A low complexity packet loss concealment algorithm for G.711 and G.722

Ashwin Kashyap, Mikael K.Rudberg

***Abstract*—In communication systems, enhancement algorithms are used to improve the perceptual quality of speech. Packet loss concealment belongs to this genre. The concept is to mask the effect of loss of a speech packet in the channel of a communication system. We present a generic packet loss concealment algorithm for speech sampled at 8kHz and 16kHz. The algorithm is coupled with both G.711 and G.722 decoders. This setup is used to enhance the perceptual quality of speech. Time-domain signal processing techniques are used to recreate the lost packet. A robust method is described to improve the quality further by updating the decoder state. A floating point model of the algorithm is considered for performance evaluation. A fixed point model is also developed for a DSP processor. Performance evaluation is done by PESQ (Perceptual Evaluation of Speech Quality) and subjective listening.**

***Index Terms*—algorithms, G.711 and G.722 codecs, packet loss concealment, speech quality enhancement**

I. Introduction

Loss of a speech packet can cause a noticeable deterioration in the quality of speech. To overcome this problem, a synthesized speech packet is substituted in the place of the lost packet. Several methods have been employed to counter the effect of packet loss. Zero stuffing and packet repetition are the most common methods in this direction.

The idea behind zero stuffing is to substitute a zero packet instead of a lost packet. Packet repetition involves the use of the last good packet received in place of the lost packet. In both of these methods, artifacts are introduced and there is a sudden noticeable transition between natural and synthesized speech. The perceptual quality of speech is not significantly improved when the above mentioned methods are employed. Due to these reasons, several algorithms which exploit the salient characteristics of speech have been developed in recent times. The algorithm we propose has also been developed with similar considerations on perceptual speech quality and speech signal characteristics.

Manuscript received June 20th, 2007. This work was supported by Infineon.
Ashwin Kashyap is with Infineon Technologies India Pvt Ltd,Bangalore 560066 India (phone: 91-8023492957; fax: 91-8041392333; (e-mail: *ashwin.kashyap@infineon.com*).
Dr. Mikael. K. Rudberg was with Infineon Nordic,Linkoping,Sweden. He is now with SP Devices,Linkoping,Sweden(e-mail: *mikael.rudberg@spdevices.com*).

The algorithm uses autocorrelation based pitch prediction to estimate the lost packet. In order to mask the discontinuity between the synthesized speech and actual speech, an overlap-add operation is performed at the packet boundary. Continuous multiple packet losses are handled by post filtering and by introducing a fade out. Performance of the algorithm is determined by PESQ[1] , a metric which models the human auditory system. PESQ evaluates the degraded signal in comparison with the reference signal based on perceptual characteristics. It also takes level normalization and delays into account during the evaluation. Further, the metric can support both narrowband and wideband speech. In our evaluation, we consider a lossless signal as the reference.

G.711[2] is a coding standard for narrowband speech and it works on the principle of companding. The codec uses A/u-law tables to code and decode speech. G.722[3] works on the principle of adaptive differential pulse code modulation. The codec works on wideband speech. The spectral band of 8 kHz is separated into two sub-bands and the bit allocation during the stage of coding is done based on the perceptual importance of the two frequency bands. The proposed algorithm works together with these codecs to support either narrowband or wideband speech.

The paper is organized into the following sections. The simulation environment is described in section II. This is followed by algorithmic description in section III. The comparison with existing algorithms along with objective scores for both narrowband and wideband speech can be found in section IV. Section V explores the possibilities of future work related to our algorithm.

II. Simulation Environment

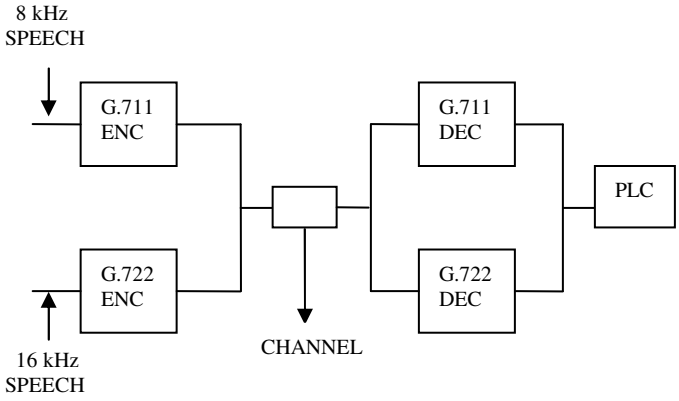

Fig. 1. Schematic of the simulation environment consisting of narrowband and wideband speech codecs with the PLC algorithm. Error patterns are introduced in the channel for simulation.

The simulation environment in Fig. 1 consists of G.722 and G.711 encoders and decoders. “Enc” and “dec” refer to encoder and decoder respectively. The channel is modeled by a binary file, where a zero corresponds to a good packet and one corresponds to a bad packet. Frame erasure patterns were generated by using a C-model for different error percentages.

