EFFICIENT HANDLING OF MODE SWITCHING AND SPEECH TRANSITIONS IN THE EVS CODEC

Václav Eksler[1,2], Milan Jelínek[1,2], Redwan Salami[2]

[1]University of Sherbrooke, QC, Canada; [2]VoiceAge Corporation, Montreal, QC, Canada

ABSTRACT

The recently standardized codec for Enhanced Voice Services (EVS) consists of a number of modes to achieve its high coding flexibility. In this paper we focus on techniques that enable a seamless switching between two linear prediction based modes running at different sampling rates within this codec. The first one deals with an efficient conversion of the linear prediction filter coefficients. The other one is based on a constrained-memory ACELP called *transition coding* (TC) that significantly limits the inter-frame long-term dependency. We show that the use of TC can be successfully extended to improve quality also in coding other transitions, e.g. strong onsets of voiced speech.

Index Terms— *Speech coding, ACELP, LP filter, transitions, EVS*

1. INTRODUCTION

The codec for Enhanced Voice Services (EVS) [1], recently standardized under the lead of the 3GPP Codec Working Group, consists of a multi-rate audio codec capable of efficiently compressing voice, music and mixed content signals. In order to keep a high audio quality for all sound material over a wide range of bit-rates, the EVS codec consists of many functionally different coding modes which need to switch artifact-free between each other on a frame-to-frame basis.

The EVS codec – similarly like other currently deployed or recently standardized speech codecs – uses the Algebraic Code-Excited Linear Prediction (ACELP) [2] to encode speech dominant content at lower bit-rates. The ACELP codecs heavily rely on prediction to achieve their high performance. This makes them strongly inter-frame dependent and thus not very flexible to well handle switching between different coding modes or abrupt events such as strong onsets. In particular, the coding of speech dominant content in the EVS codec is handled by an ACELP-based coding operating at two different internal sampling rates. While at bit-rates from 5.9 kbps to 13.2 kbps the ACELP operates at 12.8 kHz internal sampling rate, at bit-rates from 16.4 kbps to 64 kbps the ACELP operates at 16 kHz internal sampling rate. In addition, the time-domain based CNG (Comfort Noise Generation) operates at 16 kHz internal sampling rate already from 9.6 kbps.

In the EVS codec the input audio signal is processed in the 20 ms frames. Depending on the internal sampling rate these frames are further divided into four subframes (for bit-rates at internal sampling rate of 12.8 kHz) or five subframes (for bit-rates at internal sampling rate of 16 kHz) when processed using the ACELP model. In the ACELP-based coding, the speech signal is synthesized by filtering an excitation signal through an all-pole digital filter $1/A(z)$ where the excitation parameters are optimized in a perceptually weighted synthesis domain in a closed-loop manner. The filter $A(z)$ is estimated by means of linear prediction (LP) and it represents short-term correlations between speech signal samples.

The excitation signal is typically composed of two parts searched sequentially. The first part of the excitation $v(n)$ is usually selected from an adaptive codebook to exploit the quasi periodicity of voiced speech. This is done by searching in the past excitation the segment most similar to the segment being currently encoded. The second part of the excitation $c(n)$ is an innovation signal selected from a fixed algebraic codebook.

In our previous work [3] we have extended the model with a transition coding (TC) technique to attenuate strong artefacts produced by an ACELP decoder due to error propagation after lost onset frames. This has been achieved by replacing the adaptive codebook with a fixed codebook of glottal impulse shapes (glottal-shape codebook).

The ACELP encoder comprising the glottal-shape codebook, the adaptive codebook and the algebraic codebook is shown in Fig. 1, where β_1 and β_2 are the gains, and $H(z)$ denotes the weighted synthesis filter which is the cascade of the LP synthesis filter $1/A(z)$ and the perceptual weighting filter.

2. SWITCHING ACELP MODES AT DIFFERENT INTERNAL SAMPLING RATES

In the EVS codec, the switching between two ACELP modes operating at different internal sampling rates happens as a result of either 1) bit-rate switching or 2) switching between active segment and time-domain based CNG segment at 9.6 kbps and 13.2 kbps. As the ACELP is based on an inter-frame prediction, it is obvious that the ACELP internal states need a careful handling when switching between these two ACELP modes in order to avoid annoying artifacts and codec instabilities. The most critical parameters here are the LP filter coefficients, in addition to the adaptive codebook memory (past excitation) especially when switching during active speech segments as described below.

