NEW POST-PROCESSING TECHNIQUES FOR LOW BIT RATE CELP CODECS

Tommy Vaillancourt[1,2], Redwan Salami[1], Milan Jelínek[2,1]

[1]VoiceAge Corporation, Qc, Canada, [2]University of Sherbrooke, Qc, Canada

ABSTRACT

This paper presents two new post-processing techniques to address limitations of the deployed low bit rate speech codecs in case of unvoiced speech and background noise, and in case of music. Both post-processing techniques enhance the spectrum of the decoded excitation signal without increasing the codec algorithmic delay. The paper discusses how to integrate the enhancement procedure of unvoiced speech and background noise and of generic audio signals coded by low bit rate ACELP codecs. The proposed post-processing procedures are part of the AMR-WB interoperable modes of the recently standardized 3GPP EVS codec [1].

Index Terms— *s*peech coding, CELP, ACELP, speech enhancement, music enhancement

1. INTRODUCTION

Most of state-of-the-art conversational codecs are based on the code excited linear prediction (CELP) model where the speech signal is synthesized by filtering an excitation signal through an all-pole synthesis filter. The filter coefficients are estimated for each frame of typically 20 ms while the excitation signal is determined in subframes of typically 5 ms. The excitation is usually composed of two contributions selected from an adaptive and a fixed codebook, respectively. The adaptive codebook contains the past excitation signal, thus allowing to take advantage of the quasi periodicity of voiced speech. The fixed codebook excitation, also called innovative excitation, is added to model the unpredictable part of the excitation. In order to guarantee high synthesized speech quality, a large fixed codebook is usually needed. Algebraic CELP, or ACELP, uses an algebraic fixed codebook which enables efficient search of very large codebooks without the need for codebook storage. The algebraic codebook consists of a set of different combinations of pulses with amplitudes +1 and –1 where the number of pulses in a codeword, and thus the codebook size, is limited by the available bit budget. The ACELP technology has become the basis of many widely deployed speech coding standards [2][3][4].

ACELP-based codecs are extremely efficient for coding speech signals even at relatively low bit rates. Their efficiency is however limited for audio signals other than speech. First, the excited Linear Prediction (LP) model is a time-domain coding model which tends to represent better the low part of the spectrum than the higher part which is not optimum for some music signals. Second, the frequency response of the LP synthesis filter represents well the spectral envelope of the speech signals and also the LP filter order generally reflects the number of formant regions present in speech. In contrast, music signals can exhibit very variable spectral envelope, not always well fitting the LP filter model. It can be further shown that the LP analysis renders more efficiently the peaks of the spectrum than its valleys. The perceived noise between formants in CELP coding is attenuated by a LP-based perceptual weighting that makes the quantization noise follow the spectral envelope. This means that less quantization noise is present in low energy region between formants where it is audible while more noise is allowed in formant regions where they are better masked by the active signal. While this scheme is very efficient for speech, it is not sufficient to hide the quantization noise in case of tonal music sequences, and the quantization noise in the low-energy regions of the spectrum becomes audible.

Concerning the excitation construction, the adaptive codebook excitation search is well suited for voiced segments of speech signal where the spectrum exhibits a harmonic structure. However, it fails to properly model tones which are not harmonically related. Further, the search range usually assumes the periodicity corresponding to the range of vibration of human vocal cords which is not always sufficient for tonal music coding.

The fact that the algebraic codebook is composed of simple signed pulses makes the model inefficient at very low bit rates even for coding unvoiced or noisy speech. Especially for coders rendering wider than narrowband (NB) bandwidth, the decoded speech or background noise sounds coarse.

In this paper we present two frequency-domain post-processing techniques to address the limitations of the ACELP low bit rate coding of unvoiced speech and background noise, and of music. Both techniques are based on the Discrete Cosine Transform (DCT) to enhance the spectrum of the decoded excitation signal without increasing the codec algorithmic delay. As they are implemented only on the decoder side, they can be applied to existing ACELP codecs to significantly enhance their performance without affecting their interoperability. In the following description we will assume implementation in the decoder of the AMR-WB interoperable modes of EVS (AMR-WB IO), operating at 12.8 kHz internal sampling frequency with 20 ms frames.

