ADVANCES IN LOW BITRATE TIME-FREQUENCY CODING

Tommy Vaillancourt[1,2], Vladimir Malenovsky[1,2], Redwan Salami[1], Zexin Liu[3], Lei Miao[3], Jon Gibbs[3], Milan Jelinek[1,2]

[1]VoiceAge Corp., Qc, Canada, [2]University of Sherbrooke, Qc, Canada, [3]Huawei Technologies Co. Ltd, Shenzhen, China

ABSTRACT

In this paper a novel technique is presented to efficiently mix traditional ACELP time domain coding with a frequency domain coding model to improve the quality of generic audio signals coded at low bitrates without additional delay. The paper discusses how to integrate parts of a traditional Algebraic Code Excited Linear Prediction (ACELP) speech codec to create a time-domain contribution which coexists with a frequency based coding model. A mechanism to determine the value of the time-domain contribution is proposed and a method is described how the frequency-domain contribution might be added without increasing the overall delay of the codec. The proposed method forms part of the recently standardised 3GPP EVS codec.

Index Terms— *s*peech coding, ACELP, music coding, low bitrate

1. INTRODUCTION

Current state-of-the-art conversational codecs can represent clean speech signals with very good quality at bitrates up to approximately 8 kb/s and approach transparency at bitrates around 16 kb/s. To sustain this high speech quality, even at low bitrates, and with a latency below 30 ms, a time domain multi modal coding scheme is often used [1][2]. Usually the input signal is classified between different categories reflecting its characteristics and codec operational modes are optimized for speech and noisy speech content, but they do not generally address generic audio or reverberant speech inputs.

The linear prediction (LP) analysis performed in an ACELP codec makes it able to render the spectral peaks and the low frequency content efficiently. The masking property of the human ear is also exploited by shaping the quantization noise so that it has more energy close to the formant frequencies where it will be masked by the stronger signals [3] but on the other hand, the rendering of the spectral valley and high frequency content is poor. The long-term pitch filter of ACELP is very efficient for modelling the simple harmonic voiced speech segments but when the spectrum exhibits a more complex harmonic structure it fails to model the tonal structure effectively. All these characteristics make ACELP based codecs near-optimal for voiced speech signals but until now they have been sub-optimal for generic audio content, such as music, speech over music and even reverberant speech.

Generic audio is usually encoded using a frequency domain approach which often employs multi-band gain coding, different techniques to code spectral pulses and different windows shapes depending on the signal characteristics [4]. Typically, frequency domain coding methods also have longer delays when compared to time-domain codecs, to allow for high resolution time to frequency conversion. Recently, audio codecs have started to emerge which integrate both the time and frequency domain coding within a switched model; where the input signal is categorized as either speech or generic audio and then the appropriate underlying time or frequency coding model is chosen [5]. This works well but has the drawback of increasing the delay of the whole codec to ensure that the correct speech-music classification is achieved and also to accommodate the additional delay of high resolution transforms to the frequency domain. Another significant drawback is that at low bitrates, frequency domain coding techniques do not tend to be robust to classification errors, leading to poor quality coding if a signal is misclassified. This last weakness is particular important given the delay constraints imposed on the EVS codec, as there is a significant likelihood of signal misclassification.

To overcome these limitations, a hybrid technique combining the strengths of time domain coding with those of frequency domain coding at low bitrates has been developed. The technique improves the quality of music and mixed content without affecting the delay and having the benefit of being more robust to classification errors. The new approach is presented here. Section 2 describes the proposed modifications to the ACELP model and section 3 describes the efficiency of the new model. Section 4 describes the spectral model and quantization, and the specifics for the EVS codec implementation are presented in section 5. Finally, the performance of the new model is described in section 6 with conclusions in section 7.

