Flexible and Scalable Transform-Domain Codebook for High Bit Rate CELP Coders

Václav Eksler, Bruno Bessette, Milan Jelínek, Tommy Vaillancourt
University of Sherbrooke, VoiceAge Corporation
Montreal, QC, Canada

***Abstract*—The Code-Excited Linear Prediction (CELP) model is very efficient in coding speech at low bit rates. However, if the bit rate of the coder is increased, the CELP model does not gain in quality as quickly as other approaches. Moreover, the computational complexity of the CELP model generally increases significantly at higher bit rates. In this paper we focus on a technique that aims to overcome these limitations by means of a special transform-domain codebook within the CELP model. We show by the example of the AMR-WB codec that the CELP model with the new flexible and scalable codebook improves the quality at high bit rates at no additional complexity cost.**

Keywords—Speech coding, ACELP, AMR-WB

I. INTRODUCTION

The Code-Excited Linear Prediction (CELP) model [1] is widely used to encode speech signals at low bit rates. In CELP, the speech signal is synthesized by filtering an excitation signal through an all-pole digital synthesis filter where the excitation parameters are optimized in a perceptually weighted synthesis domain in a closed-loop manner. The filter 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, known as a long-term prediction (LTP), 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 is selected from an innovation codebook and it models the evolution (difference) between the past segment and the currently encoded segment.

The innovation codebook can be designed using many structures and constraints. However, in modern speech coding systems, the Algebraic CELP (ACELP) model [2] is often used. The codevectors in an algebraic innovation codebook contain only few non-zero pulses while the number of the pulses depends on the bit rate of the codebook. The task of the ACELP coder is to search the pulse positions and their signs to minimize a mean square error criterion [2]. While the ACELP coder is very efficient at low bit rates, it faces some difficulties at high bit rates.

First, the ACELP model does not gain in quality as quickly as other approaches such as transform coding and vector quantization when increasing the innovation codebook size. When measured in dB/bit/sample, the gain at higher bit rates (above approximately 16 kbps) obtained by using more pulses in an ACELP innovation codebook is not as large as the gain in dB/bit/sample of transform coding and vector quantization. At lower bit rates (below 12 kbps), the ACELP model captures quickly the essential components of the excitation. However, at higher bit rates, higher granularity and, in particular, better control over how the additional bit budget is spent across different frequency components of the signal are useful.

Then, with increased bit rate, more pulses are searched within very large codebooks and the complexity of the ACELP search algorithm becomes too high for practical implementations. Though many ACELP search algorithms have been proposed that address this problem [3][4], the computational complexity of these algorithms still represents a substantial part of the overall codec complexity. Consequently, careful control over the computational complexity is needed, which in turn limits the coding efficiency of these algorithms.

In this paper we show how the traditional CELP model can be extended at high bit rates to overcome the limitations of scalability and complexity. While the presented technique can be implemented in any CELP-based coder, we have chosen the AMR-WB codec to demonstrate the performance of the proposed technique. The rest of the paper is organized as follows. In Section II, the AMR-WB codec is described with emphasis on the innovation codebook used. In Section III we extend the traditional CELP model by a new special transform-domain codebook. Section IV shows the performance of the presented technique. In Section V we discuss some implementation considerations of the model before concluding the paper in Section VI.

II. AMR-WB CODEC BACKGROUND

The AMR-WB codec [5] is currently widely deployed in mobile communications. It delivers wideband speech with audio bandwidth of 50–7000 Hz at one of the nine specified constant bit rates: 23.85, 23.05, 19.85, 18.25, 15.85, 14.25, 12.65, 8.85 and 6.6 kbps. In the AMR-WB codec, the input audio signal is processed in frames of 20 ms while each frame is further divided into four subframes. The codec employs a band-split processing. The low band is coded by the ACELP model at the internal sampling rate of 12.8 kHz and covers frequencies up to 6.4 kHz while a band-width extension (BWE) is used to cover the rest of the spectrum. A blind BWE is used at all bit rates except of 23.85 kbps, where a guided (16 bits/frame) BWE is employed.