2.1. Conversion of LP filter coefficients

When switching between different sampling rates, the LP filter needs to be converted accordingly in order to be able to determine the interpolated LP parameters in each subframe. Let's assume in the following text that the codec needs to switch from an internal sampling rate S_1 to rate S_2. A simple downsampling or upsampling

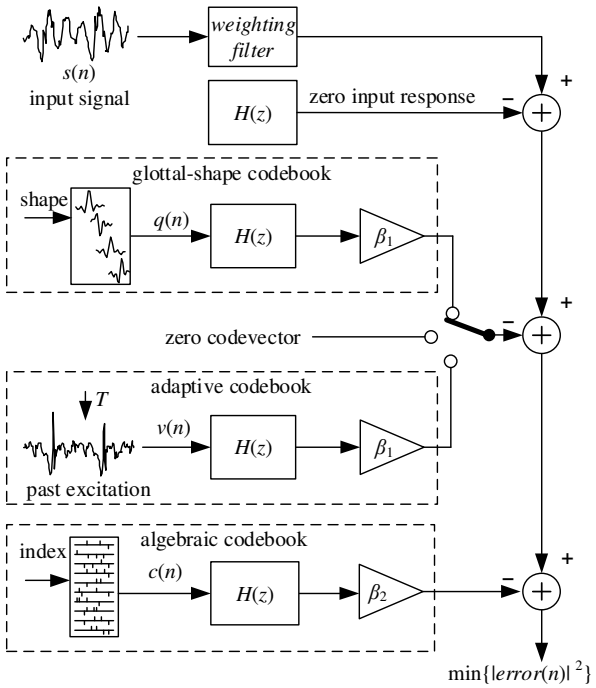

Fig. 1. Principle of the excitation coding in the extended ACELP encoder.

of the original spectrum at rate S_1 would result in a spectrum at rate S_2 that would be a warped version of the original spectrum with displaced formants. A much more accurate spectrum can be obtained when the LP filter parameters at the first frame after the switching are recomputed at rate S_2 based on the synthesis signal of the past frame. The identical synthesis signals can be available at both encoder and decoder. However this approach, presented e.g. in [4] or [5], assumes to perform a complete LP analysis at both the encoder and the decoder, and an availability of the local synthesis at the encoder. Moreover the synthesis signal of the past frame needs to be resampled from the output sampling rate to the internal sampling rate at both encoder and decoder for this approach to work, which is computationally expensive. Instead a simpler and low-complexity method is described here which comprises 1) computing the power spectrum of the LP synthesis filter at rate S_1, 2) modifying the power spectrum to change it from rate S_1 to rate S_2, 3) converting the modified power spectrum back to the time domain to obtain the filter impulse response autocorrelations at rate S_2, and 4) finally using the autocorrelations to compute the LP filter parameters at rate S_2.

The frequency response of the synthesis filter is given by

$$\frac{1}{A(\omega)} = \frac{1}{1+\sum_{i=1}^{M} a_i e^{-j\omega i}} \tag{1}$$

and the power spectrum of the synthesis filter is calculated as an energy of the frequency response of the synthesis filter given by

$$P(\omega) = \frac{1}{|A(\omega)|^2} == \frac{1}{\left(1+\sum_{i=1}^{M} a_i \cos(\omega i)\right)^2 + \left(\sum_{i=1}^{M} a_i \sin(\omega i)\right)^2}. \tag{2}$$

Initially, the LP filter is at a rate S_1. A K-sample (i.e. discrete) power spectrum is computed by sampling the frequency range from 0 to 2π. That is

$$P(k) = \frac{1}{\left(1+\sum_{i=1}^{M} a_i \cos\left(\frac{2\pi i k}{K}\right)\right)^2 + \left(\sum_{i=1}^{M} a_i \sin\left(\frac{2\pi i k}{K}\right)\right)^2} \tag{3}$$

and since the power spectrum is symmetric, it is computed for $k = 0,...,K/2$ only.

In the case where $S_1 > S_2$ the length of the truncated power spectrum is $K_2 = K(S_2/S_1)$ samples, i.e. $K(S_1 - S_2)/S_1$ samples are removed. Since the power spectrum is truncated, it can be computed for $k = 0,...,K_2/2$ only. Since the power spectrum is symmetric around $K_2/2$ we can write that

$$P(K_2/2+k) = P(K_2/2-k), k = 1,...,K_2/2-1. \tag{4}$$

Further it is known that the Fourier Transform of the autocorrelations of a signal gives the power spectrum of that signal. Thus applying the inverse Fourier Transform to the truncated power spectrum results in the autocorrelations of the impulse response of the synthesis filter at rate S_2.

