Implementing the PLC Techniques with G 729 Coded to Improving the Speech Quality for VoIP Transmission

Adil BAKRI
Speech Communication and Signal Processing Laboratory
Faculty of Electronics and Computer Science, USTHB.
Adil msilib@yahoo fr

Abderrahmane AMROUCHE
Speech Communication and Signal Processing Laboratory
Faculty of Electronics and Computer Science, USTHB.
namrouche@usthb.dz

Abstract*—Speech compression technology is widely used in digital communication systems such as wireless systems and VoIP. The compression technique described in the ITU-T G.729 Recommendation is commonly employed in speech transmission systems because of the reconstructed speech signal in reciever. To increase the quality of speech on the Voice Over Internet Protocol (VoIP), we propose in this paper the technique of Packet Loss Concealment (PLC) and analyzes the performance results of ITU-T G.729 codec with this technique. The PLC algorithms used ranged from simply inserting silence and repeated for the missing audio, to the use of the G.711 Appendix I algorithm that does a good job of generating a synthetic speech signal to cover missing data in a received bit stream, to ensure a smooth transition between the real signal and the synthetic signal.

Keywords-VoIP; OLA; PLC; G729; ITU-I G711 Appendix I; PESQ

I. INTRODUCTION

Voice over IP (VoIP) is a real time application that allows transmitting voice through the Internet network. Recently there has been amazing progress in this field, mainly due to the development of voice codecs that react appropriately under conditions of packet loss and the improvement of intelligent jitter buffers that perform better under conditions of variable inter packet delay. In addition, there are other factors that indirectly benefited VoIP. Today, computer networks are faster due to the advances in hardware and breakthrough algorithms. As a result, the quality of VoIP calls has improved considerably. However, the speech quality of VoIP calls under extreme conditions of packet loss still remains a major problem that needs to be addressed for the next generation of VoIP services. This paper concentrates in making an analysis of the effects that network impairments, such as: delay, jitter, and packet loss have in the quality of VoIP calls and approaches to solve this problem. In VoIP network packet drops, packet delay, delay variations, errors, and fragmentation are very common and may degrade voice quality. Packet loss concealment (PLC), also known as packet erasure concealment, synthesizes the missing voice samples during the packet erasures without creating the noticeable artifacts, PLC techniques are categorized as transmitter – receiver - and receiver - only based.

This work presented the PLC based on ITU-T G.711 Appendix I for improving the voice quality in the presence of packet drops. To measure the mean opinion score of VoIP calls we develop an application based on the E-Model, and utilize perceptual evaluation of speech quality (PESQ).

This paper is organized as follows: after a brief introduction, we describe the principle of VoIP networks in transmission in the section II, and a brief description of G.729 codec is given in section III. The section IV presents the network model, and specifics concerning the PLC technique are described in section V. The PESQ technique is briefly explained in section VI, and in section VII give an example for clarify the impact the PLC. Performance results obtained for the speech quality are presented and discussed in section VIII. Section IX presents the main conclusion of this paper.

II. TRANSMISSION OF SPEECH SIGNALS OVER IP NETWORKS

The speech signal is digitized and packetized at the sender at regular intervals (e.g., every 10 ms) using an encoding algorithm. The voice packet is then sent over the IP network to the receiver where it is decoded and played-out to the listener. There are many impairment types that can potentially affect voice transmission over the network. These impairments range from packet delays to actual errors that affect the data content within these packets. Using the TCP protocol usually takes care of correcting some of these errors. However, since voice and video traffic has to be transmitted in real-time, these corrective measures cannot be applied in such cases. Therefore, the transmission of voice traffic is normally done using the UDP protocol which does not have such corrective measures [1]. The drawback of this is that voice traffic will be affected by any network impairments that it encounters with no corrective measures being applied. The resulting voice, in many cases, can still be comprehended by the human ear. The same may not be true for the software modules that we use to process the voice commands in the systems that we consider in this study. There are many issues that can affect data transmission over IP networks [2]. Some packets may get lost, distorted, delayed, etc.

