TRANSITION MODE CODING FOR SOURCE CONTROLLED CELP CODECS

Václav Eksler[1,2]*, Milan Jelínek*[1,2]

[1] University of Sherbrooke, QC, Canada, [2] VoiceAge Corp., Montreal, QC, Canada
{Vaclav.Eksler, Milan.Jelinek}@USherbrooke.ca

ABSTRACT

CELP-based codecs typically rely on prediction to achieve their high coding efficiency. On the other hand, the prediction makes these codecs sensitive to frame erasures as errors propagate beyond the erased frame. We present a technique that significantly limits the error propagation by replacing inter-frame long-term prediction with a non-predictive glottal-shape codebook. The technique was implemented in the winning candidate of the EV-VBR baseline codec selection by ITU-T in March 2007. To maintain the performance in clean channel, this transition mode coding technique was used only in frames following voiced onsets frames, i.e. the frames most sensitive to frame errors.

Index Terms— Speech Coding, Speech Transitions, CELP, Frame Erasure

1. INTRODUCTION

In the fast growing market of wireless mobile communication systems and voice over packet networks, the encoded signal may be subjected to high rates of frame erasures or packet loss. To maintain good service quality, speech coding systems with high robustness against packet loss are needed.

Most recent low bit-rate speech coding standards are based on Code-Excited Linear Prediction (CELP). The CELP-based codecs rely on prediction to achieve their high coding efficiency. The most important prediction is the long-term (or pitch) prediction (LTP) usually implemented as an adaptive codebook containing excitation signal selected in past frames. Since the content of the adaptive codebook is based on the signal from past frames, this makes the coding model sensitive to frame loss.

When a frame is lost, a concealment procedure is used instead of the standard decoding. For a CELP codec it means that the content of the adaptive codebook at the decoder becomes different from its content at the encoder, and the error propagates in the frames following the erasure. The impact of a lost frame depends on the nature of the speech segment in which the erasure occurred. High quality concealment is generally achieved if the erasure occurs during a quasi-stationary speech segment [1]. On the other hand, if the frame erasure coincides with a transition, it is better to rapidly attenuate the signal as the estimated synthesis would likely be very inaccurate.

Transition from unvoiced speech segment to voiced segment (voiced onset) is the most problematic case. When a voiced onset frame is lost, the frame before the onset is typically unvoiced and no meaningful periodic excitation is stored in the buffer of the past excitation of a CELP codec. At the encoder, the past periodic excitation builds up in the adaptive codebook during the voiced-onset frame, and the following voiced frame takes advantage of this past periodic excitation. When the voiced-onset frame is lost, this periodic part of the excitation is completely missing in the adaptive codebook at the decoder.

To limit the error propagation, e.g. frame independent coding has been proposed. However, it requires significant increase in bit-rate compared to a CELP-type codec to maintain the synthesized speech quality [2]. Another technique to limit error propagation in CELP codecs is glottal pulse resynchronization [3].

The goal of the presented technique was to attenuate strong artefacts produced by a CELP decoder due to error propagation when transition frames are lost. At the same time the clean channel performance had to be maintained without increasing the bit-rate and the computational complexity. This has been achieved by replacing the inter-frame LTP with a codebook of glottal impulse shapes (glottal-shape codebook), and by using this approach in a multi-mode source-controlled model only on most sensitive frames – the frames following voiced onsets. The coding mode using this technique will be called transition mode (TM). The proposed technique was implemented in the EV-VBR winning candidate codec selected in March 2007 by ITU-T as the baseline algorithm for Q9/16 standardization [4]. Its performance was validated in formal MOS tests.

2. DESCRIPTION OF TRANSITION MODE

The glottal-shape codebook as a part of the EV-VBR codec core layer [5] is outlined in Fig. 1. The EV-VBR core layer operates at 8 kb/s with 20 ms frames and internal sampling rate of 12.8 kHz. As in a legacy CELP (e.g. [6]), the input speech signal $s(n)$ is modeled by an excitation signal filtered through the Linear Prediction (LP) synthesis filter $1/A(z)$, and the excitation is selected in two stages in a perceptually weighted domain. The weighted synthesis filter is given by $H(z) = 1/A(z)\cdot W(z)$, where $W(z)$ is the perceptual weighting filter. Typically, the first stage excitation signal is selected from the adaptive codebook and the second stage excitation signal $c_k(n)$ is selected from a fixed codebook. In our approach, the adaptive codebook is however substituted with a glottal-shape codebook for encoding of the first glottal impulse in a frame to reduce the inter-frame prediction.