ISBN 978-0-9928626-7-1

A. Innovation Codebook

In modern speech coding, very large codebooks are needed in order to guarantee a high subjective quality. In the AMR-WB codec an algebraic innovation codebook structure with codebooks as large as 88 bits is used. The codebook structure is based on interleaved single-pulse permutation design [5] where 64 positions (corresponding to 5 ms subframe at 12.8 kHz sampling rate) are divided into 4 tracks of interleaved positions, 16 positions in each track. Different codebooks at different bit rates are constructed by placing a certain number of signed pulses in the tracks, from 1 to 6 pulses per track. The codebook index, or codeword, represents the pulse positions and signs in each track.

In order to keep the computational complexity reasonable, a fast procedure known as a depth-first tree search [6] is used. The search begins with subset #1 and proceeds with subsequent subsets according to a tree structure while only 2 pulses are determined at each tree level. Obviously, the complexity increases when more pulses are searched.

To reduce complexity while testing the possible combinations of two pulses at each tree level, a limited number of potential positions of the first pulse is tested (typically 8 or fewer positions out of 16). Further, in case of large number of pulses, some pulses in the higher levels of the search tree are fixed based on a pulse-position likelihood estimate [7].

Then the search algorithm starts with placing two first pulses at two consecutive tracks and this process is iterated four times by assigning first two pulses to different starting tracks at each iteration. However, in order to further reduce the complexity of the search algorithm the number of the iterations in the AMR-WB codec is reduced to 3 (at 18.25 and 19.85 kbps) resp. 2 (at 23.05 kbps).

All these different constraints ensure that the complexity does not explode but rather saturates at higher bit rates. This is however at a price of limited quality gain when the bit rate increases, as will be shown in the next subsection.

B. Relaxed innovation codebook

To illustrate the impact of the constraints in the innovation codebook search as described in the previous subsection, we have conducted the following experiment. We have modified the innovation codebook search algorithm in the AMR-WB codec such that at least 8 potential positons are tested for each pulse and the number of iterations is fixed to 4 at all bit rates. The results showing the difference between the original AMR-WB algorithm and the *relaxed* (less constrained) algorithm are shown in Fig. 1.

In this experiment, a 188 s database consisting of 40 sentences was used. The database contained clean and noisy speech samples and was sampled at 16 kHz. The segmental SNR (segSNR) was measured in the perceptually weighted speech domain and only in the low band. The worst case (WC) complexity is measured in terms of Weighted Million Operations Per Second (WMOPS) using the ITU-T complexity evaluation tool [8]. It can be seen from Fig. 1 that as the bit rate increases, the differences between both variants of the codec get larger, which means that more constraints are

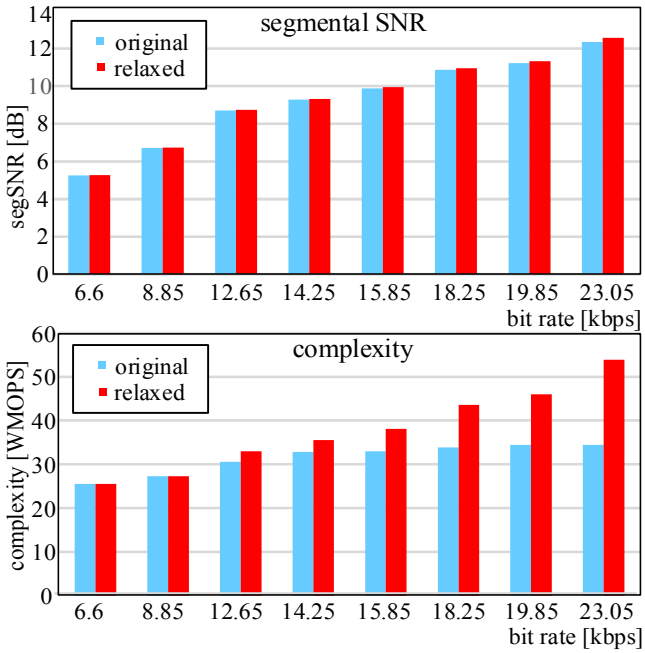

Fig. 1. Performance of the original and the relaxed innovation codebook search in the AMR-WB encoder.