The paper is organized as follows. In Section 2 we describe the enhancement procedure of unvoiced speech and background noise signals coded by low bit rate ACELP codecs. Sections 3 presents the enhancement method of generic audio signals. The performance of the proposed solutions applied to AMR-WB IO is given in Section 4, and conclusions are drawn in Section 5.

2. ENHANCEMENT OF CODED UNVOICED SPEECH AND BACKGROUND NOISE SIGNALS

As mentioned above, unvoiced speech and background noise do not sound natural when processed through a low bit rate ACELP codec. At low bit rates, a significant portion of the bit budget is used to quantize the LP filter coefficients and the adaptive codebook contribution, and only few bits are available for the algebraic codebook. This means that the algebraic codebook contribution contains only a small number of pulses, which is not

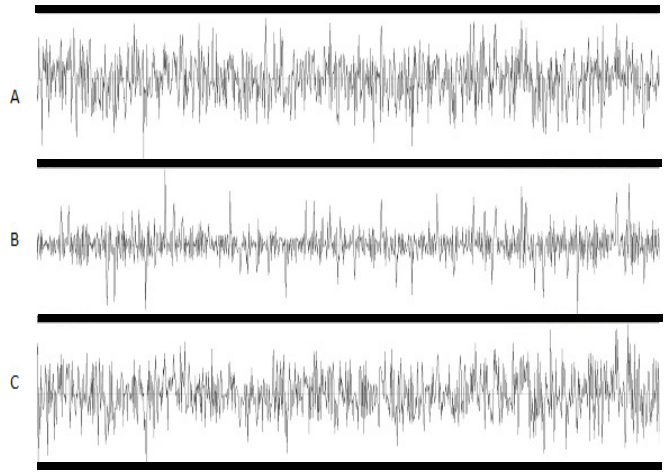

Figure 1: The enhancement effect on the coded excitation

sufficient to faithfully represent the unvoiced speech LP residual or the LP residual of background noise. The proposed enhancement makes the excitation richer by replacing its high frequency component with a random signal. This is illustrated in Figure 1 where plot A shows the LP residual of a car noise signal, plot B the excitation coded by AMR-WB IO at 6.6 kb/s, and plot C the same excitation after the proposed post-processing.

The processing steps of the proposed method are outlined in Figure 2. First, each frame is classified based on the synthesized signal. When the frame is considered as unvoiced speech or as an inactive segment containing only background noise, the unvoiced/inactive enhancement module is entered. The time domain excitation signal $u(n)$ is converted into its frequency representation $f_u(j)$ using a 256 samples DCT-II and a rectangular window. At 12.8 kHz sampling frequency, the frequency resolution of the DCT is 25 Hz. The DCT spectrum is then modified as described in the following subsections. After the spectrum modification, an inverse DCT is performed to convert the excitation back into time domain and the synthesized signal is obtained by filtering the modified excitation with the synthesis filter.

2.1. Analysis per frequency bands

To ensure keeping a similar spectral shape after the spectral modification, the spectrum energy $E_b(i)$ is computed for each frequency band i:

$$E_b(i) = \sum_{j=C_{Bb}(i)}^{C_{Bb}(i)+B_b(i)-1} f_u(j)^2 ,$$

where $C_{Bb}(i)$ is the first bin in the frequency band i, and $B_b(i)$ is the number of bins in that band. The frequency bands are defined such that at low frequencies the used bands correspond to the critical audio bands as defined in [5], but at frequencies above 3700 Hz the band widths are limited to 500 Hz to better handle potential energy variations.

2.2. Computation of a cut-off frequency

As the excitation modification affects only some signal classes, it is important to avoid artifacts when switching between frames where the excitation is modified and frames where the excitation is not modified. To achieve a transparent switching between these frames, it was found that it is better to keep the low part of the spectrum always unchanged. The spectrum is thus modified only above a cut-off frequency f_c which has a minimum value of 1.2 kHz. The cut-off frequency is variable and it corresponds to the 8th harmonics of the fundamental frequency corresponding to the lowest decoded pitch period T of all subframes of the current frame. The 8th harmonics (in Hz) is estimated as follows:

$$h_{8th} = \frac{(8 \cdot F_s)}{\min(T(k))_{k=0:3}} ,$$

where $F_s = 12800$ Hz is the sampling rate, and $T(k)$ is the pitch period in samples of subframe k. Choosing the cut-off frequency based on a multiple of the pitch period ensures a similar coding signature with and without the post-processing. While the pitch period is generally irrelevant for unvoiced speech, it is useful in transition frames or in inactive frames with a dominant quasi periodic component (e.g. car noise).

