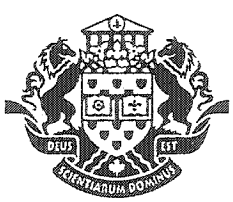
NOTE TO USERS



Université d'Ottawa · University of Ottawa

Université d'Ottawa · University of Ottawa

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES

FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

Qi LI

AUTEUR DE LA THÈSE - AUTHOR OF THESIS

M. A. Sc. (Electrical Engineering)

GRADE - DEGREE

Department of Electrical Engineering

FACULTÉ, ÉCOLE, DÉPARTEMENT - FACULTY, SCHOOL, DEPARTMENT

TITRE DE LA THÈSE - TITLE OF THE THESIS

Improved Packet Loss Concealment for PCM Voice Transmission over Internet Protocol Network

M. Bouchard

DIRECTEUR DE LA THÈSE - THESIS SUPERVISOR

CO-DIRECTEUR DE LA THÈSE - THESIS CO-SUPERVISOR

EXAMINATEURS DE LA THÈSE - THESIS EXAMINERS

R. Goubran

A. Miri

J.-M. De Koninck, Ph.D.

LE DOYEN DE LA FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES

DEAN OF THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

Improved Packet Loss Concealment for PCM Voice Transmission over Internet Protocol Network

By

QI LI

B.A.Sc., Shandong University, 2001

A thesis submitted to the

Faculty of Graduate and Postdoctoral Studies

in partial fulfilment of the requirements for the degree of

Master's of Applied Sciences in Electrical Engineering

Ottawa-Carleton Institute for Electrical and Computer Engineering

School of Information Technology and Engineering

Faculty of Engineering

University of Ottawa

July 2004

Your file Votre référence
ISBN: 0-494-01530-6
Our file Notre référence
ISBN: 0-494-01530-6

Abstract

In recent years, packet voice over the Internet Protocol (IP) network has become an attractive alternative to conventional public telephony. However, the Internet is a best-effort network, with no guarantee on its quality of service. A fundamental issue in real-time interactive voice transmissions over an unreliable Internet Protocol (IP) network is the loss, or late arrival of packets. Concealment algorithms, either transmitter or receiver based, are used to replace these lost packets. The packet loss concealment (PLC) techniques described in the standards ANSI TI.521 (Annex B) and ITU-T Rec. G.711 Appendix I produce a good performance for PCM coded speech. However, these algorithms do not use subsequent packets from the jitter buffer for reconstruction, which could further increase the PLC performance.

The goal of this work is to develop an improved PLC algorithm for PCM coded speech, using the subsequent packet information when available. A method combining pitch prediction and Linear Prediction Coding (LPC) is used to estimate the missing voice packets.

Acknowledgements

First and foremost, my heartfelt thanks to my supervisor Dr. Martin Bouchard for his invaluable guidance at every stage of my research work. I sincerely thank Dr. Rafik A. Goubran, Dr. Ali Miri and Dr. Jiying Zhao, member of the thesis committee. I also express my gratitude to the University of Ottawa staff and administration for their help throughout my studies.

I convey my special thanks to my friends Ning Ma, Qiongfeng Pan, Ying Shao, Xiaoyong Sun, Hongzhou Tan, Etienne Vincent and Wei Zhang for their constructive comments.

Last but not the least, I extend my gratitude to my father and mother for their constant support and encouragements. Without them, this thesis would have been impossible.

Contents

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION

1.1 Motivations and Challenges

1.1.1 Motivation

Traditionally, voice communication has been done through the Public Switched Telephone Network (PSTN), which is also referred to as the "Plain Old Telephone Service"(POTS). It is a time-division-multiplexing (TDM) circuit-switched network and a virtual circuit is allocated to establish a session between two end-users. In this way, bandwidth is guaranteed and so it is reliable. In a PSTN design, end points, such as telephones, are dumb devices with minimum functionalities, while its network is intelligent and handles features like access control, scheduling, and signalling.

With the exponential growth of the Internet in recent years, packet voice becomes an attractive alternative to conventional public telephony [VS02]. Generally, the basic steps to transmit voice signals over a packet-switched network at the sender side include the conversion of analog voice signals to a digital format, processing/coding/compressing the digital signals, forming packets of the resulting signals, and transmitting the packets over the network. At the other end, a receiver receives the packets, unpacks them, and decompresses/decodes/processes the signals back to their original format.

The Internet is very different from the PSTN in the sense that it moves the processing power of a network to its end points. The end points of the Internet are computers with

powerful computational capabilities and an ever-growing number of applications. The network itself is designed to be simple. The Internet Protocol (IP) delivers a packet by checking its destination address and by forwarding it to the next hop along the way, without knowing the contents. All packets competing for one outgoing link are statistically multiplexed. Functions, such as access control, scheduling, and signalling are not included. This simplicity leads to its widespread deployment and its low access cost. The constant growth of the Internet opens up great opportunities for a variety of voice applications, such as Internet Telephony, teleconferencing, and wireless voice communication. The telecommunication industry is developing a Next Generation Network (NGN) based on a common packet-based architecture for voice, data, and multimedia services [IEC02].

However, challenges such as jitter, delay, and packet loss appear when using the Internet for real-time voice applications delivering levels of quality and reliability comparable to those of the traditional telephone network.

1.1.2 Challenges

Interactive voice communications usually have strict delay requirements. It is generally accepted that one way delays of 150ms -200ms are not acceptable for VoIP. In a packet network, this requirement implies that packets delayed over a certain time limit are considered lost and cannot be used by the receiver. Consequently, from the application's point of view, those delayed packets are equivalent to lost packets.

Voice communications also have loss requirements. Comparing packet-switched networks to circuit-switched networks, loss happens in an entirely different form. In circuit-switched networks, losses happen at the bit or sample level, while in packet networks, losses happen at the packet level. Losing a packet corresponds to losing an entire interval of speech. Hence, packet losses are generally perceptible, and frequent losses

make playback intermittent and annoying. Packet losses are common in the Internet, and loss rates sometimes can be quite high. Delayed packets will further increase loss rates.

In addition, further coding complicates ways to maintain quality. In order to reduce transmission bandwidth, in certain circumstances, speech coding is employed to compress voice signals. Compression is achieved by exploiting temporal redundancies among speech signals and by realizing actual compression through quantization. Hence, speech coding is lossy. Current compression algorithms are not robust enough to transmission errors. Their sole objective is to maximize coding gain, assuming error-free channels. In particular, many low bit-rate speech coding algorithms generally incorporate complex mechanisms to remove as much redundancy as possible, which in turn introduces a great deal of time dependencies in a coded sequence.

Error resilience is not a severe problem for PSTN in which infrequent occurrences of bit errors can be corrected using error-correction codes. However, for packet networks, the loss of a speech packet can in some cases degrade playback quality not only of the lost packet itself, but also of subsequent packets [SAE96]. G.711 PCM coding is a robust coding algorithm which is deemed to be one of the most suitable coding algorithms for packet voice transmission in internet because of its low complexity, its high quality, and its good performance in tandem coding or coding of noisy speech. With PCM coding, the loss of a speech packet does not affect subsequent packets.

An advantage of voice communications over data communications is that 100 percent accuracy is not a must. This relaxation is indeed very useful because a receiver can tolerate a certain level of signal distortions without significant performance degradation. Therefore, if we can convert packet losses to voice-sample distortions, the playback quality may become tolerable.

1.2 Objective and contributions

Based on the shortcomings identified in the last section, the objective of this thesis is to design, analyze, and evaluate a robust loss concealment scheme in order to support reliable and real-time PCM voice transmissions over unreliable IP networks.

The packet loss concealment (PLC) techniques described in the standards ANSI TI.521 (Annex B) and ITU-T Rec. G.711 Appendix I produce a good performance. However, these algorithms do not use subsequent packets from the jitter buffer for reconstruction, which could further increase the PLC performance.

This research work is going to develop an improved PLC algorithm, using the subsequent packet information in the jitter buffer when available. A method based on pitch prediction and Linear Prediction Coding (LPC) will be used to estimate the missing voice packet.

The contributions of the thesis are:

· Based on the previous work from [EBA03], this research further develops schemes to improve the performance of the pitch prediction and LPC prediction concealment algorithm for PCM speech, namely the use of future packets from the jitter buffer and the tuning of the algorithm with voiced/unvoiced classification.

· Implementation of the schemes with Matlab 6.1

· Validation of the schemes with the ITU-T P.862 PESQ standard and using speech files from the ITU-T supplement P.23

· The validation of the schemes under realistic conditions such as speech distorted by additive noise or reverberation

1.3 Layout of this thesis

This thesis is organized as follows.

Chapter 2 gives a brief introduction to digital voice communications, voice transmission over IP-based networks and the problem of packet loss.

Chapter 3 first introduces some current state of the art techniques for packet loss concealment. Secondly, the ITU-T G.711 Appendix I is shown briefly for its principle and algorithm description. Finally, at the end of this chapter, the packet loss concealment scheme for PCM previously developed in [EBA03] and [EBA04] is presented, based on a pitch prediction and a LPC prediction.

In Chapter 4, a new packet loss concealment algorithm is presented. A detailed description of how the new algorithm operates in periods of normal conditions (no-loss) and in erasure periods is presented, along with a discussion of the delay consideration. The test files, test tools and test patterns are showed in a test setup section. The performance of the proposed concealment algorithm is compared to the performance of the ITU-T G.711 Appendix I, which is a PCM concealment standard, and to the algorithm previously introduced in [EBA03] and [EBA04].

In Chapter 5, the new PLC algorithm is implemented with an additional voiced/unvoiced classification. With the additional voiced/unvoiced classification, the performance can be further increased for low loss rates. Experiments and results are showed in this chapter.

In Chapter 6, the new PLC algorithm with V/UV classification is further tested for distorted speech like additive white noise distorted speech and reverberation distorted speech. Experiments show that the new PLC algorithm with V/UV classification

performs fairly well under those conditions.

Chapter 7 summarizes the work and gives some suggestions for future investigation.

Chapter 2

Background of Voice Transmission over IP Protocol

Voice transmissions over packet networks have attracted a lot of attention during recent years with the availability of fast processors and an ever increasing demand. However, their delivery quality is not satisfactory due to frequent packet losses. An active research direction is, therefore, to develop simple and robust loss concealment and coding strategies at connection end points. These approaches generally exploit redundancies in voice data and reconstruct the lost data from that received. In this chapter, the major work in this area is summarized and presented. Sections 1 and 2 present the basics of digital voice communications and networking which are relevant to this thesis. General packet voice transmission knowledge is given in section 3.

2.1 Digital Voice Communications

In this section, an overview of the production, the digitalisation and the coding of speech is presented. Sample-based and frame-based codec technologies are discussed. Here the term 'codec' refers the speech encoding/decoding system as a whole.

2.1.1 Speech Production

This section presents some basic properties of speech signals and how they are produced, especially speech properties like voiced and unvoiced sounds for their relative importance to our work.

Speech signals are non-stationary signals. At best, they can be considered as quasi-periodic over a short period of time. Thus speech signals cannot be exactly predicted. Usually, speech signals are divided into two categories: voiced and unvoiced speech.

Voiced speech are produced by pushing air from the lung through the glottis with the shape and the tension of the vocal cords adjusted so that this flow of air causes them to vibrate in a relaxation oscillation. The vibration of the vocal cords produces a sequence of quasi-periodic pulses of air that excites the vocal tract. Thus, voiced speech can be modelled by exciting a filter modelling the vocal cords. The rate of the vibration of the vocal cords' opening and closing are defined as the fundamental frequency of the phonation, whose period is called pitch period. Modifying the shape and the tension of the vocal cords can change the period of the vocal cords' vibration, i.e. the pitch. In voiced speech, certain frequency ranges are amplified by resonance within the vocal tract. Thus peaks of amplitude at some defined frequencies appear in the signal's spectra. So, the voiced speech has quasi-periodic characteristics in the time domain. The energy of voiced speech is generally higher than that of unvoiced speech. These properties make the voiced speech more important than unvoiced speech.

Unvoiced speech is generated by forcing a steady flow of air at high velocities through a constriction region in the vocal tract to produce turbulence. The location of the constriction region determines what unvoiced speech is produced. Unvoiced speech is similar to random signals and has a broad spectrum in frequency domain. Usually, random signals like white noise are used to model the excitation of unvoiced speech.

2.1.2 Digitization

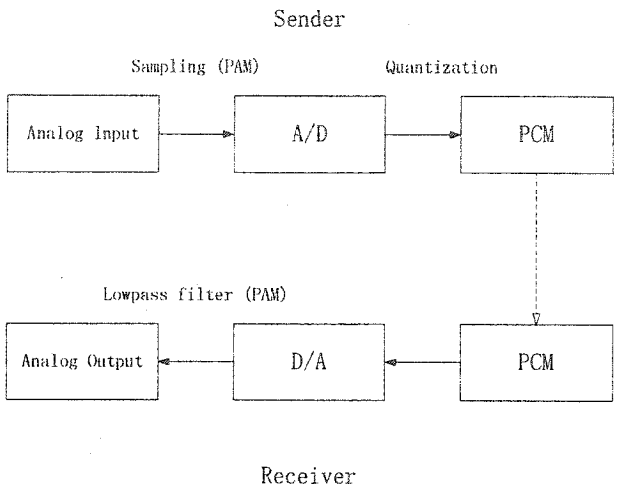


FIGURE 2-1: DIGITAL VOICE TRANSMISSION SYSTEM USING PULSE CODE MODULATION (PCM)

In Figure 2-1, an analog speech signal is converted into a digital signal at the sender. After transmission, the digital signal is converted into an analog signal at the receiver.

At the sender side, the analog signal is put through a low pass filter to avoid aliasing when sampling. Then the signal is sampled at a certain sampling frequency and thus converted into a modulated signal. Typical voice sampling frequency is 8 kHz, which is sufficient for telephone quality speech. Then the modulated signal is converted into a digital representation. This process is called quantization which implies that an analog amplitude with infinite resolution within its allowed range is mapped to one value of a discrete set of values, i.e. 16 bit quantization is a set of $2^{16} = 65536$ values.

At the receiver side, the digital representation is converted back to an analog signal. The signal is put through an interpolation low pass filter with the same cut off frequency as in the sender. Distorted by the approximation process of the A/D quantization, the original signal is then recovered.

2.1.3 Sample Based Codec

Sample based codecs try to directly encode speech signals in an efficient way by extracting redundancies and exploiting the temporal and or spectral characteristics of the speech waveform. The simplest waveform codec is Pulse Code Modulation (PCM) where the amplitude of the analog signal is quantized to one of a discrete set of values. PCM is a memory less coding. Therefore the bandwidth that it requires to transmit a speech signal is high. A first step to reduce this bandwidth while maintaining the same output quality is to employ non uniform quantization, i.e. the quantization step size varies with the signal value. This improves the quality for two reasons: first, frequently occurring amplitudes can be quantized finer and second, the human hearing exhibits logarithmic sensitivity, i.e. the perception of small amplitudes is more critical and they should be quantized finer. Typically, the non-uniform quantization is employed according to either $\mu - law$ or $A - law$ logarithmic curves [G711].

An encoding scheme which exploits the fact that the speech waveform is evolving slowly (i.e. the adjacent samples are correlated) is the differential PCM (DPCM). In its simplest form, the sender encodes the difference between two adjacent samples and the receiver restores the signal by integration. However, actual DPCM systems employ a larger predictor filter than a memory of 1 sample. The transfer function of the predictor filter in the z-domain can be computed as follows:

$$A(z) = \sum_{i=1}^{p} a_i z^{-i} \qquad (2\text{-}1)$$

where a_i are the filter coefficients.

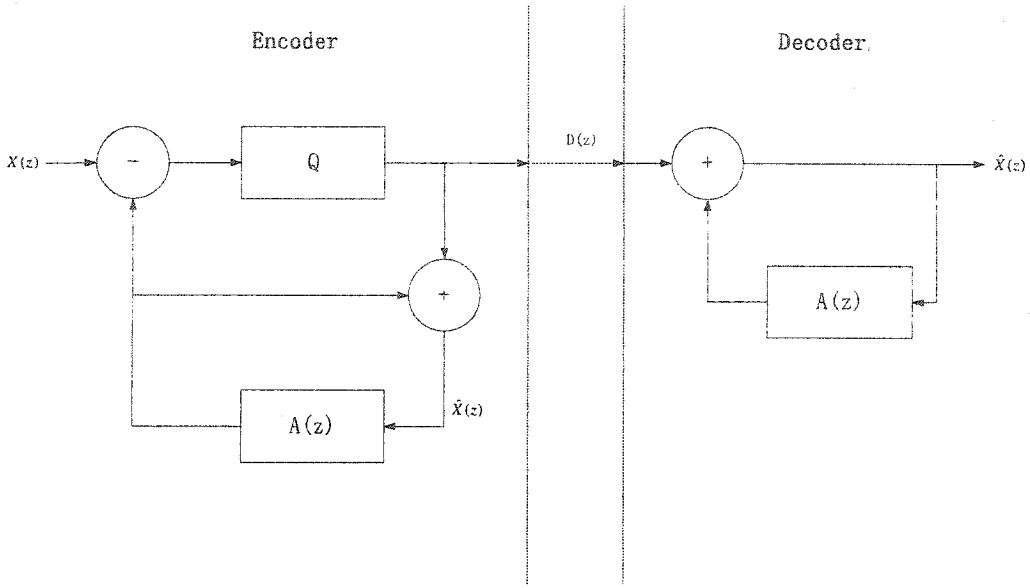


FIGURE 2-2: DIFFERENTIAL PULSE CODE MODULATION (DPCM)

Figure 2-2 shows the encoder and decoder structure of a DPCM system. At the encoder, the difference between the input speech sample $x(n)$ and its estimate $\hat{x}(n)$ is computed and transmitted to the receiver. There the signal is reconstructed using the same predictor filter loop as in the encoder.

In Adaptive Differential PCM (ADPCM), both the quantizer step size and the predictor filter coefficients are varied adaptively to the speech signal content. Typically, the predictor filter adaptation is estimated from the received signal. Thus only the quantizer step information has to be transmitted additionally.

2.1.4 Frame Based Codecs

Frame based codecs attempt to model speech signals by a set of parameters and then try to efficiently encode these parameters. They usually operate on "frames" which represent a fixed number of speech samples. They typically operate at a lower bit rate than waveform codecs at the cost of higher complexity. Examples of frame based codecs

include ITUT G.729 and G.723.1 codecs.

2.1.5 Linear Prediction Coding, Pitch Prediction and Voicing Decision

2.1.5.1 Linear Predictive Coding

Due to the fact that the new packet loss concealment algorithm to be proposed in this thesis will be based on the Linear Prediction (LP) process, this section is dedicated to this topic. Also, Linear Prediction Coding (LPC)-based coders exhibit excellent immunity to erasure thanks to their inherent concealment ability.

The analysis/synthesis method known as Linear Predictive Coding (LPC) was introduced in the sixties as an efficient and effective mean to achieve synthetic speech and speech signal communications. For the past years, LPC has been considered one of the most powerful techniques for speech analysis. In fact, this technique is the basis of other more recent and sophisticated algorithms that are used for estimating speech parameters, i.e., pitch, formants, spectra, vocal tract and low bit representations of speech.

Linear Predictive Coding (LPC) is one of the methods of compression that models the process of speech production. Actually, LPC models speech production process as a linear sum of earlier samples using a digital filter with an excitation signal as the input. An alternate explanation is that linear prediction filters attempt to predict future values of the input signal based on past signals.

A speech sample $x(n)$ is estimated by a linear combination of previous speech samples:

$$\hat{x}(n) = \sum_{i=1}^{p} a_i x(n-i) \qquad (2\text{-}2)$$

This estimation is referred to as linear prediction (LP) of order p. It amounts to filtering the signal with a filter (predictor filter) with the transfer function

$$A(z) = \sum_{i=1}^{p} a_i z^{-i} \qquad (2\text{-}3)$$

Many approaches exist for computing the coefficients a_i i.e. minimization of the mean square of the difference between the original and the estimate

$$d(n) = x(n) - \hat{x}(n) \qquad (2\text{-}4)$$

A linear prediction of sufficient order with optimally chosen coefficients yields a de-correlated difference signal $d(n)$. If the speech signal is filtered with the optimal prediction error filter (analysis filter) $1 - A(z)$ we get an output signal with a flat spectrum

$$D(z) = \frac{1}{g}(1 - A(z))X(z) \qquad (2\text{-}5)$$

where $\frac{1}{g}$ is a scaling factor computed from the variance of the difference signal $d(n)$.

If the inverse filter $H(z) = \frac{1}{1 - A(z)}$ (synthesis filter) is excited with a signal with a white spectrum, the output signal $\hat{X}(z)$ represents an approximation of the speech signal

for which the predictor coefficients have been optimized. In vocoders this behaviour is approximated by exciting the synthesis filter $H(z)$ with a periodical train of pulses.

The computation of the LPC coefficients is usually performed through the Levison-Durbin recursion [PM96], from the autocorrelation of an input speech sequence.

2.1.5.2 Pitch Determination

One of the most important parameters in speech analysis, synthesis, and coding applications is the fundamental frequency, or pitch, of voiced speech. Pitch frequency is specific to the speaker and sets the unique characteristic of a person. Voicing is generated when the airflow from the lungs is periodically interrupted by movements of the vocal cords. The time between successive vocal cord openings is called the fundamental period or pitch period.

For men, the possible pitch frequency is usually found somewhere between 50 and 250Hz, while for women the range usually falls between 120 and 500Hz. In terms of period, the range for a male is 4 to 20ms, while for a female it is 2 to 8ms.

The pitch period must be estimated at every frame. By comparing a frame with past samples, it is possible to identify the period in which the signal repeats itself, resulting in an estimate of the actual pitch period. Note that the estimation procedure makes sense only for voiced frames. Meaningless results are obtained for unvoiced frames due to their random excitation nature.

The design of a pitch period estimation algorithm is a complex undertaking due to lack of perfect periodicity, interference with formants of the vocal tract, uncertainty of the starting instance of a voiced segment, and other real-world elements such as noise and echo. In practice, pitch period estimation is implemented as a trade-off between computational

complexity and performance. Many techniques have been proposed for the estimation of the pitch period. In this thesis, considering the computational complexity, a simple pitch determination idea is applied and described later.

2.1.5.3 Voicing Decision

The normal telephone speech contains a lot of silence periods, i.e. while the user is listening, or in the case of inter-word silences and even inter-syllable silences. It has been found out that for 40% to 60% of the call duration there is no information in a given direction. This causes an ineffective usage of bandwidth.

So if voice activity detection is done at the transmitter, and packets are transmitted only when there is some information i.e. when the speech source is active, bandwidth utilization can be optimized. Voice is transmitted in the form of bursts, with silence periods in-between them.

2.2 Protocol and Network Hierarchy for Supporting Real-Time Voice Transmission

Since voice transmission over packet networks is the main interest of this thesis, some discussion of the network and protocol environments for real-time voice transmission is useful. Figure 2-3 shows the current Internet protocol stack for supporting real-time applications.

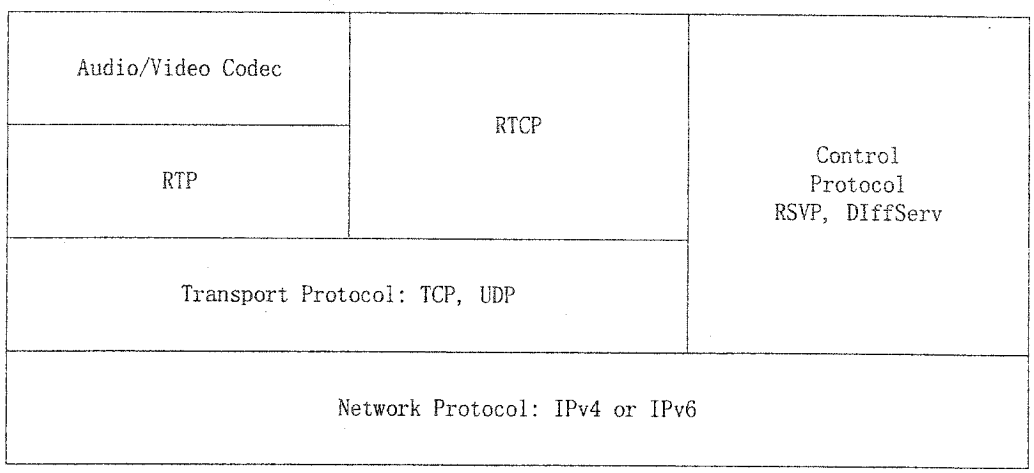


FIGURE 2-3: CURRENT INTERNET PROTOCOL

· Network-layer protocol: Currently, IP version 4 (IPv4) only supports best-effort service. Its fundamental idea is to do everything possible to deliver each packet, without guaranteeing its actual delivery or its delivery time. As indicated before, losses are common, and variations in delay further worsen the packet-loss problem. The next generation network-layer protocol is IP version 6 (IPv6) with several advanced features. However, the IPv6 specification is rather vague with regard to how packets of a given flow are treated, stating that routers may or may not treat packet regarding flow labels. A source using a flow label may, therefore, have no guarantee that its packets will receive special attention from the network. In other words, it cannot be assumed that the concept of flows in IPv6 will have significant performance impact on the network [ME00]. As for the traffic-class or type-of-service (ToS) field, the specification states that there are no common agreements on the type-of-traffic classifications that are most useful for IP packets. Currently, there are very few vendors supporting the flow and traffic-class features. Therefore, these innovations in IPv6 alone are not enough to support QoS, unless all routers and gateways implement these features and have special mechanisms to meet service requirements.