used to control the complexity of AMR-WB. At the highest bit rate with the blind BWE (23.05 kbps), there is a gain of 0.168 dB in segSNR for 20 WMOPS complexity increase when the relaxed codebook search is used. If we wanted to extend the bit rates of the AMR-WB codec even higher, the traditional CELP model would become even less efficient or enormously complex.

III. Transform-Domain Codebook

In order to overcome the trade-off between the effectiveness and the complexity of the traditional CELP, we propose an efficient, flexible and scalable extended CELP model [9]. This model introduces a transform-domain codebook that is incorporated into the traditional CELP and can be seen as a pre-quantizer of the innovation codebook. The parameters of the new codebook are set at the encoder in such a way that the subsequent innovation codebook search is applied to a target signal which has – in the residual domain – less pronounced spectral dynamics than the target after the adaptive codebook only.

The principle of the proposed extended ACELP encoder is depicted in Fig. 2. In contrast to the traditional ACELP where the excitation is composed of the adaptive excitation vector $v(n)$ and the innovation excitation vector $c(n)$ only, the extended model introduces a third part of the excitation, namely the transform-domain excitation vector $q(n)$. In Fig. 2, β_1 and β_2 are the adaptive and innovation codebook gains, $x(n)$ and $x_1(n)$ are the targets for the adaptive and innovation codebook search, 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 $W(z)$.

In a given subframe, the target in the residual domain after subtracting the adaptive codebook contribution is computed as

ISBN 978-0-9928626-7-1 © EURASIP 2017

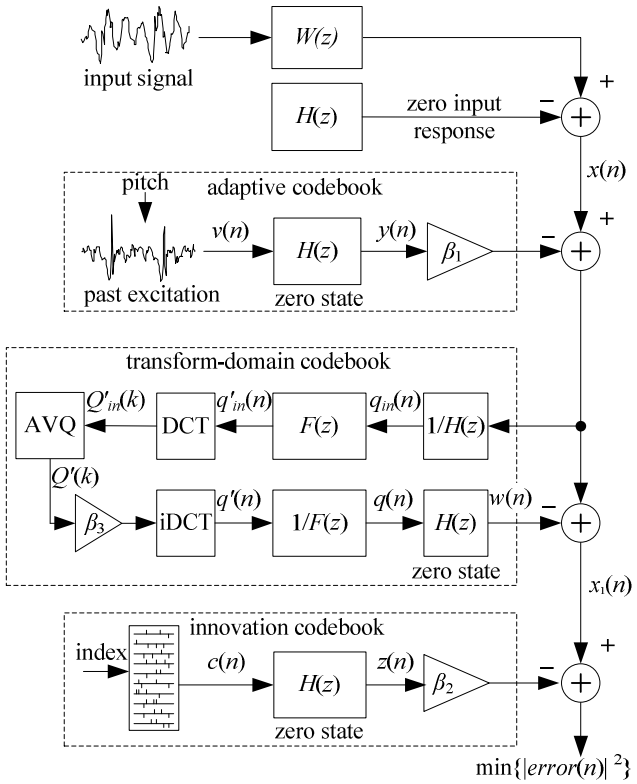

Fig. 2. Proposed transform-domain codebook in the extended ACELP encoder.

$$q_{in}(n) = r(n) - \beta_1 \cdot v(n) \tag{1}$$

where $n = 0,\ldots, N–1$ is the time-domain sample index and $N = 64$. Further, $r(n)$ is the adaptive codebook search target vector in the residual domain, i.e., signal $x(n)$ filtered through $1/H(z)$ with zero states.