2. TIME-DOMAIN CONTRIBUTION

In a regular ACELP time-domain encoder, the bit budget is split between the LPC filter quantization, the pitch filter description and its gains, and the algebraic codebook and its

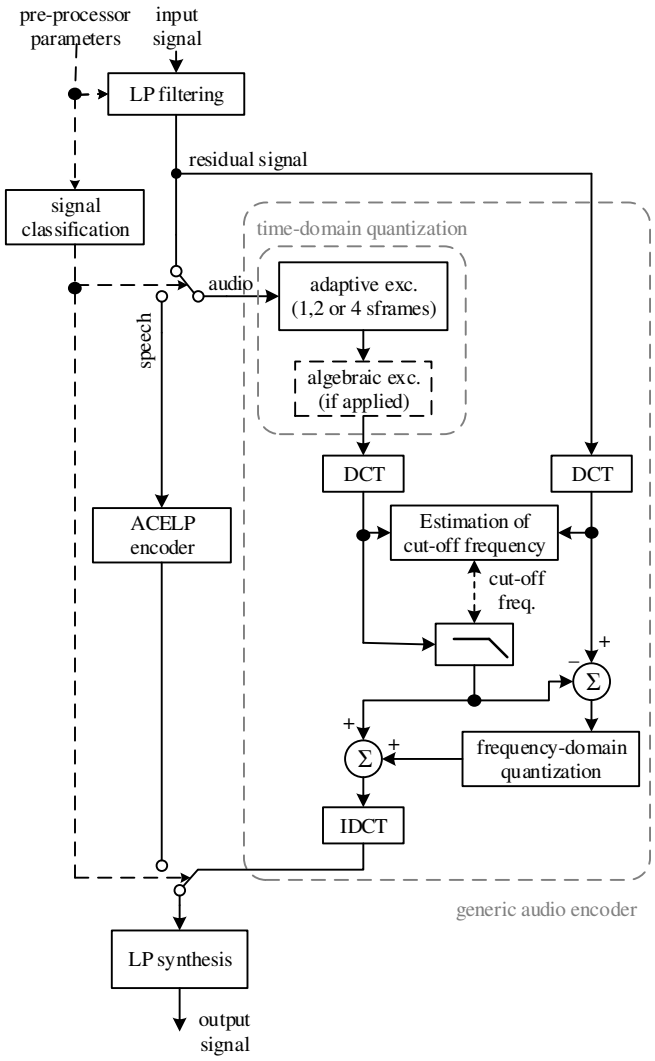

Figure 1: Simplified representation of the time/frequency model

gains. Most state-of-the-art low bitrate speech codecs use a temporal support of 20ms at an internal sampling frequency of 12.8 kHz which leads to a total of 256 samples per frame, usually split among 4 subframes of 64 samples each. The pitch information, the algebraic codebook and the gains are all computed for each subframe.

The proposed model, as depicted in Figure 1, follows a different approach by allowing the pitch filter information and its gain to be used for a restricted and preselected number of subframes. The number of subframes can be either one subframe of 256 samples, two subframes of 128 samples or the usual four subframes of 64 samples. When the number of subframes selected is four, then encoding of the algebraic codebooks and their gains is optional, otherwise the use of algebraic codebooks is disallowed.

The number of subframes selected depends upon the characteristics and the bitrate. If the signal is identified as speech, then a standard ACELP model is used to encode the