There are two major service proposals in the literature to assist IPv6 in enhancing real-time support in the Internet: integrated service (IntServ) [BCS94] and differentiated service (DiffServ) [BB98]. For either of these, further research and development is required to make them practical. It can be concluded that strict QoS is not guaranteed.

• Transport-layer protocols: Transmission Control Protocol (TCP) and User Datagram Protocol (UDP) are mainly used in the transport layer. TCP is connection oriented, or circuit switched, and is responsible for the correct delivery of data by detecting lost data and by triggering end-to-end retransmissions. Although TCP is reliable, it may result in unacceptable delays when transmitting real-time data. On the other hand, UDP is connectionless, with no guarantee of delivering packets to their destinations. With no concept of connection, packets arriving at the destination may be out of order. Hence, UDP requires the application to rearrange packets and take care of losses.

These two transport protocols are two extremes when reliability is concerned. TCP guarantees total reliability without considering delays, and is normally not a good choice for real-time applications. In contrast, UDP guarantees no reliability and, adds little extra delay. It is well recognized that UDP is preferred for handling real-time traffic. Hence, the packet-loss problem will need to be solved.

• Application-level standards: On top of the transport layer, there are application-level standards developed to support end-to-end real-time transmissions and ensure interoperability among different products. The international standards that are closely related to packet-voice transmissions are H.323 [H323], H.225.0 [H225], and H.245 [H245]. All three standards belong to ITU-T recommendations and are covered under the umbrella standard H.323.

H.323 is a standard that specifies the components, protocols, and procedures for providing multimedia communication services, including audio, video, and data

communications, over packet networks, such as IP networks. The standard describes H.323 system components, or entities, including terminals, gateways, gatekeepers, multipoint controllers, multipoint processors, and multipoint control units.

An H.323 terminal can be either a personal computer or a stand-alone device. It must support audio communications and can optionally support video or data communications. H.225.0, the standard for the call-signalling protocol and media-stream packetization for packet-based multimedia communication systems, defines how to use the Real-time Transport Protocol (RTP) [SC96] and RTP control protocol (RTCP) to packetize audio and video data and achieve synchronization. RTP is an application-layer protocol that provides functionalities like rearranging of packets, loss detection, delay adaptation by buffering, and playback-point determination.

RTP uses RTCP to monitor and convey statistics about an ongoing session. Although it is called the Real-time Transport Protocol, RTP does not contain any mechanism to support timely delivery of data, nor does it provide any recovery scheme for packet losses. Such losses will need to be handled by the applications themselves.

In short, none of the network-layer protocols, such as IPv4 and IPv6, transport-layer protocols, such as TCP and UDP, and application-level protocols, such as RTP and H.323, truly provides practical QoS support and/or can solve the fundamental packet-loss problem for real-time voice transmission. To achieve better voice transmission quality, it is therefore both necessary and natural to develop packet loss concealment schemes.

2.3 Packet Voice Transmission

This section mainly presents the general knowledge for packet voice transmission. The first part introduces the design history and motivation of voice communication over packet-switch network. The second part shows the approaches for packet voice

transmission. The quality impairments of packet voice transmission such as jitter, delay and packet loss are analysed in the third part.

2.3.1 History and motivation of voice communication over packet-switched network

In the last few decades, technology has developed faster than ever before. Among these great technology progresses, two of them are particularly relevant for this thesis. One is the great evolution of the packet-switched network, which started as an experimental platform at the beginning for scientists and developed later into the global interconnection known as the Internet. The other is the digital representation for voice signal and its further processing, storage and transmission. The two technologies began to converge around 1980s with research and experiment on *packet voice* ([COH80]). The history is shown in figure 2.4.

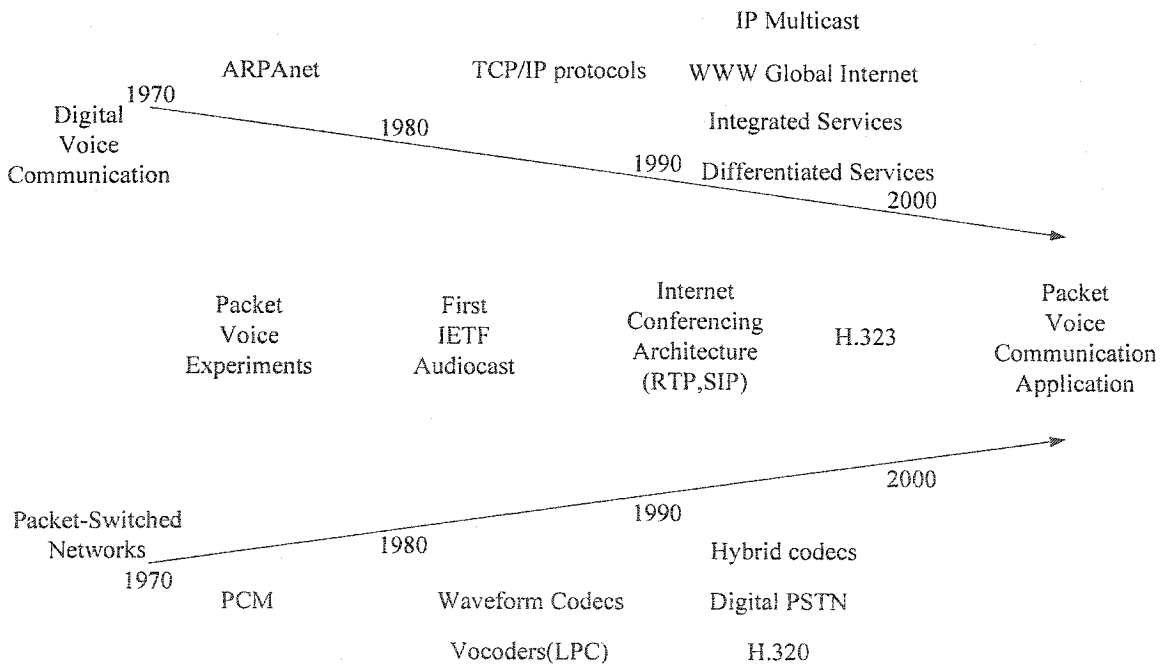


FIGURE 2-4: PACKET VOICE COMMUNICATION HISTORY

Nowadays, packet voice communication is considered as the next stage of the development of the telephone network and the first choice of incarnation for an "integrated services" network.

Despite the advantages of packet voice communication, there are still a number of barriers keeping it from more widespread usage. The packet-switched network, as indicated in the previous section, mostly guarantees the QoS for plain data transmission which does not have too much requirements on delay. It cannot meet the QoS for real time communication like voice and video perfectly. This situation thus motivates people for further developments of packet voice communications technology:

· Improving efficiency by multiplexing: Multiplexing connection can offer an increase in the number of connections per carrier. Thus the operator will be able to provide more services and applications.

· Improving efficiency by voice compression and voice activity detection: Temporal redundancies exist in the correlation between adjacent speech samples (short term correlation) and pitch periodicity (long term correlation). Additionally, the different sensitivities of the human hearing system in different frequency bands can be also exploited for compression. So, quantization of relevant samples (or coefficients) and prediction filters can be used to achieve compression of speech signal with some quality/complexity versus bit-rate tradeoffs.

· Voiced/unvoiced classification technology allows a data network carrying voice traffic to detect the absence of conversation and conserve bandwidth by reducing the transmission of silent or noise packets over the network. As previously mentioned, it is known that most connections include about 40%-60% of absence of conversation for pausing, thoughts or silence. Voiced/unvoiced decision can monitor signals for voice activity so that when absence of conversation is detected for a specified amount of time, the

application indicates to the encoder to change the transmission scheme over the network and thus improves the network efficiency.

· Private Data Networks: Private Data Networks which are used to carry data between offices within organizations or companies can also be used to transmit telephone calls or even video conferences due to the technology developments. By using private data networks for transmitting telephone calls, a large amount of toll fees can be bypassed for the organizations or companies.

2.3.2 Approaches for packet voice communication

Packet voice communications have several application forms: PC to PC, Phone to Phone, PC to Phone, web to Phone etc. The following picture shows the Phone to Phone way through gateways.

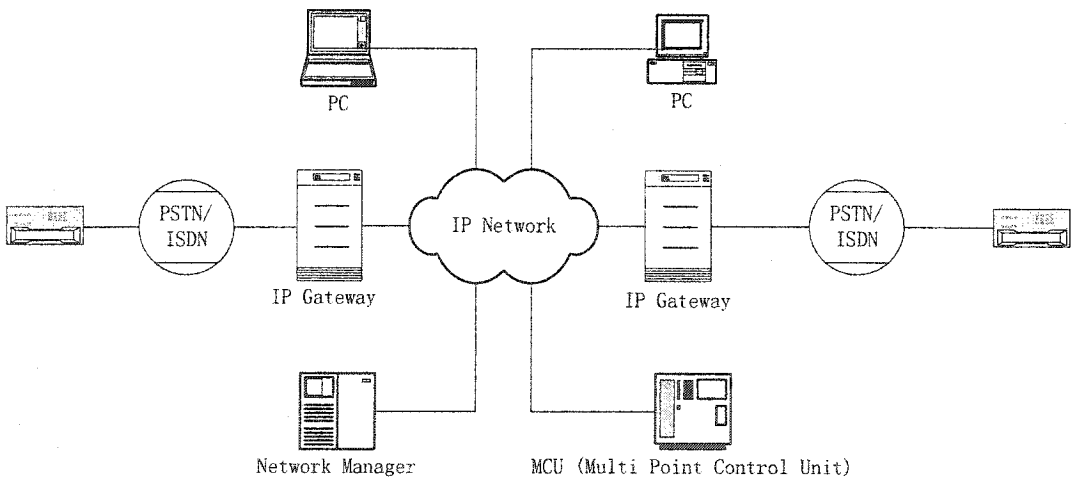


FIGURE 2-5: PHONE TO PHONE PACKET VOICE COMMUNICATION THROUGH GATEWAY

At the sender side, the sender transforms the analog voice signal from PSTN into a digital signal, codes and compresses the digital signal, packs the signal with TCP/IP protocol and transmits the packets through a TCP/IP network. At the receiver side, the

receiver rearranges the packets, decompresses the packets, decodes and transforms the digital signal into analog voice signal again.

Currently there are three categories of products for packet voice communication:

· Software solutions include *Netmeeting*™ from Microsoft, *Internetphone*™ from VOCALTEC or *CoolTalk*™ from Netscape, etc.

· DSP card solutions include *PhoneJack*™ from Quicknet Technologies, etc. Due to the fast growth of DSP chips processing ability, more and more algorithms are allowed to be implemented to improve the voice quality. The algorithms can be voice coding and decoding, echo cancellation, packet loss concealment, etc.

· Internet Telephone Gateway solution. ITG (Internet Telephone Gateway) communications can be realized for PSTN to PSTN, PSTN to PC and PC to PSTN.

2.3.3 Quality Impairments

Many quality impairments appear in packet voice transmission such as delay, jitter and packet loss. This section analyzes the properties and generating reasons of these impairments.

2.3.3.1 Delay

As previously mentioned, the International Telecommunication Union (ITU) Recommendation G.114 specifies a "one-way transmission delay" of up to 400 ms to be acceptable. In a packet network, this requirement implies that packets delayed over a certain time limit are considered lost and cannot be used by the receiver. Consequently, from the application's point of view, those delayed packets are equivalent to lost packets. A

voice packet may experience many kinds of delays:

· Propagation delay (Physical layer): The time cost for the physical signal travelling across the links along the path taken by the data packets.

· Forwarding delay (network layer): The time cost for the router to forward a packet. It can be composed of extraction of the destination address from the packet header, routing lookup and switching the packet over the router's backplane from the input to the output. Also, the time needed to send the packet completely out of the output port can be counted as forwarding delay.

· Queuing delay (link layer and network layer): The time cost for a packet waiting in the queues at the input and output before it can be processed. Specifics of the link layer can also be accounted for the queuing delay, i.e. an Ethernet collision or the segmentation/reassembly process between cells and packets in ATM (Asynchronous Transfer Mode).

· Packetization/de-packetization delay (all layers): The time cost for building up data packets at the sender as well as strip off packet headers at the receiver. The time used for waiting the arrival of a sufficient amount of data from the upper protocol layer, to compute and add headers at each layer can be seen as such kind of delay. Packetization and de-packetization delays can be minimized by using efficient protocol implementations such as avoidance of actual copy operations, proper alignment of header fields etc.

· Encoding delay (application layer): The time cost for digitizing and encoding speech signal at the sender. Usually, encoding works on a sequence of PCM samples and it has to wait for enough samples to arrive. Some codecs also need to buffer data in excess of the frame size (look ahead).

· Decoding delay (application layer): The time cost for performing decoding and

conversion of digital data into analog signal at the receiver.

2.3.3.2 Jitter

Jitter can be considered as a delay variance and is usually caused by the variation of the queuing delay component. When several packets in a router compete for the same outgoing link, only one of them can be processed and forwarded while the others have to be queued. This results that packets sent by the sender at equivalent intervals arrive at the receiver at no equivalent intervals. With a non real time operating network, all the delay components introduced above can exhibit variation. At the application layer, jitter can be modified by keeping the received packets in a play out buffer and by adding an extra amount of delay.

But there is a trade off for the additional delay: on one hand, it should be small enough to have no impact on the interactivity of voice applications; on the other hand, it should also be large enough to smooth out the jitter and to enable most of the delayed packets to arrive before their play out time. If the packet does not arrive before play out time, it is considered to be lost. A lot of research has been done on playout buffer algorithms. [RK94]

From the above, we can see that delay and jitter play very important roles to QoS in packet speech transmission. But the most important impairment for packet voice transmission is packet loss.

2.3.3.3 Packet Loss

Packet loss can happen when a router becomes congested, i.e. the router gets more packets to forward than it can process. Large loss bursts happen when network pathologies occur, i.e. a router of a link does not function. There is also packet loss caused by

transmission errors (bit errors) of the underlying media. For speech transmission over Internet Protocol (IP) networks, late packets are considered as lost packets.

Packet loss is a frequent and serious problem that packet speech transmission must solve. It can be detected by using the message sequence number of transport protocols such as RTP. In order to have an acceptable QoS, packet loss concealment should be performed. This is discussed in the next chapter.

Chapter 3

Packet Loss Concealment Schemes Review

The methods of voice packet loss concealment are presented in a more detailed way in this chapter. Recovering losses involving both the sender side and the receiver side is firstly presented. Then, methods to recover losses only involving the receiver side are addressed. Later, the ITU-T G.711 Appendix I concealment scheme for PCM coded speech is presented. Finally, a recently published concealment scheme for PCM using pitch prediction and LPC prediction [EBA04] is introduced.

3.1 Sender Based Packet Loss Concealment Schemes

This group of methods involves both the sender and the receiver. Thus it has more opportunities to influence the QoS, but it requires the sender to be involved.

3.1.1 Retransmission

Most research papers do not consider retransmission as an applicable method because of typical delay constraints. However, some work in the context of video has shown that retransmission can be used also for real time communication to avoid the effect of error propagation [RHE98]. A retransmitted packet might not be used for direct play out, but it can be used to update the internal decoder state.

3.1.2 Interleaving

Interleaving can be a simple method to increase the quality of speech packet transmission [RAM70, MYT87, PER99, PH98], i.e. sending parts of the same signal segment in different packets, thus spreading the impact of loss over a longer time period. Due to the long term correlation property, interleaving is useful in enhancing the speech quality. However, interleaving always needs buffering of the generated data at the sender and rearranging them at the receiver. By doing so, a higher latency is introduced.

An extreme case of interleaving is when the unit is equivalent to one sample. [JC81, JAY93] proposed to put consecutive samples of a waveform coder into two different packets and combine the operation with loss concealment. The speech signal is partitioned into sequences of x(n). The samples with even indices x(2n) are put into one packet. The odd samples x(2n-1) are put into another packet. If one of those two packets is lost, the missing samples can be interpolated using the samples of the respective other packet. By doing so, it is possible to recover the important low frequency parts of the signal. The overhead computation is low in this circumstance as well.

[IV95] presents a similar system using DPCM encoding. The decoder consists of three sub-decoders with different transfer functions which are used depending on if only the first, the second or both packets are received.

3.1.3 Forward Error Correction (FEC)

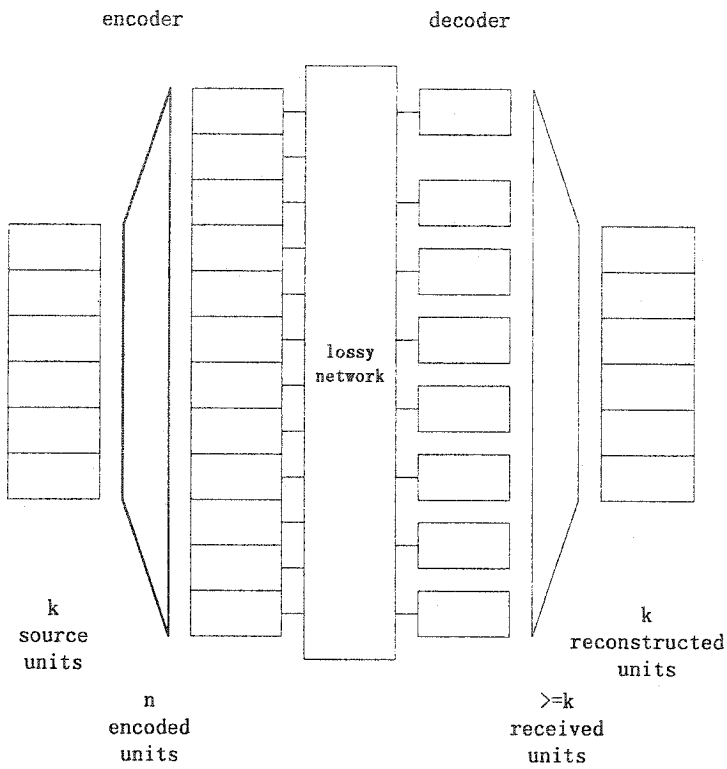


FIGURE 3-1: PRINCIPLE OF FORWARD ERROR CORRECTION

Different from interleaving methods which just change the way in which the data are transmitted to the receivers, Forward Error Correction (FEC) adds redundancy for the recovery of lost packets at the receiver. Usually, FEC can be formulated as follows: when some redundancy encoding is applied over k units resulting in n-k redundancy units and n units transmitted, the information can be fully recovered if at least k out of n units is received [RI97].

FEC has been showed to improve speech quality in VoIP by [HS95] and [PRM98]. [BCV95] have made an adaptive error control for FEC that modifies the redundancy sent based on the RTCP reports sent by the receiver. Subjective speech quality is also taken into account. The concept is tested in a conference application, with redundancy, over the

Internet.

3.1.4 Adaptive Algorithms

As an end to end intra flow QoS solution, either a rate adaptive sender or the transmission of the signal encoded in several layers with adaptive receivers can also be employed. [BG96, MJV96]

3.1.4.1 Sender Adaptive Algorithm

Rate adaptation [SCFJ96] means varying the coder output bit-rate according to loss reports by receivers. [BG96] proposed to switch between available codecs for non-continuous bit-rate adaptation. G.726 and GSM also support multi-rate transmission.

3.1.4.2 Receiver Adaptive Algorithm

Receiver-based adaptation presumes that the signal which is transmitted is decomposed into several "layers" of which at least one is decodable on its own. Furthermore, it is necessary that receivers can request the number of layers they want to receive. The IP Multicast architecture offers a suitable framework, as the individual layers can be mapped to different multicast addresses. Then, receivers can join these groups to receive the traffic. If they leave a group and nobody else requested the delivery of data belonging to that group, the multicast delivery tree is pruned back and the subnet of the receiver (and possibly upper branches of the tree) will be relieved of the traffic associated with that group, decreasing the probability of congestion. For voice, besides the problem of a suitable codec for such schemes, the gain in flexibility might not justify the layer resynchronization overhead. It should be noted that the described loss avoidance mechanism is closely related to approaches which map the layering on prioritization.

3.2 Receiver Based Packet Loss Concealment Schemes

The receiver based voice packet loss concealment schemes do not introduce packet overhead and are independent of the techniques at the sender side. It should also be noticed that sender based voice packet loss concealment schemes introduce more processing delays and look ahead delays.

Therefore, in practice, receiver based voice packet loss concealment schemes are more preferable for their property of not introducing additional implementation and processing overhead at the sender, and being well suited for heterogeneous multicast environments. This gives the sender more flexibility to choose different audio tools. Receivers can also mitigate packet loss according to their specific quality requirements. Thus backward compatibility and simple deployment can be well guaranteed.

However, to produce a high output speech quality, successfully received packets around the lost packet are necessary. This can result in additional play out delays. Usually, the delay introduced by a loss concealment algorithm is at least corresponding to one packet length for the reason that the algorithm is triggered only when a missing packet has been detected. If the packet following the missing packet is needed only for detection and not for the concealment operation itself, the concealment algorithm could be started immediately after the receipt of the previous packet and prepare a replacement packet without any indication if the packet under consideration will really be lost. This process constitutes a trade-off: on one hand, it can be a permanent computation load. On the other hand, a lower play out delay can be the benefit.

Because the fixed packetization interval is unrelated to the "importance" of the packet content and to changes in the speech signal, some parts of the signal cannot be recovered properly due to the unrecoverable loss of entire phonemes.

Some common receiver-based PLC schemes are described in the following sections.

3.2.1 Silence Substitution

This loss recovery scheme just fills the lost packet with samples of value 0 to maintain the speech timing sequence [GL86]. Due to its low complexity, this method is frequently used. But the packet loss tolerable by this method is less than 2%.

3.2.2 Noise Substitution

Noise substitution can be seen as a further development of silence substitution with the slightly increased complexity of using noise to fill the gap. Human brain tends to subconsciously repair the missing segment with the correct sound when noise substitution instead of silence substitution is filling in the gap [PHH98].

The RTP profile defines a generic comfort noise [SC03]. In addition to receiver based noise generation, it is possible to use information transmitted by the sender for appropriate noise generation. This is proposed in the context of silence detection, where during silent periods or inactive speech periods in the case of noisy environments, the sender sends "comfort noise" indication packets for appropriate noise generation during the silent periods. Several codecs like G.723.1, G.729 and GSM have codec specific comfort noise data that are triggered by specific bits in the coded data stream.

3.2.3 Packet Repetition

This scheme repeats the most recently received packet to approximate the missing waveform. All it has to do is buffering a copy of the last packet. Because the packetization interval is not related to the speech pitch period, discontinuities in the signal will occur. Reverberating sound will also occur due to repeating exactly the same speech signal.

Together with discontinuities and reverberating sounds, packet repetition schemes perform slightly better than silence substitution and can only tolerate a packet loss rate of 4% [BL92]. Packet repetition may actually produce worse performance in real networks than silence substitution, because of non-linearities that they introduce and because they can be considered as an echo for echo cancellers in the network.

3.2.4 Pattern Matching

This scheme repeats a correctly received signal segment of which maximum similarity with the lost segment is assumed [GL86, GWDP88].

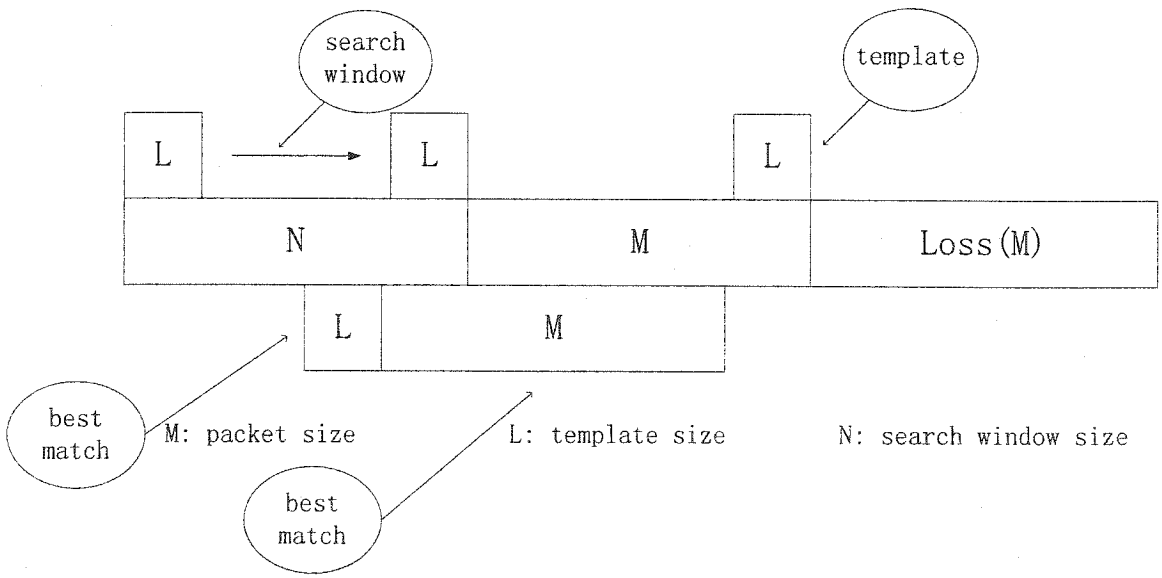


FIGURE 3-2: ILLUSTRATION OF ONE SIDED PATTERN MATCHING METHOD

The one side pattern matching method works as follows. The goal of the algorithm is to search previous packets to find a packet that most resembles the lost segment. The segment of L samples just before the missing packet is chosen as a template. Then a window containing N samples is searched for a candidate segment containing L samples that has the minimal normalized absolute difference with the template. When the candidate

segment is found, the M samples that immediately follow the candidate are extracted to fill the gap. To improve the reconstructed speech quality, the amplitude of the substitution segment is multiplied by the ratio of the average amplitude of the template to that of the candidate segment. There is also a two sided pattern matching method which is similar to the one sided pattern matching. However, this method can cause an uncomfortable clipping noise if there is a transition from unvoiced to voiced speech during the recovery. The complexity of pattern matching is higher and the tolerance of packet loss rate is about 6%.