The glottal-shape codebook consists of quantized normalized shapes of the truncated glottal impulses placed at a specific position. The codebook search consists both in the selection of the best shape and the best position.

To select the best codevector, the mean-squared error between the target signal $x_1(n)$ and the first stage contribution signal $y_1(n)$ is minimized for all candidate glottal-shape codevectors. The glottal-shape codebook search has been designed in a similar way as the fixed codebook search in the Algebraic CELP [6]. In this approach, each impulse shape is represented as an impulse response of a shaping filter $G(z)$. This impulse response can be

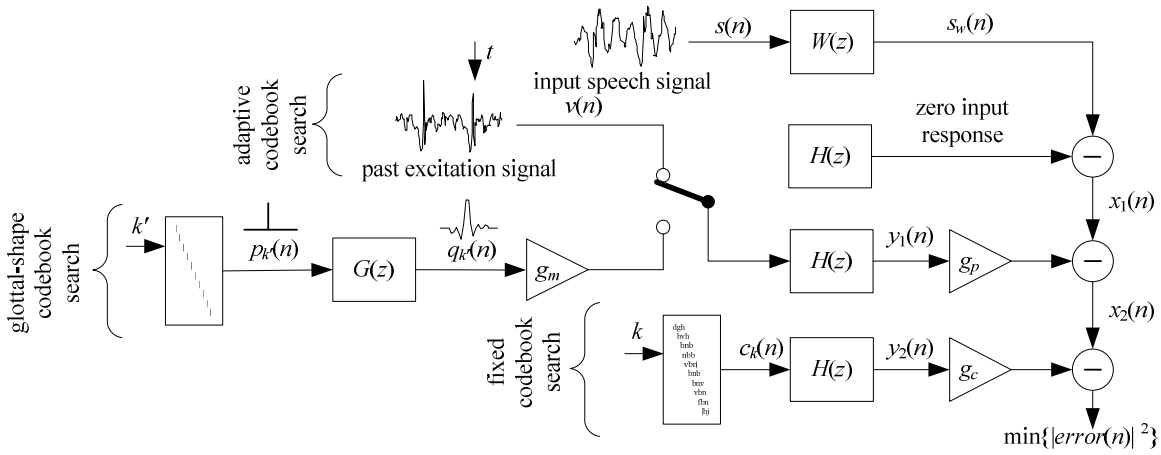

Fig. 1. Principle of the proposed technique as a part of a CELP encoder.

integrated in the impulse response of the weighted synthesis filter $H(z)$ prior to the search of the optimum impulse position. The searched codevectors can be then represented by vectors containing only one non-zero element corresponding to candidate impulse positions, and they can be searched very efficiently. Once selected, the position codevector is convolved with the impulse response of the shaping filter. This procedure needs to be repeated for all the candidate shapes and the best shape-position combination will form the first stage excitation signal.

In the following all vectors are supposed column vectors. Let $\mathbf{p}_{k'}$ is a position codevector with one non-zero element at position k', and $\mathbf{q}_{k'}$ is the corresponding glottal-shape codevector with index k' representing the middle of the glottal shape. Index k' is chosen from the range $[0, N-1]$, N being the subframe length. Note that due to the non-causal nature of the shaping filter, its impulse response is truncated for positions in the beginning and at the end of the subframe. The glottal-shape codevector $\mathbf{q}_{k'}$ can be expressed in a matrix form as $\mathbf{q}_{k'} = \mathbf{G}\cdot\mathbf{p}_{k'}$, where $\mathbf{G}$ is a Toeplitz matrix representing the glottal impulse shape. Similar to the fixed codebook search we can write