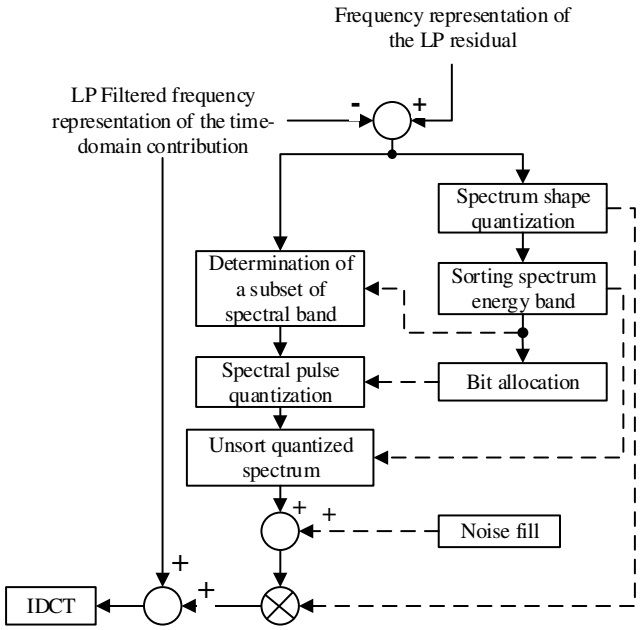

Figure 2: Frequency-domain encoder

frame. When the frame is not purely speech, the generic audio encoder is used. In [6], a similar model is presented, though the spectral quantization was performed on the target and the time-domain contribution was always present for speech and never for music. In our approach, the time-domain contribution does not depend on the content type and is low-pass filtered as explained in Section 3. Furthermore, the spectral quantization is not performed on the target, but on the difference between the residual and the time-domain contribution. If the spectrum of the frame entering the generic audio coding mode exhibits high frequency tones, then a lower number of subframes is favoured to increase the bit budget of the frequency domain coding mode. On the other hand, for signal spectra with either noise or speech-like characteristics a higher number of subframes is favoured. At low bitrates, only one or two subframes are permitted in order to conserve sufficient bits for the second step of frequency domain quantization.

3. TEMPORAL CONTRIBUTION CUT-OFF FREQUENCY

The excitation obtained after the first quantization step, taking into account the pitch filter information, and possibly the algebraic codebooks, corresponds to the time-domain contribution of the proposed model. The time-domain contribution and the input residual are transformed into the frequency domain using a DCT with a rectangular window. When coding generic audio signals, the temporal contribution (the combination of adaptive and/or algebraic codebook) rarely contributes much to the coding gain. Occasionally, however, it does increase the coding gain of the lowest part

of the signal spectrum, but the coding gain in the higher frequency part of the spectrum is often minimal.

To obtain the cut-off frequency where the coding gain of the time-domain contribution becomes very low, the spectrum is split into different frequency bands. Then, an analysis of the normalized cross-correlation between the LP residual and time-domain excitation contribution is performed in each of the frequency bands. The resultant cross-correlation vector is then offset and normalized between 0 and 1 for each band. The average cross-correlation is then used to identify the last frequency band f_c where the coding gain is considered to be beneficial and above which the temporal contribution should stop. All the frequencies above the cut-off frequency f_c are then gradually faded to zero. This operation may be considered as mimicking that of a low pass filter, at the frequency f_c, applied to the time-domain contribution.

Sometimes, particularly for noisy content, the temporal contribution offers little benefit and its coding gain is very low for all frequency bands. In this case the temporal contribution is forced to zero and the bit budget reallocated to encoding just the frequency domain contribution.

4. FREQUENCY-DOMAIN CONTRIBUTION

Once the cut-off frequency of the time-domain contribution is defined, the frequency domain encoding is performed. First a new vector is formed from the difference between the spectral representations of the residual and the low pass filtered time-domain contribution. The resulting spectrum difference vector f_d is then quantized using the frequency domain coding module. Before the spectral quantization is performed, the spectrum is split into 16 bands, where the number of bins B_b per band is defined as:

$$B_b(i) = \{8,8,16,16,16,16,16,16,16,16,16,16,16,16,16,32\} \quad (1)$$

The gain per band on the difference signal G_{bd} is computed as the logarithm of the cumulative energy in each band, i:

$$G_{bd}(i) = \log_{10}\left(\sqrt{\sum_{k=0}^{B_b(i)-1} f_d(k)^2}\right), \quad for\ 0 \le i < 16 \quad (2)$$

The gain per band vector is quantized using a split vector quantizer. In total, between 21 and 26 bits are used to quantize $\hat{G}_{bd}$ depending on the bitrate. Given the limited number of bits available, the quantized gain per band is used to sort the frequency bands before spectral quantization so that only the more energetic spectral bands are allocated quantizer bits. Spectrum quantization is performed using a pulse vector quantizer, PVQ, scheme as described in [7].