3.2.5 Pitch Waveform Replication

Pitch waveform replication can avoid echoes because it repeats only one pitch period found in the most recently received packets. This method measures the pitch period of the signal content immediately preceding the gap and copies a sequence of samples from a previous pitch period until the gap is filled.

An extension to this technique called Phase Matching [VA89] provides for synchronization on both edges of the substitute. The patch period is measured both before and after the gap. Then the repetition of the samples sequence is compressed or expanded in time to be in phase with the following signal segment. The amplitude of the repeated segments is adapted according to the difference of the amplitudes before and after the gap.

Reverse Order Replicated Pitch Periods Algorithm (RORPP) can be seen as a combination of pitch waveform replication and pattern matching [TEL99]. It is basically similar to pitch waveform replication for short signal segments. For longer missing segments, the search algorithm uses earlier segments of pitch period length for recovery. To avoid discontinuities, the algorithm uses the "overlap-add" technique, also called "packet merging". Some very short segments at the edge of a correctly received speech packet are multiplied with complementary windows. By doing so, the transition between

the received and the replacement speech is smoothed.

Beside the above schemes for packet loss concealment, yet another scheme is to modify the underlying network and thus provide a more robust channel. New technologies such as Resource Reservation Protocol (RVSP) and the Differentiated Service architecture have been proposed for this direction. But these schemes require further investments and for now they are only applicable in private networks. IPv6 is not applied widely. In the future, they may become available for public internet, but they will not be further discussed in this thesis.

3.3 G.711 Appendix I Concealment Scheme for PCM Coded Speech

For the rest of this thesis, the focus will be on receiver-based packet loss concealment schemes for PCM coded speech. This section describes a benchmark for such concealment: the ITU-T G.711 Appendix I concealment scheme.

ITU-T G.711 Appendix I [G.711I] provides a method for recommendation G.711 (PCM) with the objective of generating a synthetic speech signal to cover missing data packet in a received bit stream. In ideal circumstances, the synthesized signal should have the same timbre and spectral characteristics as the missing signal and no unnatural artefacts should be heard. Due to the local stationary property of speech signals, it is possible to achieve an approximation of the missing segment by using the signals' past history. In circumstances of not too long erasure and a region where the speech signal does not change too fast, the recreated signal can produce a very good quality.

In the G.711 Appendix I, the G.711 coded audio data is assumed to be partitioned into 10 ms frames (80 samples with the 8 kHz sampling rate used in the G.711). With some

parameter modifications, it can be adjusted to operate under other circumstances.

3.3.1 Good Frames

For the operation with good frames, the receiver decodes the received packet and sends its output to the audio port. With the PLC algorithm implemented, some changes are made to the receiver:

· G.711 Appendix I creates a circular history buffer of 48.75ms (390samples) long to save a copy of the decoded output. No delay is introduced by this buffering and it is only used to calculate the current pitch period and extract waveforms during an erasure.

· G.711 Appendix I also introduces a delay of 3.75ms (30samples) before the output is sent to the audio port. This algorithmic delay is used for an Overlap Add (OLA) at the start of an erasure and it allows the PLC code to make a smooth transition between the real and synthesized signal.

3.3.2 First Bad Frame

At the start of an erasure, a non-circular buffer *pitch buffer* is created and it copies the content of the circular history buffer.

Also, a non-circular buffer *lastq buffer* is created for the circumstance where the erasure lasts longer than 10ms. An additional copy of the most recent 1/4 pitch period is put into this buffer.

3.3.3 Pitch Detection

The pitch information is used in the G.711 Appendix I and it is abstracted by finding the peak of the normalized cross-correlation of the most recent 20ms of speech in the

circular history buffer with the previous speech at taps from 5 ms (40 samples) to 15 ms (120samples). This corresponds to frequencies of 200 Hz to 66 Hz.

3.3.4 Synthetic Signal Generation for the First 10 ms

In the first 10 ms of the erasure, the best results can be obtained by generating the synthesized signal from the last pitch period with no attenuation. Only the most recent 1.25 pitch periods of the pitch buffer are used during the first 10 ms. A transition between the real and synthetic signal may exist and can result in an audible effect. To smooth the transition, an Overlap Add (OLA) is performed using a triangular window on 1/4 of the pitch period between the last and the next to last pitch period. The 1/4 wavelength of the signal starting at 1.25 pitch periods from the end of the pitch buffer is multiplied by an up-sloping ramp and is added to the last 0.25 pitch period in the *lastq* buffer multiplied by a down-sloping ramp.

The result of the OLA replaces both the tail of the pitch buffer and the tail of the history buffer. It is also output by the receiver during the tail of the last good frame, replacing the original signal. The algorithm delay is introduced here: the tail of the last frame cannot be output until it is known whether the next frame is erased or not. If an erasure occurs, the signal in the tail of the last good frame is modified by the OLA to ensure a smooth transition to the synthesized signal.

The synthesized signal for the 10 ms during the erasure is generated by placing a pointer one pitch period back from the end of the pitch buffer, and copying the samples to the output. If the pitch period is shorter than 10 ms, when the pointer rolls off the end of the pitch buffer the pointer is set back exactly one pitch period before continuing. If the pitch period is short (the frequency is high), the last pitch period in the pitch buffer is repeated multiple times during the 10 ms erasure.

While the erasure progresses, the history buffer is updated with the synthesized output. This way, the history buffer always has a smooth, continuous signal in it.

3.3.5 Synthetic Signal Generation after 10 ms

For continuous packet losses, the erasure will be at least 20 ms long and further action is required. Repeating a single pitch period works well for short erasures, but it introduces unnatural harmonic artefacts on long erasures. This is quite obvious if the erasure lands in an unvoiced region of speech, or in a region of rapid transition such as a stop. It was discovered by experimentation that these artefacts are significantly reduced by increasing the number of pitch periods used to synthesize the signal as the erasure progresses. Playing more pitch periods increases the variation in the signal. Although the pitch periods are not played in the order they occurred in the original signal, the resulting output still sounds natural. At 10 ms into the erasure the number of pitch periods used to synthesize the speech is increased to two, and at 20 ms a third pitch period is added.

For erasures longer than 20 ms no additional modifications to the pitch buffer are made.

When the number of pitch periods used in the pitch buffer increases, it is important that the transition in the synthesized signal be smooth. This is accomplished by continuing the output of the existing pitch buffer for 1/4 of a pitch period at the start of the second and third erased frame, updating the pitch buffer, keeping the buffer pointer synchronized with the correct phase, and then doing an OLA with the output from the new pitch buffer.

The pitch buffer is updated exactly as during the first erased frame, except that the number of pitch periods is increased.

The result of the OLA replaces the last 1/4 wavelength in the pitch buffer. To maintain

the phase of the current output pointer, pitch periods are subtracted from the pointer until it is in the first pitch period used.

3.3.6 Attenuation

As with other PLC algorithms, such as G.729 and G.728 Annex I, with long erasures it is necessary to attenuate the signal as the erasure progresses. As the erasure gets longer, the synthesized signal is more likely to diverge from the real signal. Without attenuation strange artifacts are created by holding certain types of sounds too long, even if the synthesized signal segment sounds natural in isolation. For the first 10 ms of an erasure the signal is not attenuated. At the start of the second 10 ms, the synthesized signal is linearly attenuated with a ramp at the rate of 20% per 10 ms. After 60 ms, the synthesized signal is zero.

3.3.7 First Good Frame after an Erasure

At the first good frame 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 length of the OLA depends on both the pitch period and the length of the erasure. For short, 10 ms erasures, a 1/4 wavelength window is used. For longer erasures the window is increased by 4 ms per 10 ms of erasure, up to a maximum of the frame size, 10 ms.

3.4 Concealment Algorithm Based on Pitch Prediction and Linear Prediction Coding (LPC)

The PLC schemes to be introduced in the next chapters will be based on the PLC

method described in the previous section and the work in [EBA04] which uses a Linear Prediction (LP) concealment algorithm for packet loss. This receiver based packet loss concealment algorithm produced better objective PESQ scores compared with other methods such as ITU-T G.711 Appendix I, packet repetition etc.

The main equation of the Linear Prediction (LP) for loss concealment is:

$$S_1[n] = \sum_{i=1}^{p} a_i \times S[n-i] + b[n] \qquad (3\text{-}1)$$

where $S[n-i]$ is replaced by $S_1[n-i]$ when it is not available, $S[n]$ is the original speech, $S_1[n]$ is the LPC predicted speech and a_i are the LPC coefficients.

When the packets are lost, the input excitation $b[n]$ is very difficult to estimate. In model based coders, b is usually a multi-pulse or a white noise component. As this is a receiver-based algorithm, $b[n]$ can only be based on the information at the receiver side.

In the algorithm from [EBA04], the excitation input $b[n]$ is replaced with a small fraction of the long-term prediction of the lost frame, which is referred to as a Reverse Order Pitch Period Replication of the lost frame as in ITU-T G.711 Appendix I. Thus, the input excitation to the model is found as:

$$b[n] = \hat{S}[n] \times G \qquad (3\text{-}2)$$

Where $\hat{S}[n]$ is the Reverse Order Pitch Period Replication (ROPPR) of the lost frame as in G.711 Appendix I, which is the long term pitch prediction of speech. G=0.01 is found to give the best results.

So, equation (3-1) becomes:

$$S_1[n] = \left(\sum_{i=1}^{p} a_i \times S[n-i] \right) + \left(\hat{S}[n] \times G \right) \qquad (3\text{-}3)$$

Then, the algorithm is further modified by a weighted summation of the short-term prediction $\left(\sum_{i=1}^{p} a_i \times S[n-i] + \hat{S}[n] \times G \right)$ and the long-term pitch prediction $\hat{S}[n]$.

Then the final prediction algorithm becomes:

$$S[n] = S_1[n] \times \alpha + \hat{S}[n] \times \beta \qquad (3\text{-}4)$$

Based on [EBA04], the best performance is achieved as

$\alpha = 0.7$ and $\beta = 0.3$.

For 5%, 10% and 25% loss rates, it was reported in [EBA04] that this PLC algorithm for PCM outperforms the ITU-T G.711 Appendix I method and the packet repetition method. The performance of those PLC algorithms (and the PLC algorithms to be introduced in the next chapters) will be evaluated experimentally later in the thesis.

Prediction from both the future packet and past packets will be introduced in Chapter 4. Then, the tuning of the algorithm with voiced/unvoiced classification will be addressed in Chapter 5. The resulting final algorithm will be validated under realistic conditions such as speech distorted by additive noise and reverberation. The results and discussions will be shown in Chapter 6.

Chapter 4

Packet Loss Concealment Algorithm with Future Packets

This chapter presents a new packet loss concealment algorithm. A detailed implementation of the new algorithm is shown. The test environment and the test results are introduced in section 3 and section 4 respectively.

4.1 Introduction

The methods for packet loss concealment described in this chapter are applicable to packetized speech transmission systems that use ITU-T Recommendation G.711 [G711] as the coding mechanism.

Unlike CELP-based coders, G.711 has no model of speech production. Hence, the concealment algorithm for G.711 is entirely receiver-based. G.711 has the advantage that the signal returns to the original signal at the first sample in the first good packet after an erasure. With CELP-based coders, the decoder's state variables take time to recover after a lost packet. Thus, PLC for G.711 has the ability to recover rapidly after a lost is over.

Most voice over IP equipments have jitter buffers. A jitter buffer is a shared data area where voice packets can be collected, stored, and sent to the voice processor in evenly spaced intervals. Variations in packet arrival time, called jitter, can occur because of network congestion, timing drift, or route changes. The jitter buffer, which is located at the receiving end of the voice connection, intentionally delays the arriving packets so that the

end user experiences a clear connection with very little sound distortion. There are two kinds of jitter buffers, static and dynamic. A static jitter buffer is hardware-based and is configured by the manufacturer. A dynamic jitter buffer is software-based and can be configured by the network administrator to adapt to changes in the network's delay. In "best effort" IP networks, future packets can be arriving early or during the intentionally delayed time slot. Thus it becomes possible to use future packets from the jitter buffer to predict a lost packet.

4.2 Proposed New Algorithm

The PLC technique described in this chapter uses the linear predictive model of speech production to estimate the vocal tract characteristics from the received past packets and the future packet, to reconstruct the signal contained in the missing packet.

Basing on the discussion in section 3.4, the previous algorithm is thus further developed with the information from a future packet. In the "best effort" IP network, future packets often arrives early. This makes it possible to use future packets from the jitter buffer to predict the lost packet.

Then the proposed prediction algorithm becomes:

$$S_1^p(n) = \sum_{i=1}^{P} \left(a(i) \times S_1^p(n-i)\right) + \hat{S}(n) \times G \qquad (4\text{-}1)$$

$$S_1^f(n) = \sum_{i=1}^{P} \left(a(i) \times S_1^f(n+i)\right) + \hat{S}(n) \times G \qquad (4\text{-}2)$$

$$S^c(n) = \begin{pmatrix} \left(\alpha \times S^p(n) + \beta \times \hat{S}(n)\right) \times \text{Hamming win. } 2^{\text{nd}} \text{ half} \\ + \left(\alpha \times S^f(n) + \beta \times \hat{S}(n)\right) \times \text{Hamming win. } 1^{\text{st}} \text{ half} \end{pmatrix} \qquad (4\text{-}3)$$

where $S^c(n)$ is the final concealed signal, $S^p(n)$ and $S^f(n)$ are the linear prediction result from the past and future samples, respectively.

4.2.1 Implementation of the Algorithm

The algorithm was implemented to work with packet sizes of 10 ms. Default sampling rate was 8 kHz. However, with a modification of the parameters, the algorithm could support other packet sizes or sampling rates.

4.2.1.1 State variables used in this algorithm

· **History Buffer:** The most recent 30 ms of good speech stream is stored to be used in the prediction of the lost packet.

· **Future Packet Buffer:** The algorithm checks the jitter buffer to see if there is a future packet available to be used for the reconstruction of missing packet. If so, it will be stored in the future packet buffer.

· **Forward Overlap Buffer:** This buffer contains a 10 samples extension of the generated concealed signal multiplied by a down slopping ramp, to be added to the 10 samples at the start of the next good packet multiplied by an uprising ramp. This addition is necessary to ensure smooth transition between the concealed packets and the correct packet after erasure. Because the good packets before erasure were used in the speech model to predict the concealed packet with linear prediction, a smooth transition between the last good speech samples and the next predicted samples is realized directly.

· **LP Coefficients:** the LP coefficients are used to form the poles of the all-pole synthesis speech filter. The order of the prediction filter implemented is set to be 50. The coefficients are calculated for the first packet of a lost speech segment and stored for use

with consecutive lost packets.

· **Pitch Period Buffer:** The pitch period estimated segment for the first packet of erasure is stored and used with consecutive lost packets.

4.2.1.2 Good packets

When there is no packet loss occurring, the receiver decodes the received packets and sends the output to the audio port. Due to the PLC algorithm, a circular history buffer of 30 ms (240 samples) long is created to save a copy of the decoded output. The content in this buffer is used to calculate the autocorrelation function, to estimate both the pitch and the LP coefficients, extract the long term excitation and provide the past samples $S[n-i]$ $1 \le i \le p$ where p is the order of the prediction filter.

4.2.1.3 First bad packet

The majority of the computation load is in the first 10 ms of the erasure. Figure 4-1 shows a block diagram of the principal blocks of the algorithm.

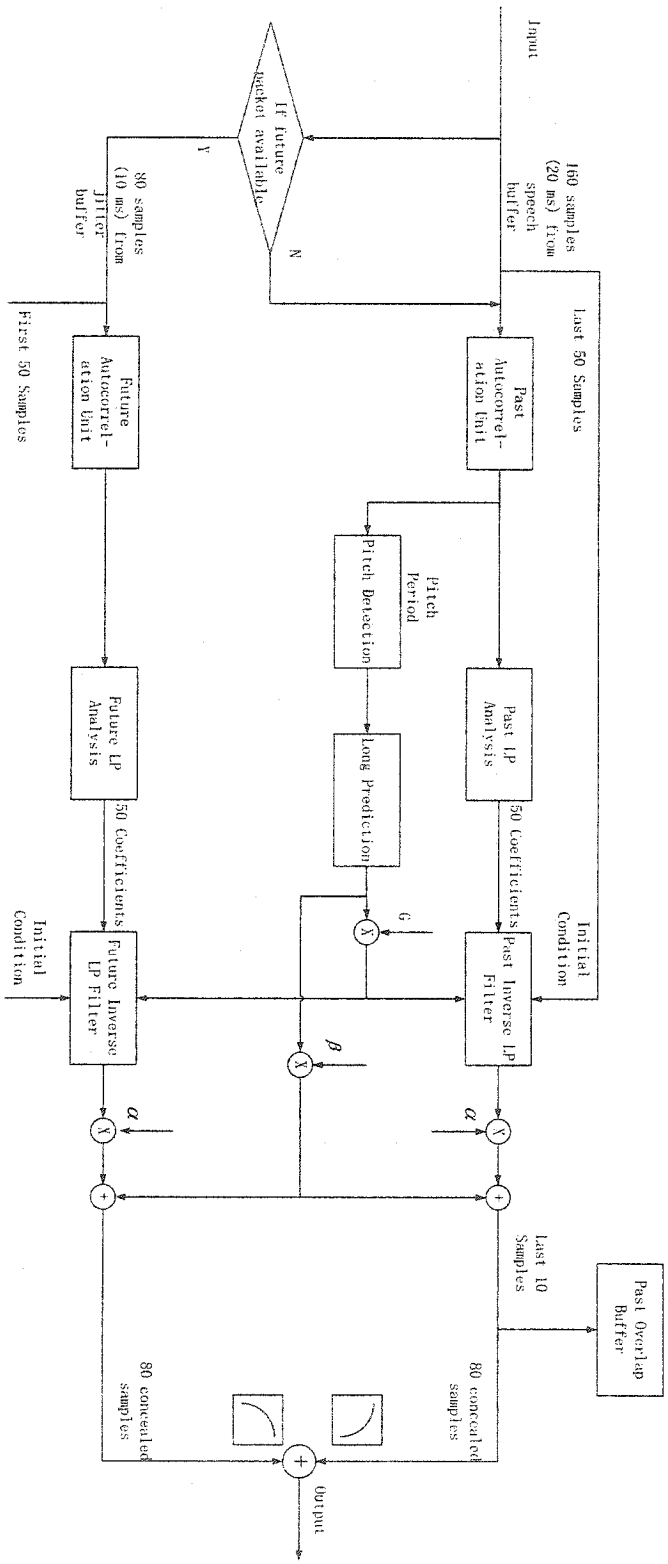


FIGURE 4-1: BLOCK DIAGRAM OF THE NEW ALGORITHM FOR THE FIRST LOST PACKET

· **Past Autocorrelation Unit:** When the loss of a packet is detected, the past autocorrelation unit uses the most recent 20 ms (160 samples) of correctly received speech for the pitch detection and the computation of the Linear Prediction (LP) coefficients.

· **Future Autocorrelation Unit:** An indicator is used here to show if the jitter buffer contains a future packet of the current lost one. If so, this future autocorrelation unit uses the 10 ms (80 samples) of future packet for the computation of the Linear Prediction (LP) coefficients.

· **Past/Future LP analysis:** The LP-analysis unit is implemented with the Levinson-Durbin algorithm of LP estimation. It uses the autocorrelation to calculate the LP coefficients. The prediction order is chosen to be 50. These 50 coefficients are used as the poles of the LP-synthesis filter which is the model of the speech production.

· **Past/Future Inverse LP Filter:** This LP filter models the speech production at the current period of erasure. It is an all-pole filter working on the same principle of model-based coders and follows equations:

For past inverse LP filter:

$$S_1[n] = \left(\sum_{i=1}^{p} (a_i \times S\ [n-i]) \right) + \hat{S}[n] \times G \qquad (4\text{-}1)$$

G=0.01, $S[n-i]$ is replaced by $S_1[n-i]$ when not available.

For future inverse LP filter:

$$S_1[n] = \left(\sum_{i=1}^{p} (a_i \times S[n+i]) \right) + \hat{S}[n] \times G \qquad (4\text{-}4)$$

G=0.01, $S[n+i]$ is replaced by $S_1[n+i]$ when not available.

Where $\hat{S}[n]$ is the Reverse Order Pitch Period Replication (ROPPR) of the lost frame as in G.711 Appendix I, $S[n]$ is the original speech, $S_1[n]$ is LPC predicted speech and a_i are LPC coefficients.

· **Past Overlap Add Unit:** The past synthesis model actually produces 10 more samples per lost packet instead of the actual packet size. The last 10 samples are the predicted samples of the packet next to the lost packet. If the next is lost, then these samples are played as the concealed samples of that lost packet. If the next packet is not lost, then the samples are multiplied by a decaying ramp and added to the corresponding first 10 samples in the new correct speech sequence that are to be multiplied by an uprising ramp. The output of the addition is played instead of the first 10 good samples after erasure. This cross fading process guarantees a smooth transition from the concealed speech segment to the good packet condition.

If the future packet is available for the concealment algorithm, the fact that a linear prediction is performed from the future samples will produce a smooth transition between the concealed lost packet and the next good packet. Thus the cross-fading mechanism is no longer required and the results of the forward overlap add unit will simply be dropped.

· **Process of Concealment of the First Bad Packet:**

When the loss of a packet is detected, the past autocorrelation unit uses the most recent 20 ms (160 samples) of correctly received speech packet for the pitch detection and the computation of the Linear Prediction (LP) coefficients. The algorithm also checks if the jitter buffer contains a future packet of the current lost one. If so, the future autocorrelation unit uses the 10 ms (80 samples) of future packet for the computation of the Linear

Prediction (LP) coefficients for the prediction from future samples.

The past/future LP-analysis unit implemented with the Levinson-Durbin algorithm of LP estimation uses the autocorrelation from past/future autocorrelation unit to calculate the LP coefficients. The prediction order is chosen to be 50. These 50 coefficients are used as the poles of the LP-synthesis filter which is the model of the speech production.

The past/future inverse LP filters model the speech production at the current period of erasure. They are all-pole filters working on the same principle as described earlier and they follow the equations below.

The results of the past synthesis unit and the future synthesis unit are multiplied with the second half and first half of a Hamming window, respectively, and added up.

$$S^p(n) = \sum_{i=1}^{P} \left(a_i \times S^p(n-i)\right) + \hat{S}(n) \times G \tag{4-5}$$

$$S^f(n) = \sum_{i=1}^{P} \left(a_i \times S^f(n+i)\right) + \hat{S}(n) \times G \tag{4-6}$$

$$S^c(n) = \begin{pmatrix} \left(\alpha \times S^p(n) + \beta \times \hat{S}(n)\right) \times \text{Hamming win. } 2^{\text{nd}} \text{ half} \\ + \left(\alpha \times S^f(n) + \beta \times \hat{S}(n)\right) \times \text{Hamming win. } 1^{\text{st}} \text{ half} \end{pmatrix} \tag{4-7}$$

where $S^c(n)$ is the final concealed signals, $S^p(n)$ and $S^f(n)$ are the linear prediction result from the past and future samples, respectively.

α and β are coefficients weighting the long term pitch prediction and the short term LPC prediction of speech.

The above is the whole process for the first lost packet. It can be seen that the majority of the computation load is in the first 10 ms of the erasure.

4.2.1.4 Consecutive Bad Packet

There won't be any new parameters calculated if the number of consecutive bad packets is more than one. The parameters used for the first lost packet concealment are re-used with a slight modification. The results are multiplied by a decaying ramp starting at the initial value of 1 and decaying at a rate of 0.2 per 10 ms. This ramp multiplication introduces smooth decay increasing along the loss period. It should also be noticed that no future packet is used from the jitter buffer. The future autocorrelation unit, future LP analysis unit and future inverse LP filter unit do not operate.

4.2.1.5 First Good Packet after Erasure

The first 10 samples of the first good packet after an erasure are multiplied with an up-sloping ramp and added to the 10 extra predicted samples produced from the past inverse LP filter multiplied by a decaying ramp.

4.2.2 Delay of the New Algorithm

The algorithmic delay is determined by the jitter buffer. During the intentional delay time slot for the jitter buffer, if a future packet arrived, it will be added in the new algorithm for the prediction.