$$\Im_{k'} = \frac{\left(\mathbf{x}_1^{\mathrm{T}}\mathbf{y}_1\right)^2}{\mathbf{y}_1^{\mathrm{T}}\mathbf{y}_1} = \frac{\left(\mathbf{x}_1^{\mathrm{T}}\mathbf{H}\mathbf{q}_{k'}\right)^2}{\mathbf{q}_{k'}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{H}\mathbf{q}_{k'}} = \frac{\left(\mathbf{x}_1^{\mathrm{T}}\mathbf{H}\mathbf{G}\mathbf{p}_{k'}\right)^2}{\mathbf{p}_{k'}^{\mathrm{T}}\mathbf{G}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{H}\mathbf{G}\mathbf{p}_{k'}} = \frac{\left(\mathbf{x}_1^{\mathrm{T}}\mathbf{Z}\mathbf{p}_{k'}\right)^2}{\mathbf{p}_{k'}^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\mathbf{p}_{k'}} = \frac{\left(\mathbf{d}_g^{\mathrm{T}}\mathbf{p}_{k'}\right)^2}{\mathbf{p}_{k'}^{\mathrm{T}}\mathbf{\Phi}_g\mathbf{p}_{k'}}, \tag{1}$$

where $\mathbf{H}$ is the lower triangular Toeplitz convolution matrix of the weighted synthesis filter. The rows of the convolution matrix $\mathbf{Z}^{\mathrm{T}}$ correspond to the filtered shifted version of the glottal impulse shape or its truncated representation as shown in Fig. 2.

Because of the fact that the position codevector $\mathbf{p}_{k'}$ has only one non-zero sample, the computation of the criterion (1) is very simple and can be expressed as

$$\Im_{k'} = \frac{\left(d_g(k')\right)^2}{\Phi_g(k',k')}. \tag{2}$$

As it can be seen from (2), only the diagonal of the correlation matrix $\mathbf{\Phi}_g$ from (1) needs to be computed.

The codebook has been designed using the *k-means* algorithm [7] and a database of more than three hours of a LP residual of speech signal. As a result we have chosen 8 prototype glottal impulse shapes of length $L = 17$ samples, shown in Fig. 3. The number and the length of these prototype impulses are a compromise of a good codec performance on one side, and bit budget constraints and computational requirements on the other side. Note that as L is shorter than the subframe length N ($N = 64$ samples in EV-VBR), the remaining samples in the subframe are set to zero.

In general, the coding efficiency of the glottal-shape codebook is lower than the efficiency of the LTP, and more bits are generaly needed to assure good synthesized speech quality. However, the glottal-shape codebook does not need to be used in all subframes. First, there is no reason to use this codebook in subframes that do not contain any significant glottal impulse in the LP residual signal. Second, the glottal-shape codebook search is important only in the first pitch-period in a frame. The following pitch periods can be encoded using the more efficient standard adaptive codebook search as it does not use the excitation of the past frame anymore. To satisfy the constant bit-rate requirement, we have therefore chosen to use the glottal-shape codebook search only in one of the four subframes in a frame. This has led to a highly structured coding mode where the bit allocation is dependent on the position of the first glottal impulse and the pitch period. The subframe where the glottal-shape codebook is used is chosen as the subframe with maximum sample in the LP residual signal in the range $[0, T_{op}+2]$, where T_{op} is the open-loop pitch period estimated over the first half of the frame.

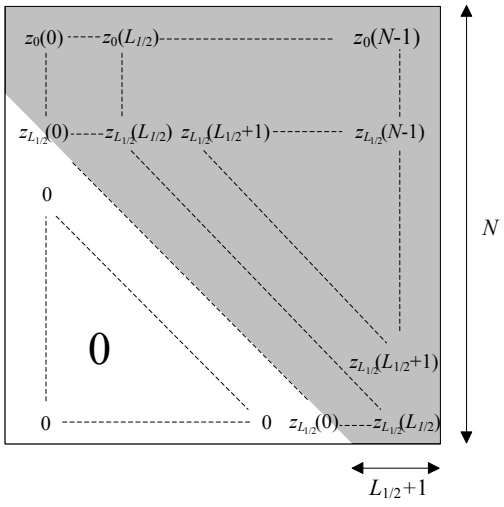

Fig. 2. Structure of the convolution matrix $\mathbf{Z}^{\mathrm{T}}$.

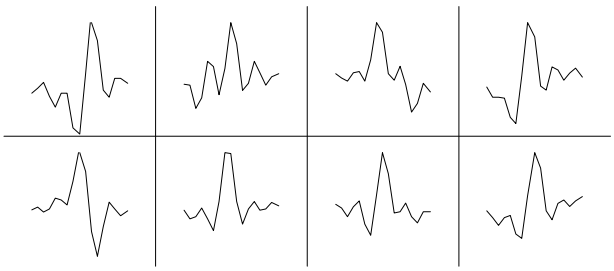

Fig. 3. Eight prototype glottal impulse shapes.