Then the transform-domain codebook target vector in the residual domain $q_{in}(n)$ is pre-emphasized with a filter $F(z)$ to amplify lower frequencies of this vector:

$$q'_{in}(n) = q_{in}(n) + \alpha \cdot q'_{in}(n-1) \tag{2}$$

where the coefficient $\alpha = 0.3$ controls the level of the pre-emphasis.

Next, a Discrete Cosine Transform (DCT) is applied to the pre-emphasized vector $q'_{in}(n)$ using a rectangular non-overlapping window resulting in 8 DCT blocks, each covering one 800 Hz DCT band. These blocks of DCT coefficients $Q'_{in}(k)$ are then quantized using a vector quantizer. However, depending on the bit budget, some of the blocks might not be transmitted. Those blocks, corresponding usually to higher frequencies, are set to zero.

A. DCT Quantization

In our implementation we have chosen the Algebraic Vector Quantizer (AVQ) [10] to quantize the DCT coefficients $Q'_{in}(k)$. The AVQ encoder thus produces quantized DCT coefficients $Q'(k)$ and the AVQ indices are transmitted as transform-domain codebook parameters to the decoder.

In every subframe, the bit budget allocated to the AVQ is composed of a fixed bit budget and a floating number of bits. The fixed AVQ bit budget is directly derived from the codec bit rate. Giving the scalability of the AVQ, practically arbitrary bit budget can be allocated to the transform-domain codebook. The effectiveness of the codebook scales up with a higher bit budget as both the number of the quantized DCT blocks and consequently the SNR increase. In other words, with increasing the bit budget the level of the coding noise (shaped to follow the frequency response of the inverse of the weighting filter) decreases.

Then, depending on the used AVQ sub-quantizers [10], the AVQ usually does not consume all of the allocated bit budget, leaving a small variable number of bits available in each subframe. These bits are floating bits employed in the following subframe within the same frame. The floating number of bits is equal to 0 in the first subframe, and the floating bits resulting from the AVQ in the last subframe in a given frame remain unused or could be used by another coding module.

B. Transform-Domain Codebook Gain

The inner mechanism of the AVQ scales down in amplitude the DCT coefficients [10]. Consequently, the transform-domain codebook gain, β_3, is estimated to compensate for the AVQ scaling as follows

$$\beta_3 = \frac{\sum_{k=0}^{K-1} \left[Q'_{in}(k) \cdot Q'(k) \right]}{\sum_{k=0}^{K-1} \left[Q'(k) \cdot Q'(k) \right]} \tag{3}$$

where $k = 0, \ldots, K–1$ is the transform-domain coefficient index and $K = 64$ is the number of the DCT coefficients in the current subframe.

Subsequently, the gain is quantized. In our experimental implementation a 6-bit scalar quantizer is used whereby the quantization levels are uniformly distributed in the log domain. The index of the quantized gain is transmitted to the decoder once per subframe as another transform-domain codebook parameter.

C. Reconstruction

To obtain the target vector for the innovation codebook search, a time-domain contribution from the transform-domain codebook, $q(n)$, is reconstructed as follows. First, the quantized DCT coefficients $Q'(k)$ are scaled up using the quantized gain β_3. Next, the scaled DCT coefficients are inverse transformed using inverse DCT (iDCT). Finally, a de-emphasis filter $1/F(z)$ is applied to obtain the time-domain contribution $q(n)$. The same operations are also done in the decoder.