4.3 Testing Environment

A test was executed on an Intel Pentium 4 processor operating at 2.4GHz, in a system with 512 MB RAM running Microsoft Windows XP. The applications used for testing were built using Matlab 6.1.

4.3.1 Testing Distribution Pattern

To simulate packet loss caused by the network, a Gaussian distribution was used to simulate the packet loss pattern. In practice, the loss distribution may tend to be more bursty. However, some degree of interleaving could be used to make the distribution of missing samples more Gaussian.

4.3.2 Testing tool

The result of the concealment techniques is assessed by the Perceptual Estimation of Speech Quality (PESQ) standard P.862, which was developed by the ITU-T. It is a recent and accurate tool that has proved to give reliable estimation of the subjective quality test. The PESQ is implemented to take proper account of variable delays, filtering and burst distortion effects, transfer function equalization, time alignment, etc. The key process of the PESQ operation is to use a perceptual model analogous to the psychological representation of the original and degraded signals in the human auditory system, taking into account perceptual frequencies "Barks" and "Sons".

The PESQ algorithm compares the distorted subject file to the original file and generates a score indicating the Absolute Category Rating (ACR) score that would be estimated if subjective tests were applied. The score is given in the range [-0.5 4.5]. A lower score indicates lower speech quality and a higher score indicates higher quality of the investigated file.

4.4 Performance and Discussion

Several experiments were set up to reach the final form of the new PLC algorithm. Some of those experiments are described in this section.

4.4.1 Windows to Be Used for Combining the Prediction from Past Samples and Prediction from Future Samples

This experiment was performed on a set of speech files from four speakers (two males and two females) referred to in the results as femaleN_01, femaleN_02, maleN_01 and maleN_02. Each speaker gives out a sentence in English. Frames of 80 samples (10ms) were used. The format of the files was linear PCM, with 8 kHz sampling rate.

The parameters are set as follows to give the best results:

$\alpha = 0.7$, $\beta = 0.3$ and $G = 0.01$

Different values of α, β and G were tested. The set of parameters above gives the best performance.

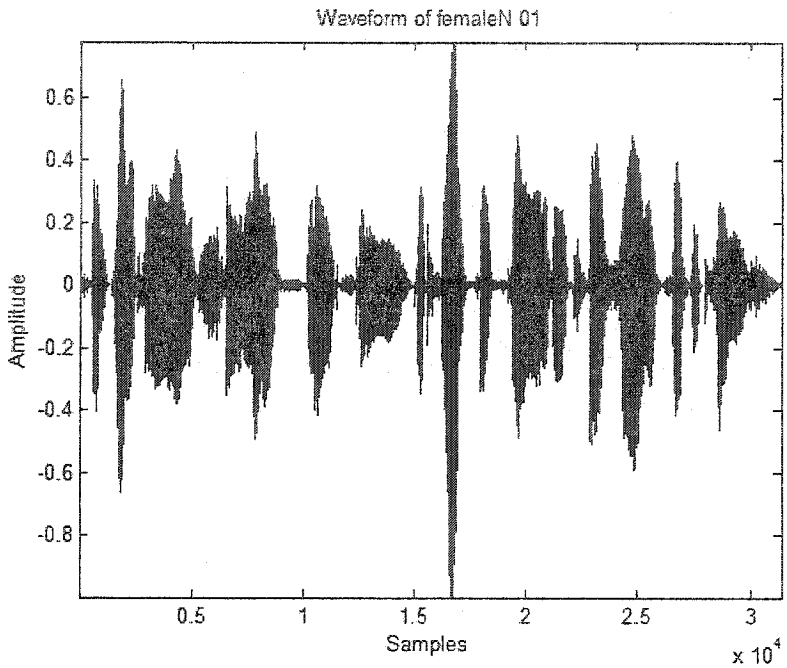


FIGURE 4-2: WAVEFORM OF FEMALEN_01

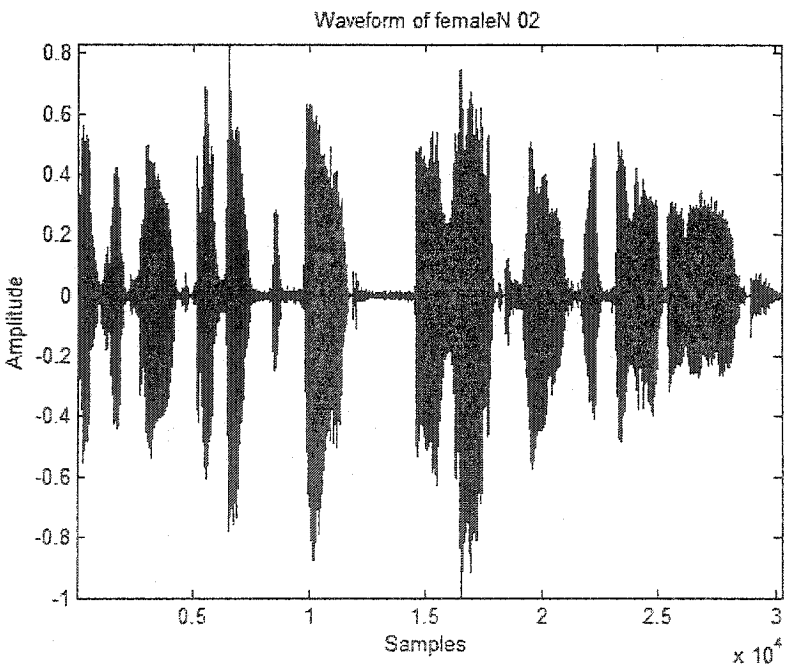


FIGURE 4-3: WAVEFORM OF FEMALEN_02

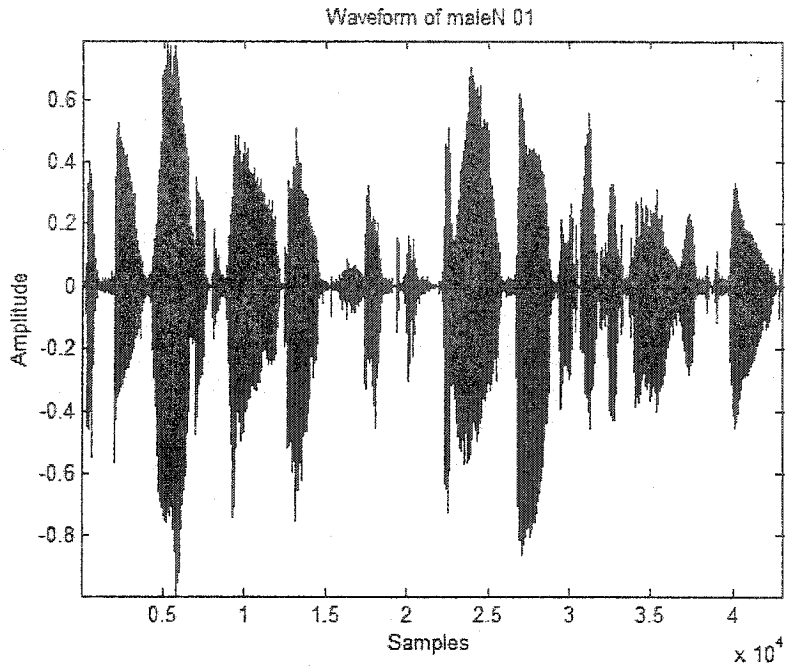


FIGURE 4-4: WAVEFORM OF MALEN_01

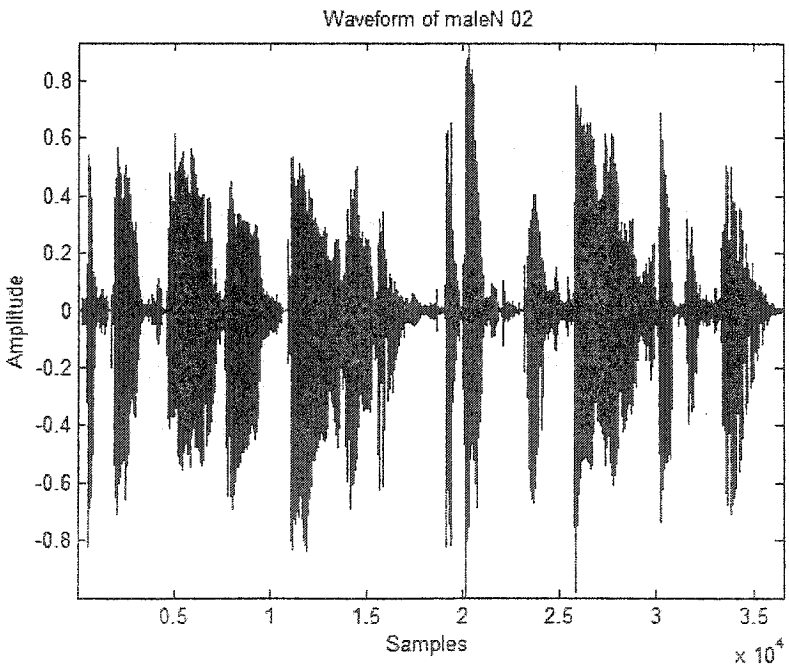


FIGURE 4-5: WAVEFORM OF MALEN_02

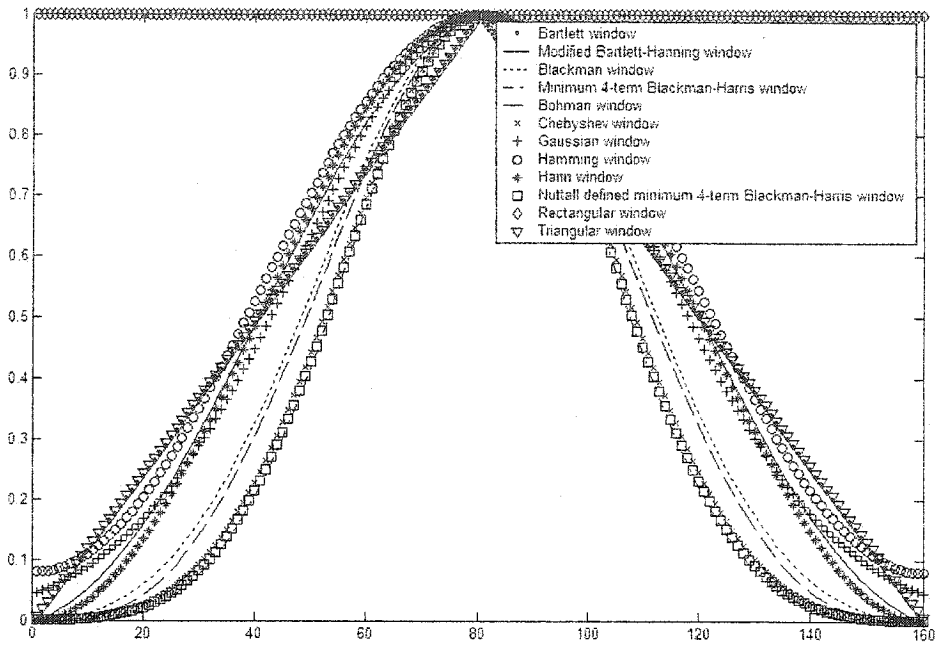


FIGURE 4-6: DIFFERENT WINDOWS FOR COMBINING PREDICTION FROM PAST PACKETS AND FUTURE PACKETS

In this experiment, all future packets are supposed to be available. The average shown is the average PESQ result of femaleN_01, femaleN_02, maleN_01 and maleN_02. As described in Figure 4-1 and Equation 4-7, the first half of a window is used to weight the prediction from the future packet, while the second half of a window is used to weight the prediction from the past packet.

Window	femaleN_01	femaleN_02	maleN_01	maleN_02	Average
Bartlett window	3.80	3.86	3.66	3.64	3.74
Modified Bartlett-Hanning window.	3.76	3.80	3.61	3.64	3.70
Blackman window	3.63	3.65	3.57	3.57	3.61
Minimum 4-term Blackman-Harris window	3.49	3.55	3.54	3.51	3.52
Bohman window	3.60	3.63	3.56	3.56	3.59
Chebyshev window	3.53	3.57	3.54	3.52	3.54
Gaussian window	3.78	3.79	3.65	3.64	3.72
Hamming window	***3.81***	***3.85***	***3.68***	***3.69***	***3.76***
Hann window	3.74	3.78	3.60	3.63	3.69
Nuttall defined minimum 4-termBlackman-Harris window	3.50	3.56	3.54	3.51	3.53
Rectangular window	3.62	3.81	3.67	3.73	3.71
Triangular window	3.80	3.87	3.66	3.64	3.74

TABLE4-1: COMPARISON OF PERFORMANCE OF DIFFERENT WINDOWS IN COMBINATION OF PREDICTION FROM PAST SAMPLES AND PREDICTION FROM FUTURE SAMPLES IN PESQ SCORE (5% LOSS RATE)

Window	femaleN_01	femaleN_02	maleN_01	maleN_02	Average
Bartlett window	3.56	3.65	3.47	3.43	3.52
Modified Bartlett-Hanning window.	3.52	3.56	3.46	3.42	3.49
Blackman window	3.43	3.35	3.39	3.39	3.39
Minimum 4-term Blackman-Harris window	3.34	3.20	3.32	3.36	3.31
Bohman window	3.40	3.31	3.36	3.38	3.36
Chebyshev window	3.35	3.24	3.33	3.36	3.32
Gaussian window	3.53	3.54	3.47	3.46	3.50
Hamming window	***3.58***	***3.64***	***3.50***	***3.52***	***3.56***
Hann window	3.51	3.54	3.45	3.42	3.48
Nuttall defined minimum 4-termBlackman-Harris window	3.34	3.22	3.32	3.36	3.31
Rectangular window	3.48	3.40	3.41	3.63	3.48
Triangular window	3.57	3.66	3.48	3.44	3.54

TABLE4-2: COMPARISON OF PERFORMANCE OF DIFFERENT WINDOWS IN COMBINATION OF PREDICTION FROM PAST SAMPLES AND PREDICTION FROM FUTURE SAMPLES IN PESQ SCORE(10% LOSS RATE)

Window	femaleN_01	femaleN_02	maleN_01	maleN_02	Average
Bartlett window	3.30	3.33	3.19	3.26	3.27
Modified Bartlett-Hanning window.	3.27	3.29	3.17	3.24	3.24
Blackman window	3.14	3.16	3.11	3.22	3.16
Minimum 4-term Blackman-Harris window	3.05	3.06	3.07	3.18	3.09
Bohman window	3.11	3.14	3.11	3.21	3.14
Chebyshev window	3.07	3.08	3.08	3.19	3.11
Gaussian window	3.27	3.29	3.18	3.26	3.25
Hamming window	***3.27***	***3.33***	***3.23***	***3.30***	***3.28***
Hann window	3.25	3.27	3.16	3.24	3.23
Nuttall defined minimum 4-termBlackman-Harris window	3.06	3.07	3.07	3.18	3.09
Rectangular window	3.21	3.38	3.22	3.32	3.28
Triangular window	3.30	3.33	3.19	3.27	3.27

TABLE4-3: COMPARISON OF PERFORMANCE OF DIFFERENT WINDOWS IN COMBINATION OF PREDICTION FROM PAST SAMPLES AND PREDICTION FROM FUTURE SAMPLES IN PESQ SCORE(25% LOSS RATE)

From the above results, it is very clear that the Hamming window always produced the best results although by a small margin. In the final algorithm, the Hamming window is thus implemented for the combination of prediction from past samples and prediction from future samples.

4.4.2 Presence of Future Packet in the Performance of the New Algorithm

This experiment is performed on a set of speech files from four speakers (two males and two females) referred to in the results as M1, M2, F1 and F2. For each of those speakers, 10 speech files were used, each containing two sentences in English with a duration of 8 sec. Frames of 80 samples (10 ms) were used. The format of the files was linear PCM, with 8 kHz sampling rate. The files were taken from the ITU-T supplement P.23.

The parameters are again set as follows:

$\alpha = 0.7$, $\beta = 0.3$ and $G = 0.01$.

	M1	M2	F1	F2
Combination of Prediction from Past Packet and 100% of Future Packets available	3.81	3.85	3.68	3.69
Combination of Prediction from Past Packet and 50% of Future Packets available	3.72	3.72	3.52	3.64
Prediction from Past Packet	3.66	3.51	3.42	3.44
G.711 Appendix I	3.45	3.41	3.36	3.31
Packet Repetition	2.99	3.00	2.73	3.13
Prediction from Future Packet Using Coefficients Calculated from Past Packet	3.37	3.41	3.07	3.16

TABLE 4-4: AVERAGE RESULTS OF 5% RANDOM PACKET LOSS

	M1	M2	F1	F2
Combination of Prediction from Past Packet and 100% of Future Packets available	3.58	3.64	3.50	3.52
Combination of Prediction from Past Packet and 50% of Future Packets available	3.43	3.46	3.33	3.36
Prediction from Past Packet	3.19	3.27	3.05	3.01
G.711 Appendix I	3.09	3.12	2.93	2.87
Packet Repetition	2.69	2.84	2.51	2.45
Prediction from Future Packet Using Coefficients Calculated from Past Packet	3.07	2.98	2.85	2.73

TABLE 4-5: AVERAGE RESULTS OF 10% RANDOM PACKET LOSS

	M1	M2	F1	F2
Combination of Prediction from Past Packet and 100% of Future Packets available	3.27	3.33	3.23	3.30
Combination of Prediction from Past Packet and 50% of Future Packets available	3.03	3.18	3.03	3.08
Prediction from Past Packet	2.74	2.79	2.66	2.60
G.711 Appendix I	2.60	2.63	2.58	2.43
Packet Repetition	2.41	2.47	2.15	2.24
Prediction from Future Packet Using Coefficients Calculated from Past Packet	2.74	2.72	2.47	2.60

TABLE 4-6: AVERAGE RESULTS OF 25% RANDOM PACKET LOSS

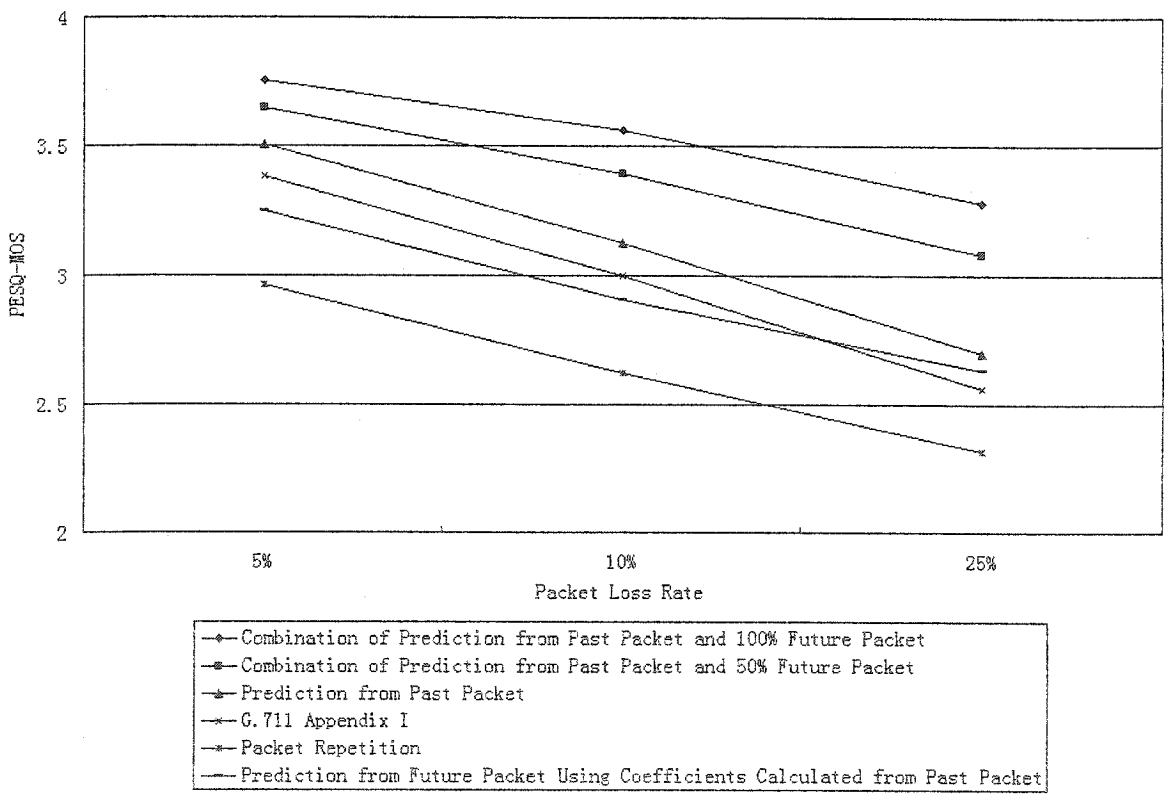


FIGURE 4-7 PERFORMANCE OF ALGORITHMS

There are several conclusions can be drawn from the above tables and figures:

First, the linear prediction concealment algorithm predicting from the past packet performs superior to both the existing ITU-T G.711 Appendix I standard and the packet repetition method. This algorithm first appears in [EBA03] as a 'New Hybrid Long-term and Short-term Prediction Algorithm for Packet Loss Erasure over IP-Network'. Actually, the performance of the packet repetition method is much worse than both the algorithm from [EBA03] and the ITU-T G.711 Appendix I concealment standard. A significant and almost steady margin appears as a difference between the algorithm from [EBA03] and the ITU-T standard. The margin presents the performance gain of incorporating the LP model with the plain long-term pitch-repetition-based concealment standard. This is shown in Figure 4-10.

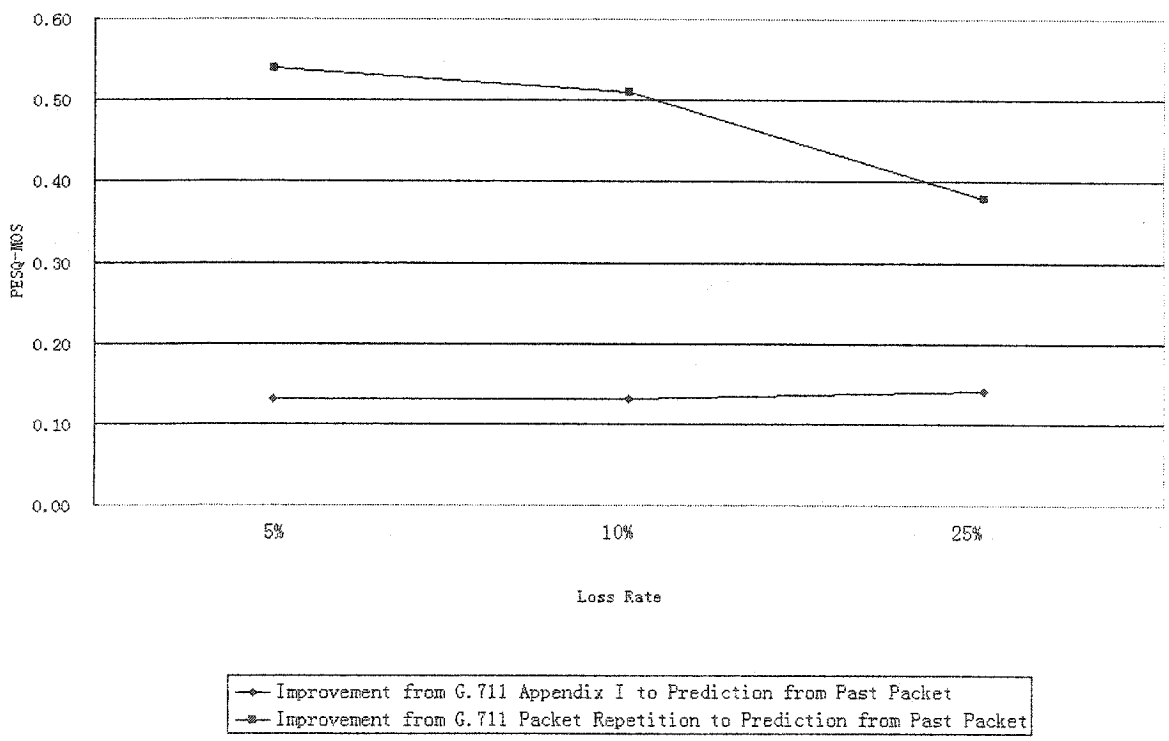


FIGURE 4-8 IMPROVEMENT OF LINEAR PREDICTION CONCEALMENT ALGORITHM

Second, the use of a future packet can enhance the performance of the prediction

significantly. 50% presence of future packets and 100% future packets presence were tested. An increase of PESQ-MOS results can clearly be seen, as shown in the following figures and tables.