The other subframes in TM frame (frame coded using the TM technique) are processed as follows. If a subframe does not contain any significant glottal impulse in the LP residual signal, the first stage excitation signal in this subframe is zero, i.e. only the fixed codebook contribution is computed in this subframe. If a subframe following the glottal-shape codebook subframe contains a significant glottal impulse, the standard CELP adaptive and fixed codebook search is used. The standard CELP coding is also used in all subframes following the glottal-shape codebook subframe if this is the 2nd or 3rd subframe in the TM frame.

The error propagation would be greatly limited if TM technique were used in all active speech frames. However the synthesized speech quality as well as the coding efficiency in error-free conditions would drop also significantly. As a compromise, the TM technique can be used only in frames following transitions. In our implementation in EV-VBR framework, we have used the TM only in frames following voiced onsets. To detect the onsets, a transition detector based on the classification of VMR-WB [1] was used. This way about 6% of active speech frames are selected for TM encoding.

An illustration of the TM technique is shown in Fig. 4 where the first important glottal impulse appears in the 2nd subframe. The first stage excitation signal is thus zero in the 1st subframe, it is built using a glottal-shape codevector in the 2nd subframe, and the adaptive codebook is used in the last two subframes.

3. COMPUTATIONAL ASPECTS

Criterion (2) is typically used in the ACELP algebraic codebook search by precomputing the backward filtered target vector $\mathbf{d}_g$ and the correlation matrix $\mathbf{\Phi}_g$. Given the non-causal nature of the filter $G(z)$, the matrix $\mathbf{Z}$ is not triangular and Toeplitz anymore, and this approach cannot be efficiently applied for the first $L_{1/2}$ positions in the glottal-shape codebook search.

Let us denote $\mathbf{z}_{k'}$ to be the $(k'+1)^{\text{th}}$ row of the matrix $\mathbf{Z}^{\mathrm{T}}$, where $\mathbf{Z}^{\mathrm{T}} = \mathbf{G}^{\mathrm{T}} \cdot \mathbf{H}^{\mathrm{T}}$ is computed in two steps to minimize the computational complexity. In the first step we compute the first $L_{1/2} + 1$ rows of this matrix $\mathbf{Z}^{\mathrm{T}}$ that correspond to the positions k' from the range $[0, L_{1/2}]$. In the second step, the criterion (2) is used in similar way as in the ACELP fixed codebook search for the remaining part of $\mathbf{Z}^{\mathrm{T}}$ (the last $N - L_{1/2} - 1$ rows of the matrix $\mathbf{Z}^{\mathrm{T}}$).

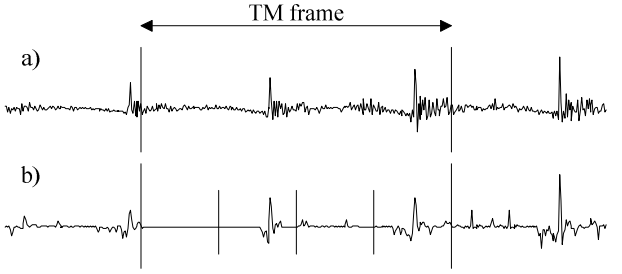

Fig. 4. Comparison of (a) LP residual signal, (b) first stage excitation signal using TM technique.

In the first step the convolution between the glottal-shape codebook entry for position $k' = 0$ and the impulse response $h(n)$ is first computed using

$$z_0(n) = \sum_{i=0}^{n} g(n-i)h(i) \quad \text{for} \quad n = 0,...,N-1, \tag{3}$$

where we take advantage of the fact that the filter $G(z)$ has only $L_{1/2} + 1$ non-zero coefficients.

Next the convolution $z_1(n)$ between the glottal-shape codebook entry for position $k' = 1$ and the impulse response $h(n)$ is computed reusing values of $z_0(n)$. For the following rows, the recursion is reused again resulting in

$$\begin{aligned} z_{k'}(0) &= g(-k')h(0), \\ z_{k'}(n) &= z_{k'-1}(n-1) + g(-k')h(n) \quad \text{for} \quad n = 1,\ldots,N-1. \end{aligned} \tag{4}$$

The recursion (4) is repeated for all $k' \le L_{1/2}$.