2.3. Spectrum modification above the cut-off frequency

To overcome the sparseness of the excitation signal, the values of the spectrum above f_c are replaced with a random noise, that is:

$$f_u'(j) = f_u(j) \qquad j = 0,..., j_c$$
$$f_u'(j) = 0.75 \cdot s_r(j) \quad j = j_c + 1,..., L-1 ,$$

where $s_r(j)$ are random numbers limited between -1 and 1, $f_u'(j)$ represents the modified excitation spectrum, j_c is the last frequency bin lower than or equal to f_c, and $L = 256$ is the DCT spectrum length.

2.4. Energy matching

After the spectrum modification, the per-band energy E_b' of the modified spectrum is recalculated using the same method as described in 2.1. An energy matching is then performed by adjusting the energy of each band of the modified excitation spectrum to its initial value. For each band i, the gain $G_b(i)$ to be applied to all bins in that band is defined as:

$$G_b(i) = \sqrt{\frac{E_b(i)}{E_b'(i)}} ,$$

And the final excitation spectrum f_u'' is obtained as follows:

$$f_u''(j) = G_b(i) \cdot f_u'(j), \quad j = C_{Bb}(i),..., C_{Bb}(i) + B_b(i) - 1$$

Finally, the inverse DCT is applied and the synthesis filtering is performed again to overwrite the initial synthesis.

3. MUSIC ENHANCEMENT

The second post-processing method is dedicated to improving generic audio content quality of signals coded with low bit rate ACELP codecs. Previous music post-processing techniques usually work in the synthesis domain and require addition of an extra delay to perform the overlap and add operation to smooth transitions [6]. Our proposal is based on [7], but similarly to the unvoiced speech post-processing method described in Section 2 and shown in Figure 2, it works in the excitation domain.

3.1. Excitation signal extrapolation

While the unvoiced/inactive post-processing method works well with a transform length equal to the frame length, and thus with a relatively low frequency resolution, the reduction of quantization noise in tonal music sequences often requires better frequency resolution to get satisfactory results. To get a better resolution one