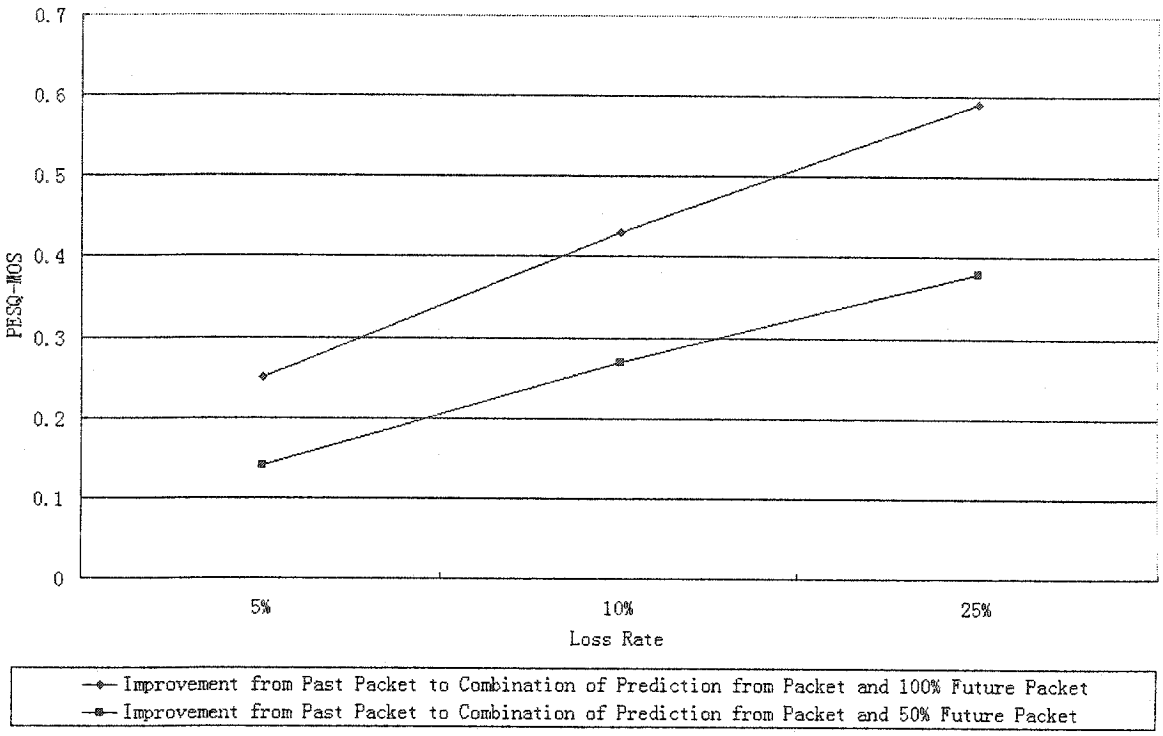


FIGURE 4-9: IMPROVEMENT BROUGHT BY THE PREDICTION FROM FUTURE PACKET

Algorithms	Average Improvement in PESQ-MOS
Prediction from Past Packet to Combination of Prediction from Past Packet and Future Packets 100% Available	0.25
Prediction from Past Packet to Combination of Prediction from Past Packet and Future Packets 100% Available	0.14

TABLE 4-7: AVERAGE IMPROVEMENT IN THE PRESENCE OF FUTURE PACKET IN THE PREDICTION (5% LOSS RATE)

Algorithms	Average Improvement in PESQ-MOS
Prediction from Past Packet to Combination of Prediction from Past Packet and Future Packets 100% Available	0.43
Prediction from Past Packet to Combination of Prediction from Past Packet and Future Packets 100% Available	0.27

TABLE 4-8: AVERAGE IMPROVEMENT IN THE PRESENCE OF FUTURE PACKET IN THE PREDICTION (10% LOSS RATE)

Algorithms	Average Improvement in PESQ-MOS
Prediction from Past Packet to Combination of Prediction from Past Packet and Future Packets 100% Available	0.59
Prediction from Past Packet to Combination of Prediction from Past Packet and Future Packets 100% Available	0.38

TABLE 4-9: AVERAGE IMPROVEMENT BY THE PRESENCE OF FUTURE PACKET IN THE PREDICTION (25% LOSS RATE)

Third, there is also a point to be noticed that a prediction from a future packet but using coefficients calculated from past packets is also implemented during the experiments. The idea was to save some computations in calculating the coefficients. But the experimental results showed that this algorithm does not perform as good as the prediction from a future packet using coefficients obtained from the future packet. Also, the presence of the prediction from a future packet using coefficients from past packets even distorted the performance contributed by the prediction from past packets.

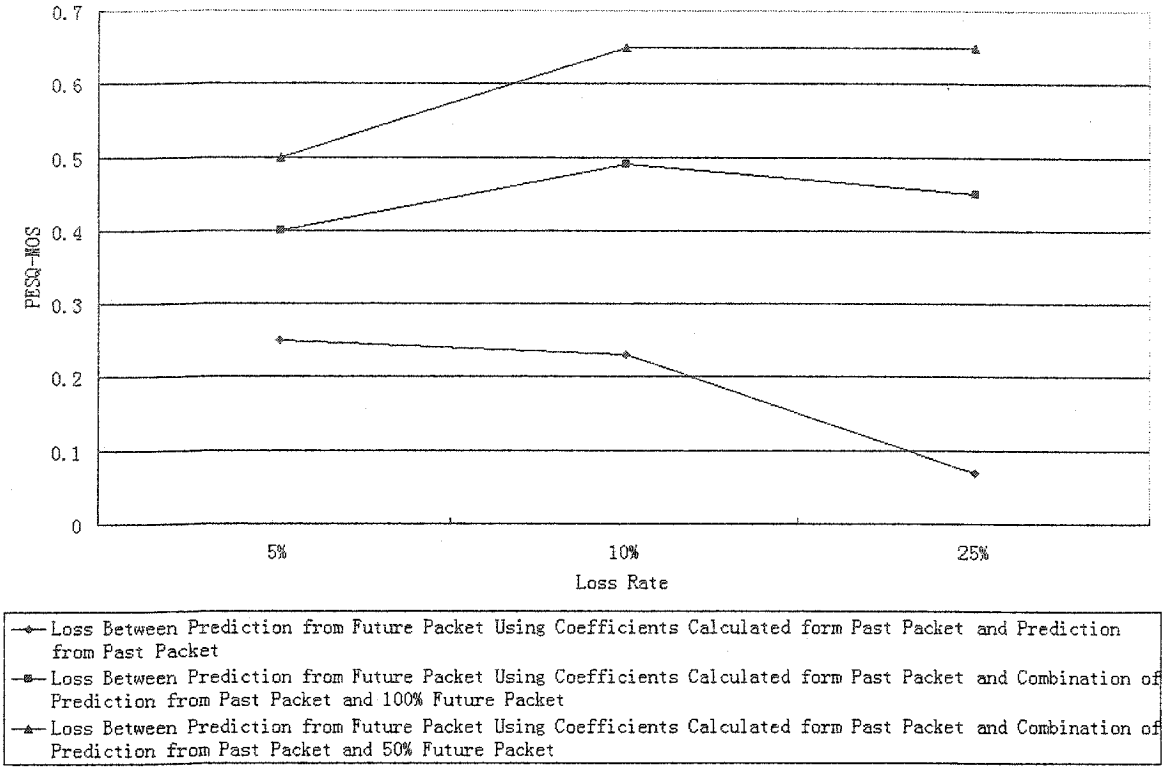


FIGURE 4-10: LOSS BROUGHT BY PREDICTION FROM A FUTURE PACKET USING COEFFICIENTS CALCULATED FROM PAST PACKETS

4.5 Summary

This chapter presented a new PLC algorithm for packet loss concealment using prediction from both past packets and one future packet available at a jitter buffer. The research started with the results from [EBA04] which is an LP-based prediction of lost packets using past packets. Due to the fact that future packets may also be available in the jitter buffer in a real network, new research was performed on a new PLC algorithm with uses of both past packets and a future packet. The input excitation to the synthesis filter is replaced by a small fraction (0.01) of the pitch prediction replica of the missed speech packet. The pitch replica is extracted in the same manner as in the ITU-T G.711 Appendix I concealment algorithm. The resulting predicted packet of the synthesis filter is then multiplied by a factor (0.7) and then added to the long term prediction (the pitch replica

estimated as in the ITU-T G.711 Appendix I standard) multiplied by a factor of 0.3. Finally, the concealed packets from past packets and concealed packets from the future packet are joined together using a Hamming window. This sum proved to provide a better prediction of the concealed packet as shown in the results section.

Chapter 5

New PLC Algorithm with Voiced/Unvoiced Classification

To further improve the performance of the new packet loss concealment algorithm, information of the voiced/unvoiced classification property of a given voice packet is to be abstracted from the packet. Thus voiced/unvoiced (V/UV) classification is introduced in this chapter and implemented with the new PLC algorithm presented in Chapter 4. The first section of this chapter introduces principles of voiced/unvoiced classification algorithms. Then some single feature classification algorithms are presented in section 2. Sections 3 presents the test results of new packet loss concealment algorithm with V/UV classification.

5.1 Voiced/Unvoiced Classification

The PLC scheme of the previous chapter follows the equations (shown here with the prediction from past samples only):

$$S_1[n] = \left(\sum_{i=1}^{p} (a_i \times S\ [n-i]) \right) + \hat{S}[n] \times G \qquad (5\text{-}1)$$

G=0.01, $S[n-i]$ is replaced by $S_1[n-i]$ when not available.

$$S^c[n] = S_1[n] \times \alpha + \hat{S}[n] \times \beta \qquad (5\text{-}2)$$

Instead of having constant α and β parameters, there should be different concealment parameters for different speech packets. So, the modified new algorithm is generated by modifying α and β for different speech packets, more specifically based on the voiced or unvoiced property of a speech packet. An important problem is the determination to which category a given packet belongs. The decision is made based on an observation vector, frequently called feature vector, which serves as the input to a decision rule that assigns a sample vector to one of the given classes.

There exists a variety of approaches described in the literature for voiced/unvoiced classification. It can be divided mainly into two categories.

- Single feature methods
- Multiple feature vector methods

For single feature methods, only one measurement is taken from the signal and thus the feature dimension is one. A simple threshold is to be figured out for voiced/unvoiced classification. Single feature methods are mostly fast methods for their low computation complexity.

For multiple features vector methods, the features are multidimensional. This gives more possibility to determine more complex shaped regions of voiced and unvoiced speech.

5.2 Voiced/Unvoiced Classification Methods Selected for the Experiments

In this research, since the emphasis is on loss concealment and not on voiced/unvoiced (V/UV) classification, only simple V/UV classification methods have

been considered such as single feature methods. These methods have a low complexity. Silence of inactive speech frames are all considered as unvoiced frames.

5.2.1 Selected Voiced/Unvoiced Classification Methods

- RMS value of speech samples

$$C = \sqrt{\frac{\sum_{n=1}^{N} s^2(n)}{N}} \tag{5-3}$$

The RMS value of the speech samples is calculated for the voiced/unvoiced classification. Based on the fact that the voiced speech tends to be more energetic, this parameter of voiced speech is often higher than that of unvoiced speech.

- Lag one normalized autocorrelation coefficient

$$C = \frac{\sum_{n=1}^{N} s(n)s(n-1)}{\sum_{n=0}^{N-1} s^2(n)} \tag{5-4}$$

Due to the concentration of low frequency energy for voiced sounds, adjacent samples of voiced speech waveforms are highly correlated and thus this parameter is close to 1. On the other hand, this autocorrelation value for unvoiced speech is close to 0.

- Lag one normalized autocorrelation coefficient of LPC prediction residue.

$$C = \frac{\sum_{n=1}^{N} \mathrm{Res}(n)\mathrm{Res}(n-1)}{\sum_{n=0}^{N-1} \mathrm{Res}^2(n)} \tag{5-5}$$

In this method, the autocorrelation coefficient of the LPC residue is calculated. Based on the fact that voiced speech tends to concentrate on low frequency energy, this autocorrelation coefficient of the LPC residue is close to 1 for voiced speech, and for unvoiced speech it is close to 0.

5.2.2 Performance of Selected Voiced/Unvoiced Classification Methods

The methods mentioned in the last section were implemented and their performance is discussed here.

The experiment was performed on one speech file of a female. In this speech file, a sentence in English lasts 3.92s. Frames of 80 samples (10ms) were used. The format of the files was linear PCM, with 8 kHz sampling rate. For the computation of the LPC residual signal, two different LPC order were used: order 10 and order 50. Order 50 corresponds to the same LPC coefficients as the ones used for the concealment, where such a high LPC order is used to include in the LPC prediction the case of low pitch values, which was particularly a concern for concealment performance. Order 10 corresponds to the LPC order normally used in speech codecs, where the purpose of the LPC modeling is only to model the vocal-tract, and not the pitch. Since voiced/unvoiced classification can be related to the detection of a strong pitch component in the LPC residual, in principle a LPC prediction of order 10 could be more suitable (i.e. no pitch component removed from LPC residual).

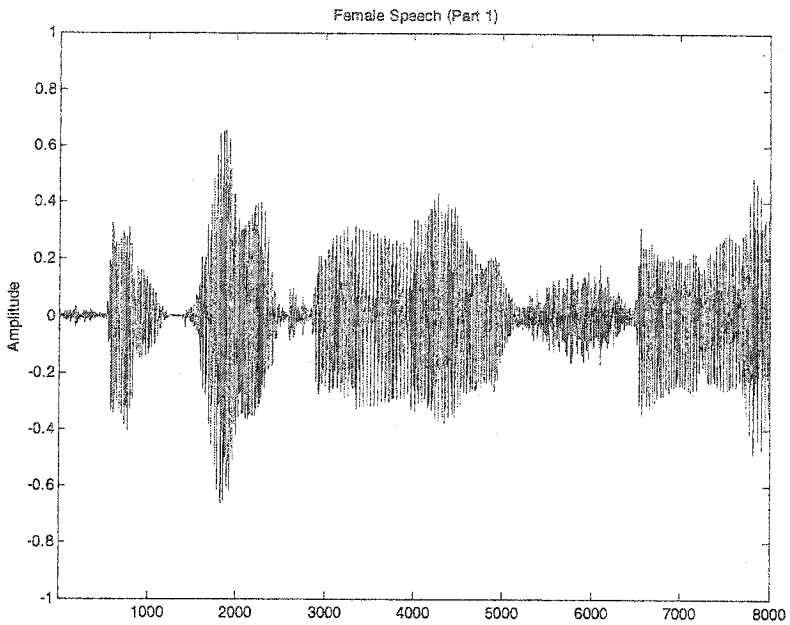


FIGURE 5-1: SPEECH FILE (PART 1)

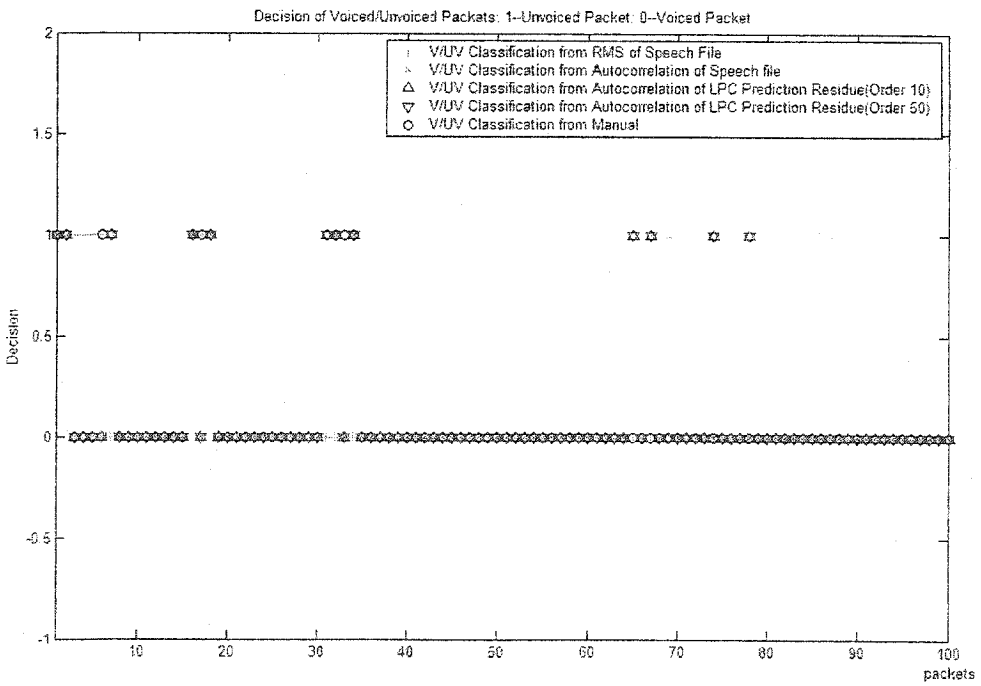


FIGURE 5-2: ENERGETIC V/UV CLASSIFICATION RESULTS (PART 1)

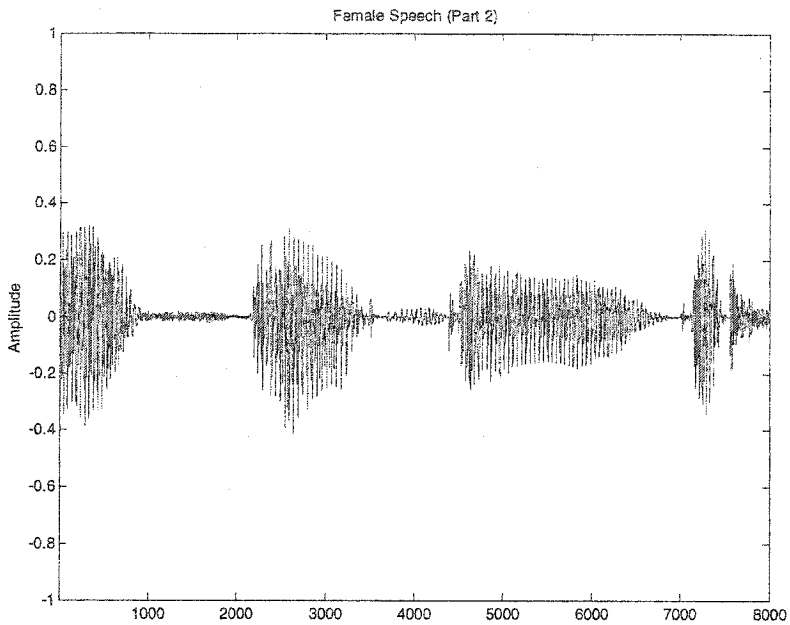


FIGURE 5-3: SPEECH FILE (PART 2)

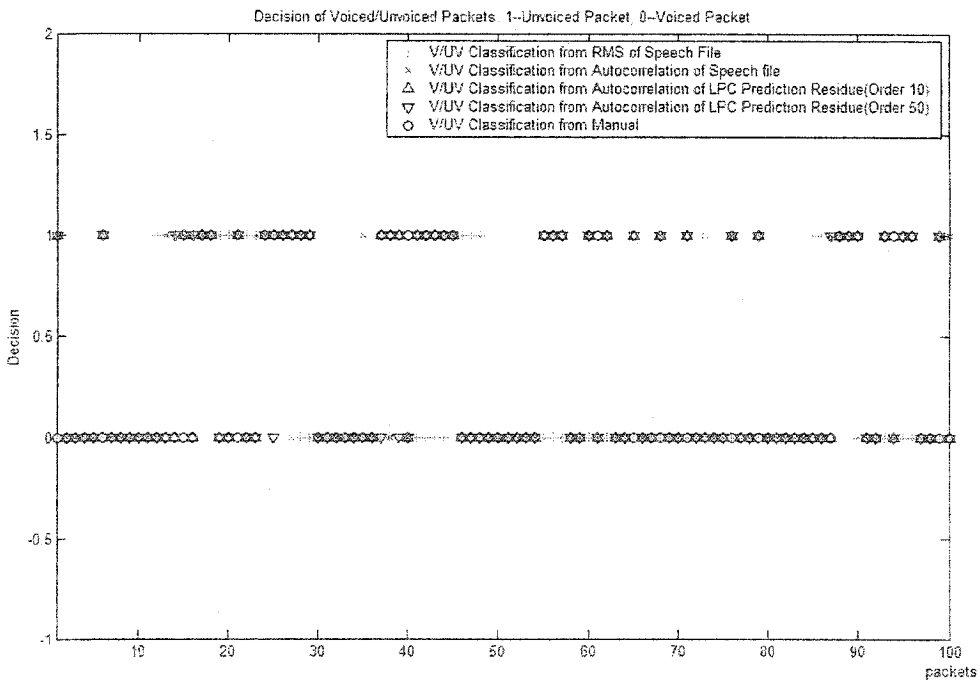


FIGURE 5-4: ENERGETIC V/UV CLASSIFICATION RESULTS (PART 2)

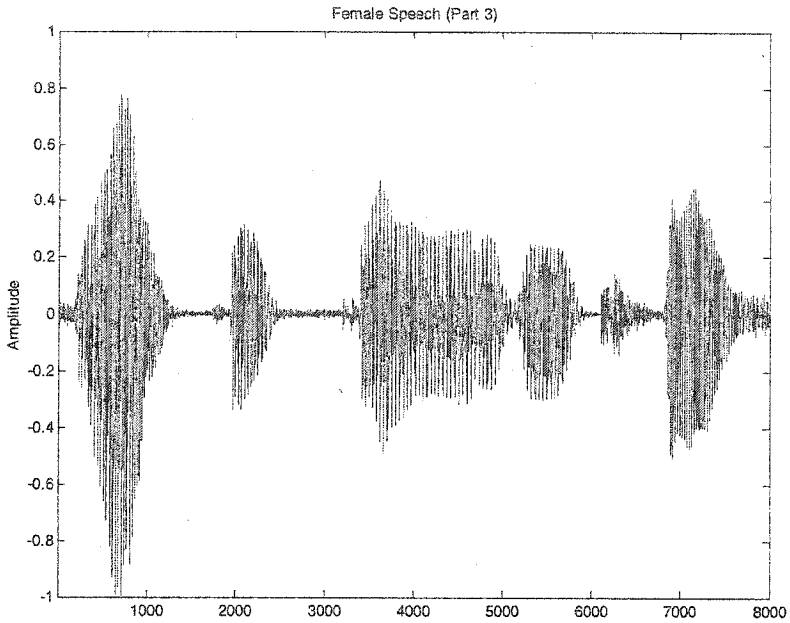


FIGURE 5-5: SPEECH FILE (PART 3)

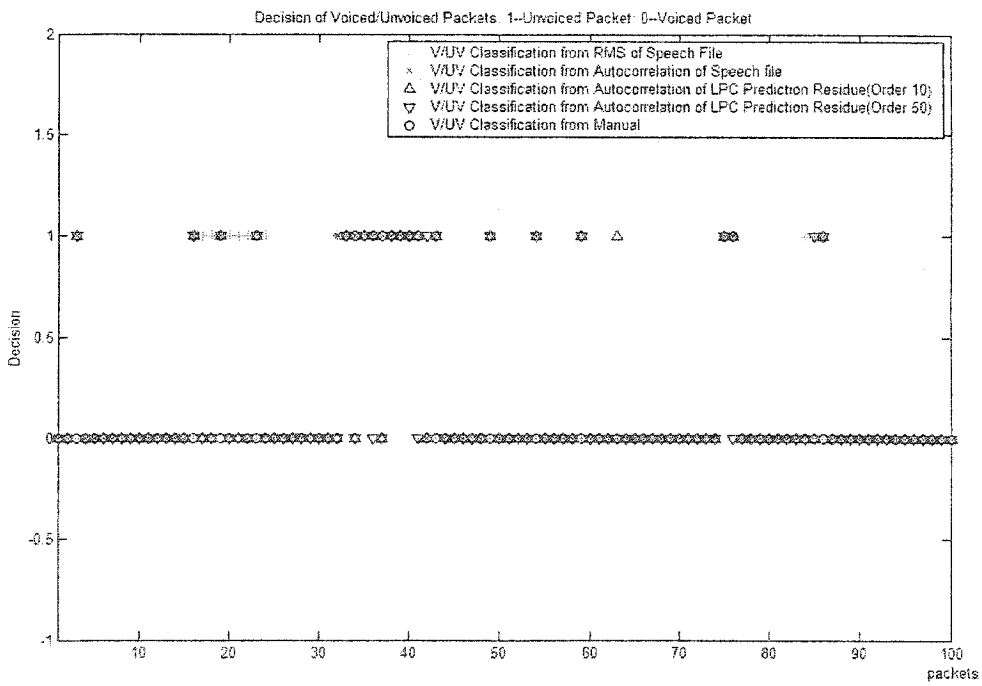


FIGURE 5-6: ENERGETIC V/UV CLASSIFICATION RESULTS (PART 3)

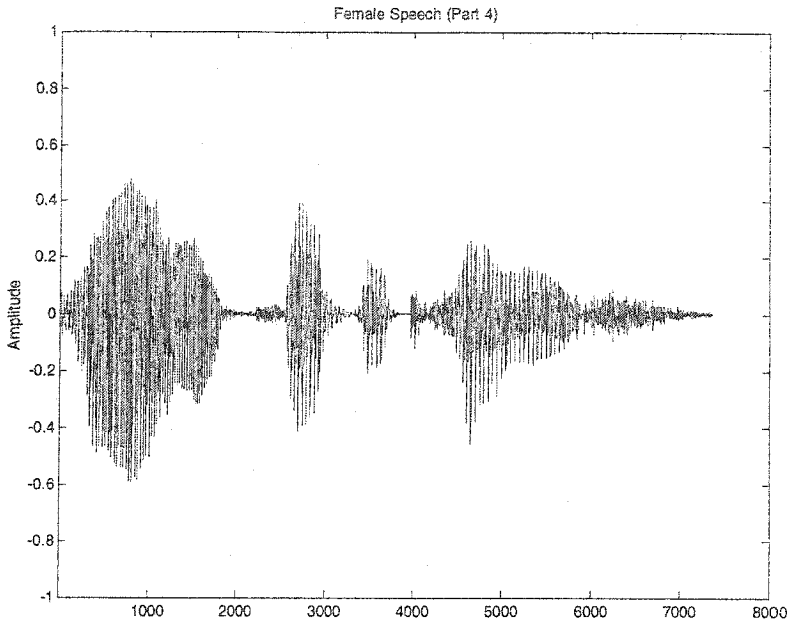


FIGURE 5-7: SPEECH FILE (PART 4)

