}(1)$	–
$\hat{g}_{\mathrm{SF}}(k, k_{\mathrm{SF,odd}})$	–	$\hat{g}_{\mathrm{SF}}(2,1)$ $\hat{g}_{\mathrm{SF}}(2,3)$ $\ldots$	–
$\hat{g}_{\mathrm{SF}}(k-1, k_{\mathrm{SF,even}})$	–	–	$\hat{g}_{\mathrm{SF}}(2,0)$ $\hat{g}_{\mathrm{SF}}(2,2)$ $\ldots$

First, the temporal structure of the input signal frame is analyzed and it is classified as either "transient" ($t(k) = 1$) or "stationary" ($t(k) = 0$). We propose a simple yet effective transient detector which determines if the maximum rising and/or falling slopes within the subframe gains $g_{\mathrm{SF}}(k, k_{\mathrm{SF}})$ exceed certain pre-specified thresholds. Then, for *stationary* frames, we simply quantize and transmit the gain $g(k)$ whereas for *transient* frames, the "temporal envelope" consisting of the N_{SF} logarithmic subframe gains $g_{\mathrm{SF}}(k, k_{\mathrm{SF}})$ is encoded. It is important to note that the subframe gains for even and odd indices k_{SF} are encoded *separately*. To enable a reuse of this information for FEC purposes we propose a special (variable) bitstream arrangement for a given frame k:

- The mode bit $t(k)$
- The repeated bit $t(k-1)$
- If $t(k) = 0$: the quantized gain $\hat{g}(k)$
- If $t(k-1) = 0$: the repeated quantized gain $\hat{g}(k-1)$
- If $t(k) = 1$: the encoded subframe gains with *odd* indices corresponding to the *current* frame, i.e.: $\hat{g}_{\mathrm{SF}}(k, k_{\mathrm{SF}})$ with $k_{\mathrm{SF}} \in \{1, 3, \ldots, N_{\mathrm{SF}} - 1\}$
- If $t(k-1) = 1$: the encoded subframe gains with *even* indices corresponding to the *previous* frame, i.e.: $\hat{g}_{\mathrm{SF}}(k-1, k_{\mathrm{SF}})$ with $k_{\mathrm{SF}} \in \{0, 2, \ldots, N_{\mathrm{SF}} - 2\}$

To further illustrate this concept, Tab. 1 shows an exemplary PEC/FEC bitstream for a stationary/transient/stationary frame sequence. The information is either transmitted redundantly (for stationary frames, e.g. $k = 1$) or distributed across neighboring frames (for transients, $k = 2$). Thereby, the amount of redundancy is kept to a minimum and, in particular for transient frames, there is no redundant information (except for the repeated mode bit). Instead, the available bits can be used both for PEC and FEC by the decoder as described in Sec. 4.2.

With the quantized temporal (subframe) gains, the encoder can construct an interpolated "temporal gain function" (TGF) which is used to normalize the (subband) input signal $s(n)$ before the frequency transform:

$$s^{\mathrm{T}}(kL+n) = s(kL+n) \cdot \mathrm{TGF}^{-1}(kL+n) \tag{3}$$

Such an operation reduces the signal dynamics, but it has to be taken care that the TGF exhibits a pronounced low-pass characteristic such that spectral leakage is avoided as much as possible. For instance, the TGF can be constructed by an overlap-add of scaled Hann windows.

4.2 Decoder

After inverse transform and overlap-add at the decoder (cf. Fig. 1), pre-echos are suppressed or, for lost frames, concealment is carried out by restoring the temporal signal characteristics.

Let $\tilde{s}^{\mathrm{T}}(kL+n)$ with $n \in \{0, \ldots, L-1\}$ be the k-th decoded signal frame after overlap-add, but before PEC/FEC. Then, for each received frame, the temporal envelope is restored by multiplication of the decoded signal with the TGF[1]:

$$\hat{s}(kL+n) = \tilde{s}^{\mathrm{T}}(kL+n) \cdot \mathrm{TGF}(kL+n)\,. \tag{4}$$

This temporal denormalization procedure effectively suppresses pre-echos that are particularly strong at low bit rates or in the case of a purely parametric signal. Nevertheless, due to the stationary/transient distinction, spectral details in stationary signal segments can be preserved.