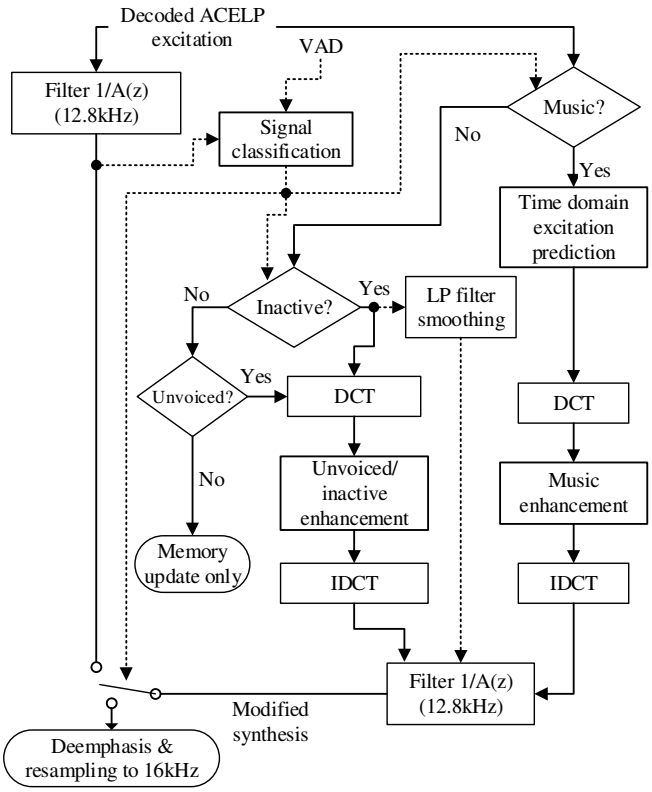

Figure 2: Music and unvoiced/inactive enhancement techniques

usually needs to increase the size of the frequency transform, which leads to a delay increase. In the proposed solution, to avoid increasing the delay, the frequency resolution is artificially increased by extrapolating the current frame excitation signal into the future. We thus build a concatenated excitation vector $u_c(n)$ which contains the 192 samples of the previous frame excitation, the decoded excitation of the current frame $u(n)$, and an extrapolation of 192 excitation samples of the future frame $u_x(n)$. This increases the total excitation length L_c to 640 samples giving a frequency resolution of 10 Hz.

The extrapolation of the future excitation samples $u_x(n)$ is computed by periodically extending the current frame excitation signal $u(n)$ using the decoded fractional pitch of the last subframe of the current frame. Given the fractional resolution of the pitch lag, an upsampling of the current frame excitation is performed using Hamming windowed sinc function.

The concatenated excitation is then windowed using a trapezoidal window defined such that the gain applied to the excitation part corresponding to the current frame is equal to 1, and the past and the extrapolated parts of the excitation are faded to 0 using a Hanning window. A DCT-II is performed on the resulting windowed concatenated excitation to get its frequency representation f_u.

3.2. Inter-tone noise reduction

The proposed music post-processing method is based on inter-tone noise reduction, performed in 2 steps. The first step consists of a SNR-based noise reduction similar to [7] giving an enhanced spectrum $f_u^{'}(k)$ with the difference that here the post-processing is performed in the excitation domain on the artificially extended signal and the quantization noise is estimated differently. In the proposed technique, the quantization noise of a specific frequency band is computed as the average energy of that band excluding the maximum energy bin.

The second step of noise removal consists in applying a weighting mask based on the normalized energy of the concatenated excitation spectrum. The energy spectrum $E_{BIN}(k)$ is normalized between 0.925 and 1.925 to get the normalized spectrum $E_n(k)$:

$$E_n(k) = \frac{E_{BIN}(k)}{\max(E_{BIN})} + 0.925, \qquad k = 0,...,639,$$

where $E_{BIN}(k)$ represents the bin energy. The offset of 0.925 has been chosen such that only a small part of the normalized energy bins would have a value below 1.0. Once the normalization is done, a power of 8 is applied to the normalized energy spectrum $E_n(k)$ to get the resulting scaled normalized energy spectrum $E_p(k)$. The scaled spectrum is then upper limited to 5.

Then, the scaled energy spectrum is smoothed along the frequency axis from low frequencies to high frequencies with an averaging filter. The smoothing can be described with following function:

$$\overline{E}_p(k) = \frac{E_p(k) + E_p(k+1)}{2}, \qquad k = 0$$

$$\overline{E}_p(k) = \frac{E_p(k-1) + E_p(k) + E_p(k+1)}{3}, \quad k = 1,...,638$$

$$\overline{E}_p(k) = \frac{E_p(k-1) + E_p(k)}{2}, \qquad k = 639,$$

where $\overline{E}_p(k)$ is the scaled energy spectrum smoothed along the frequency axis.

Finally, the spectrum is filtered along the time axis to smooth the bin values from frame to frame. The smoothing along the time axis results in a time-averaged amplification/attenuation weighting mask $G_m(k)$ to be applied to the spectrum $f_u^{'}(k)$. The weighting mask is described with the following equation:

$$G_m^t(k) = 0.95 \cdot G_m^{(t-1)}(k) + 0.05\overline{E}_p(k), \quad k = 0,...,319$$

$$G_m^t(k) = 0.85 \cdot G_m^{(t-1)}(k) + 0.15\overline{E}_p(k), \quad k = 320,...,639,$$

where t is the frame index and $G_m^t(k)$ is the time-averaged weighting mask of the frame *t* at the bin *k*. A slower adaptation rate has been chosen for the lower frequencies to substantially reduce gain oscillations. A faster adaptation rate is allowed for higher frequencies given that the positions of the tones are more likely to change rapidly in the higher part of the spectrum. The weighting mask $G_m^t(k)$ is then applied directly on the enhanced spectrum $f_u^{'}(k)$ of the concatenated excitation to get the modified excitation spectrum as:

$$f_u^{''}(k) = f_u^{'}(k) \cdot G_m^t(k).$$

Finally, the modified spectrum of the excitation goes through an inverse DCT. The part of the temporal excitation corresponding to the flat window segment is extracted and passed through the synthesis filter to replace the decoded synthesis.