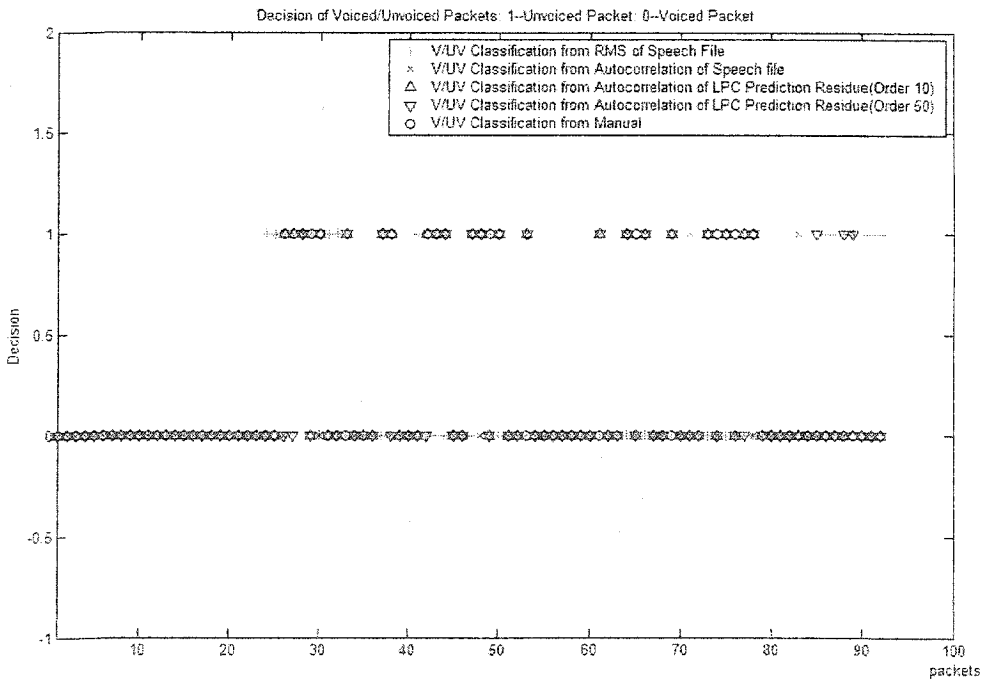


FIGURE 5-8: ENERGETIC V/UV CLASSIFICATION RESULTS (PART 4)

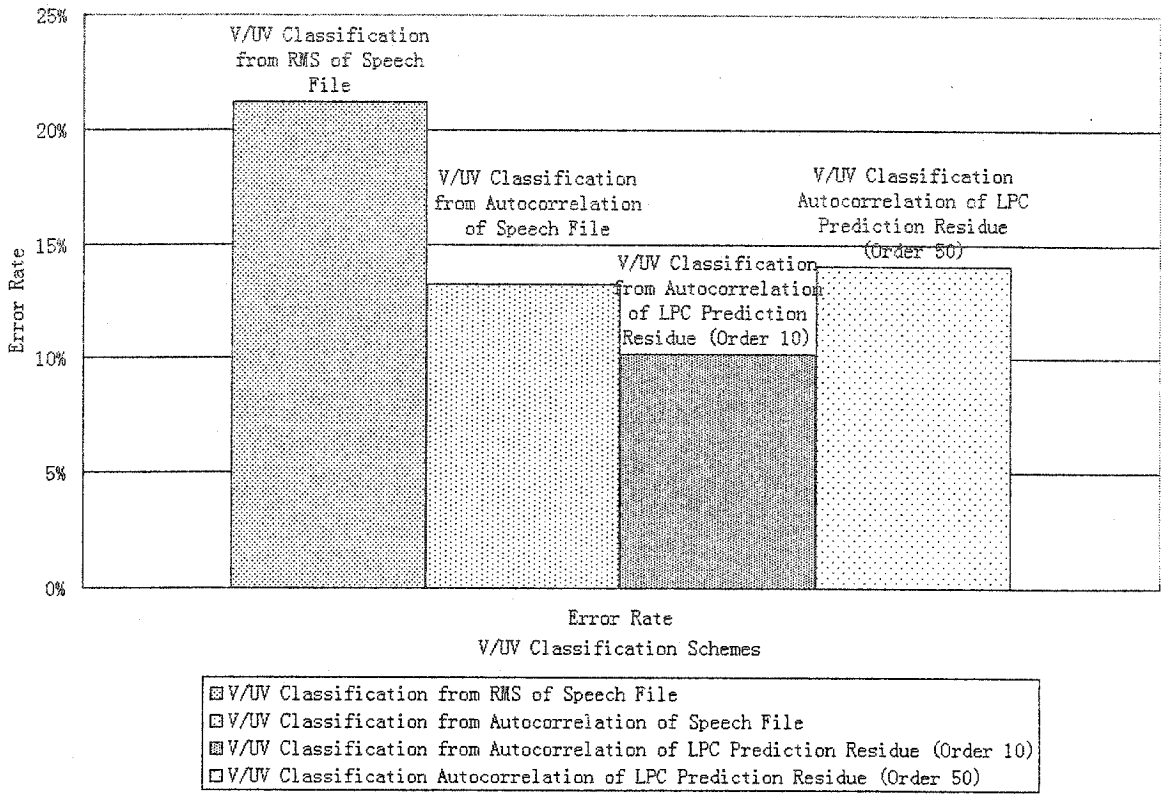


FIGURE 5-9: ERROR RATE OF DIFFERENT V/UV CLASSIFICATION SCHEMES

	Voiced Error	Unvoiced Error	Total Error	Error Rate
V/UV Classification from RMS of Speech File	43	40	83	21%
V/UV Classification from Autocorrelation of Speech File	38	14	52	13%
V/UV Classification from Autocorrelation of LPC Prediction Residue (Order 10)	28	12	40	10%
V/UV Classification Autocorrelation of LPC Prediction Residue (Order 50)	34	21	55	14%

TABLE 5-1: ERROR RATE OF DIFFERENT V/UV CLASSIFICATION SCHEMES

From the figures above, it is shown that V/UV classification from the autocorrelation of the LPC Prediction residue (Order 10) produced the least error. It was thus chosen to be implemented in the new PLC algorithm with V/UV classification.

5.3 New PLC Algorithm with V/UV Classification

5.3.1 Implementing Voiced/Unvoiced Classification into the New Packet Loss Concealment Algorithm

To provide a better approximation of the original signal, the concealment algorithm is further modified to perform a weighted summation of the speech predicted by linear prediction (i.e. equation (5-6) for the case where future samples are not considered) and the pitch-based prediction, using a voiced/unvoiced speech classification to tune the parameters. In our experiments, the voiced/unvoiced classification is a simple scheme based on the linear prediction residual correlation. The proposed packet loss concealment algorithm for PCM speech thus becomes:

$$S^{p}(n)=\sum_{i=1}^{P}\left(a_i\times S^{p}(n-i)\right)+\hat{S}(n)\times G \tag{5-6}$$

$$S^{f}(n)=\sum_{i=1}^{P}\left(a_i\times S^{f}(n+i)\right)+\hat{S}(n)\times G \tag{5-7}$$

$$S_{v}(n)=\begin{pmatrix}\left(\alpha_{v}\times S^{p}(n)+\beta_{v}\times\hat{S}(n)\right)\times\text{Hamming win. }2^{\text{nd}}\text{ half}\\+\left(\alpha_{v}\times S^{f}(n)+\beta_{v}\times\hat{S}(n)\right)\times\text{Hamming win. }1^{\text{st}}\text{ half}\end{pmatrix} \tag{5-8}$$

$$S_{uv}(n)=\begin{pmatrix}\left(\alpha_{uv}\times S^{p}(n)+\beta_{uv}\times\hat{S}(n)\right)\times\text{Hamm. win. }2^{\text{nd}}\text{ half}\\+\left(\alpha_{uv}\times S^{f}(n)+\beta_{uv}\times\hat{S}(n)\right)\times\text{Hamm. win. }1^{\text{st}}\text{ half}\end{pmatrix} \tag{5-9},$$

where $S_{v}(n)$ and $S_{uv}(n)$ are the final concealed signals for voiced and unvoiced frames, respectively, $S^{p}(n)$ and $S^{f}(n)$ are the linear prediction results from the past and future samples, respectively, α_{v} and β_{v} are summation weights for voiced frames, and

α_{uv} and β_{uv} are summation weights for unvoiced frames. The best results were obtained with $\alpha_v = 0.9$ and $\beta_v = 0.1$, $\alpha_{uv} = 0.6$ and $\beta_{uv} = 0.4$, which were optimized for a packet loss rate of 5%.

5.3.2 Performance of the New PLC Algorithm with V/UV Classification

The new PLC algorithm with V/UV classification was compared to the packet repetition method, to the ITU-T G.711 Appendix I concealment tool and to the new PLC algorithm presented in Chapter 4. The test was performed on a set of speech files from four speakers (two males and two females) referred to in the results as M1, M2, F1 and F2. For each of those speakers, 10 speech files were used, each containing two sentences in English with duration of 8 seconds. Frames of 80 samples (10 ms) were used. The format of the files was linear PCM, with 8 kHz sampling rate. The files were taken from the ITU-T supplement P.23.

The assessment tool used to evaluate the results of the concealment techniques was again the Perceptual Estimation of Speech Quality (PESQ) standard P.862 developed by the ITU-T [P862].

	M1	M2	F1	F2
New PLC Algorithm with V/UV Classification with 100% of Future Packets available	3.99	3.89	3.84	3.77
Combination of Prediction from Past Packet and 100% of Future Packets available	3.81	3.85	3.68	3.69
Combination of Prediction from Past Packet and 50% of Future Packets available	3.72	3.72	3.52	3.64
Prediction from Past Packet	3.66	3.51	3.42	3.44
G.711 Appendix I	3.45	3.41	3.36	3.31
Packet Repetition	2.99	3.00	2.73	3.13

TABLE 5-2: AVERAGE RESULTS OF 5% RANDOM PACKET LOSSS

	M1	M2	F1	F2
New PLC Algorithm with V/UV Classification with 100% of Future Packets available	3.61	3.55	3.53	3.53
Combination of Prediction from Past Packet and 100% of Future Packets available	3.58	3.64	3.50	3.52
Combination of Prediction from Past Packet and 50% of Future Packets available	3.43	3.46	3.33	3.36
Prediction from Past Packet	3.19	3.27	3.05	3.01
G.711 Appendix I	3.09	3.12	2.93	2.87
Packet Repetition	2.69	2.84	2.51	2.45

TABLE 5-3: AVERAGE RESULTS OF 10% RANDOM PACKET LOSS

	M1	M2	F1	F2
New PLC Algorithm with V/UV Classification with 100% of Future Packets available	3.20	3.02	3.05	2.97
Combination of Prediction from Past Packet and 100% of Future Packets available	3.27	3.33	3.23	3.30
Combination of Prediction from Past Packet and 50% of Future Packets available	3.03	3.18	3.03	3.08
Prediction from Past Packet	2.74	2.79	2.66	2.60
G.711 Appendix I	2.60	2.63	2.58	2.43
Packet Repetition	2.41	2.47	2.15	2.24

TABLE 5-4: AVERAGE RESULTS OF 25% RANDOM PACKET LOSS

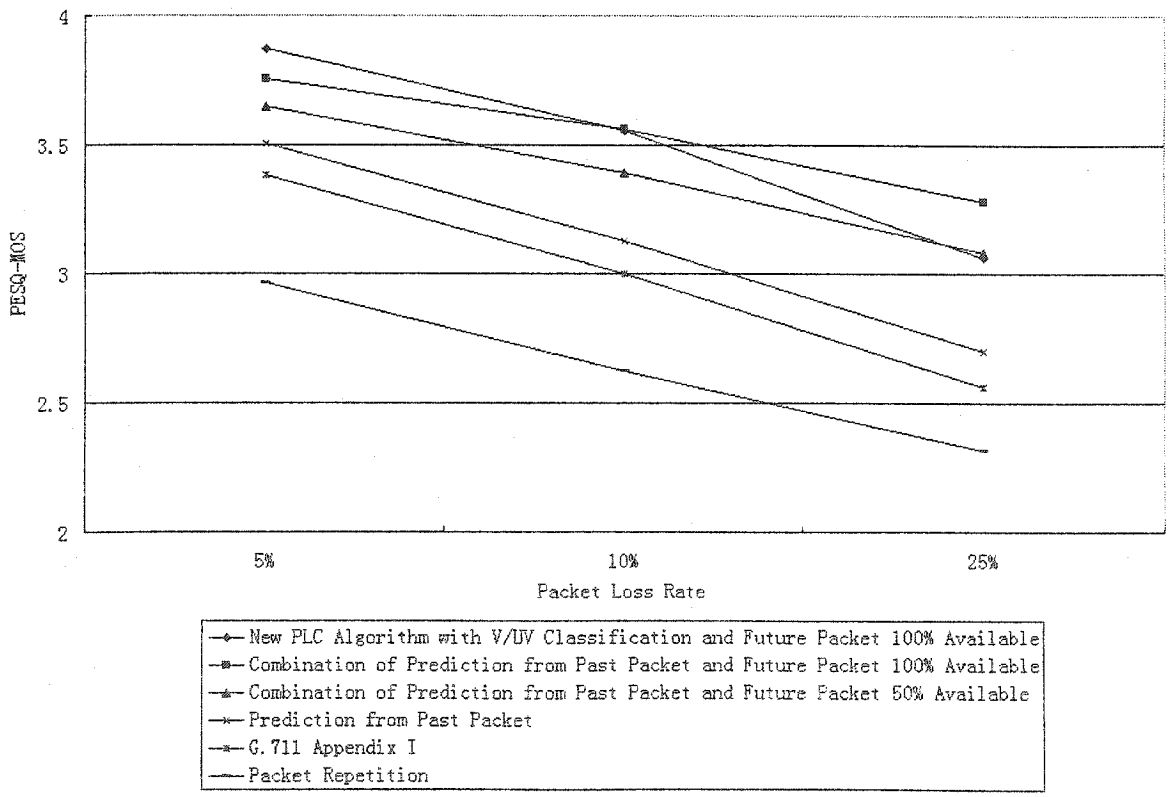


FIGURE 5-10: PERFORMANCE OF ALGORITHMS

Random loss pattern tests were performed for loss rates of 5%, 10 % and 25%. It can be seen from the figures that for the loss rate of 5%, the performance of the new PLC algorithm with V/UV classification is better than the new PLC algorithm presented in Chapter 4. And obviously, the performance of the new PLC algorithm with V/UV classification is better than the existing ITU-T G.711 Appendix I method and the packet repetition method.

For a loss rate of 10%, the performance of the new PLC algorithm with V/UV classification is almost the same as the new PLC algorithm presented in Chapter 4. And the performance of the new PLC algorithm with V/UV classification is thus better than the existing ITU-T G.711 Appendix I method and the packet repetition method. It can be observed that the new PLC algorithm with V/UV classification does not work as well for 10 % packet loss as for 5 % packet loss (previous paragraph), this is because the parameters were tuned for a loss rate of 5 %.

For the same reason, for a loss rate of 25% the performance of the new PLC algorithm with V/UV classification drops and its performance is not better than the new PLC algorithm presented in Chapter 4. But the performance of the new PLC algorithm with V/UV classification is still better than the existing ITU-T G.711 Appendix I method and the packet repetition method.

These results are emphasized and summarized in the following figure and tables.

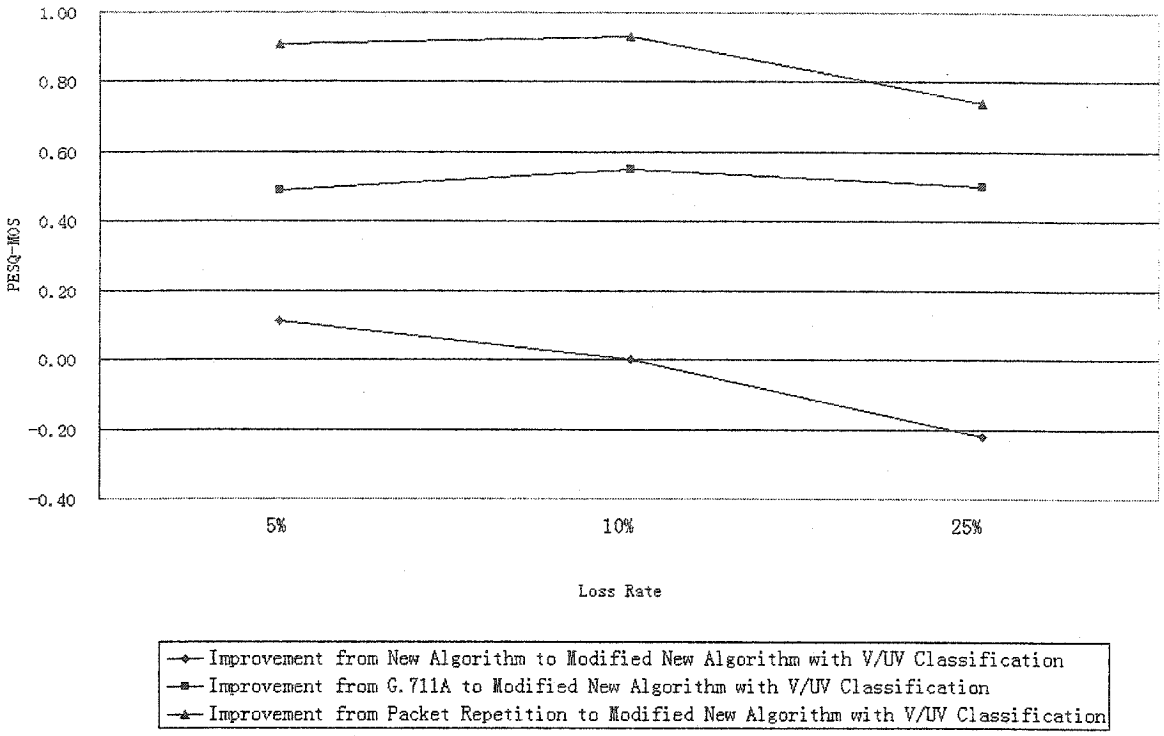


FIGURE 5-11: IMPROVEMENT OF NEW PLC ALGORITHM WITH V/UV CLASSIFICATION

Algorithms	Average Improvement in PESQ
New PLC Algorithm to New PLC Algorithm with Voiced/Unvoiced Classification	0.11
G.711 Appendix I to New PLC Algorithm with Voiced/Unvoiced Classification	0.49
Packet Repetition to New PLC Algorithm with Voiced/Unvoiced Classification	0.91

TABLE 5-5: IMPROVEMENT OF VOICED/UNVOICED DECISION NEW ALGORITHM (5% LOSS RATE)

Algorithms	Average Improvement in PESQ
New PLC Algorithm to New PLC Algorithm with Voiced/Unvoiced Classification	0.00
G.711 Appendix I to New PLC Algorithm with Voiced/Unvoiced Classification	0.55
Packet Repetition to New PLC Algorithm with Voiced/Unvoiced Classification	0.93

TABLE 5-6: IMPROVEMENT OF VOICED/UNVOICED DECISION NEW ALGORITHM (10% LOSS RATE)

Algorithms	Average Improvement in PESQ
New PLC Algorithm to New PLC Algorithm with Voiced/Unvoiced Classification	-0.22
G.711 Appendix I to New PLC Algorithm with Voiced/Unvoiced Classification	0.50
Packet Repetition to New PLC Algorithm with Voiced/Unvoiced Classification	0.74

TABLE 5-7: IMPROVEMENT OF VOICED/UNVOICED DECISION NEW ALGORITHM (25% LOSS RATE)

5.4 Summary

In this chapter, the new concealment algorithm for PCM packetized speech was implemented with an adaptation of the weighting coefficients α and β based on a V/UV classification. The performance of the modification provides encouraging results,

especially for low loss rates for which the parameters were optimized.

Chapter 6

Test of New PLC Algorithm with V/UV Classification on Additive Noise Distorted Speech and Reverberation Distorted Speech

This chapter is dedicated to test our new PLC algorithm with V/UV classification in more realistic environments. The new PLC algorithm with V/UV classification is tested here with additive noise distorted speech and reverberation distorted speech. It should be noticed that neither enhancement of speech nor de-reverberation algorithms are applied before or after the algorithm.

6.1 Noise

Noise in the input signal can degrade performance of speech significantly. This is generally dominated by the background acoustic noise picked up by the microphone. The primary sources are fans in the ventilation system or nearby equipment hum from fluorescent lighting and chatter from adjoining work areas. While usually much less significant, electronic noise from amplifiers, quantization noise from codecs and finite precision arithmetic can be lumped into this category. Active Noise Cancellation may be employed to improve the signal-to-noise ratio. However, this subject is beyond the scope of the thesis.

6.1.1 Additive Noise Model

In this research, we assume that a sequence of independent, identically distributed Gaussian noise signal d(n) has been added to a speech signal s(n), with their sum denoted by x(n). Then

$$x(n) = s(n) + d(n) \tag{6-1}$$

The noise amplitude was adjusted to achieve specified S/N (speech-to-noise) ratio for testing. The S/N ratio is defined as:

$$\text{S/N in dB} = 10\log_{10}\left(\frac{\sum_n s^2(n)}{\sum_n d^2(n)}\right) \tag{6-2}$$

where s(n) is the speech waveform, d(n) is the noise, and the summation is over the length of the test sentence.

The following figures show the original speech file (voiced parts or unvoiced parts) being distorted by additive noise with different levels of 20dB, 10dB and 0dB SNR respectively.

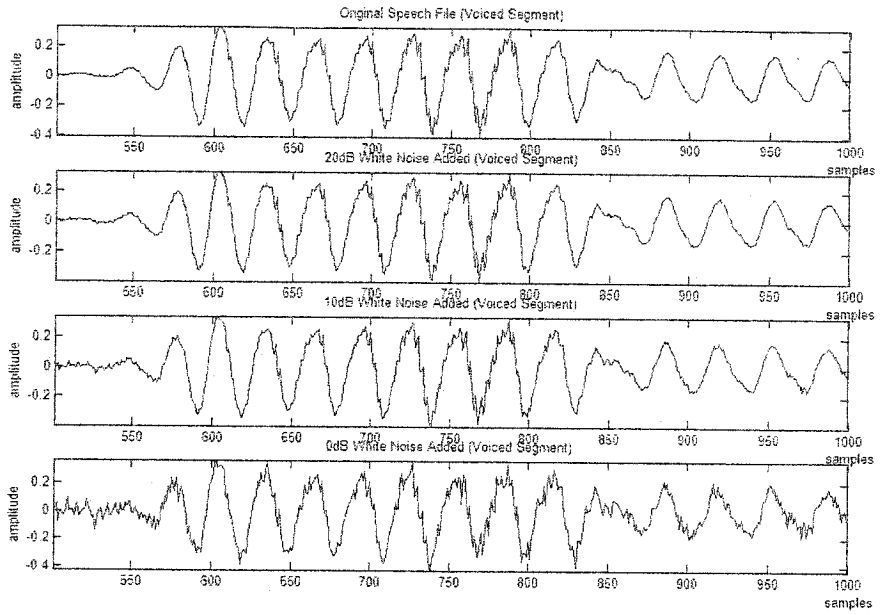


FIGURE 6-1: ADDITIVE WHITE NOISE DISTORTED SPEECH (VOICED SEGMENT)

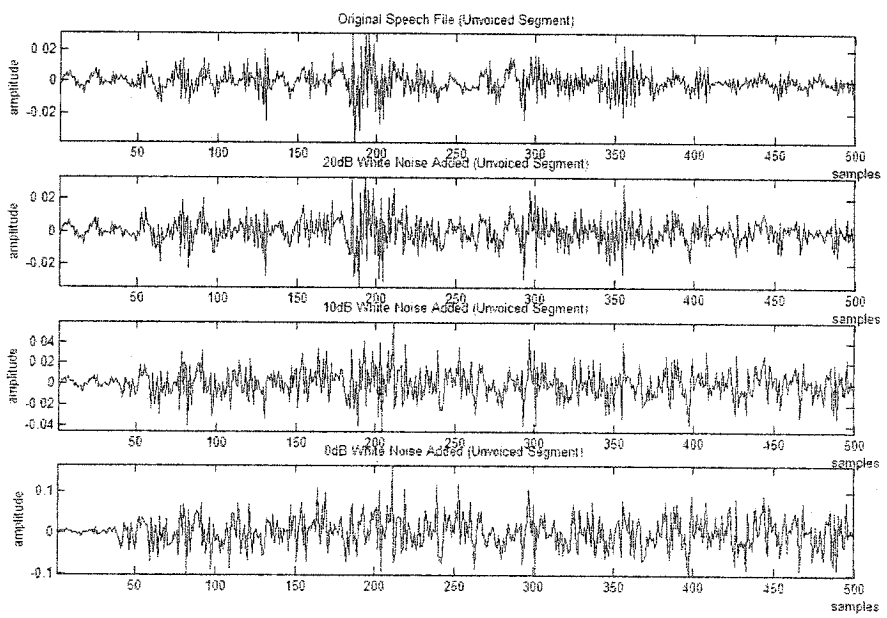


FIGURE 6-2: ADDITIVE WHITE NOISE DISTORTED SPEECH (UNVOICED SEGMENT)

6.2 Reverberation

Several examples of hands-free telephones are quite commonplace today. Many desktop phones in an office environment have a loud speaker capability. Corporate conference rooms are often equipped with some sophisticated version of these speakers phones. Video conference equipment also includes those components. All of these communication devices are distinguished from conventional telephone systems by the fact that the loudspeaker and the microphone are located at some distance from the users' ear and mouth. This has several implications, such as increased noise pick up by the microphone, including additive noise but also room effects such as reverberation. Another implication is coupling between the loudspeaker and the microphone, which is not addressed in this chapter. It should be noted that some hands-free terminals incorporate multiple loudspeakers, multiple microphones or both. For example, video conference system often provides a stereophonic communication channel.