If a *frame erasure* is signaled to the decoder, concealment is performed. In that case, the decoded signal frame $\tilde{s}^{\mathrm{T}}(kL+n)$ is unavailable and has to be estimated. The simplest approach is frame repetition: $\tilde{s}^{\mathrm{T}}(kL+n) = \tilde{s}^{\mathrm{T}}((k-1)L+n)$. Though, a better alternative is to repeat the transform coefficients instead. The inverse transform and the overlap-add operation will then smooth the transition between the (possibly correct) previous frame and the missing frame. Now, with the estimated signal frame $\tilde{s}^{\mathrm{T}}(kL+n)$ and the available PEC/FEC information, the concealed output can be produced.

For each lost frame, the transient/stationary flag $t(k)$ is still available from the following frame ($k+1$). For missing *stationary* frames, the FEC redundancy additionally contains the overall gain factor $\hat{g}(k)$. If, in the example from Tab. 1, the decoder wanted to reconstruct signal frame $k = 1$ but the corresponding frame in the bitstream was lost, the flag $t(1) = 0$ and the respective gain $\hat{g}(1)$ would still be available from the second bitstream frame ($k = 2$).

When a *transient* frame is lost, the information about the subframe gains of subframes with *odd* indices $\hat{g}_{\mathrm{SF}}(k, k_{\mathrm{SF}})$ with $k_{\mathrm{SF}} \in \{1, 3, \ldots, N_{\mathrm{SF}} - 1\}$ is missing. Therefore, they have to be interpolated from the subframe gains of subframes with *even* indices:

$$\tilde{g}_{\mathrm{SF}}(k, k_{\mathrm{SF}}) = \frac{\hat{g}_{\mathrm{SF}}(k, k_{\mathrm{SF}} - 1) + \hat{g}_{\mathrm{SF}}(k, k_{\mathrm{SF}} + 1)}{2} \tag{5}$$

with $k_{\mathrm{SF}} \in \{1, 3, \ldots, N_{\mathrm{SF}} - 1\}$, before the TGF for denormalization can be constructed. If, in the example from Tab. 1, the decoder wanted to reconstruct signal frame $k = 2$, but the corresponding frame in the bitstream was lost, the subframe gains with even indices would still be available from the third frame ($k = 3$) and the gains with odd indices could be obtained through interpolation.

Likewise, if a frame after a transient frame is lost, the even subframe gains (describing the preceding transient) are unavailable. Again, the complete TGF cannot be reconstructed and only the subframe gains with *odd* indices, i.e., $\hat{g}_{\mathrm{SF}}(k, k_{\mathrm{SF}})$ with $k_{\mathrm{SF}} \in \{1, 3, \ldots, N_{\mathrm{SF}} - 1\}$, are used to

[1] To compensate for a possibly poorly coded (or disturbed) signal $\tilde{s}^{\mathrm{T}}(n)$, denormalization can alternatively be carried out by a gain function TGF$'$ instead. TGF$'$ is based on (subframe) gain *differences* between the desired gains $\hat{g}(k)$ or $\hat{g}_{\mathrm{SF}}(k, k_{\mathrm{SF}})$ and the respective *measured* gains $g'(k)$ or $g'_{\mathrm{SF}}(k, k_{\mathrm{SF}})$ of the decoded signal $\tilde{s}^{\mathrm{T}}(kL+n)$. If the encoder side normalization according to (3) has been omitted, it is mandatory to use TGF'.