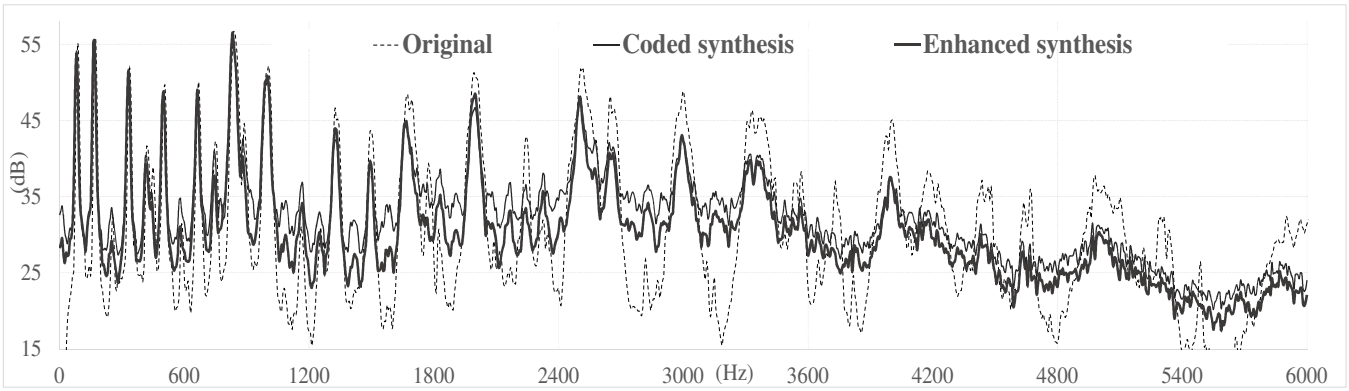

Figure 3: Spectral effect of the music post-processing

The combined effect of both enhancement steps is shown in Figure 3 for a spectrum of a music signal. The dashed line represents the non-coded spectrum, the thin curve corresponds to the coded spectrum and the thick curve corresponds to the enhanced coded spectrum. Compared to [7], the combined inter-tone noise reduction is much more efficient to remove noise between tones that are close in frequency, particularly in low frequency region.

4. PERFORMANCE

The performance of the proposed post-processing methods was evaluated in a Reference-A-B test using expert listeners and in a formal DCR listening test using naïve listeners.

Figure 4 shows the results of the expert listening tests for both post-processing methods. In the tests, both the reference (REF) and the enhanced version (CuT) used the same legacy AMR-WB bitstream. The left side of the figure compares the results for the AMR-WB IO mode at 6.60 kb/s with and without the unvoiced/inactive post-processing. The test was performed on 24 sequences of speech with street noise at 20 dB SNR. Overall, listeners preferred the unvoiced/inactive post-processed version in 92% of the votes. The right side shows the results of the expert listening test for the music post-processing for the AMR-WB IO mode at 8.85 kb/s. The test included 24 different music items, comprising vocal and instrumental sequences. Overall, listeners preferred the music post-processed version in 84% of the votes.

Figure 5 shows the results of the formal DCR listening test, meeting the ITU-T P800 [8] requirement, with 16 naïve listeners. The results for the unvoiced/inactive post-processing are shown on the left side. The test was run with office noise at 20 dB SNR. Both the reference and the enhanced version used the same legacy AMR-WB encoder at 6.60 kb/s. Then the AMR-WB IO decoder with the proposed post-processing (CuT) was compared to the legacy AMR-WB decoder (REF). The results show a clear positive tendency for the proposed unvoiced/inactive post-processing. The right side of the figure shows the results for the music post-processing applied to the AMR-WB IO modes of EVS at 14.25 kb/s and 23.05 kb/s. Again, only the decoder was different and the legacy AMR-WB encoder was used in both cases. Again, a clear preference can be seen for the proposed music post-processing, where the quality of the 14.25 kb/s mode using the music post-processing is close to the quality of the 23.05 kb/s mode without the music post-processing. Very similar results were observed during the selection phase of the EVS codec.