Given that some of the bits are allocated to the LP filter description, and to the time-domain contribution, it is very probable that at low bitrates several frequency bands will not be allocated any spectrum quantizer bits. To prevent musical noise and important distortion in speech, noise is therefore added in the bands where no bits have been allocated. The noise level corresponds to a fraction of the quantized pulse level with the level being higher at low bitrates and in the high frequency region.

After noise filling, the quantized and the noise filled spectral bands are gain-adjusted, using the gains computed above, to obtain the quantized frequency contribution. This contribution is then added to the frequency representation of the time-domain contribution derived earlier; as depicted in Figure 2. Once the two contributions have been added together, an inverse DCT, without overlap-add, is performed to obtain the final time-domain excitation coded by the generic audio encoder. The absence of overlap-add windows is possible, as explained in [8], due to the fact that the frequency domain quantization is performed in the excitation domain, the synthesis filter smoothing any potential time domain discontinuities.

5. EVS SPECIFIC

When applying the technique to the EVS codec, some design choices have had to be made. Firstly, the model is used at bitrates below 16.4 kb/s to code inactive background noise signals. In EVS, a signal classifier [9] is applied to establish signals that will be considered as speech-like. At low bitrates, all signals not considered as speech-like, such as music, or reverberant speech, are encoded with the model described. At the lowest bitrates, i.e. 7.2 and 8.0 kb/s, due to the very limited number of bits available, only 1 subframe is permitted to describe the pitch filter but more configurations are permitted at 13.2 kb/s. Finally, in case of music at 13.2 kb/s, the model is used in conjunction with an MDCT coding paradigm, thus this model is used only for some specific signals. More details can be found in [10].

At 8 kb/s the bit allocation is sometimes suboptimal. When the index of the highest frequency band with bits allocated, I_{high_bit} is higher than a threshold, the bit allocation for the frequency excitation bands is adjusted by decreasing the number of bits allocated to the bands having the largest numbers of bits, and increasing the number of bits of the band I_{high_bit} and the bands near to it. The threshold is determined from the number of available bits and the resolution of the frequency excitation signal. This procedure helps to improve the quality at the sub-bands around I_{high_bit}.

The audio bandwidth of the proposed model extends up to 6.4 kHz. However at low bitrates, the audio signal above 6 kHz is never coded but noise filled. To extend the audio bandwidth from 6 kHz to 8 kHz for wideband (WB) inputs, a dual mode bandwidth extension (BWE) scheme is employed which is based on the multi-mode BWE described in [11]. Two signal classes are identified, HARMONIC and NORMAL, according to the degree of spectral fluctuation present.

At 13.2 kb/s, a bit budget of 6 bits is allocated to the WB BWE. One bit reflects the signal class and five bits are allocated to two spectral envelopes, which are based upon the 80 MDCT coefficients of the higher band (HB) signal. At 7.2

kb/s or 8 kb/s, only blind BWE is used with no bits allocated. At the decoder, the excitation de-normalization, envelope adjustment and noise filling are directed depending on the band classification, the current frame is smoothed, based upon the previous frame, and is then applied to the high frequency band excitation to reconstruct the high band.

For the blind BWE at 7.2 kb/s and 8 kb/s, two frequency envelopes of the high frequency band are estimated from the low frequency band. Two average energy values $E_L(0)$ and $E_L(1)$, are calculated based on 32 MDCT coefficients in consecutive sub-bands.

$$E_L(0) = \sum_{k=192}^{223} \left(X_M(k)\right)^2 \quad (3)$$

$$E_L(1) = \sum_{k=224}^{255} \left(X_M(k)\right)^2 \quad (4)$$

From these a base frequency envelope is calculated

$$f_{env\_base} = \sqrt{(E_L(0) + E_L(1))/64} \quad (5)$$

Finally, two frequency envelopes for the high frequency band are derived from the base frequency envelope with the ratio of $E_L(0)$ and $E_L(1)$ which indicates the trend of the envelopes.