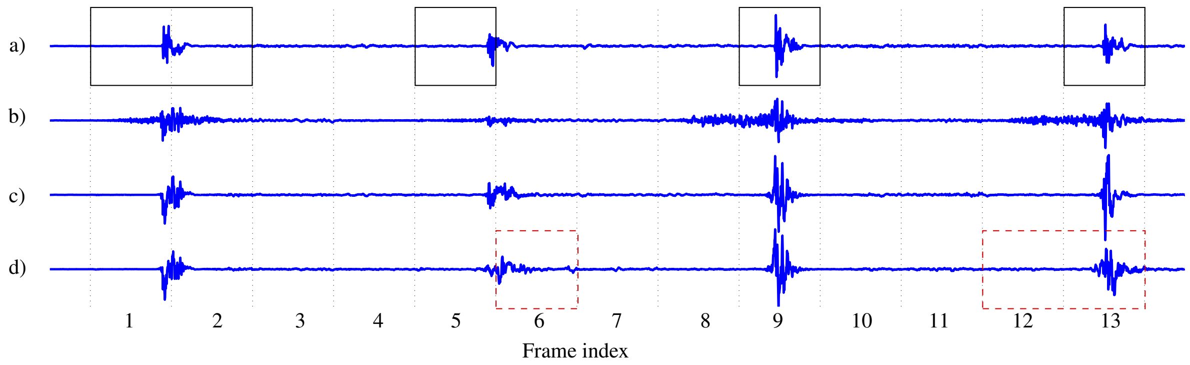

Figure 2: Example waveforms for the proposed PEC/FEC method based on purely parametric coding of the 8–14 kHz subband. a) original signal (castanets), solid boxes: frames classified as transient — b) synthesized signal without PEC — c) synthesized signal with PEC — d) synthesized signal under frame erasures (all bits from the frames with dashed boxes have been discarded)

form the TGF. Therefore, the subframe gains with *odd* indices are interpolated according to (5), whereby $k_{\mathrm{SF}} \in \{0, 2, \ldots, N_{\mathrm{SF}} - 2\}$ in this case. If, again in the example from Tab. 1, the decoder wanted to reconstruct signal frame $k = 2$ and the *third* bitstream frame ($k = 3$) was lost, the subframe gains with odd indices would already be available from bitstream frame $k = 2$ and the missing gains with even indices could be obtained through interpolation.

In case of *bursty frame erasures*, i.e., if two consecutive bitstream frames are lost, there is no information available about the current decoder frame. Therefore, we have to resort to an extrapolation based approach where the currently missing frame is assumed to be stationary and the (averaged) gain factor from the previous (reconstructed) frame is decreased by a certain amount before applying the gain denormalization.

5. EVALUATION

This section summarizes the codec framework in which our PEC/FEC proposal has been implemented and tested. Objective and subjective test results are presented.

5.1 Experimental Framework

The proposed PEC/FEC scheme has been implemented in the Huawei/ETRI candidate [6] for the super-wideband extensions of ITU-T Rec. G.729.1 [2,3] and G.718 [7,8]. The actual implementation of the candidate coder is based on ITU-T G.729.1 which has a total bit rate of 32 kbit/s and encodes wideband signals (16 kHz sampling rate).

In the encoder, the 32 kHz sampled input signal is split into two critically sampled (16 kHz) subband signals $s_{\mathrm{wb}}(n)$ and $s'_{\mathrm{swb}}(n)$ by means of an infinite impulse response quadrature mirror filter (IIR QMF) bank. The wideband signal $s_{\mathrm{wb}}(n)$ is then encoded by the G.729.1 core codec as described in [2] and the 8–16 kHz signal $s'_{\mathrm{swb}}(n)$ is preprocessed by a 6 kHz lowpass filter. The resulting 8–14 kHz signal $s_{\mathrm{swb}}(n)$ is then processed by the PEC/FEC unit as described in Sec. 4.1, whereby the frame length is 20 ms ($L = 320$) and each frame is divided into $N_{\mathrm{SF}} = 8$ subframes of length 2.5 ms ($L_{\mathrm{SF}} = 40$). The bit allocation for PEC/FEC is as follows:

- 2 mode bits $t(k)$ and $t(k-1)$,
- 5 bits for the gain $\hat{g}(k)$ if $t(k) = 0$,
- 5 bits for the gain $\hat{g}(k-1)$ if $t(k-1) = 0$,
- 17 bits for the differentially coded subframe gains $\hat{g}_{\mathrm{SF}}(k, k_{\mathrm{SF,odd}})$ if $t(k) = 1$; concretely 5 bits for $\hat{g}_{\mathrm{SF}}(k, 1)$ and 4 bits to quantize each of the subsequent differential subframe gains $g_{\mathrm{SF}}(k, k_{\mathrm{SF,odd}}) - \hat{g}_{\mathrm{SF}}(k, k_{\mathrm{SF,odd}} - 2)$ with $k_{\mathrm{SF,odd}} \in \{3, 5, 7\}$,
- 17 bits for the differentially coded subframe gains $\hat{g}_{\mathrm{SF}}(k-1, k_{\mathrm{SF,even}})$ if $t(k-1) = 1$; concretely 5 bits for $\hat{g}_{\mathrm{SF}}(k, 0)$ and 4 bits to quantize each of the subsequent differential subframe gains $g_{\mathrm{SF}}(k, k_{\mathrm{SF,even}}) - \hat{g}_{\mathrm{SF}}(k, k_{\mathrm{SF,even}} - 2)$ with $k_{\mathrm{SF,even}} \in \{2, 4, 6\}$.