Room reverberation consists of a series of echoes of the original speech, delayed from a few milliseconds to several seconds. The earliest echoes are single reflections from table tops and walls near the talker and the microphone. Later echoes are multiple reflections from the walls of the room which gradually become weaker. In a typical office their level is attenuated by 60dB after about one third of a second. This time is usually referred to as the reverberation time.

For speech sounds heard through a single microphone, early and late echoes seem to be perceived differently. The most important perceptual effect of early echoes is to change the frequency spectrum of the speech sounds, giving the speech a hollow quality. The effect is especially annoying in small, hard-walled rooms. Late reflections can be heard as distinct echoes of the speech sounds. This effect can be heard in an auditorium, in which speech generally does not have the hollow sound caused by early echoes but is

smeared out in time by later echoes.

For our experiments, the reverberation distortion is simply generated following the equation:

$$H(Z) = \frac{1}{\left(1 - 2r\cos(\omega_o)Z^{-1} + r^2 Z^{-2}\right)} \tag{6-3}$$

where the parameters are chosen as

$r = 0.95$ and $\omega_o = \frac{3\pi}{4}$. The effect of this reverberant system can be observed in the following two figures, for a voiced and an unvoiced speech segment.

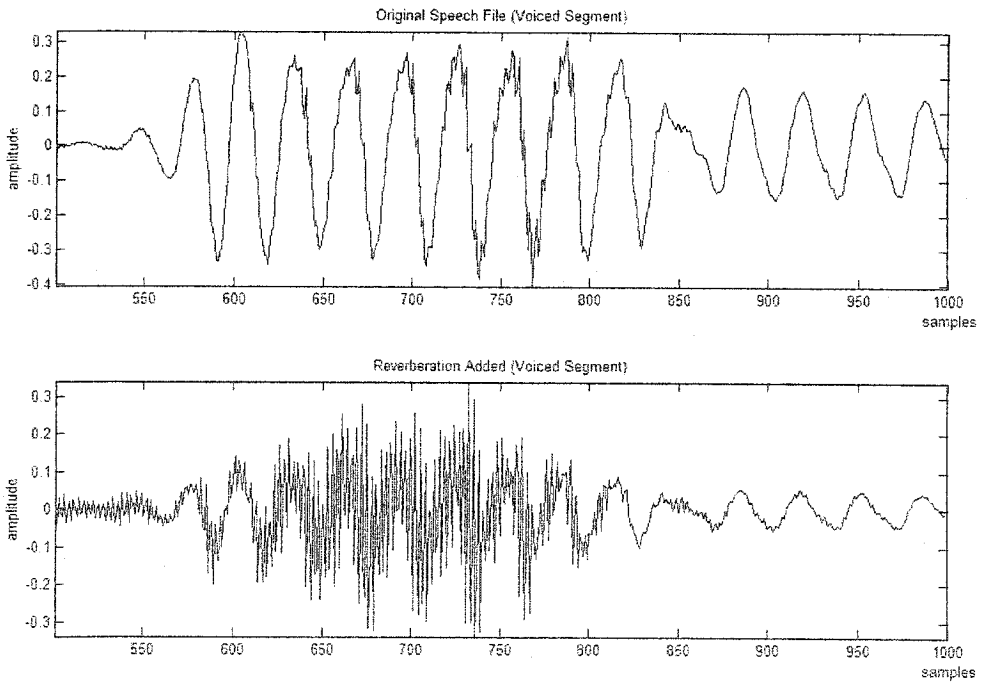


FIGURE 6-3: REVERBERATION DISTORTED SPEECH (VOICED SEGMENT)

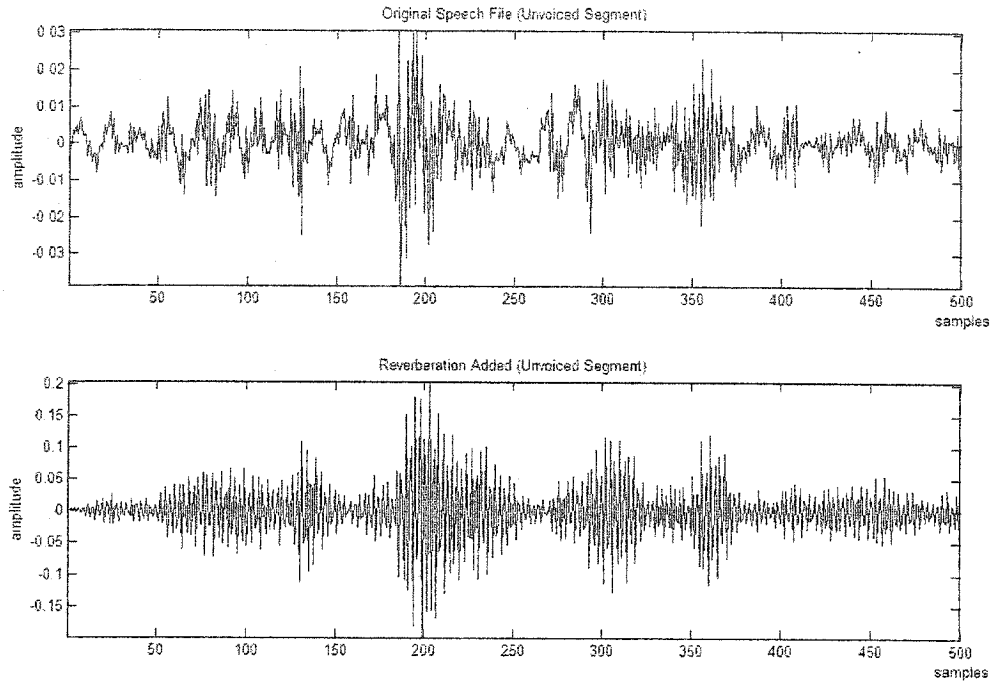


FIGURE 6-4: REVERBERATION DISTORTED SPEECH (UNVOICED SEGMENT)

6.3 Experiment Results and Discussion

The test was performed on a set of speech files from four speakers (two males and two females) referred to in the results as M1, M2, F1 and F2. For each of those speakers, 10 speech files were used, each containing two sentences in English with a duration of 8 sec. Frames of 80 samples (10 ms) were used. The format of the files was linear PCM, with 8 kHz sampling rate. The files were taken from the ITU-T supplement P.23.

For each speech file, a distortion is performed first. The speech files are distorted by additive white noise distortion or reverberation distortion.

Then the new PLC algorithm with V/UV classification is performed on the clean speech and the distorted speech.

The assessment tool used to evaluate the results of the concealment techniques was again the Perceptual Estimation of Speech Quality (PESQ) standard P.862 developed by the ITU-T [P862].

The results are presented in the following figures and tables.

	M1	M2	F1	F2
Additive White Noise Distortion (20db)	3.33	3.45	3.41	3.56
Additive White Noise Distortion (10db)	2.56	2.62	2.59	2.77
Additive White Noise Distortion (0db)	1.88	1.88	1.78	2.06
Reverberation Distortion	2.93	2.93	2.62	2.60

TABLE 6-1: DISTORTED SPEECH FILES WITHOUT PACKET LOSS

	M1	M2	F1	F2
New Algorithm with V/UV Classification	3.73	3.66	3.66	3.75
G.711 Appendix I	3.06	3.09	2.84	3.24
Packet Repetition	2.83	2.80	2.53	2.93
Silence Insertion	2.77	2.69	2.46	3.07

TABLE 6-2: DIFFERENT PLC ALGORITHMS FOR ADDITIVE WHITE NOISE DISTORTION 20DB SNR AND 5% PACKET LOSS

	M1	M2	F1	F2
New Algorithm with V/UV Classification	3.35	3.42	3.38	3.53
G.711 Appendix I	2.94	2.99	2.70	2.85
Packet Repetition	2.56	2.70	2.39	2.39
Silence Insertion	2.51	2.56	2.44	2.45

TABLE 6-3: DIFFERENT PLC ALGORITHMS FOR ADDITIVE WHITE NOISE DISTORTION 20DB SNR AND 10% PACKET LOSS

	M1	M2	F1	F2
New Algorithm with V/UV Classification	3.03	3.01	2.86	2.95
G.711 Appendix I	2.52	2.54	2.26	2.42
Packet Repetition	2.33	2.41	2.08	2.18
Silence Insertion	2.27	2.28	2.03	2.18

TABLE 6-4: DIFFERENT PLC ALGORITHMS FOR ADDITIVE WHITE NOISE DISTORTION 20DB SNR AND 25% PACKET LOSS

	M1	M2	F1	F2
New Algorithm with V/UV Classification	3.38	3.35	3.48	3.62
G.711 Appendix I	2.48	2.49	2.39	2.68
Packet Repetition	2.38	2.40	2.18	2.53
Silence Insertion	2.40	2.29	2.15	2.64

TABLE 6-5: DIFFERENT PLC ALGORITHMS FOR ADDITIVE WHITE NOISE DISTORTION 10DB SNR AND 5% PACKET LOSS

	M1	M2	F1	F2
New Algorithm with V/UV Classification	3.07	3.19	3.13	3.26
G.711 Appendix I	2.45	2.45	2.32	2.50
Packet Repetition	2.24	2.35	2.12	2.18
Silence Insertion	2.22	2.29	2.14	2.22

TABLE 6-6: DIFFERENT PLC ALGORITHMS FOR ADDITIVE WHITE NOISE DISTORTION 10DB SNR AND 10% PACKET LOSS

	M1	M2	F1	F2
New Algorithm with V/UV Classification	2.79	2.88	2.68	2.89
G.711 Appendix I	2.38	2.39	2.22	2.12
Packet Repetition	2.12	2.18	1.89	2.02
Silence Insertion	2.07	2.12	1.86	2.03

TABLE 6-7: DIFFERENT PLC ALGORITHMS FOR ADDITIVE WHITE NOISE DISTORTION 10DB SNR AND 25% PACKET LOSS

	M1	M2	F1	F2
New Algorithm with V/UV Classification	2.82	2.76	2.91	3.08
G.711 Appendix I	1.85	1.86	1.72	2.04
Packet Repetition	1.80	1.86	1.62	1.98
Silence Insertion	1.86	1.79	1.65	2.03

TABLE 6-8: DIFFERENT PLC ALGORITHMS FOR ADDITIVE WHITE NOISE DISTORTION 0DB SNR AND 5% PACKET LOSS

	M1	M2	F1	F2
New Algorithm with V/UV Classification	2.54	2.60	2.53	2.71
G.711 Appendix I	1.85	1.84	1.72	1.98
Packet Repetition	1.77	1.79	1.61	1.81
Silence Insertion	1.79	1.78	1.62	1.83

TABLE 6-9: DIFFERENT PLC ALGORITHMS FOR ADDITIVE WHITE NOISE DISTORTION 0DB SNR AND 10% PACKET LOSS

	M1	M2	F1	F2
New Algorithm with V/UV Classification	2.27	2.37	2.23	2.45
G.711 Appendix I	1.80	1.83	1.66	1.91
Packet Repetition	1.72	1.73	1.53	1.71
Silence Insertion	1.71	1.73	1.52	1.71

TABLE 6-10: DIFFERENT PLC ALGORITHMS FOR ADDITIVE WHITE NOISE DISTORTION 0DB SNR AND 25% PACKET LOSS

	M1	M2	F1	F2
New Algorithm with V/UV Classification	3.01	2.95	2.81	2.92
G.711 Appendix I	2.80	2.74	2.39	2.36
Packet Repetition	2.67	2.66	2.30	2.38
Silence Insertion	2.59	2.58	2.29	2.40

TABLE 6-11: DIFFERENT PLC ALGORITHMS FOR REVERBERATION DISTORTION AND 5% PACKET LOSS

	M1	M2	F1	F2
New Algorithm with V/UV Classification	2.76	2.68	2.54	2.59
G.711 Appendix I	2.65	2.73	2.31	2.25
Packet Repetition	2.51	2.55	2.19	2.12
Silence Insertion	2.48	2.50	2.21	2.14

TABLE 6-12: DIFFERENT PLC ALGORITHMS FOR REVERBERATION DISTORTION AND 10% PACKET LOSS

	M1	M2	F1	F2
New Algorithm with V/UV Classification	2.46	2.53	2.22	2.17
G.711 Appendix I	2.55	2.56	2.25	2.17
Packet Repetition	2.31	2.39	2.00	1.97
Silence Insertion	2.25	2.30	1.91	1.93

TABLE 6-13: DIFFERENT PLC ALGORITHMS FOR REVERBERATION DISTORTION AND 25% PACKET LOSS

	M1	M2	F1	F2
New Algorithm with V/UV Classification	3.99	3.89	3.84	3.77
G.711 Appendix I	3.45	3.41	3.36	3.31
Packet Repetition	2.99	3.00	2.73	3.13
Silence Insertion	2.96	2.85	2.64	3.26

TABLE 6-14: DIFFERENT PLC ALGORITHMS FOR CLEAN SPEECH AND 5% PACKET LOSS

	M1	M2	F1	F2
New Algorithm with V/UV Classification	3.61	3.55	3.53	3.53
G.711 Appendix I	3.09	3.12	2.93	2.87
Packet Repetition	2.69	2.84	2.51	2.45
Silence Insertion	2.65	2.66	2.57	2.53

TABLE 6-15: DIFFERENT PLC ALGORITHMS FOR CLEAN SPEECH AND 10% PACKET LOSS

	M1	M2	F1	F2
New Algorithm with V/UV Classification	3.20	3.02	3.05	2.97
G.711 Appendix I	2.60	2.63	2.58	2.43
Packet Repetition	2.41	2.47	2.15	2.24
Silence Insertion	2.33	2.32	2.09	2.23