Now the criterion (2) can be computed for all positions k' from the range $[0, L_{1/2}]$ in the form

$$\Im_{k'} = \frac{\left(\sum_{i=0}^{N-1} z_{k'}(i) \cdot x_1(i) \right)^2}{\sum_{i=0}^{N-1} z_{k'}(i) \cdot z_{k'}(i)}. \tag{5}$$

In the second step we take advantage of the fact that rows $L_{1/2} + 1, \ldots, N - 1$ of the matrix $\mathbf{Z}^{\mathrm{T}}$ are built using coefficients of the convolution $z_{L_{1/2}}(n)$ that are already computed as described by (4) for $k' = L_{1/2}$ (see also Fig. 2). That is, each row corresponds to the previous row shifted to the right by 1 with a zero added at the beginning

$$\begin{aligned} z_{k'}(0) &= 0, \\ z_{k'}(n) &= z_{k'-1}(n-1) \quad \text{for} \; n = 1,\ldots,N-1 \end{aligned} \tag{6}$$

and this is repeated for k' from the range $[L_{1/2} + 1, N - 1]$.

Next the target vector $\mathbf{d}_g$ and the diagonal of the matrix $\mathbf{\Phi}_g$ need to be computed. First we evaluate the numerator and the denominator of the criterion for the last position $k' = N - 1$

$$d_g(N-1) = \sum_{i=0}^{L_{1/2}} x(N-1-L_{1/2}+i) z_{L_{1/2}}(i) \tag{7}$$

and

$$\Phi_g(N-1, N-1) = \sum_{i=0}^{L_{1/2}} z_{L_{1/2}}(i) z_{L_{1/2}}(i). \tag{8}$$

For the remaining positions the numerator is computed using equation (7), only with the summation index changed. In the denominator computation some of the previously computed values can be reused. I.e. for the position $k' = N - 2$ the denominator of the criterion (2) is computed using

$$\begin{aligned} \Phi_g(N-2, N-2) = \Phi_g(N-1, N-1) + \\ + z_{L_{1/2}}(L_{1/2}+1) z_{L_{1/2}}(L_{1/2}+1). \end{aligned} \tag{9}$$

Similarly, we can continue to compute the numerator and the denominator of (2) for all positions $k' > L_{1/2}$.

The search continues using the previously described procedure for all other glottal impulse shapes and the codevector corresponding to the best combination of glottal-shape and position is selected. To maintain the complexity low the computation described above is further reduced by limiting the position search to ± 4 samples around the maximum absolute value of the LP residual signal.

The last parameter to be determined is the gain of the glottal-shape codebook excitation. The gain is quantized in two steps. First a roughly quantized gain g_m of the glottal-shape codevector is found. Then, after both first stage and second stage contributions of the excitation signal are found, the gain g_p of the first stage contribution signal is further jointly quantized with the second stage contribution gain g_c. This is done using EV-VBR memoryless gain vector quantization [5].

It should be noted that the close-loop pitch period value T_0 does not need to be transmitted anymore in a glottal-shape codebook search subframe except if the subframe contains more than one glottal impulse, i.e. when $(k' + T_0) < N + L_{1/2}$.

A TM bit-allocation for the example of Fig. 4 is summarized in Table I. For comparison a bit allocation for a generic EV-VBR frame without TM technique is also shown. In Table I, coder type refers to the signal classification based encoding type, ISF to Immitance Spectral Frequencies representation of the LP filter coefficients [8], and gains represent joint g_p and g_c gains. TM parameters represent position and shape of the glottal impulse, signed gain g_m, and an identification of the subframe where the glottal-shape codebook search is used. Note that Table I presents only one of seven possible configurations of the TM coding used in core layer of the EV-VBR framework.

Parameter	Generic frame	TM frame
Coder type	3	3
ISF	36	36
Pitch delay	26	13
Algebraic code	72	72
Gains	23	20
TM parameters	–	16
Total per frame	160	160

Table 1. An example bit-allocation for Generic frame and TM frame in core layer of the EV-VBR codec.

4. PERFORMANCE

Figure 5 shows a comparison of the synthesized speech in presence of a frame-error (FE). When frame with voiced onset is missing and the TM coding technique is not used, the output signal is very different from the error-free synthesis for several frames (Fig. 5 b)). When the TM technique is used in the frame following the onset frame, the output signal is much more similar (Fig. 5 c)).

