Automatic Speech Recognition for VoIP with Packet Loss Concealment

Adil BAKRI[1],Abderrahmane AMROUCHE[1], Mourad ABBAS[2],Lallouani BOUCHAKOUR[2]
Speech Communication and Signal Processing Laboratory Faculty of Electronics and Computer Science, USTHB[1]
Scientific Research and Technical Center for the Development of Arabic Language, CRSTDLA[2]

***Abstract*—This paper proposes a Packet loss concealment (PLC) techniques for recognition of speech coded with the G729 codec, which is widely used in VoIP networks. PLC at a receiver has a substantial effect on the performance of automatic speech recognition (ASR) systems in VoIP (Voice over IP). Many of the standard ITU-T CELP based speech coders, such as the G.723.1, G.728, and G.729, model speech reproduction in their decoders. These decoders have enough state information to integrate PLC algorithms directly in the decoder, and are specified as part of their standards in particular by PLC based ITU-T G711 Apendix I. Speech is transmitted with source and channel codes optimized, this channel is simulated by two states Markov model to modeled loss packets. The objective of PLC based ITU-T G711 Apendix I is to generate a synthetic speech signal to cover missing data or loss packets in a received bit stream for the ASR application, i.e., to minimize word error rate.**

***Index Terms*—VoIP, ASR, OLA, PLC, G729, ITU-I G711 Appendix I, HMM.**

I. Introduction

In a VoIP system, at the receiver, some packages may be missing, because of delays, congestion or transfer errors. In communication networks, such losses are caused by several factors at different stages of the chain of transmission, particularly congested nodes (routers).We also know the packet loss causes loss of synchronization between the encoder and decoder. Packet loss degrades the voice quality and affects the quality of speech. It results in breaks in the conversation and a sense of the speech hatching. It is therefore essential to establish a mechanism for packet loss concealment. Several packet loss concealment algorithms (PLC) are used, both at the transmitter at the receiver.

In this work, we are interested in studying the effect of packet loss on the system performance of automatic speech recognition (ASR). As such we have implemented the technique of packet loss concealment PLC based on ITU G.711 Appendix I. We used the database ARADIGIT8K which was passed through the G.729 codec, to obtain the database transcoded by G.729 codec, it is called the database ARADIGIT_G729. This paper is organized as follows: after a brief introduction, we describe the principle of VoIP networks in transmission in the section II. In the section III presents the adaptive PLC technique. Performance results obtained for the speech recognition and discussed in section IV. Section V presents the main conclusion of this paper.

II. Dissimulation Technique of Packet Loss

A. Transmission of the Speech in VoIP

Transmission networks VoIP using the codec mainly G711. But, because of its high-rate (64 Kbits / s), it begins to be gradually supplanted by the much lower rate G.729. The voice codec G.729 is based on the prediction algorithm CS-ACELP (Conjugate-Structure Algebraic Code-Excited Linear Prediction) and operates on speech frames of 10 ms which correspond to 80 samples digitized in 16-bit for a sampling frequency of 8 kHz [2]. The speech signal is analyzed to extract the parameters of encoder packet and sent through the IP network [3]. The decoder uses these parameters to reconstruct a synthetic speech signal as shown Figure(1)

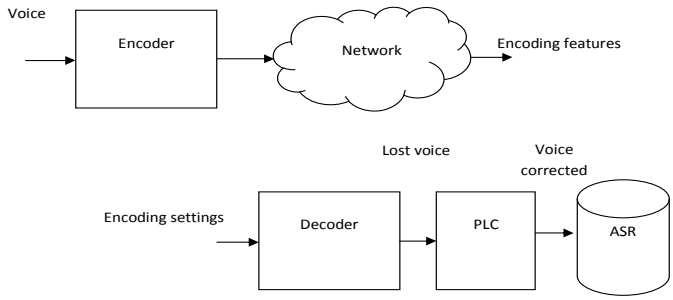

Fig. 1. Transmission of voice over IP network [4].

B. Network Model

We used a simple network model called two-state Markov process to model point-to-point packet loss on the IP network. State 0 indicates that the packet is received and state 1 that is lost. Figure (2) shows the packet loss modeled by a Markov random process with two states.