The G.729 family of codecs is popularly used in most VoIP deployments speech compression algorithm. G.729 makes use of human vocal tract models suitable for voice signals unlike the G.711 and G.726 codecs, which use waveform - based compression. The voice codec G.729 is operates on speech frames of 10 ms which correspond to 80 samples digitized in 16-bit for a sampling frequency of 8 kHz [3]. The speech signal is analyzed to extract the features of encoder packet and sent through the IP network. The decoder uses these features to reconstruct a synthetic speech signal as shown “Fig. 1”.

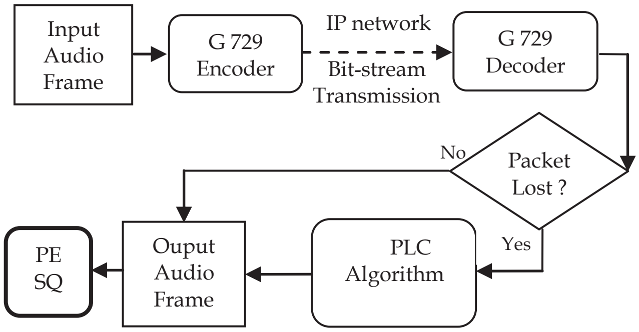

Figure.1 Transmission of voice over IP network.

III. G. 729 ALGORITHM

G.729 codec is part of the CELP (Code-Excited Linear Prediction) family of speech compression algorithms. These types of algorithms take advantage of the model of speech production such as the one presented in “Fig. 2” [3].

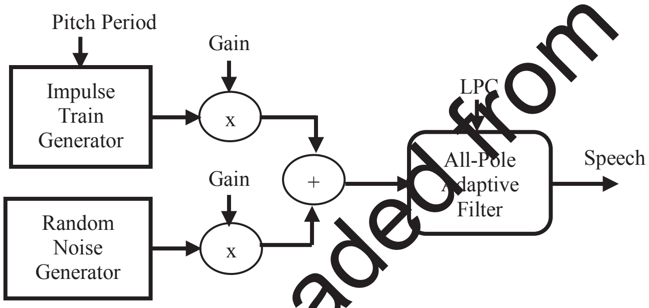

Figure. 2 Block diagrams of speech production model.

This model essentially stipulates that voice can be synthesized by filtering an excitation through a linear prediction synthesis filter, where the filter represents the short-term predictability in the speech. The excitation is either formed of a periodic impulse for voiced speech where the period is the pitch in the voice or by random noise for unvoiced speech. A gain is applied to the excitation in order to obtain the proper speech level. G.729 uses this model to extract from the speech the required parameters such as the linear prediction coefficients (LPC), the gain, the pitch and the best random noise needed to synthesize the speech. The overall excitation is always formed by the sum of the adaptive codebook vector (voiced speech) and the fixed codebook vector (unvoiced speech) in order to better represent speech that is partly voiced or unvoiced. The fixed codebook vector is chosen through analysis-by-synthesis where different values are chosen and the one that leads to the synthesized speech that is perceptually closer to the input speech is used.

G.729 divides the speech into frames of 10 ms, which are further subdivided into two 5 ms sub-frames. The required parameters are transmitted as 80 bits per frame, which results in a required bit rate of 8 kbps. For more details on the exact implementation of the G.729 algorithm, refer to [4].

IV. NETWORK MODEL

The communication channel is modeled by method of two-state Markov shown in “Fig. 3”. This model is characterized by the transition probabilities p and q between the state 0 (packet received correctly) and the state 1 (packet lost) [5]. From these probabilities is easy to derive that:

$$p_0 = \frac{1-q}{p+1-q} \tag{1}$$

$$p_1 = \frac{p}{p+1-q} \tag{2}$$

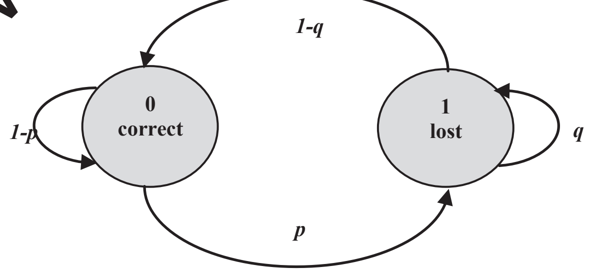

Figure. 3 Markov Model of the IP channel.