6. PERFORMANCE

The performance of the proposed algorithm has been subjectively assessed by naïve listeners according to the ITU-T P.800 methodology and compared to the AMR-WB codec operating at higher bitrates. Subjective tests were performed for mixed content and music in error-free conditions and for data channels exhibiting frame losses. The results are presented in Figures 3 and 4 with error-bars indicating the 95% confidence interval.

Figure 3 shows a comparison for clean channel conditions of AMR-WB, operating at 8.85 and 12.65 kb/s, with the EVS Codec, operating at 7.20 kb/s and 8.0 kb/s. It can be seen that the EVS Codec operating points are statistically better than AMR-WB at 8.85 kb/s. EVS at 7.20 kb/s is equivalent to AMR-WB at 12.65 kb/s and EVS at 8.0 kb/s was slightly better than AMR-WB at 12.65 kb/s.

In Figure 4, AMR-WB at 12.65 and 15.85 kb/s are compared with the EVS Codec operating at 7.20 and 8.0 kb/s for 6% frame loss. In this case, it is clear that the proposed method is very robust to frame loss and that the quality improvement is very significant as the EVS Codec operating points are statistically better than AMR-WB at 12.65 kb/s. EVS operating at 8.0 kb/s is also statistically better than AMR-WB at 15.85 kb/s. Similar results were observed during the 3GPP EVS selection phase.

Figure 5 shows the results per samples for a comparison between EVS without the proposed mode (CuT 1) at 7.2 kb/s, EVS integrating the proposed coding mode as in [10] (CuT 2) and a 7.8 kb/s ACELP/TCX based on [4] (CuT 3) but limited to a 20 ms frame as in the EVS. The MUSHRA methodology (ITU-R BS.1534) was used with 9 expert listeners to compare 20 generic audio sequences coded with the above configurations. In CuT 2, the proposed scheme is