The Inverse Discrete Fourier Transform (IDFT) of the truncated power spectrum is given by

$$R(i) = \frac{1}{K_2}\sum_{k=0}^{K_2-1} P(k) e^{j2\pi i k/K_2}. \tag{5}$$

Since the order of the LP filter is M, the IDFT may be computed only for $i = 0,...,M$. As the power spectrum is real and symmetric and that only $M+1$ autocorrelation coefficients are needed, the inverse transform of the power spectrum is given as

$$R(i) = \frac{1}{K_2}\left(P(0) + (-1)^i P(K_2/2) + 2(-1)^i \Phi\right) \tag{6}$$

where

$$\Phi = \sum_{k=1}^{K_2/2-1} P(K_2/2-k)\cos(2\pi i k/K_2) \tag{7}$$

and the computation of (6) can be further elaborated as

$$R(0) = \frac{1}{K_2}\left(P(0) + P(K_2/2) + 2\sum_{k=1}^{K_2/2-1} P(k)\right), \tag{8}$$

$$R(i) = \frac{1}{K_2}\left(P(0) - P(K_2/2) - 2\Phi\right) \text{ for } i = 1,3,...,M-1, \tag{9}$$

$$R(i) = \frac{1}{K_2}\left(P(0) + P(K_2/2) + 2\Phi\right) \text{ for } i = 2,4,...,M. \tag{10}$$

In the context of the EVS codec, $S_1 = 16$ kHz and $S_2 = 12.8$ kHz in this case. Then K has been chosen to be 100 as explained later and the length of the truncated power spectrum is $K_2 = 80$ samples. The power spectrum is thus computed for 41 samples using (3), and then the autocorrelations are computed using (8) – (10) with $K_2 = 80$.

In the other case where $S_1 < S_2$, the length of the extended power spectrum is $K_2 = K(S_2/S_1)$ samples. After computing the power spectrum from $k = 0,...,K/2$, the power spectrum is extended to $K_2/2$, i.e. by $K(S_2 - S_1)/S_1$ samples. Since there is no original spectral content between $K/2$ and $K_2/2$, a simple approach is used in the EVS codec that repeats the sample at $K/2$ up to $K_2/2$.

In the context of the EVS codec, $S_1 = 12.8$ kHz and $S_2 = 16$ kHz in this case. Then $K = 80$ and the length of the extended power spectrum is $K_2 = 100$ samples. The power spectrum is computed for 51 samples using (3), and then the autocorrelations are computed using (8) – (10) with $K_2 = 100$.

In either case, the IDFT is computed next to obtain the autocorrelations at sampling rate S_2 and the Levinson-Durbin algorithm is used to compute the LP filter parameters at sampling rate S_2. Finally the filter coefficients are transformed to the LSP domain for interpolation with the current frame LSPs in order to obtain the LP parameters at each subframe.

Note that the conversion of the LP filter parameters between different internal sampling rates is applied to the quantized LP parameters in order to be able to determine the interpolated synthesis filter parameters in each subframe at the same way in the encoder and the decoder. The number of samples K used to compute the discrete power spectrum was found experimentally. While the accuracy of the conversion increases with a higher sampling points, it also increases the computational complexity. Consequently in the EVS codec the length was set to $K = 100$ at 16 kHz internal sampling rate and 80 at 12.8 kHz internal sampling rate as the best compromise between the conversion accuracy and its complexity. An example of a conversion for different spectrum lengths when $S_1 = 16$ kHz and $S_2 = 12.8$ kHz is shown in Figure 2.

2.2. Handling of excitation memory

The most important inter-frame prediction in the ACELP model is the excitation memory employed in the adaptive codebook. When switching between different internal sampling rates, one can consider to resample the past excitation or, alternatively, completely replace the adaptive codebook by a fixed one in order to avoid the inter-frame excitation memory dependency. In the EVS codec we have chosen the latter approach by using the glottal-shape codebook of the TC technique [3]. It is used in the first frame after switching between ACELP modes running at different sampling rates. As the adaptive codebook memory of the previous frame is not employed at this frame due to the use of TC, it can be simply reset to zeros.