V. PLC BASED ON THE ITU-T G.711 APPENDIX I

PLC based on ITU-T G711 algorithm conceals the missing packet by generating synthetic speech that has similar characteristics to the speech in the history buffer. The idea is as follows. If the signal is voiced we assume the signal is quasi-periodic and locally stationary as shown as “Fig. 4”. We estimate the pitch using an autocorrelation technique and repeat the last pitch period in the history buffer a few times. This is a low-complexity time-domain algorithm that uses the most recent 48.75 ms history of the decoded output signal to estimate what the signal should be in the missing frames. The algorithm delays the output by 3.75 ms to allow the synthetic speech signal to be mixed with the tail of the last good packet using an Overlap-Add (OLA) at the start of a loss [6,7].

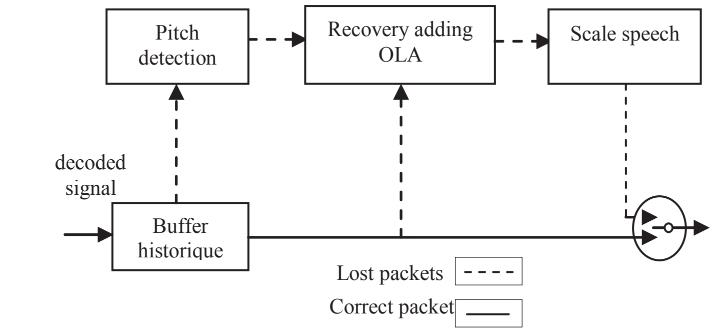

Figure. 4 PLC scheme based on G.711 Appendix I.

A. History Buffer

The technique PLC makes a copy of the decoded output and is saved in a circular history buffer that is 48.75 ms (390 samples) long as shown as “Fig. 5”. The history buffer is used to calculate the current pitch period and extract waveforms during an erasure. This buffering does not introduce any delay into the output signal [6].

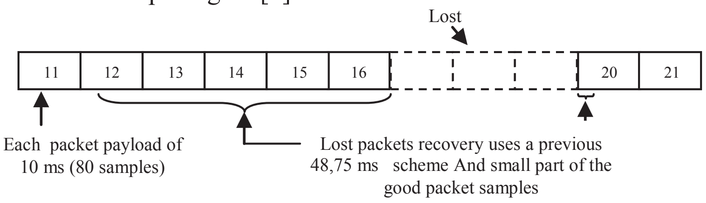
Lost
11
12
13
14
15
16
20
21
Each packet payload of 10 ms (80 samples)
Lost packets recovery uses a previous 48,75 ms scheme And small part of the good packet samples

Figure. 5 Diagram is shown for three packets loss and the dependencies on previous and future samples.

B. Pitch detection

the pitch period is estimated by finding the peak of the normalized cross-correlation of the most recent 20 ms of speech in the history buffer with the previous speech at taps from 5 (40 samples) to 15 ms (120 samples) [8].

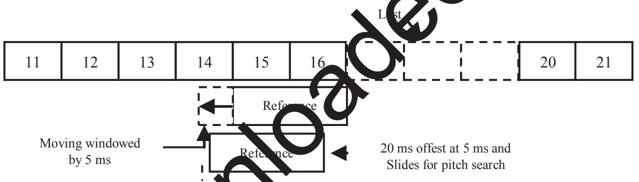

Figure. 6 Representation of correlation windows for pitch detection.

