EFFICIENT FRAME ERASURE CONCEALMENT IN PREDICTIVE SPEECH CODECS USING GLOTTAL PULSE RESYNCHRONISATION

Tommy Vaillancourt[1,2], Milan Jelinek[1,2], Redwan Salami[1], and Roch Lefebvre[2]

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

ABSTRACT

Error propagation after a frame loss is an important factor in quality degradation for predictive speech coders. This is mainly due to the lack of synchronization in the adaptive codebook in the good frames following a frame erasure. This article presents a method for resynchronizing the glottal pulse after an erased frame. The method uses an extra frame delay at the decoder, and can be applied with or without additional side information. The approach has been implemented in the G.729.1 standard and results in improved decoder convergence after erased frames. Subjective tests have demonstrated that this improves perceived quality in the presence of frame erasures.

Index Terms— Speech coding, CELP, frame erasure concealment, packet loss concealment

1. INTRODUCTION

Improving the robustness of low bit rate speech codecs in case of frame erasures is of significant importance. The main applications of low bit rate speech encoding are in wireless mobile communication systems and voice over packet networks, where the encoded signal may be subjected to high rates of frame erasures (or packet loss). This necessitates the use of efficient frame erasure concealment in order to maintain good service quality.

Most recent low bit rate speech coding standards are based on Code-Excited Linear Prediction (CELP), where the pitch predictor, or the adaptive codebook, plays an important role in maintaining high speech quality at low bit rates. However, since the content of the adaptive codebook is based on the signal from past frames, this makes the coding model sensitive to frame loss. In case of erased or lost frames, the content of the adaptive codebook at the decoder becomes different from its content at the encoder. Thus, after a lost frame is concealed and subsequent good frames are received, the synthesized signal in the received good frames is different from the intended synthesis signal since the adaptive codebook contribution has been changed. The impact of a lost frame depends on the nature of the speech segment in which the erasure occurred. If the erasure occurs in a speech onset or a transition, the effect of the erasure can propagate through several frames. For instance, if the beginning of a voiced segment is lost, then the first pitch period will be missing from the adaptive codebook content. This will have a severe effect on the pitch predictor in consequent good frames, resulting in longer time before the synthesis signal converges to the one intended at the encoder.

Frame independent coding has been proposed to limit the error propagation in future frames after concealed frames. However, it requires significant increase in bit rate compared to a CELP-type codec to maintain the synthesized speech quality [1]. Another approach was proposed where the adaptive codebook gain in a CELP encoder is constrained to reduce the error propagation at the decoder [2], and where the remaining periodicity in the innovative excitation resulting from constraining the encoder is used to resynchronize the adaptive codebook after frame erasures. An efficient approach has been introduced in the VMR-WB codec standard [3,4], where side information at little bit rate overhead is transmitted to improve the frame erasure concealment and reduce the error propagation in future frames. The parameters used in VMR-WB consist of frame classification, energy, and phase information. In this article, this approach is further improved by using the phase information to synchronize the content of the adaptive codebook (the pitch pulse) with its content at the encoder. This significantly limits the error propagation in future frames. Further, the phase information is also used for the reconstruction of lost voiced onsets. This approach has been applied in Recommendation G.729.1 standardized by ITU-T [5].

The article is organized as follows. In section 2, the issue of adaptive codebook mismatch is discussed and the resynchronization of the adaptive codebook content is described. A method for artificial onset reconstruction is also explained. In Section 3, the concealment in G.729.1 is briefly described. Section 4 discusses the performance of the proposed approach and finally Section 5 gives the conclusions.

2. GLOTTAL PULSE RESYNCHRONIZATION IN FRAME ERASURE CONCEALMENT

In the concealment of erased voiced frames, the pitch from the past frame is repeated. This may result in a drift in the glottal pulse position, since the pitch period used to build the excitation can be different from the encoder pitch period. This will cause the adaptive codebook (or past CELP excitation) to be desynchronized from the actual CELP excitation. Thus, in case a good frame is received after an erased frame, the pitch excitation (or adaptive codebook excitation) will have an error which may persist for several frames. This will affect the performance of the correctly received frames since the fixed codebook excitation is no more optimized with the adaptive codebook content. Figure 1 shows an example that illustrates this issue. Figure 1-a shows the original synthesis, Figure 1-b shows the excitation signal without frame loss and Figure 1-c shows the excitation when the first frame is lost and concealed. Figure 1-d shows the difference between the signals in Figures 1-c and 1-b. The pitch desynchronization problem is evident from Figure 1-d (the excitation difference). Further, the signal-to-noise ratio between synthesis with and