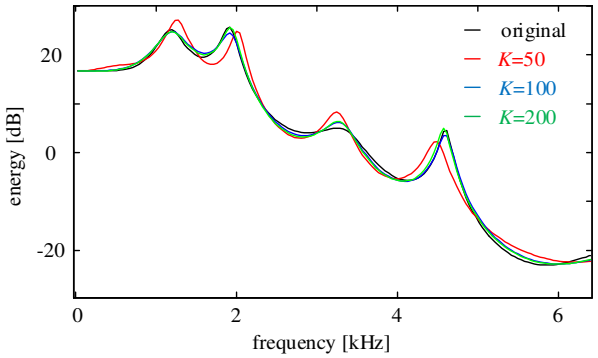

Fig. 2. Comparison of LP filter conversion accuracy for several spectrum lengths.

In the frame coded by TC, the glottal-shape codebook is employed only in the subframe containing the first glottal impulse of the frame in order to take advantage of the intra-frame prediction and to optimize the bit allocation. The other subframes are coded using a combination of a standard adaptive codebook and algebraic codebook, or using only an algebraic codebook if a subframe does not contain any significant portion of a glottal impulse.

3. CODING SPEECH TRANSITIONS

In section 2.2 we have discussed an excitation memory handling in case of ACELP switching. A similar problem happens when the codec processes transitions where the previous and current frame excitation are very different. This situation usually happens in coding of transitions from unvoiced or inactive (CNG) segments to voiced speech or transitions between two voiced segments. In these situations the long-term prediction is not very efficient and the adaptive codebook contribution to the total excitation is limited.

In addition, in several recent codecs, e.g. G.718 [6] or EVS [7], the adaptive and algebraic codebook gains β_1 and β_2 are quantized jointly in each subframe. While the adaptive codebook gain is quantized directly, the algebraic codebook gain is quantized indirectly using a predicted energy of algebraic codebook vector estimated per whole frame. When there is an abrupt transition in a current frame, it is obvious that the gain quantization efficiency decreases. This is in particular critical when a transition like a strong voiced speech onset happens towards the frame end where the gain quantizer is not able to adequately react to an abrupt energy increase.

In order to overcome the above limitations, it is advantageous to use the TC in speech transition frames, in particular when a transition or onset is located towards the frame end. These frames can be detected using an attack tracking algorithm as follows.

3.1. Transition detection

The current frame input audio signal at 12.8 kHz internal sampling rate $s(n)$ is divided into 32 segments. Given the frame length of 20 ms, this results in 8 samples per segment. In the next step, energy is calculated in each segment as

$$E_{seg}(k) = \sum_{i=0}^{7} s^2(8k+i), \quad \text{for } k = 0,..,31 \cdot \tag{11}$$

The segment with the maximum energy is then found using

$$k_{att} = \max_k \left(E_{seg}(k)\right), \tag{12}$$

and this is the segment position of the candidate attack. In all active frames that are not within an unvoiced or a stable voiced segment, the following logic is executed to eliminate false attacks, i.e. attacks that are not sufficiently strong. First, the mean energy in the first three or four sub-frames is calculated as

$$E_{SF} = \frac{1}{N_{SF}} \sum_{k=0}^{N_{SF}-1} E_{seg}(k) \cdot \tag{13}$$

The number of segments N_{SF} where the mean energy is calculated depends on the EVS bit-rate and the corresponding number of subframes per frame. More specifically, the mean energy is calculated in the first 24 segments at bit-rates with the internal sampling rate of 12.8 kHz, and in the first 26 segments at bit-rates with an internal sampling rate of 16 kHz.

Then the mean energy after the detected candidate attack is defined as

$$E_{after} = \frac{1}{32 - k_{att}} \sum_{k=k_{att}}^{31} E_{seg}(k) \qquad (14)$$

and the ratio of the two energies is compared to a certain threshold. That is if $E_{after} / E_{SF} < 8$ then k_{att} is set to 0. In other words the candidate attack position is reset if the attack is not sufficiently strong. In addition, k_{att} is also set to 0 if the last frame was classified as a stable voiced speech frame and if $E_{after} / E_{SF} < 20$.

To further reduce the number of falsely detected attacks, the segment with maximum energy is also compared to other segments. If $E_{seg}(k_{att}) / E_{seg}(k) < 2$ for $k = 2,..,N_{SF} - 2$, $k \neq k_{att}$, then k_{att} is again set to 0. In other words, if the energy in any of the above defined segments, other than k_{att}, is close to that of the candidate attack, the candidate attack is eliminated by setting k_{att} to 0.