The simulation was done for one, three, five and ten percent packet losses. A sanity check is done to ensure that the PLC algorithm has been correctly integrated in the simulation platform.

Frames of 10ms were used in the simulation environment. However, frames of any duration could be used by modifying a few parameters inside the PLC algorithm. Choice of frame duration depends on the spectral characteristics of speech. Chosen speech duration should ensure that the signal is stationary for all practical purposes during that window. Due to this constraint, the upper bound is 20ms.

III. ALGORITHMIC DESCRIPTION

A. *Reception of a good frame:* For a frame received correctly, the packet loss concealment algorithm stores the samples in an internal buffer of length 256, which corresponds to roughly three frames of narrowband data. The buffer is designed to ensure that the pitch information can be extracted correctly when there is a packet loss.

B. *Reception of a bad frame:* For a frame not received correctly, we exploit the salient characteristics of the speech signal. Time domain signal processing techniques are used to recreate the lost frame and they are described below.

1) *Pitch prediction*: The autocorrelation function is used to estimate the pitch of the speech signal. The correlation window is designed to include the range of frequencies from 88 Hz to 222 Hz and this covers male and female pitch frequencies. The correlation is performed in two stages – coarse pitch search and fine pitch search.

The course pitch search uses a grid of every third sample for narrowband data and every sixth sample for wideband data. The first peak of the autocorrelation function, which corresponds to the energy of the signal is computed for future use. The pitch information lies in the proximity of the second peak in the autocorrelation function. This is roughly estimated by the coarse pitch search and followed by the fine pitch search which estimates the pitch accurately around the second peak. The fine pitch search can be skipped without compromising the quality of speech significantly if a faster algorithm is required. As less than one percent of the energy falls within the 4 kHz-8 kHz frequency band, every alternate sample in the correlation window can be used for estimating the pitch in wideband speech.

2) *Voiced and unvoiced components of speech*: The algorithm uses a hard threshold to determine if the lost frame is voiced or unvoiced. The energy, which corresponds to the first peak in the autocorrelation function is weighted and compared with the second peak corresponding to the pitch period. If the two quantities are comparable, we conclude that the frame is strongly periodic. This is one of the characteristics of voiced speech. If the frame is classified as unvoiced, we fade out the samples in synthesized speech and thereby limit the energy level in the recreated frame.

3) *Recreation of the lost frame:* The speech samples from the first pitch period are used to synthesize the lost frame. The frame is usually bigger than the length of the pitch period. Repeating the samples in the pitch period continuously introduces an artificial periodicity in the synthesized signal. To overcome this problem, we also use the speech samples from the second pitch period. For this purpose, we use a constraint to estimate the similarity of the signals in the first and second pitch periods. And this is analogous to comparing the second and third peaks in the autocorrelation function.

4) *Enhancement at the frame boundaries:* When a frame is lost and recreated, there is a transition from natural speech to synthesized speech. To make this transition as smooth as possible, we use overlap-add windows. This operation is carried out on 10 percent of the frame. A linear ramp is used for this purpose. The coefficients chosen for the overlap-add window are normalized and result in an efficient implementation in fixed point processors where shifts are better supported than division.

5) *Post-filtering*: The algorithm has a provision to handle continuous frame erasures. When multiple frames are lost, we would introduce artifacts in the synthesized speech unless we fade out the recreated signal. The fade out will typically start after one to four frames are lost and the signal will gradually fade out during these one to four frames. The fade out rate is parametric and can be chosen depending on the volatility of the channel. The signal level is restored gradually when good frames are received. A first order low-pass post filter is introduced to the synthesized speech frame. The bandwidth limitation introduced by the post filter reduces the subjective distortion introduced in the signal.

6) *Updating the decoder states:* The PLC algorithm works in tandem with the decoders used in the communication system. We explore two approaches to deal with packet loss from the purview of the decoders.

The first approach would be to “freeze” the decoder. In this approach, whenever a frame is lost, the signal flow bypasses the decoder completely. The decoder would be active only when a good frame is received. This approach would be optimal when computational complexity is a key issue as it avoids an invocation of the decoder during the reception of a bad frame.

The second approach is to update the states of the decoder by using the samples in the last good frame whenever a frame erasure occurs. This approach is suitable for adaptive decoders as the adaptive predictors are sensitive to the discontinuity in the signal. The use of the last good frame to update the decoder states maintains information about the signal level and hence the adaptive

predictors function better in this approach rather than the first approach of freezing the decoder. The overhead for this approach over the first one would be an invocation of the decoder during frame erasure and a memory requirement of a buffer equivalent to one full frame. The output generated by the decoder is not to be used as the PLC algorithm recreates the lost frame.

The G.711 standard works on the principle of A-law or u-law. The decoder is memoryless. Hence the decoder can be frozen during frame erasure without any deterioration in speech quality.