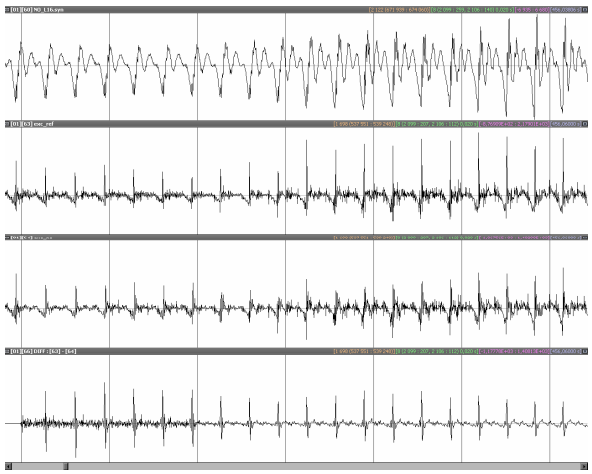

Figure 1:
a) Original synthesis.
b) Excitation without frame erasure.
c) Excitation with the first frame erased and concealed.
d) Difference between (c) and (b).

without frame erasure is very low even after good received frames (it goes down from 16 dB to 2 dB for the segment in Figure 1).

In this article, we describe the glottal pulse resynchronization to address this issue in the context of the ITU-T G.729.1 Recommendation, whereby one frame delay is available at the decoder. Thus, when a single frame is erased, the parameters of the future frame are available and can be used in the concealment of the erased frame.

The resynchronization will be addressed for three different cases: the case where extra information related to position and sign of the glottal pulse in the previous frame is transmitted to further improve the synchronization, the case where no extra information is available, but the future frame is available (i.e. only G.729.1 lower layers are received – see Section 3), and the case where neither extra information nor future frame is available (the case of multiple frame erasures).

The resynchronization consists of forcing the position of the last glottal pulse in the concealed frame to be aligned with the actual pulse. In the first case, the actual pulse position is estimated based on the pitch period in the future and past frames. In the second case the position and sign of the maximum pulse at the end of the erased frame are available from the future frame.

2.1. Determination of phase information

In the case where the glottal pulse information is transmitted, the search of the maximum pulse is performed on a low-pass filtered LP residual, which is given by

$$r^{lpf}(n) = 0.25r(n-1) + 0.5r(n) + 0.25r(n+1)$$

The position of the last glottal pulse τ is searched by looking for the sample with the maximum absolute amplitude among the T_0 last samples of the low-pass filtered residual in the frame, where T_0 is the rounded closed-loop pitch lag of the last subframe.

The position of the last glottal pulse τ is coded using 6 bits (with more precision for lower values of T_0). The sign of the maximum absolute pulse amplitude is also transmitted with 1 bit. The sign is important for phase resynchronization since the glottal pulse shape often contains two large pulses with opposite signs. Ignoring the sign may result in a small drift in the position and reduce the performance of the resynchronization procedure.

2.2. Concealment of the periodic part of the excitation

The described concealment is dependent on signal classification [4]. For the concealment purpose, each frame is classified as unvoiced, voiced, voiced onset or transition. If side information is transmitted, this information is coded using 2 bits. Otherwise it is estimated in the decoder.

In case of erasure following unvoiced frames, no periodic part of the excitation signal is generated. For other classes, the periodic part of the excitation signal is constructed in the following manner.

The last pitch cycle of the previous frame is repeatedly copied. The pitch period T_c used to select the last pitch cycle is defined so that pitch multiples or submultiples can be avoided or reduced. The period T_c is then maintained constant during the concealment for the whole erased block.

For erased frames following a correctly received frame other than "unvoiced", the CELP excitation is updated with this periodic part only. This update will be used to construct the pitch codebook excitation in the next frame.