The total number of PEC/FEC bits per 20 ms-frame is 12 for consecutive stationary frames, 24 for transient-stationary transitions and 36 for consecutive transient frames.

After applying temporal normalization according to (3), the resulting normalized signal is encoded in an embedded fashion: For low bit rates (e.g. 4 kbit/s on top of G.729.1), a parametric coding method is used. At higher bit rates (up to 32 kbit/s on top of G.729.1), vector quantization (VQ) of MDCT coefficients is employed, cf. [6]. In Sec. 5.3 the quality impact of the PEC/FEC solution will be assessed for parametric as well as VQ coding modes.

On the *decoder* side of the SWB candidate codec, first, the G.729.1 bitstream layer is decoded as described in [2]. The high frequency band is synthesized by the SWB decoder which includes temporal denormalization (and FEC to conceal lost frames) according to (4). Finally, the synthesized 0–8 kHz and 8–16 kHz bands are combined by IIR QMF synthesis with integrated phase equalization.

5.2 Example

Fig. 2 illustrates the performance of the proposed PEC/FEC method based on the 8–14 kHz components of the EBU SQAM castanet signal [12], whereby the codec has been run at its lowest bit rate (4 kbit/s on top of G.729.1[2]) to produce considerable pre-echos as seen in Fig. 2-b). When applying PEC, these artifacts are clearly reduced (Fig. 2-c)). A signal with *concealed* frames is shown in Fig. 2-d), where the bits from Frames 6, 12 and 13 have been discarded. Note that also Frame 5 is a “partially concealed” output since the subframe gains with even indices from Frame 6 are missing.

[2] This bit rate includes the PEC/FEC bits.

Table 2: PEAQ improvements (Δ-PEAQ) obtained through PEC/FEC. Evaluation is based on the G.729.1-SWB candidate [6].

Test item	Average number of bits per frame for PEC/FEC	0% FER 36 kbit/s (param.)	0% FER 64 kbit/s (quant.)	5% FER 36 kbit/s (param.)	5% FER 64 kbit/s (quant.)	10% FER 36 kbit/s (param.)	10% FER 64 kbit/s (quant.)	10% FER SWB only 36 kbit/s
Castanets	14.49	+0.17	+0.31	+0.14	+0.25	+0.11	+0.19	+0.20
German Male	13.58	+0.13	+0.10	+0.04	+0.07	+0.04	+0.05	+0.20
German Female	14.33	+0.27	+0.17	+0.05	+0.06	+0.06	+0.05	+0.27
Pop Music (ABBA)	13.34	+0.33	+0.18	+0.12	+0.07	+0.05	+0.03	+0.31
∅	13.94	+0.23	+0.19	+0.09	+0.11	+0.07	+0.08	+0.25

5.3 Objective Quality Assessment

Tab. 2 quantifies the quality gain achieved by PEC/FEC in terms of a PEAQ score improvement [13] compared to a codec version without temporal normalization (3) and denormalization (4). Both the parametric (32+4 kbit/s) and the VQ coding modes (32+32 kbit/s) of the codec have been used. The four test items have been taken from the EBU SQAM corpus [12]. Moreover, the average number of bits per frame used by the PEC/FEC module is tabulated. Note that, if the FEC functionality is not desired, 6 (or 1) redundant bits can be saved for each stationary (or transient) frame without compromising PEC performance.

From the Δ-PEAQ values for 0% FER it becomes obvious that a mechanism for PEC remains very important even for quantized MDCT coefficients (64 kbit/s). However, in frame erasure conditions, it can be observed that the PEC/FEC quality gain decreases with increasing frame erasure rate. This behavior can be explained by the overall dominance of the errors that are introduced by the low band (0–8 kHz) FEC module. We verify this proposition by limiting frame erasures to the SWB bits (8–14 kHz) only. We have introduced such "partial" erasures at the same (pseudo-random) positions as before. This way, the performance gain through the proposed FEC module can be measured independently from the 0–8 kHz FEC. Some results are shown in the last column of Tab. 2 which lists Δ-PEAQ values for a bit rate of 36 kbit/s and 10% SWB erasure. Obviously, the quality gain through PEC/FEC at 36 kbit/s and 0% FER (+0.23) can be maintained even at 10% SWB erasure (+0.25). Hence, the proposed PEC/FEC module is also useful for the concealment of short-term *bandwidth switchings* from SWB (36 kbit/s) down to WB (32 kbit/s).