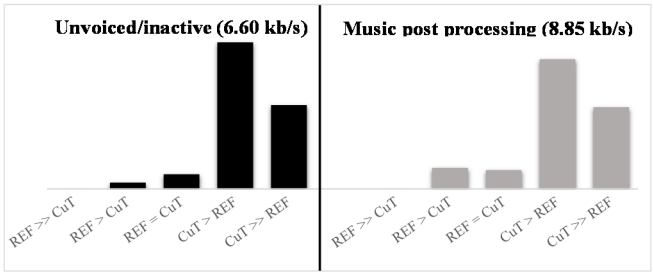

Figure 4: Reference-A-B test results using expert listeners

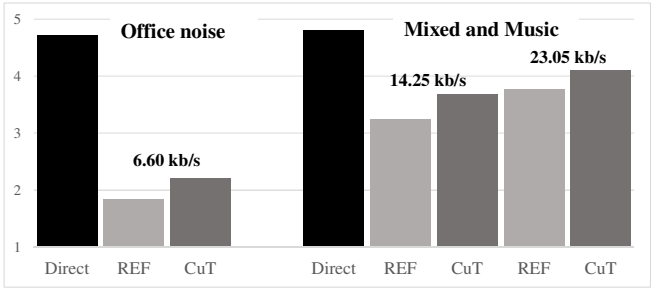

Figure 5: Post-processing evaluation using DCR listening test

5. CONCLUSION

In this paper we have presented two novel techniques to address the limitations of the deployed low bit rate speech codecs in case of processing unvoiced speech, background noise and music. These post-processing techniques are based on modification of the excitation spectrum. They are part of the interoperable mode of the recently standardized 3GPP EVS codec. It has been shown that the quality of unvoiced/inactive and generic audio signals coded at low bit rates can be improved significantly without affecting the algorithmic delay. The quality benefits of the proposed method have been shown for both expert and naïve listeners.

6. REFERENCES

[1] 3GPP Spec., Codec for Enhanced Voice Services (EVS); Detailed Algorithmic Description, TS 26.445, v.12.0.0, Sep 2014.

[2] Bessette, B., Salami, R., Lefebvre, R., Jelinek, M., Rotola-Pukkila, J., Vainio, J., Mikkola, H., and Järvinen, K., "The Adaptive Multi-Rate Wideband Speech Codec (AMR-WB)," *Special Issue of IEEE Trans. Speech and Audio Proc.*, Vol. 10, pp.620-636, November 2002.

[3] Salami, R., Laflamme, C., Bessette, B., and Adoul, J-P., "ITU-T G.729 Annex A: reduced complexity 8 kbit/s CS-ACELP codec for digital simultaneous voice and data", *IEEE Communication Magazine,* vol. 35, no. 9, pp. 56-63, September 1997.

[4] Järvinen, K., Vainio, J., Kapanen, P., Honkanen, T., Haavisto, P., Salami, R., Laflamme, C., and Adoul, J-P., "GSM Enhanced Full Rate Codec", *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, Munich, Germany, pp. 20-24, April 1997.

[5] Johnston, J. D., "Transform coding of audio signal using perceptual noise criteria," *IEEE J. Select. Areas Communication*, vol. 6, pp. 314–323, February 1988.

[6] Fuchs, G. and Lefebvre, R., "A speech coder post-processor controlled by side-information", *Proc. of IEEE International Conference on Acoustics*, Speech and Signal Processing, Philadelphia, USA, pp. 433-436, March 2005.

[7] Vaillancourt, T., Jelinek, M., Salami, R., Malenovsky, V., and Lefebvre, R.,, "Inter-tone noise reduction in a low bit rate CELP decoder", *Proc. of IEEE International Conference on Acoustics*, Speech and Signal Processing, Taipei, Taiwan, pp. 4113-4116, April 2009.

[8] ITU-T Rec. P.800, "Methods for Subjective Determination of Transmission Quality," Geneva, Switzerland, May 1996.