The pitch excitation of the entire lost frame is built by repeating the last pitch cycle of length T_c. If the current frame is the first erased frame after a good frame, this pitch cycle is first low-pass filtered. The filter used is a simple 3-tap linear phase FIR filter with filter coefficients {0.18, 0.64. 0.18}. This is done as follows

$$u(n) = 0.18u(n-T_c-1) + 0.64u(n-T_c) + 0.18u(n-T_c+1), \quad n = 0,\cdots,T_c-1$$
$$u(n) = u(n-T_c), \quad n = T_c,\cdots,199$$

where $u(n)$ is the excitation signal. If this is not the first erased frame, the concealed excitation is simply built as

$$u(n) = u(n-T_c), \qquad n = 0,...,L+N-1$$

Note that the concealed excitation is also computed for an extra subframe to help in the resynchronization as will be shown below.

2.3. Glottal pulse resynchronization

The procedure described above may result in a drift in the glottal pulse position, since the pitch period used to build the excitation can be different from the encoder pitch period.

The resynchronization procedure is performed as follows. If the future frame is available and contains the glottal pulse information, then this information is decoded.

Then the position of the maximum pulse in the concealed excitation $u(n)$ from the beginning of the frame with the same sign as the decoded sign is determined (based on a low past filtered excitation). If the decoded maximum pulse position is positive then the maximum positive pulse in the concealed excitation from the beginning of the frame is determined, otherwise the negative maximum pulse is determined. If $T(0)$ is the first maximum pulse in the concealed excitation, the positions of the other maximum pulses are given by

$$T(i) = T(0) + iT_c, \qquad i = 1,...,N_p - 1$$

where N_P is the number of pulses (including the first pulse in the future frame).

The error in the pulse position of the last concealed pulse in the frame is found by searching for the pulse $T(i)$ closest to the actual pulse P_{last}. The error is given by:

$$T_e = P_{last} - T(k)$$

where k is the index of the pulse closest to P_{last}.

If, T_e=0 then no resynchronization is required. If $T_e \geq 0$ then T_e samples need to be inserted. If $T_e \leq 0$ then T_e samples need to be removed.

The samples that need to be added or deleted are distributed across the pitch cycles in the frame. The minimum energy regions in the different pitch cycles are determined and the sample deletion or insertion is performed in those regions. The minimum energy regions are determined by computing the energy using a sliding 5-sample window.

The sample deletion or insertion is performed around $T_{min}(i)$, where $T_{min}(i)$, i=0, …, N_{min}-1 are the minimum energy positions and N_{min}= N_p-1 is the number of minimum energy regions. The samples to be added or deleted are distributed across the different pitch cycles as follows.

If N_{min}=1, then there is only one minimum energy region and all pulses T_e are inserted or deleted at $T_{min}(0)$.

For N_{min}>1, a simple algorithm is used to determine the number of samples to be added or removed at each pitch cycle whereby less samples are added/removed at the beginning and more towards the end of the frame. If the total number of pulses to be removed/added is T_e, and the number of minimum energy regions is N_{min}, then the number of samples to be removed/added per pitch cycle, $R(i)$, i=0,…,N_{min}-1, is found using the following recursive relation:

$$R(i) = round\left(\frac{(i+1)^2}{2} f - \sum_{k=0}^{i-1} R(k) \right)$$

where $f=2|T_e|/N^2_{\min}$.

Note that at each stage if $R(i) < R(i-1)$ then the values of $R(i)$ and $R(i-1)$ are interchanged. Since $R(i)$ are in increasing order, then more samples are added/removed towards the cycles at the end of the frame.

Using the above procedure, the last maximum pulse in the concealed excitation is forced to be aligned with the actual maximum pulse position at the end of the frame which is transmitted in the future frame.

If the side information for concealment is not transmitted, the pitch value of the missing frame is first estimated. If the future frame is not available, the pitch value is predicted based on past pitch values. If the future frame is available, the missing frame pitch value is estimated using also the future frame pitch. The pitch in the missing frame is then interpolated to find an estimated pitch lag per subframe. Then the total delay of all pitch cycles in the concealed frame is computed for both the pitch used in concealment and the estimated pitch lag per subframe. The difference between these two total delays gives an estimation of the difference between the last concealed maximum pulse in the frame and the estimated pulse. The pulses are then resynchronized as described above.

2.4. Artificial onset construction

The most complicated situation related to the use of the long-term prediction in CELP decoding is when a voiced onset is lost. The lost onset means that the voiced speech onset happened somewhere during the erased block. In this case, the last good received frame was unvoiced and thus no periodic excitation is found in the excitation buffer. The first good frame after the erased block is however voiced, the excitation buffer at the encoder is highly periodic and the adaptive excitation has been encoded using this periodic past excitation. As this periodic part of the excitation is completely missing at the decoder, it can take up to several frames to recover from this loss.