Finally, if a transition has been found by the attack tracking algorithm described above, the position of this attack is tested. If this position is in the last $N_{SF} - 2$ segments of the current frame, the frame is encoded using the TC technique and the glottal-shape codebook is employed at the last subframe of that frame.

3.2. Performance in transitions coding

Figure 3 compares a segment of input speech signal containing a strong plosive with two synthesized signals. The first synthesis is obtained using a transition coded by legacy ACELP while the second one shows the same segment coded using the TC. The signal excerpt was obtained from the EVS codec running at 13.2 kbps. It can be seen from the Figure that the coding efficiency has greatly increased with the use of TC in this frame. In general, we have observed that the use of TC greatly improves the intelligibility of coded speech in such transition frames, in particular in frames containing strong plosives.

Table 1 compares the coding efficiency of the TC technique with the legacy ACELP measured in the perceptual domain at three EVS codec bit-rates. It is measured in terms of segmental SNR on the frames where a transition was detected using the attack tracking algorithm from Section 3.1, i.e. only on the transition frames where TC is used. The numbers were obtained from more than 7 min long American English clean speech database (22,804 frames) with the active frame ratio being about 60-65% depending on the EVS codec bit-rate.

4. CONCLUSION

We have presented efficient techniques that help to provide a seamless frame-to-frame mode switching between ACELP modes running at different internal sampling rates and improve speech transitions coding. These techniques form part of the recently standardized 3GPP EVS codec.

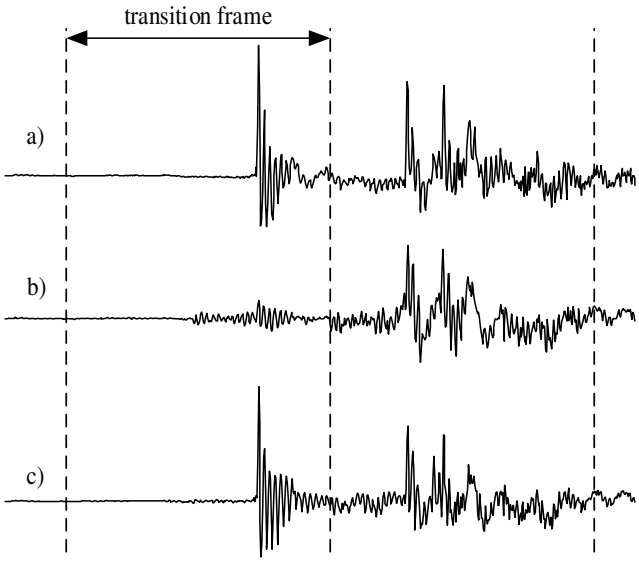

Fig. 3. Comparison of (a) input signal, (b) synthesis without TC in transition frame, and (c) with TC in transition frame.

bit-rate [kbps]	segmental SNR [dB]		# of TC transition frames
	without TC	with TC	
7.2	5.027	5.439	655
13.2	5.652	7.455	702
32	11.825	12.721	561

Table 1. Impact of using TC in transition frames.

5. REFERENCES

[1] M. Dietz *et al.*, "Overview of the EVS codec architecture," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Brisbane, Australia, 2015.

[2] J. P. Adoul *et al.*, "Fast CELP coding based on algebraic codes," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Dallas, TX, 1987, pp. 1957 – 1960.

[3] V. Eksler and M. Jelínek, "Glottal-Shape Codebook to Improve Robustness of CELP Codecs," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 18, no. 6, pp. 1208 – 1217, Aug. 2010.

[4] ISO/IEC 23003-3:2012, "MPEG-D (MPEG audio technologies), Part 3: Unified speech and audio coding," 2012.

[5] U. Mittal *et al.*, "Encoder for Audio Signal Including Generic Audio and Speech Frames," US Patent No. 8,423,355, Apr. 2013.

[6] M. Jelínek, T. Vaillancourt, J. Gibbs, "G.718: A new embedded speech and audio coding standard with high resilience to error-prone transmission channels," *IEEE Commun. Mag.,* vol. 47, no. 10, pp.117-123, Oct. 2009.

[7] 3GPP Spec. TS 26.445: "EVS Codec Detailed Algorithmic Description," v.12.0.0, Sep. 2014.