P is the probability that the network model drops a packet, given the previous packet is delivered i.e. the probability of transition from state 0 to state 1 and q is the probability that the network model drop a packet given that the previous packet is dropped i.e. the probability that the model remains in state 1. This probability is also known as the conditional probability of loss. The probabilities to stay in state 0 and state 1 respectively are:

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

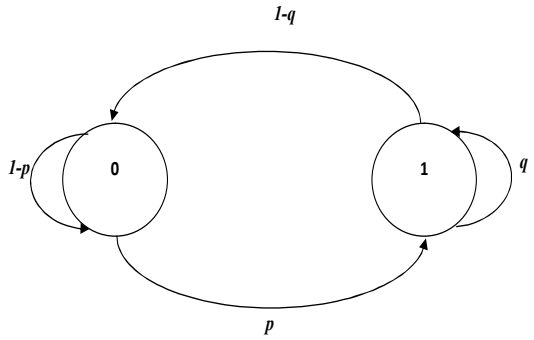

Fig. 2. Packet loss modeled by a Markov random process.

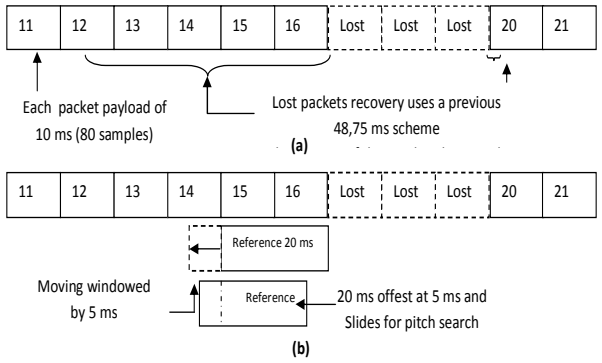

Fig. 4. (a) Diagram is shown for three packets loss and the dependencies on previous and future samples. (b) Representation of correlation windows for pitch detection

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

C. Packet loss Concealment based on the ITU-T Recommendation G.711 Appendix I

The objective of PLC is to generate a synthetic speech signal to cover missing data in a received bit stream as shown in Figure (3). Ideally, the synthesized signal will have the same timbre and spectral characteristics as the missing signal, and will not create unnatural artifacts. Since speech signals are often locally stationary, it is possible to use the signals past history to generate a reasonable approximation to the missing segment [5, 6].

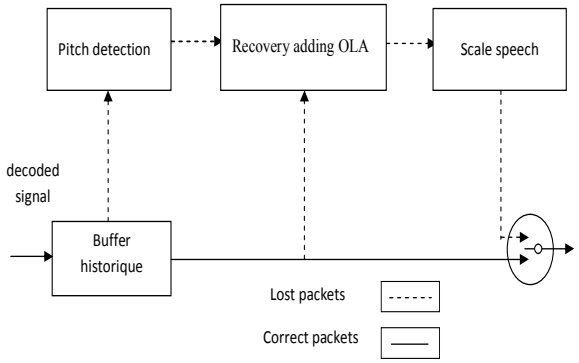

Fig. 3. Scheme based on G.711 Appendix I.

1) History Buffer: The technique PLC a copy of the decoded output is saved in a circular history buffer that is 48.75 ms (390 samples) long as shown as Figure (a .4). 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.

2) 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)as shown in Figure (b .4) [7].

3) 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 [8]. 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. The Figure (5) shows the improvement of speech quality (PESQ) with the use of PLC technology based on ITU-T G.711 Appendix I to the loss rate from 5% to 20%.

III. Automatic Speech Recognition

The automatic speech recognition is a process that converts the acoustic signal of speech in a set of words or phrases. The ASR system comprises the following steps:

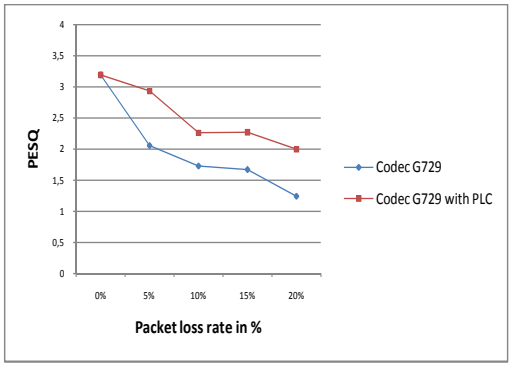

Fig. 5. Speech quality with packets loss

A. Creation of the Database ARADIGIT_G729

The speech database used in this work is the original ARADIGIT. The database ARADIGIT8K (sub sampled at 8 kHz) is then passed through the G.729 codec, which we have simulated to reach the database transcoded to G.729, Named ARADIGIT_G729.