D. Target Vectors for Innovation Codebook Search

It was experimentally found that the transform-domain codebook contribution can be used to refine the previously computed adaptive codebook gain, β_1, to enhance the coding

ISBN 978-0-9928626-7-1 © EURASIP 2017

efficiency. Consequently, the adaptive codebook gain is recomputed as

$$\beta_1' = \frac{\sum_{n=0}^{N-1}\left[\{x(n)-w(n)\}\cdot y(n)\right]}{\sum_{n=0}^{N-1}\left[y(n)\cdot y(n)\right]} \quad (4)$$

where $w(n)$ is the filtered transform-domain codebook contribution, i.e., the zero-state response of the weighted synthesis filter $H(z)$ to the vector $q(n)$. Similarly $y(n)$ is the filtered adaptive codebook contribution. Note that in traditional ACELP the vector $x(n)$ is used instead of $\{x(n)-w(n)\}$ in (4).

Then the computation of the target vector for innovation codebook search, $x_1(n)$, is done using

$$x_1(n) = x(n) - \beta_1' \cdot y(n) - w(n). \quad (5)$$

Finally, the innovation codebook search is performed using a moderate innovation codebook size. We have experimentally found that it is advantageous to allocate more bits to the AVQ quantization of the transform-domain codebook than to the innovation codebook. On the other hand, a too small innovation codebook would not be able to capture all the remaining essential components of the residual signal. Consequently, the 36-bits innovation codebook was found as the best choice, and it was used in our experimental implementation within the AMR-WB codec, regardless of the codec bit rate. Note that the 36-bits codebook is used in AMR-WB at the 12.65-kbps bit rate. It places 8 pulses in a given subframe and it has a computational complexity of about 4 WMOPS below the worst case complexity of the complete encoder (see Fig. 1). This difference in complexity can be thus spent to perform the transform-domain codebook search without affecting the encoder's WC complexity.

IV. Performance

As mentioned previously, we have implemented the transform-domain codebook as described in Section III in the AMR-WB codec. We have chosen several bit rates starting from 30 kbps that extend the AMR-WB bit rate coverage and tested if the performance scales-up proportionally with increased bit rate. Our goal was also to optimize the model such that the complexity at the extended bit rates does not increase the WC complexity of the AMR-WB encoder. In our experimental implementation the traditional AMR-WB internal sampling rate of 12.8 kHz and the blind BWE were used.

Using the same database as described in subsection II.B, we have computed complexity, segSNR and Perceptual Evaluation of Audio Quality (PEAQ) scores. The PEAQ [11] is an objective metric that automatically assesses the audio quality degradation in terms of the Objective Difference Grade (ODG) and the Distortion Index (DI). The results are shown in Fig. 3. They also contain the original higher AMR-WB bit rates to better observe the trends.

It can be seen from the results that the performance of the AMR-WB codec scales effectively at the extended bit rates, both in terms of segSNR and PEAQ. We have also verified that the encoder's complexity at the extended bit rates indeed does not exceed the WC complexity of the original AMR-WB encoder.

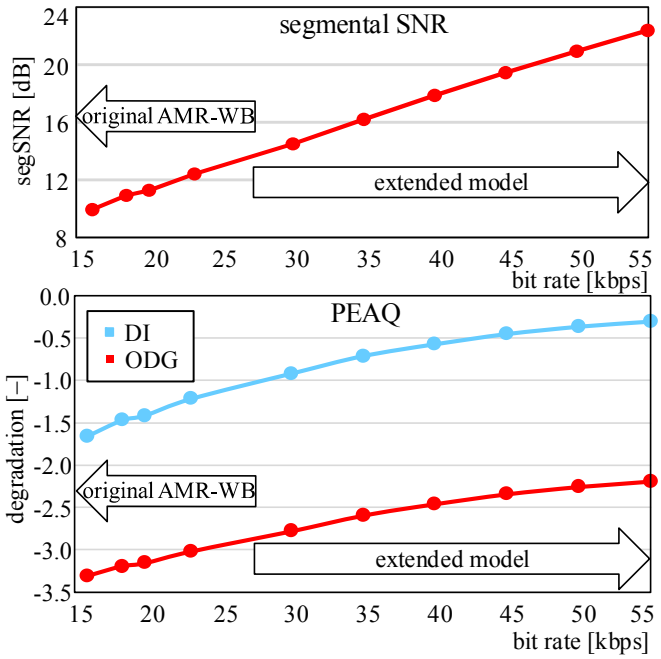

Fig. 3. Performance of the transform-domain codebook implemented in the AMR-WB codec.