If an onset frame is lost (i.e. a voiced good frame arrives after an erasure, but the last good frame before the erasure was unvoiced, a special technique has been introduced in VMR-WB to artificially reconstruct the lost onset and to trigger the voiced synthesis [4]. Here, the technique has been modified to exploit the fact that an extra frame is available at the decoder for single frame erasures. Thus, the position of the last glottal pulse in the concealed frame can be available from the future frame. In this case, the concealment of the erased frame is performed as usual. However, the last pulse of the erased frame is artificially reconstructed based on the position and sign information available from the future frame. This information consists of the position of the maximum pulse from the end of the frame and its sign. The last glottal pulse in the erased frame is constructed artificially as a low-pass filtered pulse. If the pulse sign is positive, the low-pass filter used is a simple linear phase FIR filter with the impulse response h_{low}={-0.0125, 0.109, 0.7813, 0.109, -0.0125}. If the pulse sign is negative, the low-pass filter used is a linear phase FIR filter with the impulse response - h_{low}.

The low-pass filtered pulse is realized by placing the impulse responses of the low-pass filter in the memory of the adaptive excitation buffer (previously initialized to zero), after proper scaling. In the decoding of the next good frame, normal decoding is resumed. Placing the low-pass filtered glottal pulse at the proper position at the end of the concealed frame significantly improves the performance of the consecutive good frames and accelerates the decoder convergence to actual decoder states.

3. FRAME ERASURE CONCEALMENT IN G.729.1

The new G729.1 standard is an embedded CELP codec where the core layer is interoperable with G.729 at 8 kbit/s [6]. The second layer is a narrowband layer at 4 kbit/s (total 12 kbit/s). Then 2 kbit/s layers are added to result in bit rates from 14 to 32 kbit/s with wideband rendering.

The concealment strategy in G.729.1 is based on VMR-WB in addition to the innovations described above in Section 2. Side information consists of 14 bits and it is distributed in three layers.

The classification information requires 2 bits which are transmitted in Layer 2 (at 12 kbit./s). The phase information is quantized using 7 bits and transmitted in Layer 3 (at 14 kbit/s). The determination of the phase information has been described in Section 2.1. The energy information is quantized using 5 bits and transmitted in Layer 4 (at 16 kbit/s). Frame erasure concealment is performed at the decoder by making use of the transmitted concealment/recovery parameters and by exploiting the extra frame delay at the decoder.

Efficient concealment and recovery techniques are used including glottal pulse resynchronization and artificial onset reconstruction as was described in Section 2. Proper energy control is also applied similar to VMR-WB standard.

At the decoder, if the output is at a layer where no side information is available then this information is estimated. For example, when decoding at 8 kbit/s, no side information is

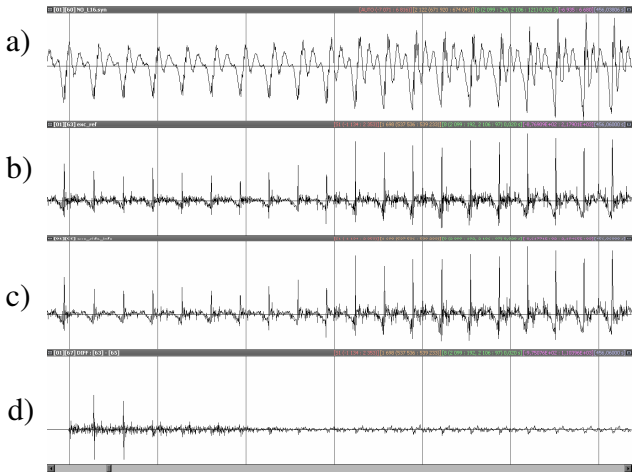

Figure 2:
a) Original synthesis.
b) Excitation without frame easure.
c) Excitation with the first frame erased and concealed using pitch resynchronization.
d) Difference between (c) and (b).

available. In this case, the class of the erased frame is estimated at the decoder. Similarly, at 8 and 12 kbit/s decoding, the phase information is estimated at the decoder based on the pitch delay in future and past frames.