B. Extraction of Acoustic Parameters

The extraction of signal parameters consists in associating to the speech signal a set of acoustic parameter vectors. This step involves cutting the frames during which it is assumed quasi-stationary, frame has duration of 25 ms, with an overlap between two consecutive frames of 15 ms. To reduce the side effects produced by the segmentation, frames are then multiplied by a weighting window (the Hamming window in our case). From a bank of 24 filters in Mel frequency scale, 12 parameters MFCCs (Mel-Frequency Cepstral Coefficients) are calculated for each frame. With these coefficients, the differential coefficients of the first and second order are added to form a vector of dimension equals 36 (12 MFCC +12 MFCC + 12 MFCC).

C. Learning Step

A 3-state HMM transmitters is estimated for each word (digits). The emission probability of each state is modeled by a multi-Gaussian distribution with diagonal covariance matrix, HTK uses the Viterbi algorithm to initialize the prototype models and next the Baum-Welch algorithm for driven.

D. Recognition Step

Recognition is performed by the Viterbi algorithm that calculates the likelihood between the sequence of acoustic observations (the word to be recognized) and all acoustic models re-estimated in learning step. The recognized message is the one corresponding to the acoustic model that generates the maximum likelihood.

IV. Experimental Results

We present in this part the results of the evaluation of the influence of lost frames on the speech recognition. For this we use the open source platform based on the HTK HMM. To minimize the influence of packet loss on the recognition rate, we use the mechanism for recovering lost frames PLC-based ITU G.711 Appendix I. We varied the packet loss rate from 0% to 20%, for both cases, that is to say, before and after applying masking technique packet loss PLC based ITU-T G.711 Appendix I. For a loss rate ranging from 0 to 20%, the influence of losses on the recognition rate is more observable. With PLC technology, the influence of losses on the recognition rate is less observable.

Overall, the recognition system used HTK, based on HMMs, takes as a reference during the learning step a speech signal without loss. We also note that the reconstructed signal by PLC technology is more similar to the original signal as the lost signal, although the distortions are present and can adversely affect speech recognition. By comparing the results of Figure.8, we note that the significant improvement in recognition rates using the masking technique of packet loss PLC-based ITU-T G.711 Appendix I. The results obtained in this second phase have shown that the use of PLC technology improves recognition performance in case of packet loss. The results obtained with our method show a large increase of the recognition threshold, thus the effectiveness of the implemented method is significant.

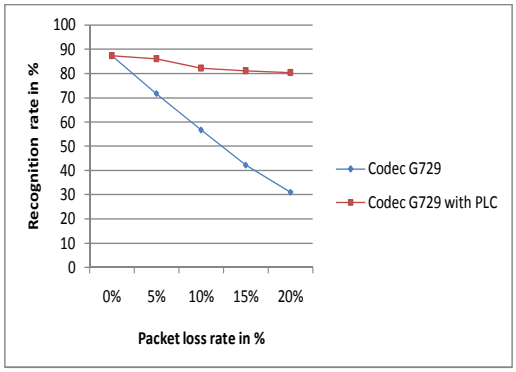

Fig. 6. Evolution of recognition rate as a function of packet loss.

V. 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 ASR in VoIP networks. After implementing the system HTK, we proposed the introduction of the PLC in the recognition 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 ASR in VoIP and make recognition 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] HTK "Hidden Markov Model ToolKit, Speech Recognition Toolkit", avalaible at:http://www.htk.eng.cam.ac.uk.

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

[3] H.Yong, Z.Jiang, "Imlementation of ITU-T G729 Speech Codec in IP Telephony Gateway", Wuhan University Journal of Natural Sciences, vol. 5, pp.159-163, 2000.

[4] B.Milner, B , S .Semnani, "Robust Speech Recognition Over Networks", IEEE International ConferenceAcoustics, Speech, and Signal Processing, pp. 1791 - 1794, vol.3, 2000.

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

[6] J. Wiley, "VoIP voice and fax signal processing", Published simultaneously in Canada, p.592, 2008.

[7] 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.

[8] 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.