C. Technical recovery adding OLA

The recovery technique adding OLA (Over Lap and Add), assure transition is needed between the synthesized erasure speech and the real signal. To create the pitch buffer, the 1/4 wavelength from before the erasure is OLAed with a triangular window to the 1/4 wavelength from the previous pitch period [9]. The results of this OLA replace the 1/4 wavelength of signal before the erasure. During the first 10 ms (80 samples) of an erasure, the synthetic signal is generated from the last pitch period with no attenuation. The most recent pitch periods of the history buffer are used during the first 10 ms. An OLA is performed using a triangular window on one quarter of the pitch period between the last and the next - to - last period. If the erasure is 20 ms long, the number of pitch periods used to synthesize the speech is increased to two, and if erasure is 30 ms long, a third pitch is added, the synthesized signal is attenuated by 20%.Beyond 30 ms of erasure, no changes are made to the history buffer, the number of pitch periods used to synthesize the speech is third pitch periods, the synthesized signal is linearly attenuated with a ramp at a rate of 20% per 10 ms. After 60 ms, the synthesized signal is zero. The synthetic signal is attenuated with a linear ramp by a call to scalespeech. At the first good frame (10 ms) after an erasure, a smooth transition is needed between the synthesized erasure speech and the real signal. To do this, the synthesized speech from the pitch buffer is continued beyond the end of the erasure, and then mixed with the real signal using an OLA.

VI. Perceptual Evaluation of Speech Quality (PESQ)

The PESQ (Perceptual Evaluation of Speech Quality) is an objective method for the evaluation of speech quality speech coders used in narrowband telephone networks. The evaluation of the PESQ compares the original signal with the degraded after passing through the communication system signal, in our case after encoding / decoding [10]. “Fig. 7” shows the general diagram simulation PESQ.

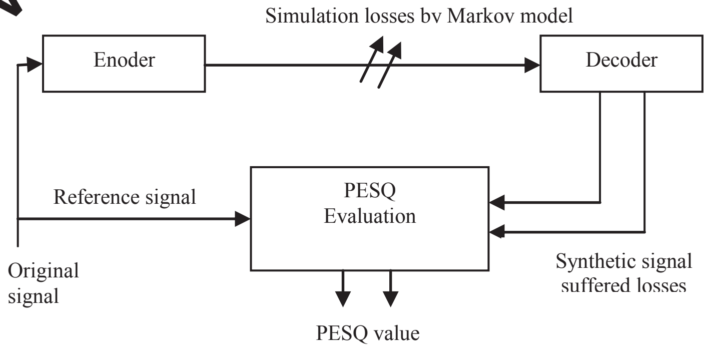

Figure. 7 Diagrame of the simulation.

The rate of packet loss is given by the following formula:

$$RATE(\%) = \frac{Number\ of\ lost\ frames}{Total\ number\ of\ frames} \times 100$$

VII. Comparison between waveforms

We give an example of the results which shows the reconstructed after a packet loss signal. The correction made by use of the PLC based on ITU-T G.711 Appendix I. It is observed a “Fig. 8”.

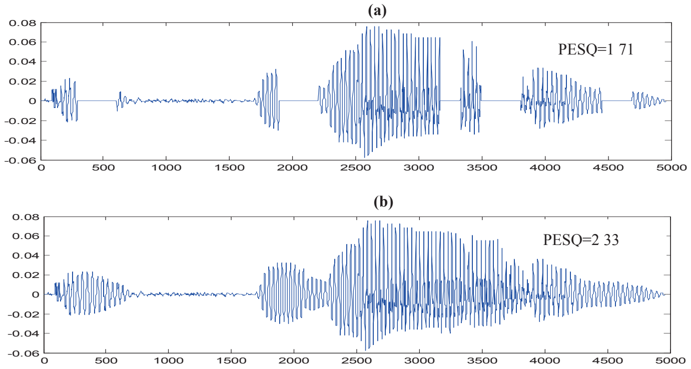

Figure. 8 Shows a result of the male figure '2 ‘speaker :

(a) Synthetic signal obtained by codec G.729 lossy. synthetic

(b) Signal obtained by codec G.729 with PLC.

VIII. EXPERIMENTAL RESULTS

In this experiment the objective is to study the influence of the frames lost on the reconstructed speech by G.729 decoder. We varied the packet loss rate from 0% to 20%, in both cases, that is to say before and after the application of PLC technique based on ITU-T G.711 Appendix I. We took twenty voice files uttered by male speakers and twenty voice files made by women female speakers. We give each of PESQ average value for each value of loss rates. “TAB. I” and “Fig .9” show the variations of speech quality (PESQ) based on the rate of packet loss with and without PLC.