V. Other Considerations

In Section IV we have demonstrated how the new transform-domain codebook enhances the performance of a CELP codec by increasing the segSNR and PEAQ scores without affecting the WC complexity. While the codec performance scales effectively at the extended bit rates, it is not possible to reach transparent quality without introducing further tunings into the codec. An example would be a need of an improved coding of the high band either by increasing the internal sampling rate or by introducing a high-quality BWE.

Another example is the gain quantization. In AMR-WB, a 7-bit vector quantizer (VQ) is used to quantize the adaptive and innovation codebook gains β_1 and β_2. While this VQ works sufficiently well at AMR-WB bit rates, its performance is not good enough at the extended bit rates. It would thus be advantageous to increase the bit budget of the gain quantizer.

One needs to be also very careful about enhancers and post-processing techniques used in a codec, e.g., the enhancement of the excitation signal used in AMR-WB [7]. At the extended bit rates these techniques are usually not needed and could actually have a negative effect on the synthesised speech quality. In general, it thus seems beneficial to disable them.

A. Variable Bit Rate Audio Coding

In Section IV the performance of the proposed model has been assessed for the case of constant bit rate coding. This is typical in low delay applications such as telephony where the total bit budget is usually fixed. In higher delay applications like audio streaming, a bit reservoir is often used

ISBN 978-0-9928626-7-1 © EURASIP 2017

in conjunction with a variable bit rate encoder. In those applications, the proposed technology can be further tuned to profit from the variable bit allocation. Consequently, the audio quality can be perceptually further improved as the bit allocation can be source controlled.

It is known that coding of lower frequencies is more critical for the overall perceptual quality, and thus a higher SNR is usually targeted in lower frequencies than in higher frequencies. In the CELP model, this higher SNR is mainly achieved by the LTP. However, the LTP is not efficient in every case, such as in voice onsets, transients or in the context of background noise or speech over music. These scenarios require a much better control of quality which cannot be simply achieved by either the LTP or the innovation codebook. To guarantee a uniform quality in case of a variable bit rate coding, we have experimented with different tunings of the transform-domain codebook from those described in Section III.

To apply the technology in the context of a variable bit rate coding, the coefficient α of the pre-emphasis filter $F(z)$ could be increased in order to emphasize even more the low frequencies. E.g., $\alpha = 0.9$ can be chosen to obtain an emphasis around 20 dB at 20 Hz, 11 dB at 800 Hz and 3 dB at 2400 Hz. Relative to the coded bandwidth (6400 Hz), this area contains 70% of the auditory critical bands [12]. This configuration has thus the benefit to control the SNR in the area which is perceptually most important and contributes in the frequency range which is more likely to be profitable by the LTP. For example, many vowels are voiced (predictable) in low frequencies and noisy (unpredictable) in higher frequencies. Using this configuration, the DCT spectrum is quantized only in the concerned area ($\leq$ 2400 Hz) meaning that only 3 DCT blocks are quantized in every subframe. In low frequencies, the transform-domain codebook gain then controls the SNR, and its bit consumption depends on the targeted SNR and the LTP efficiency. The weaker the LTP contribution is, the stronger and richer is the transform-domain codebook contribution and vice versa.

If the transform-domain codebook is applied only in low frequencies, a larger algebraic innovative codebook is preferable to get a sufficient quality in higher frequencies. The 64-bits innovation codebook (with 16 pulses per subframe) has been found perceptually optimal. Based on the fact that the CELP model was theoretically imagined to produce a spectrally flat excitation, a high emphasis in the filter $F(z)$ may be seen inadequate. However in practice, the ACELP, with its ability to permute pulses, is very efficient to produce an emphasized excitation. In fact, also the LTP conditions the target in the same way as the transform-domain codebook does when the prediction is limited to the low frequencies.