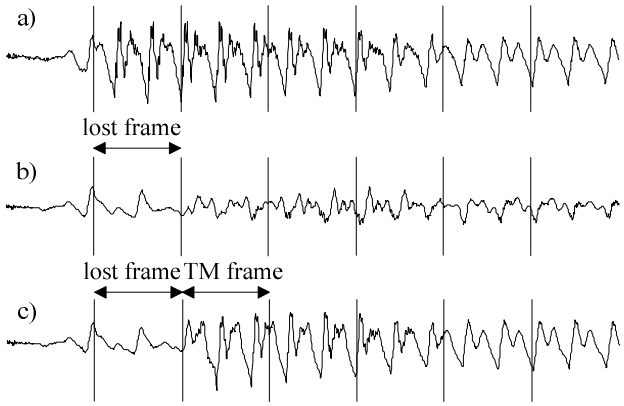

Fig. 5. Comparison of three synthesized speech signals: (a) error-free signal, (b) synthesized signal without TM and second frame lost, (c) synthesized signal with TM and second frame lost.

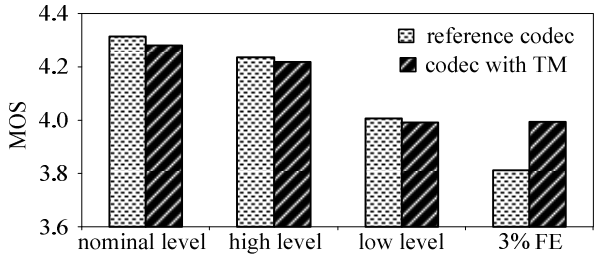

Fig. 6. Example of the TM technique performance.

The performance of the TM technique as a part of the EV-VBR framework has been assessed with formal Mean Opinion Scores (MOS) tests using 32 listeners. Some of the test results are shown in Fig. 6 for the core layer of EV-VBR (operating at 8 kb/s). In the figure the reference signal was obtained by EV-VBR without using the transition mode. It can be seen that while the clean channel performance remains practically the same, a significant improvement was obtained in the FE condition. Note that the relative performance is generally better for higher FE rates.

5. CONCLUSION

We presented a technique to reduce frame error propagation in CELP-based speech codecs. This coding technique eliminates inter-frame long-term prediction by using a non-predictive glottal-shape codebook to encode the first glottal impulse in a frame. It has been implemented in the ITU-T EV-VBR baseline codec core layer to encode frames following voiced onsets. Formal MOS tests shown similar performance in clean channel conditions while the performance was significantly improved in FE conditions. The algorithm requires no extra delay, negligible additional complexity, and no increase in bit-rate compared to the EV-VBR generic encoding.

6. REFERENCES

[1] M. Jelínek 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.

[2] S.V. Andersen *et al.*, "ILBC - A Linear Predictive Coder with Robustness to Packet Losses," in Proc. *IEEE Speech Coding Workshop*, Tsukuba, JAPAN, pp. 23-25, October 2002.

[3] T. Vaillancourt *et al.*, "Efficient Frame Erasure Concealment in Predictive Speech Codecs using Glottal Pulse Resynchronisation," in Proc. *IEEE Int. Conference on Acoustic, Speech and Signal Processing (ICASSP)*, Honolulu, Hawaii, USA, vol. 4, pp. 1113-1116, April 2007.

[4]"*Global Analyses for the Selection Phase for the Embedded-VBR Speech Codec*", ITU-T Q7/SG12 Cont. AH-07-03R1, March 2007.

[5] "*Extended high-level description of the Q9 EV-VBR baseline codec*," ITU-T SG16 Tech. Cont. COM16–C199R1–E, June 2007.

[6] R. Salami *et al.*, "A toll quality 8 kb/s speech codec for the personal communications system (PCS)," In *IEEE Transactions on Vehicular Technology*, Vol. 43, No. 3, pp. 808-816, August 1994.

[7] C. W. Chui, "*Speech Coding Algorithms. Foundation and Evolution of Standardized Coders*", John Wiley & Sons, 2003.

[8] Y. Bistritz and S. Pellerm, "Immittance Spectral Pairs (ISP) for speech encoding," in *Proc. IEEE Int. Conference on Acoustic, Speech and Signal Processing (ICASSP)*, Minneapolis, MN, USA, vol. 2, pp. 9-12, April 1993.