TABLE I. SPEECH QUALITY AS A FUNCTION OF PACKET LOSS RATE WITH AND WITHOUT PLC

Rate of packet loss %	***Codec G729***	***Codec G729 with PLC***
0%	***3.197***	***3.197***
5%	***2.056***	***2.935***
10%	***1.736***	***2.650***
15%	***1.670***	***2.273***
20%	***1.247***	***1.998***

We note that the reconstructed signal by PLC technique is more similar to the original signal as the lost signal, although the distortions are present and can adversely affect speech quality.

By comparing the results of “Fig. 9”, we note that the significant improvement in quality rates (PESQ) using the PLC-based ITU-T G.711 Appendix I. The results obtained in this second phase have shown that the use of PLC technique improves speech quality in case of packet loss. The results obtained with our method show a large increase of the quality threshold, thus the effectiveness of the implemented method is significant.

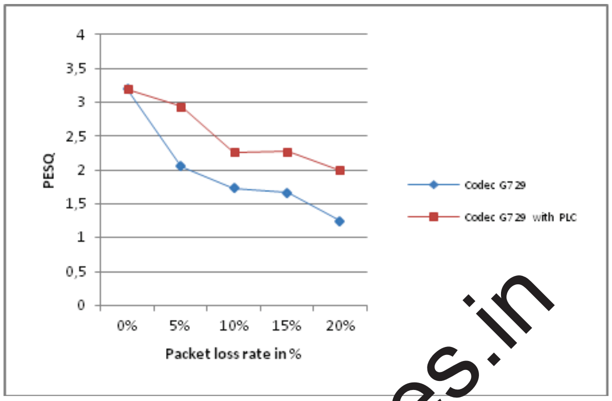

Figure. 9 Speech quality with packets loss.

IX. CONCLUSION

In this work, we adapted the technique to conceal the packet loss in the Recommendation ITU-G.711 Appendix I to the G729 codec dedicated to VoIP. Our main objective was the improvement of speech quality in VoIP networks. We proposed the introduction of the PLC in the transmission of speech in VoIP networks. Experimental results show that our method based on the inclusion of loss concealment technique can be applied effectively for application in speech recognition using VoIP networks. The proposed solution can help improve the speech quality in VoIP and make systems more robust when packet losses.

There is a clear need to bring the results of our work to the specifications of networks, including the protocols used. Thus, a significant part, and often dominant, in which the frames and packets consist of headers. We must find a relationship between the loss of packets in the communication network and the effective portion of the speech signal. These future works are considered in order to enhance the work on QoS.

REFERENCES

[1] R. G. Cole, J. H. Rosenbluth “Voice over IP performance monitoring” SIGCOMM Comput. Commun. Rev, vol.31, pp. 9–24, 2001.

[2] Roychoudhuri L, Al-Shaer E, “Real-time audio quality evaluation for adaptive multimedia protocols,” Proceedings of Multimedia Networks and Services Spain, 2005.

[3] ITU-T Recommandation G.729, “ Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP)”, 1996.

[4] Salami R. "ITU-T G.729 Annex A: Reduced Complexity 8 kb/s CS-ACELP Codec for Digital Simultaneous Voice and Data". IEEE Communications Magazine, Sept. 97, pp.56- 63

[5] Zhicheng Li; Chakareski, J “Hidden Markov model-based packet loss concealment for voice over IP,”,Audio, Speech, and Language Processing, IEEE Transactions on, vol.19, pp. 1609 – 1623, 2006.

[6] Recommandation UIT-T G.711, “A high quality low-complexity algorithm for packet loss concealment with G.711 ”, Septembre 1999.

[7] J. Wiley, “VoIP voice and fax signal processing”, Published simultaneously in Canada, p.592, 2008.

[8] K. Nakamura, "An Improvement of G.711 PLC Using Sinusoidal model", Proceedings of the IEEE The International Conference on Computer as a Toll, pp.1670-1673, 2005.

[9] P.C.W. Sommen and J.A.K.S. Jayasinghe, "On Frequency Domain Adaptive Filters using the Overlap-add Method", IEEE Philips Research Laboratories, pp.28-30, 1988.

[10] E.Conway, "Output-based method of applying PESQ to measure the perceptual quality of framed speech signals", IEEE Wireless Communications and Networking Conference, vol. 4, pp. 2521–2526, 2004.