In order to assess the quality of the tuned transform-domain codebook as described in this subsection, we have implemented it into the MPEG USAC codec [13]. USAC is a hybrid audio codec, which consists of a time-domain coding mode and a frequency-domain coding mode. The time-domain coding mode is in general used to encode speech segments (with or without music), transients and attacks. It has a similar structure as the AMR-WB codec and was extended by the proposed technology. A bit reservoir was used so that the model was fully flexible and source controlled in order to obtain the desired SNR in low frequencies. We have tested a database consisting of speech, speech over music and music items where the segSNR was measured only in the frames operating in the time-domain coding mode. Objectively the overall performance of the tuned transform-domain codebook was similar to what is presented in Fig. 3. However, an informal subjective assessment showed a clear advantage in controlling the quality in the context of source controlled variable bit rate coding of a general audio content.

VI. Conclusion

A new transform-domain codebook implemented in the CELP model was introduced. The new codebook is efficient, flexible and scalable and can easily extend any existing CELP codec to provide coding at high bit rates. We demonstrated by the example of the AMR-WB codec that the proposed model gains in quality at the extended bit rates while not increasing the computational complexity of the codec.

References

[1] M. R. Schroeder and B. S. Atal, "Code-excited linear prediction (CELP): High-quality speech at very low bit rates," in Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP), Tampa, FL, 1985, pp. 937–940, vol. 10.

[2] J. P. Adoul, P. Mabilleau, M. Delprat, and S. Morissette, "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] R. Salami, C. Laflamme, J. -P. Adoul, and D. Massaloux, "A toll quality 8 kb/s speech codec for the personal communications system (PCS)," IEEE Trans. on Vehicular Technology, vol.43, no.3, pp. 808-816, August 1994.

[4] E.-D. Lee, M. S. Lee, an D. Y. Kim, "Global pulse replacement method for fixed codebook search of ACELP speech codec," in Proc. IASTED CIIT 2003, Scottsdale, AZ, 2003, pp. 372-375.

[5] B. Bessette, R. Salami, R. Lefebvre, M. Jelinek, J. Rotola-Pukkila, J. Vainio, H. Mikkola, K. Jarvinen, "The Adaptive Multirate Wideband Speech Codec (AMR-WB)," IEEE Trans. Speech Audio Process., vol. 10, no. 8, pp. 620–636, November 2002.

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

[7] 3GPP TS 26.190: Speech codec speech processing functions; Adaptive Multi-Rate – Wideband (AMR-WB) speech codec; Transcoding functions.

[8] Recommendation ITU-T G.191: Software Tools for Speech and Audio Coding Standardization (STL), 2010.

[9] B. Bessete, "Flexible and scalable combined innovation codebook for use in CELP coder and decoder," US Patent 9,053,705, June 2015.

[10] M. Xie and J.-P. Adoul, "Embedded algebraic vector quantization (EAVQ) with application to wideband audio coding," In Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP), Atlanta, GA, 1996, pp. 240-243, vol 1.

[11] Recommendation ITU-R BS.1387: Method for objective measurements of perceived audio quality (PEAQ), 2001.

[12] J. D. Johnston, "Transform Coding of Audio Signals Using Perceptual Noise Criteria," IEEE Journal on Selected Areas in Communications, vol. 6, no. 2, pp. 314-323, February 1988.

[13] M. Neuendorf et al., "MPEG Unified Speech and Audio Coding - The ISO/MPEG Standard for High-Efficiency Audio Coding of All Content Types," in Proc. 132nd AES Convention, Budapest, Hungary, 2012, pp. 248–269.

ISBN 978-0-9928626-7-1 © EURASIP 2017