On the other hand, the G.722 decoder is strongly backward adaptive and hence the introduction of the decoder update as outlined in our second approach achieves a better subjective speech quality.

Updating the decoder states with the reconstructed frame has been explored in [4] and [5]. The algorithm described in [4] uses the recreated frame to be fed back into the encoder to update its states. This in turn would also introduce an additional call to the decoder. The recreated frame has to be transmitted to the encoder and this process would involve the channel and introduce a small delay.

IV. PERFORMANCE

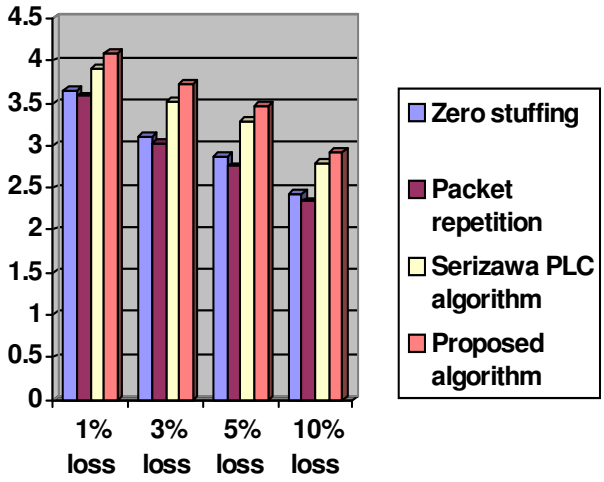

Fig. 2. Comparison of different PLC algorithms for wideband speech (16 kHz) using PESQ. The x-axis shows the frame erasure rate and the y-axis shows the MOS values.

Fig. 2 and Fig. 3 show the quality of speech synthesized for different PLC algorithms for different frame erasure rates. The proposed algorithm achieves a Mean Opinion Score (MOS) of 3 at ten percent frame erasure rate for both narrowband and wideband speech. The MOS values are slightly higher for narrowband speech than wideband speech due to the number of samples lost in a frame and the step sizes used for correlation. Even at high frame erasure rates, the proposed algorithm synthesizes speech of perceptually moderate quality.

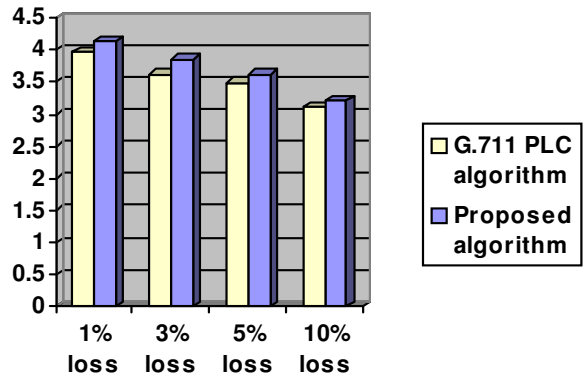

Fig. 3. Comparison of the G.711 PLC algorithm with the proposed algorithm for narrowband speech (8 kHz) using PESQ. The x-axis shows the frame erasure rate and the y-axis shows the MOS values.

V. FUTURE WORK

The algorithm currently uses a hard threshold for voiced-unvoiced classification. In order to have a higher accuracy in predicting the nature of the frame, we can introduce a signal dependent threshold. Time domain parameters like zero crossing rate weighted RMS energy [6] and Kaiser-Teager frame energy [6] can be used for multiple frames to have a signal dependent threshold.

VI. CONCLUSION

The proposed packet loss concealment algorithm gives a significant improvement in perceptual quality of speech over conventional methods like zero stuffing and packet repetition as shown in the improvement of PESQ-MOS values over all frame erasure rates. There is also an improvement in the performance over the algorithm defined in [4]. The decoder update further enhances the speech quality without introducing delay, which is an important factor in real-time systems.

The algorithm has the advantage of working with G.711 and G.722 codecs simultaneously and this is useful in a multirate system where both narrowband and wideband speech are supported. The framework of the algorithm facilitates its use in conjunction with other codecs and also for different frame lengths.

REFERENCES

[1] ITU-T Rec. P.862, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Feb. 2001.

[2] ITU-T Rec. G.711 Appendix I, "A high quality low-complexity algorithm for packet loss concealment with G.711," Sep. 1999.

[3] ITU-T Rec. G.722,"7 kHz audio coding within 64 kbps," Nov. 1988.

[4] Serizawa M. , Nozawa Y.,"A packet loss concealment method using pitch waveform repetition and internal state update on the decoded speech for the sub-band ADPCM wideband speech codec," IEEE speech coding workshop, pp. 68-70, 2002.

[5] Niranjan Shetty, Jerry D.Gibson, "Improving the robustness of the G.722 wideband speech codec to packet losses for voice over WLANs," ICASSP 2006.

[6] Shahnaz,C., Zhu, W.P., Ahmad,M.O.,"A multi-feature voiced/unvoiced decision algorithm for noisy speech," ISCAS 2006.