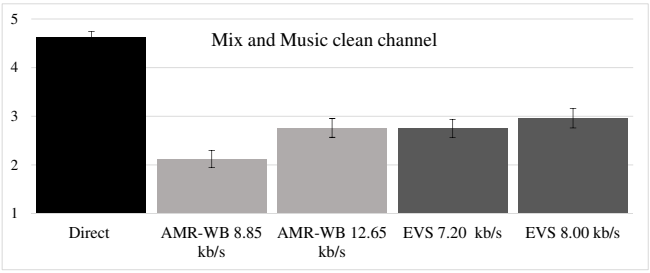

Figure 3: Mix and music performance on clean data channel

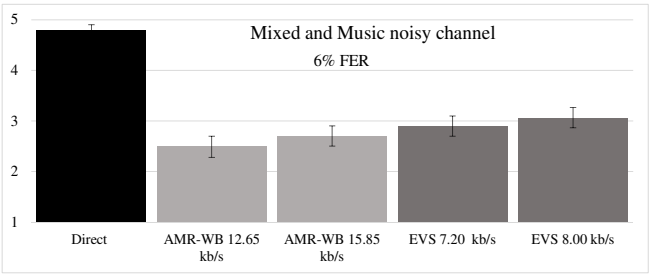

Figure 4: Mix and music performance with 6% frame lost

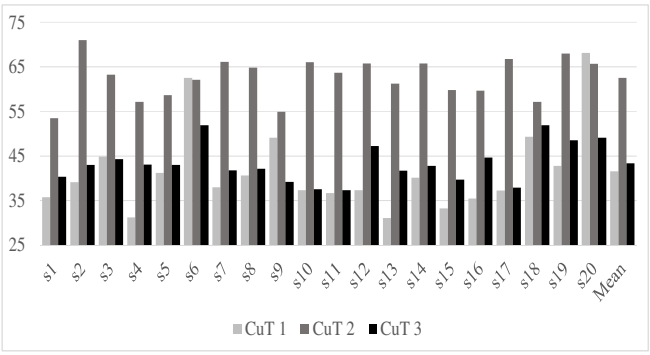

Figure 5: Comparison between different coding models

used for 85 % of the coded frame, while in CuT 3, the TCX is chosen for 73 % of the coded frames. The results highlight that CuT 3 performs only slightly better than a state of the art ACELP codec (CuT 1), while the proposed technique (CuT 2) in combination to the state of the art ACELP brings a major quality improvement of the coded generic audio signal at low bitrates.

7. CONCLUSION

In this paper a novel technique to efficiently mix the traditional ACELP time domain coding with a frequency domain coding model has been presented. A significant advantage of the proposed model is its coexistence with the traditional ACELP speech codec; allowing both time-domain contribution and a frequency based coding models to be combined. This approach forms part of the recently standardized 3GPP EVS codec and it has been shown that the quality of generic audio signals coded at low bitrates can be improved significantly compared to previous generation codecs for both clean and error-prone data channels.

8. REFERENCES

[1] M. Jelinek, T. Vaillancourt, A. Erdem Ertan, J. Stachurski, A. Rämo, L. Laaksonen, J. Gibbs and S. Bruhn, "ITU-T G.EV-VBR Baseline Codec", *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 4749-4752, March 2008.

[2] M. Jelinek and R. Salami, "Wideband Speech Coding Advances in VMR-WB standard", *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 15, no. 4, pp. 1167-1179, May 2007.

[3] B. Bessette, R. Salami, R. Lefebvre and M. Jelinek, "Efficient methods for high quality low bit rate wideband speech coding", *IEEE Speech Coding WorkShop,* pp. 114-116, October 2002.

[4] B. Bessette, R. Lefebvre, and R. Salami, "Universal speech / audio coding using hybrid ACELP/TCX techniques", *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Vol. 3, pp. 301-304,* March 2005.

[5] M. Neuendorf, M. Multrus, N. Rettelbach, G. Fuchs, J. Robillard, J. Lecompte, S. Wilde, S. Bayer, S. Disch, C. Helmrich, R. Lefevbre, P. Gournay, et al., "The ISO/MPEG Unified Speech and Audio Coding Standard - Consistent High Quality for All Content Types and at All Bit Rates", *J. Audio Eng. Soc., vol. 61, no. 12, pp. 956-977, Dec. 2013.*

[6] R. Lefevbre, R. Salami, C. Laflamme, J.-P. Adoul, "High quality coding of wideband audio signals using transform coded excitation (TCX)", *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. 1, pp. 193-196, April 1994.

[7] J. Svedberg, V. Grancharov, S. Sverrisson, E. Norvell, T. Toftgard, H. Pobloth and S. Bruhn, "MDCT Audio Coding with Pulse Vector Quantizers", accepted for publication, *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, April 2015.

[8] T. Backstrom, "Comparison of windowing in speech and audio coding", Application of Signal Processing to Audio and Acoustics (WASPAA), IEEE Workshop, October 2013.

[9] V. Malenovsky, T. Vaillancourt, W. Zhe, K. Choo and V. Atti, "Two-stage speech/music classifier with decision smoothing and sharpening in the EVS codec", accepted for publication, *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, April 2015.

[10] 3GPP Spec., Codec for Enhanced Voice Services (EVS); Detailed Algorithmic Description, TS 26.445, v.12.0.0, Sep 2014.

[11] L. Miao, Z.X. Liu, C. Hu, V. Eksler, S. Ragot, C. Lamblin, B. Kovesi, J. sung, M. Fukui, S. Sasaki and Y. Hiwasaki, "G.711.1 ANNEX D AND G.722 ANNEX B – NEW ITU-T SUPERWIDEBAND CODECS," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing,* pp. 5232-5235, May 2011.