5.4 Subjective Quality Assessment

Within the ITU-T qualification phase for G.729.1-SWB, extensive subjective listening tests have been conducted in order to compare the proposed "codec under test" (CuT) with the existing ITU-T SWB codec G.722.1 Annex C [14, 15]. Some test results are discussed in [6]. Summarizing, for clean speech, the proposed coder is clearly better than G.722.1C at comparable bit rates. For clean speech at a bit rate of 48 kbit/s and with 3% frame erasures, the CuT outperforms G.722.1C by 0.5 MOS. For music input, also coded with 48 kbit/s, the codec is not worse than G.722.1C at the same bit rate. The good performance for both speech and music can, in part, be attributed to the proposed PEC/FEC mechanism for the 8–14 kHz frequencies.

6. CONCLUSIONS

We have proposed a method to jointly implement pre-echo control and frame erasure concealment in speech/audio codecs for Voice over IP communication based on a single side information stream. Good performance could be shown within the framework of the Huawei/ETRI G.729.1-SWB candidate codec. Clearly, the question remains whether the proposed method is applicable to other frequency bands than 8–14 kHz. For the PEC case, a similar method [4] which is based on temporal energy information from other codec layers, has shown effectiveness for 0–7 kHz signals. Though, a high quality FEC for lower frequencies (e.g. 0–4 kHz) might require additional measures to be able to deal with strongly periodic signals such as voiced speech, cf. [2, 7, 10].

REFERENCES

[1] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proceedings of the IEEE*, vol. 88, no. 4, pp. 451–513, Apr. 2000.

[2] ITU-T Rec. G.729.1, "G.729 based embedded variable bit-rate coder: An 8-32 kbit/s scalable wideband coder bitstream interoperable with G.729," 2006.

[3] S. Ragot *et 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. of IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Honolulu, Hawai'i, USA, Apr. 2007.

[4] B. Kövesi, S. Ragot, M. Gartner, and H. Taddei, "Pre-echo reduction in the ITU-T G.729.1 embedded coder," in *Proc. of European Signal Processing Conf. (EUSIPCO)*, Lausanne, Switzerland, Aug. 2008.

[5] B. Geiser, P. Jax, P. Vary, H. Taddei, S. Schandl, M. Gartner, C. Guillaumé, and S. Ragot, "Bandwidth extension for hierarchical speech and audio coding in ITU-T Rec. G.729.1," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2496–2509, Nov. 2007.

[6] B. Geiser, H. Krüger, H. W. Löllmann, P. Vary, D. Zhang, H. Wan, H. T. Li, and L. B. Zhang, "Candidate proposal for ITU-T super-wideband speech and audio coding," in *Proc. of IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei, Taiwan, Apr. 2009.

[7] ITU-T Rec. G.718, "Frame error robust narrowband and wideband embedded variable bit-rate coding of speech and audio from 8-32 kbit/s," 2008.

[8] T. Vaillancourt *et al.*, "ITU-T EV-VBR: A robust 8-32 kbit/s scalable coder for error prone telecommunications channels," in *Proc. of European Signal Processing Conf. (EUSIPCO)*, Lausanne, Switzerland, Aug. 2008.

[9] ITU-T Rec. G.722, "7 kHz audio coding within 64 kbit/s," in Blue Book, vol. Fascicle III.4 (General Aspects of Digital Transmission Systems; Terminal Equipments), 1988.

[10] F. Mertz and P. Vary, "Packet Loss Concealment with Side Information for Voice over IP in Cellular Networks," in *ITG-Fachtagung Sprachkommunikation*, Kiel, Germany, Apr. 2006.

[11] V. K. Goyal, "Multiple description coding: Compression meets the network," *IEEE Signal Processing Magazine*, vol. 18, no. 5, pp. 74–93, Sept. 2001.

[12] European Broadcast Union, EBU Tech 3253, "Sound quality assessment material — recordings for subjective tests," 1988.

[13] ITU-R Rec. BS.1387, "Method for objective measurements of perceived audio quality," 1998.

[14] ITU-T Rec. G.722.1, "Low complexity coding at 24 and 32 kbit/s for hands-free operation in systems with low frame loss," 2005.

[15] M. Xie, D. Lindbergh, and P. Chu, "ITU-T G.722.1 Annex C: A new low-complexity 14 kHz audio coding standard," in *Proc. of IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Toulouse, France, May 2006.