4. PERFORMANCE

Figure 2 shows the example given in Figure 1 but with glottal pulse resynchronization applied. It is clear that in the good frames after the erased frame, the signal converges quickly to the actual state at the encoder.

An A-B test was performed to compare the performance with and without glottal pulse resynchronization. Figure 3 shows the result of the A-B test with 6% random frame erasure. Twelve expert listeners participated in the test. The improvement due to resynchronizing the pitch is evident even without transmitting the phase information.

Figure 4 shows the results of an A-B test where pitch synchronization was performed but with and without the use of side information. Again it is evident that clear improvement is obtained when the position and sign of the glottal pulse are transmitted compared to predicting the position at the decoder.

In a recent formal MOS test, G.729.1 was tested with and without the extra frame lookahead at the decoder at 14 kbit/s. With the extra frame delay resynchronization was performed using the side information and resulted in an improvement of 0.15 in the MOS scale compared to resynchronization using a predicted pitch lag.

5. CONCLUSION

An efficient frame erasure concealment technique was presented whereby the adaptive codebook content is synchronized with the actual content at the encoder after erased frames. This improves the decoder convergence when good frames are received after the erased frame. The technique has been implemented in the recently selected ITU-T G.729.1 Recommendation.

The improvement due to the adaptive codebook synchronization, both with and without the transmission of the side information, has been demonstrated in A-B tests. Further, G.729.1 characterization test results, where this technique has been implementer, showed very good performance in the presence of frame erasures.

6. REFERENCES

[1] R. Lefebvre, P. Gournay, R. Salami, “A study of design compromises for speech coders in packet networks”, Proc. ICASSP-2004, pp. 256-268.

[2] M. Chibani, R. Lefebvre, and P. Gournay, "Resynchronization of the adaptive codebook in a constrained CELP codec after a frame erasure," Proc. ICASSP-2006, pp. 13-16.

[3] 3GPP2 C.S0052-0 V1.0, “Source-Controlled Variable-Rate Multimode Wideband Speech Codec (VMR-WB)”, July 2004.

[4] M. Jelinek and R. Salami, “Wideband Speech Coding Advances in VMR-WB standard”, Accepted for publication in IEEE Trans. on Audio, Speech and Language Processing.

[5] ITU-T Recommendation G.729.1, “G.729 based Embedded Variable bit-rate coder: An 8-32 kbit/s scalable wideband coder bitstream interoperable with G.729,” ITU-T, Geneva, 2006

[6] R. Salami, et al., “Design and description of CS-ACELP: a toll quality 8 kb/s speech coder,” *IEEE Trans. on Speech and Audio Processing*, vol.6, no. 2, pp. 116–130, March 1998.

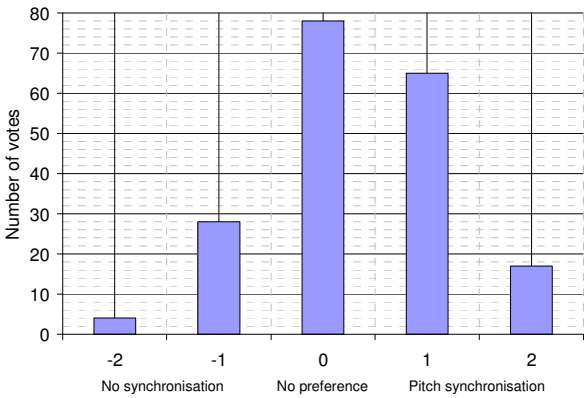

Figure 3: A-B test with and without pitch synchronization (without side information).

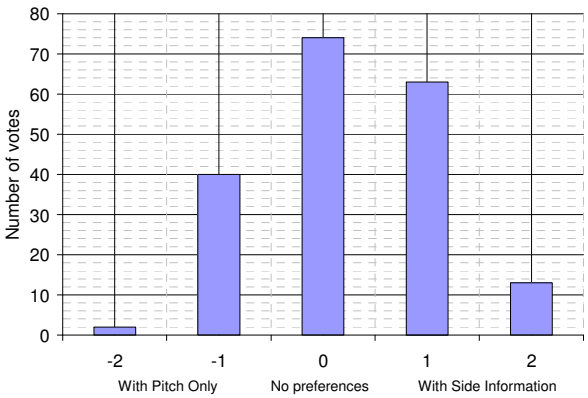

Figure 4: A-B test with pitch synchronization with and without side information.