TABLE 6-16: DIFFERENT PLC ALGORITHMS FOR CLEAN SPEECH AND 25% PACKET LOSS

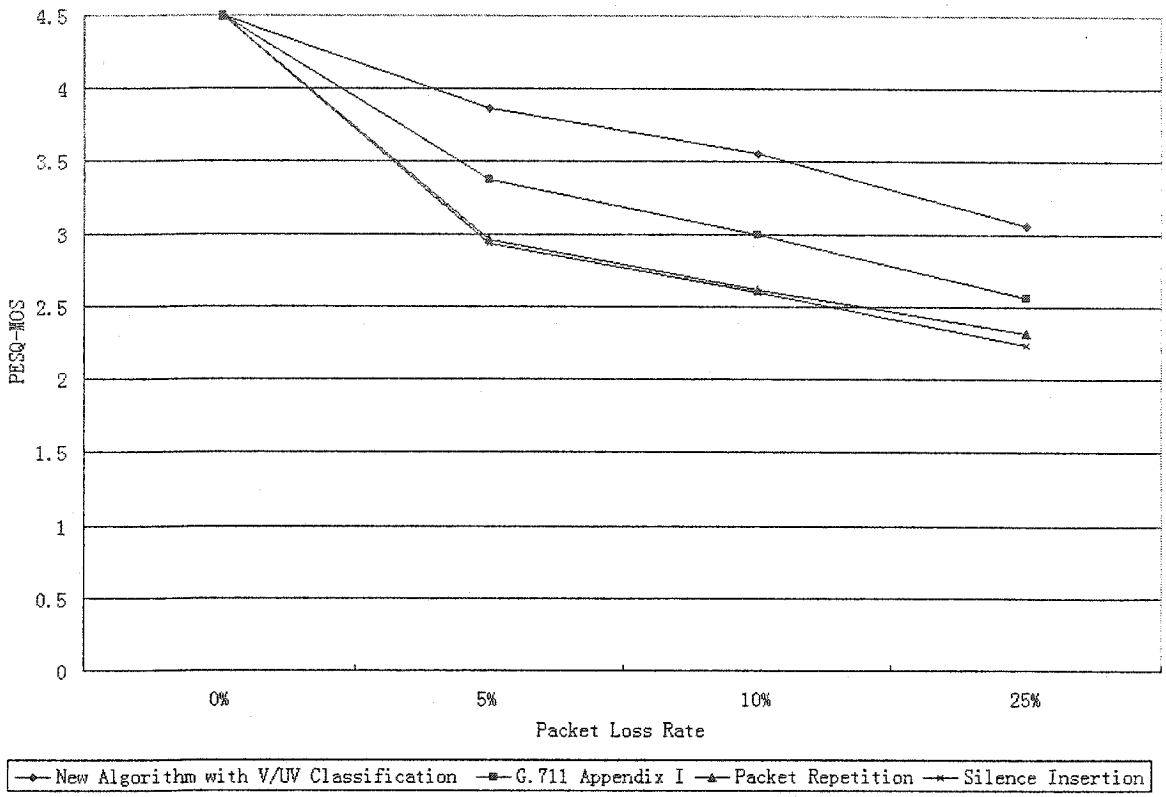


FIGURE 6-5: ALGORITHMS PERFORMED ON CLEAN SPEECH FILES

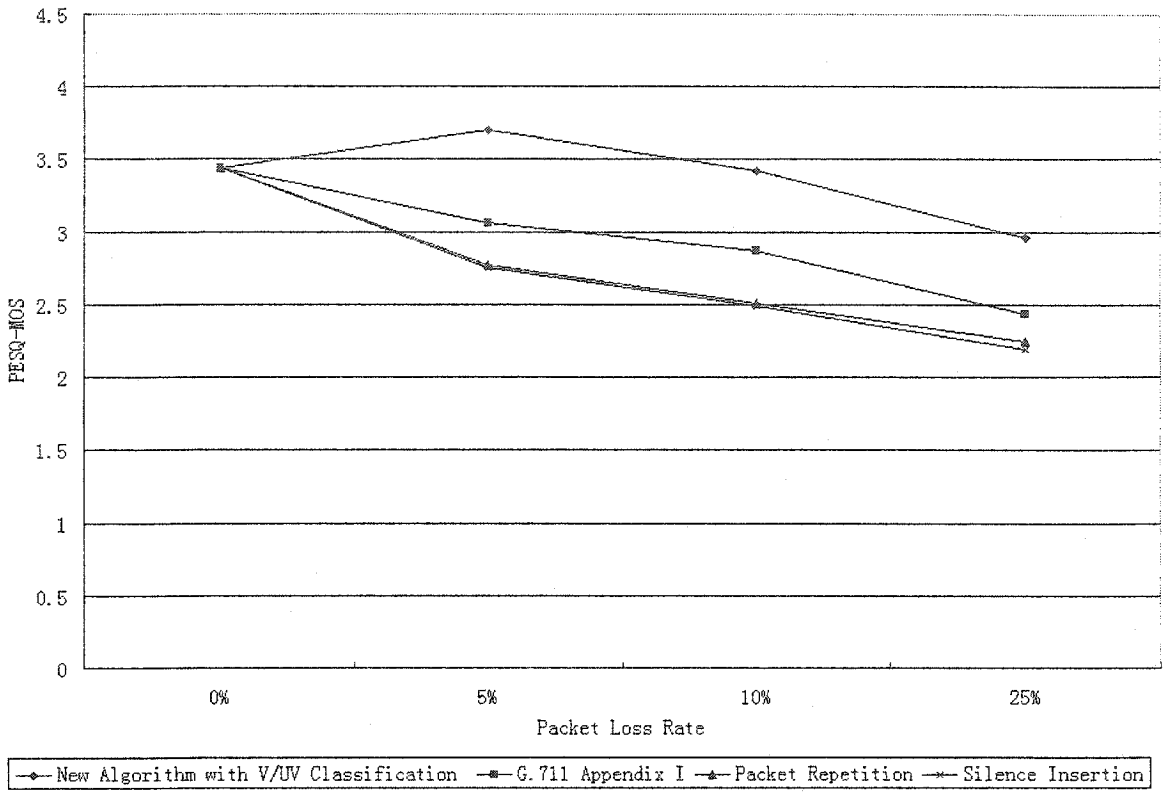


FIGURE 6-6: ALGORITHMS PERFORMED ON 20DB ADDITIVE WHITE NOISE DISTORTED SPEECH FILES

FIGURE 6-7: ALGORITHMS PERFORMED ON 10DB ADDITIVE WHITE NOISE DISTORTED SPEECH FILES

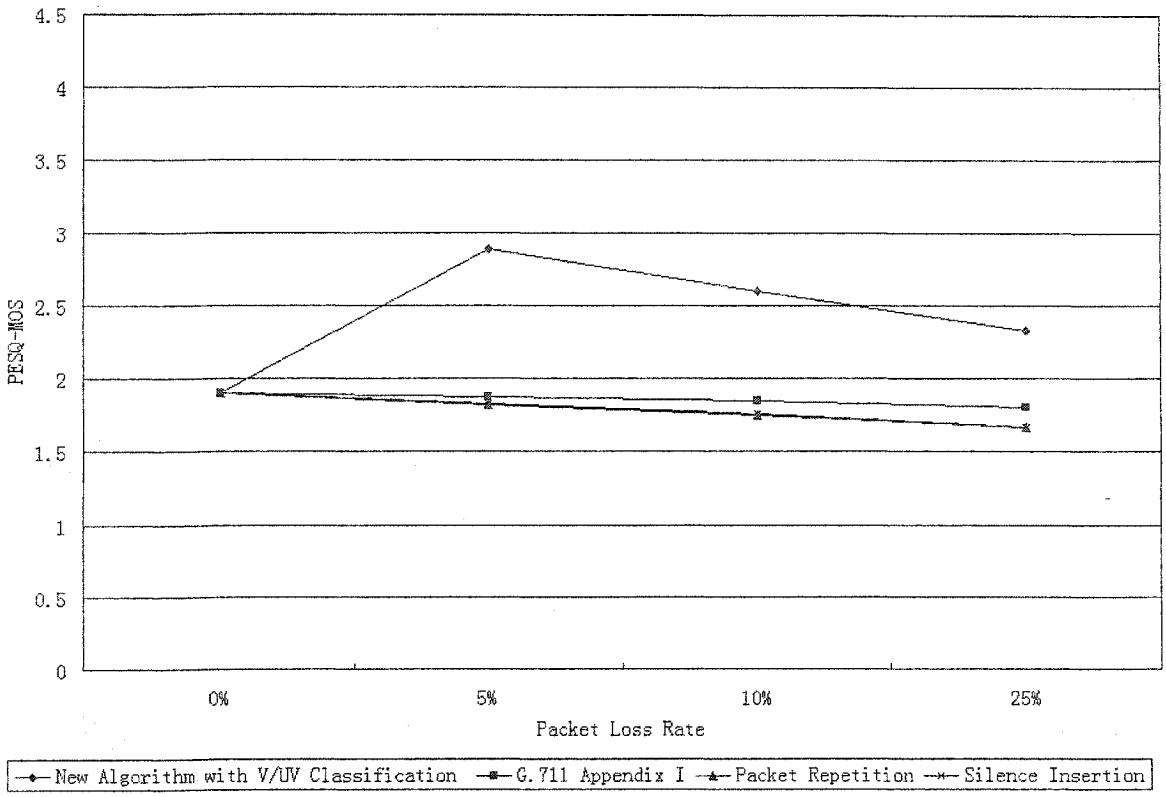


FIGURE 6-8: ALGORITHMS PERFORMED ON 0DB ADDITIVE WHITE NOISE DISTORTED SPEECH FILE

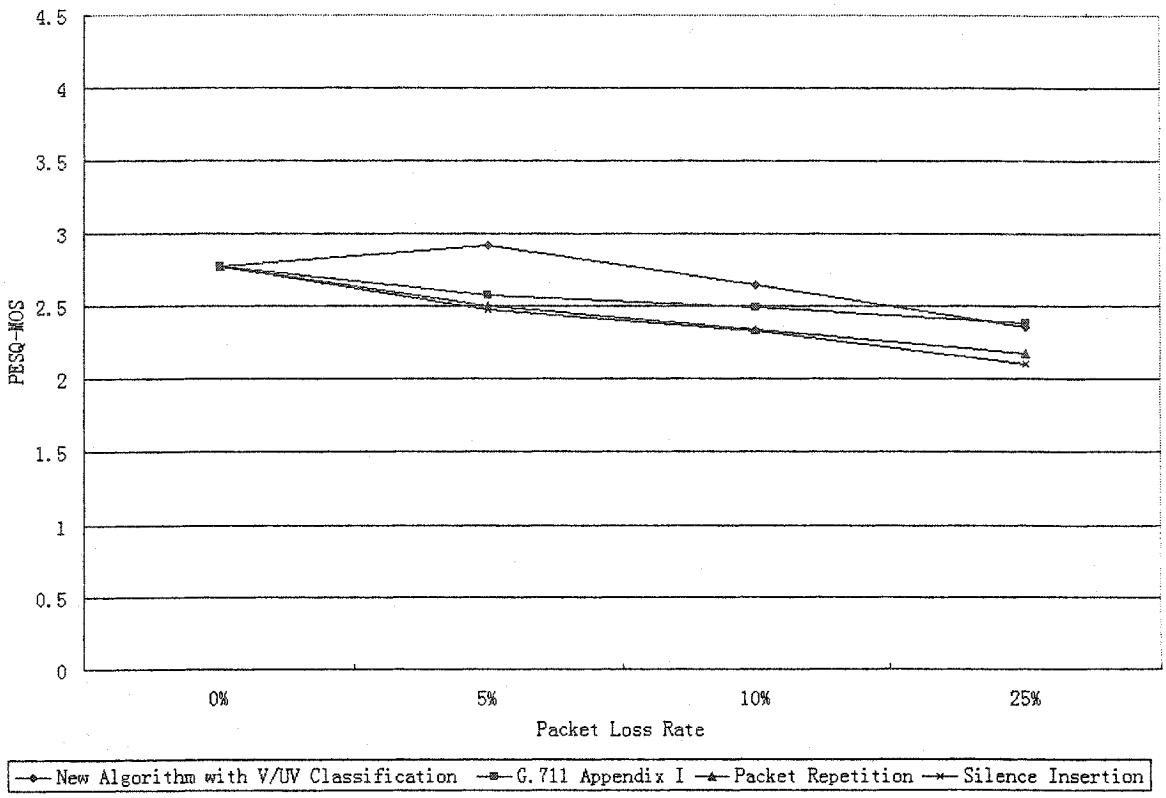


FIGURE 6-9: ALGORITHMS PERFORMED ON REVERBERATION DISTORTED SPEECH FILES

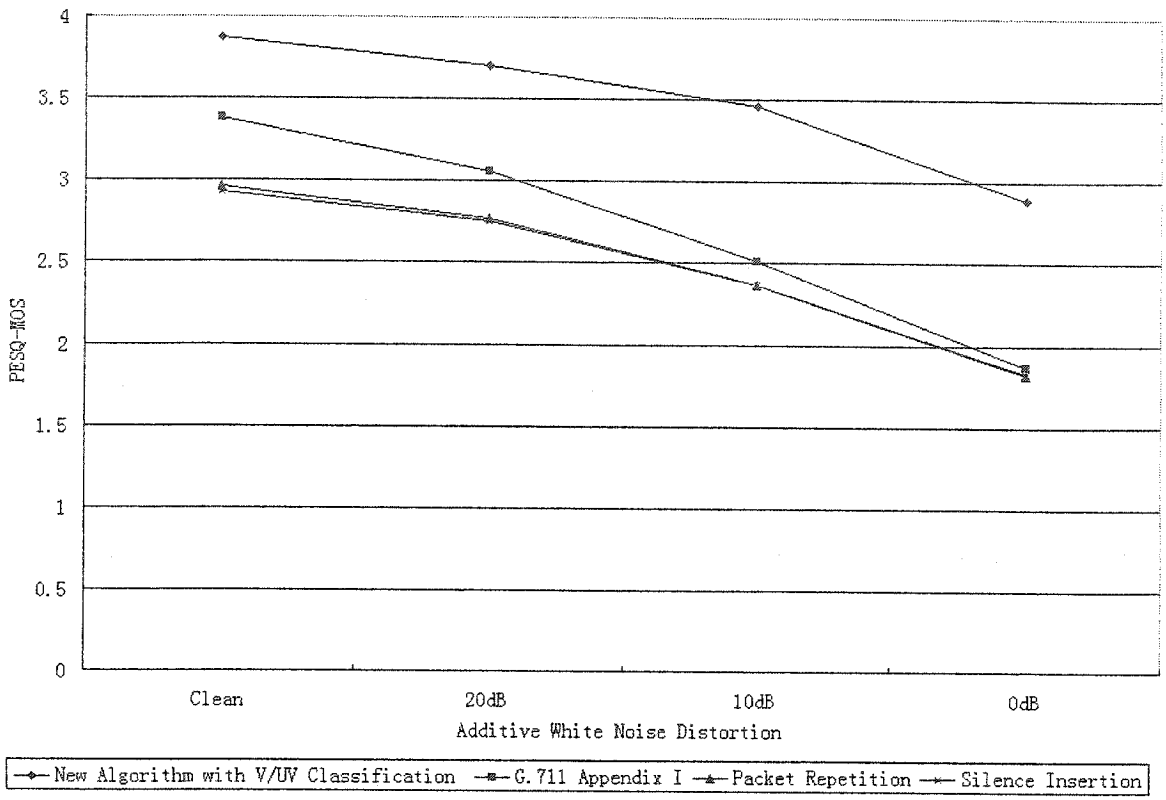


FIGURE 6-10: ALGORITHMS PERFORMED ON SPEECH FILES WITH 5% PACKET LOSS

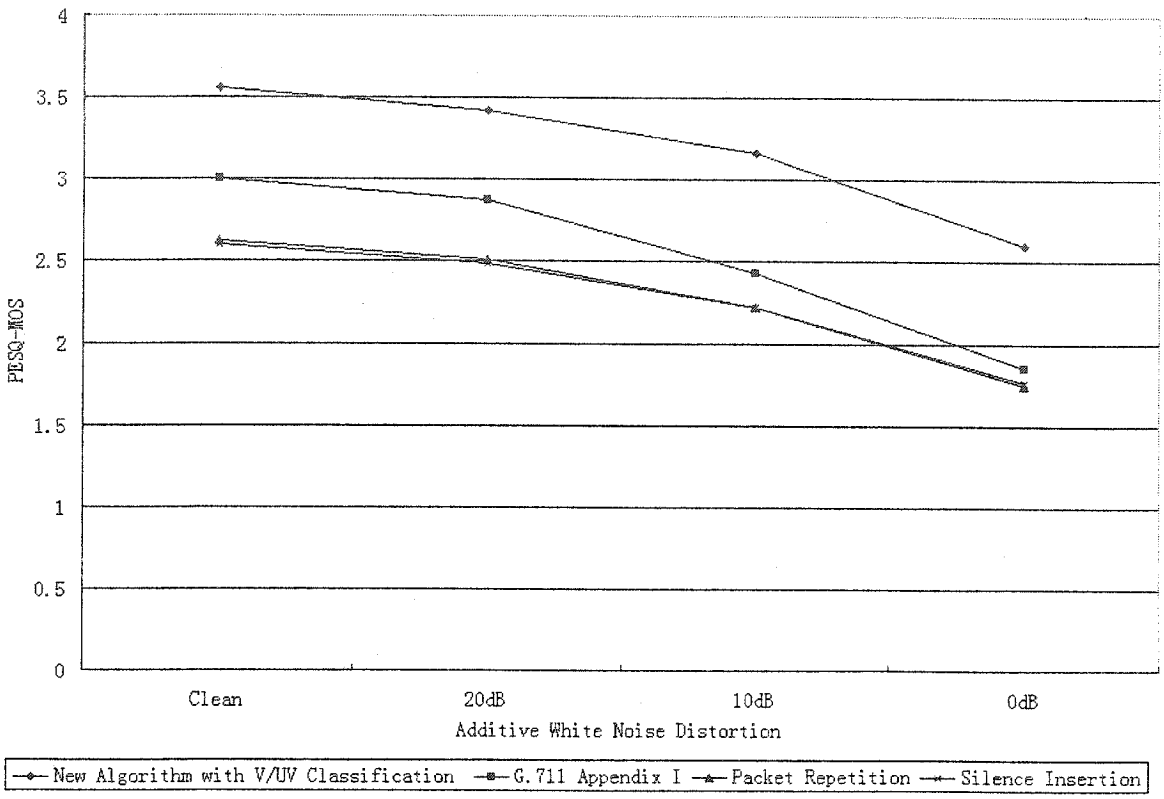


FIGURE 6-11: ALGORITHMS PERFORMED ON SPEECH FILES WITH 10% PACKET LOSS

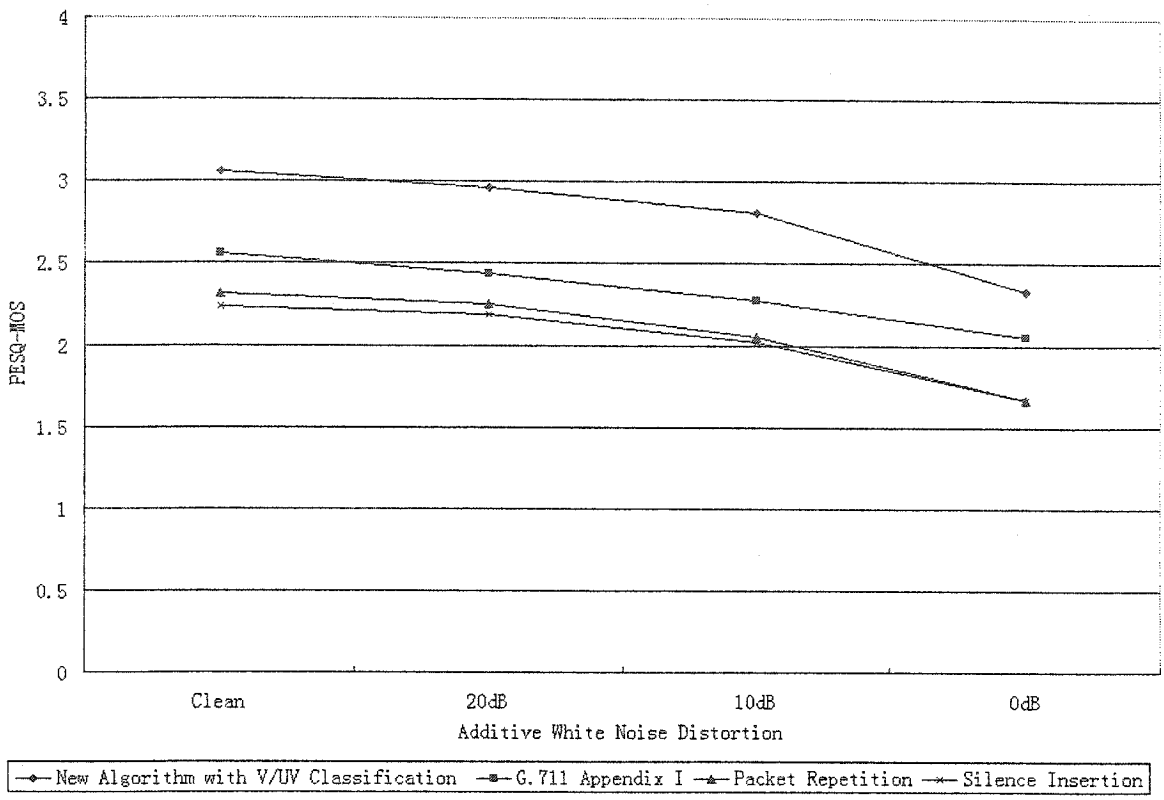


FIGURE 6-12: ALGORITHMS PERFORMED ON SPEECH FILES WITH 25% PACKET LOSS

Several conclusions can be drawn from these results. For similar distortion conditions and loss conditions, the order of performance observed in the previous chapters for the different PLC algorithms has not changed: the proposed V/UV classification algorithm outperforms the ITU-T G.711 Appendix I method, which outperforms the packet repetition method. An additional method was compared in this chapter: the silence substitution method (i.e. absence of concealment), produced the worse results. It should noted that in practice D/A interpolation would smooth the discontinuities in the signal produced by the silence substitution algorithm. Therefore there would be less clicks sound effects and the performance of silence substitution may thus be a bit better than reported here. But the results also show that the proposed PLC algorithm with V/UV classification operating under distortion conditions can even outperform other algorithms operating under less

severe distortion conditions.

First, for a given noise distortion the algorithm can often produce a better performance at higher loss rates than other algorithms at lower loss rates:

- the proposed PLC algorithm with V/UV classification produces a better performance for clean speech at 10 % loss rate than any other algorithm operating on clean speech at 5% loss rate
- the proposed PLC algorithm with V/UV classification produces a better performance for clean speech at 25 % loss rate than any other algorithm operating on clean speech at 10% loss rate
- the proposed PLC algorithm with V/UV classification produces a better performance for speech distorted with 20 dB SNR additive noise + 10 % loss rate than any other algorithm with 20 dB SNR additive noise + 5 % loss rate
- the proposed PLC algorithm with V/UV classification produces a better performance for speech distorted with 20 dB SNR additive noise + 25 % loss rate than any other algorithm with 20 dB SNR additive noise + 10 % loss rate
- the proposed PLC algorithm with V/UV classification produces a much better performance for speech distorted with 10 dB SNR additive noise + 25 % loss rate than any other algorithm with 10 dB SNR additive noise + 5 % loss rate
- the proposed PLC algorithm with V/UV classification produces a much better performance for speech distorted with 0 dB SNR additive noise + 25 % loss rate than any other algorithm with 0 dB SNR additive noise + 5 % loss rate
- the proposed PLC algorithm with V/UV classification produces a better average performance for speech distorted with reverberation noise + 10 % loss rate than any other algorithm with reverberation noise + 5 % loss rate.

Also, for a given loss rate, the proposed algorithm can often produce a better

performance under a higher noise distortion condition than other algorithms under a lower noise distortion condition:

- the proposed PLC algorithm with V/UV classification produces a better average performance for speech distorted with 10 dB SNR additive noise + 5 % loss rate than any other algorithm with clean speech + 5 % loss rate
- the proposed PLC algorithm with V/UV classification produces a better performance for speech distorted with 0 dB SNR additive noise + 5 % loss rate than any other algorithm with 10 dB SNR additive noise + 5 % loss rate
- the proposed PLC algorithm with V/UV classification produces a better average performance for speech distorted with 10 dB SNR additive noise + 10 % loss rate than any other algorithm with clean speech + 10 % loss rate
- the proposed PLC algorithm with V/UV classification produces a better performance for speech distorted with 0 dB SNR additive noise + 10 % loss rate than any other algorithm with 10 dB SNR additive noise + 10 % loss rate
- the proposed PLC algorithm with V/UV classification produces a better performance for speech distorted with 10 dB SNR additive noise + 25 % loss rate than any other algorithm with clean speech + 25 % loss rate
- the proposed PLC algorithm with V/UV classification produces a slightly better average performance for speech distorted with 0 dB SNR additive noise + 25 % loss rate than any other algorithm with 10 dB SNR additive noise + 25 % loss rate.

The most surprising and unexpected results may be the following, comparing to the original speech with no packet loss:

- the proposed PLC algorithm with V/UV classification produces a better performance for speech distorted with 20 dB SNR additive noise + 5 % loss rate than the original speech with 20 dB SNR additive noise and no packet loss

- the proposed PLC algorithm with V/UV classification produces a better performance for speech distorted with 10 dB SNR additive noise + 25 % loss rate than the original speech with 10 dB SNR additive noise and no packet loss
- the proposed PLC algorithm with V/UV classification produces a better performance for speech distorted with 0 dB SNR additive noise + 25 % loss rate than the original speech with 0 dB SNR additive noise and no packet loss
- the proposed PLC algorithm with V/UV classification produces a better performance for speech distorted with reverberation noise + 5 % loss rate than the original speech with reverberation noise and no packet loss.

The explanation for these results is that the proposed concealment method with V/UV classification performs some enhancement of the noise distorted speech (either in terms of SNR or perceptually), by computing a LPC prediction of the missing speech samples from the past samples and the futures samples. Since only the LPC-predictable components are reproduced in the reconstructed speech, it therefore has less noise distortion in it.

It should be noted that for speech distorted with reverberation noise, the difference of performance between the proposed PLC algorithm and the ITU-T G.711 appendix I method was found to be smaller. While the proposed method produces better results for 5 % and 10 % loss rates, for 25 % loss rates the ITU-T G.711 appendix I method produces slightly better average results.

6.4 Summary

A new PLC algorithm with V/UV classification was tested on distorted speech. The results show that the new PLC algorithm with V/UV classification produces a fairly good concealment performance on distorted speech, outperforming other concealment algorithms even if they operate on less distorted speech.

Chapter 7

Conclusion and Future Work

7.1 Summary

In this thesis, a new concealment algorithm for PCM packetized speech was presented. The model provides very encouraging results for the idea of combining a pitch prediction along with a high-order linear prediction to produce the concealed speech samples. Future samples are taken into account (optionally), as well as the adaptation of the weighting coefficients α and β based on a V/UV classification. For clean speech, the PESQ-MOS scores obtained for the random loss tests have shown that the proposed algorithm exhibits a superior high-quality concealment performance in all cases, when compared to an existing commercial methods (silence substitution, packet repetition) or to the ITU-T G.711 Appendix I concealment technique. The new PLC algorithm with V/UV classification was further tested in distorted speech like additive white noise distorted speech and reverberation distorted speech. Experiments showed that the new PLC algorithm with V/UV classification can produce a fairly good performance under these conditions.

7.2 Motivation for Future Work

For distorted speech, to improve the performance of the speech transmission system including the PLC algorithm, the combination of the proposed algorithm with speech enhancement and de-reverberation algorithms should be investigated. Also, the coupling between loudspeakers and microphones should be considered, with possibly an echo

canceller to mitigate it.

Investigating the interaction of the concealment algorithm with network or acoustic echo cancellers would be helpful to provide an idea about the applicability of the algorithm in real transmission situations.

Bibliography:

BB98 S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "*An Architecture for Differentiated Services*". Internet requests for comments 2475, December 1998, http://www.cis.ohio-state.edu/htbin/rfc/rfc2475.html.

BCS94 R. Braden, D. Clark, and S. Shenker, "*Integrated services in the Internet architecture: An overview*". Internet requests for comments 1633, June 1994, http://www.cis.ohio-state.edu/htbin/rfc/rfc1633.html

BCV95 J-C.Bolot, H. Crepin and A.Vega-Garcia "*Analysis of Audio packet loss in the Internet*". Proceeding of 5th International Workshop on Networking and Operating System Support for Digital Audio and Video, Durham, New Hampshire, USA, pp. 163-174, 1995

BG96 J.-C. Bolot and A.V. Garcia. "*Control Mechanisms for Packet Audio in the Internet*". Proceedings of IEEE INFOCOM, San Francisco, CA, pp. 232-239, April 1996.

BG98 J.C.Bolot and A.Vega-Farcia, "*Control Mechanisms for Packet Audio in the Internet*". IEEE workshop on Signal Processing Systems, San Francisco, CA, USA, pp. 220-229, 1998

BL92 M.M.Lara-Barron and G.B.Lockhart. "*Speech Encoding and Reconstruction for Packet Networks Using Redundancy*". IEE Colloquium on Coding for Packet Video and Speech Transmission, pp. 1-4, February 1992.

COH80 D. Cohen. "*On Packet Speech Communications*". In Proceedings of the Fifth International Conference on Computer Communications, Atlanta, GA, pp. 271--274, October 1980.

EBA03 M.M. Elsabrouty, M. Bouchard and T. Aboulnasr, "*A new hybrid long-term and short-term prediction algorithm for packet loss erasure over IP-networks*". Proceedings of IEEE-EURASIP Seventh International Symposium on Signal Processing and its Applications (ISSPA) 2003, Paris, France, vol. 1, pp.361-364, July 2003.

EBA04 M.M. Elsabrouty, M. Bouchard and T. Aboulnasr, "*Receiver-based packet loss concealment for pulse code modulation (PCM G.711) coder*". Signal Processing, vol. 84, n. 3, pp. 663-667, March 2004.

G711 ITU-T, "*Pulse Code Modulation (PCM) of voice frequencies*". Nov. 1988. ITU-T Recommendation G.711.

G711I ITU-T, "*A High Quality Low-Complexity Algorithm for Packet Loss Concealment with G.711*". Sept. 1999. ITU-T Recommendation G.711Appendix I.

GL86 D.J. Goodman, G.B. Lockhart, O.J. Wasem, and W. Wong. "*Wave Form Sumstitution Techniques for Recovering Missing Speech Segments in Packet Voice Communications*". IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP-34, pp. 1449-1464, December 1986

GWDP 88 D.J.Goodman, O.J.Wasem, C.A.Dvorak and H.G.Page."*The Effect of Waveform Substitution on the Substitution on the Quallity of PCM Packet Communications*". IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP-36, pp. 342-348, March 1988.

H225 International Telecommunication Union, "*ITU-T H.225.0, Call signaling protocol and media stream packetization for packet-based multimedia communication systems*". 2001.

H245 International Telecommunication Union, "*ITU-T H.245, Control protocol for multimedia communications*". 2000.

H323 International Telecommunication Union, "*ITU-T H.323, Packet-based multimedia communications systems*". 2000.

HS95 V.Hardman, M.A.Sasse, M.Hardley and A.Watson, "*Reliable Audio For Use over the Internet*". Proceedings of INET'95, Honolulu, Hawaii, pp.171-178, June 1995

IEC02 International Engineering Consortium, "*Voice portal solutions: An introduction to Next-Generation Network services; the next big opportunity on the web*". 2002, http://www.iec.org/online/tutorials/voice portal/topic01.html.

IV95 A. Ingle and V. Vaishampayan. "*DPCM System Design for Diversity Systems With Applications to Packetized Speech*". IEEE Transactions on Speech and Audio Processing, pp.48-58, January 1995.

JAY93 N.S. Jayant. "*High Quality Networking of Audio-Visual Information*". IEEE Communications Magazine, pp. 84-95, September 1993.

JC81 N.S. Jayant and S.W. Christensen. "*Effects of Packet Losses in Wave form Coded Speech and Improvements due to an Odd-Even Sample Interpolation Procedure*". IEEE Transactions on Communications, Vol. COM-29(2), pp.101-109, February 1981.

ME00 L. Mathy, C. Edwards, D. Hutchison, and L. University, "*The Internet: A global telecommunications solution?*" IEEE Network, vol. 14, pp. 46-57, July-Aug. 2000.

MJV96 S. McCanne, V. Jacobson, and M. Vetterli. "*Receiver-driven Layered Multicast*". In Proceedings ACM SIGCOMM, pp. 117-130, Stanford, CA, September 1996.

MYT87 N. Matsuo, M. Yuito, and Y. Tokunaga. "*Packet Interleaving for Reducing Speech Quality Degradation in Packet Voice Communications*". In Proceedings GLOBECOM, Tokyo, Japan, pp. 1787-1791, 1987.

P862 ITU-T P.862 (2000). "*Perceptual Evaluation of Speech Quality (SEQP), An Objective Method for End-toned Speech Quality Assessment of Narrow-band Telephone Network and Speech Codecs*". ITU-T Recommendation

PER99 C. Perkins. "*RTP Payload Format for Interleaved Media*". Internet Draft, IETF Audio/Video Transport Working Group, February 1999.
ftp://ftp.ietf.org/internet-drafts/draft-ietf-avt-interleaving-01.txt.

PH98 C. Perkins and O. Hodson. "*Options for the Repair of Streaming Media*". RFC 2354, IETF, June 1998.
ftp://ftp.ietf.org/rfc/rfc2354.txt.

PHH98 C.Perkins, O.Hodson and V.Hardman, "*A Servay of Packet Loss Recovery Techniques for Streaming Audio*", IEEE Network, vol.12, no.5, pp.40-48, Sept-Oct 1998.

PM96 J. G. Proakis and D. G. Manolakis, "*Digital Signal Processing Principles, Algorithms, and Applications*". Third Edition, Macmillan, New York, 1996.

PRM98 M. Podolsky, C. Romer, S. McCanne, "*Simulation of FEC-based error control for packet audio on the Internet*". Proceeding of IEEE Infocom'98, San Francisco, CA, pp. 505-515, April 1998

RAM70 J.L. Ramsey. "*Realization of Optimum Interleavers*". IEEE Transactions on Information Theory, IT-16, pp. 338-345, May 1970.

RHE98 I.Rhee. "*Error Control Techniques for Interactive Low bit Rate Video Transmission Over the Internet*", Proceeding of ACM SIGCOMM, Vancouver, B.C., Canada, pp. 290-301, September 1998

RI97 D. Reininger and R. Izmailov. "*Soft Quality of Service with VBR Video". In Proceedings of 8th International Workshop on Packet Video (AVSPN97)*". Aberdeen, Scotland, pp.207-211, September 1997.

RK94 R.Ramjee, J.Kurose, D.Towsley and H. Schulzrinne. "*Adaptive Playout Mechanism for Packetized Audio Applications in Aide Area Networks*". In proceedings IEEE INFOCOM, pp. 680-688, Toronto, ON, Canada, 1994

SAE96 A. Shah, S. Atungsiri, A. Kondoz, and B. Evans, "*Lossy multiplexing of low bit rate speech in thin route telephony*". Electronics Letters, vol. 32, pp. 95-97, January 1996.

SC96 H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, "*RTP: A Transport Protocol for Real Time Applications*". Internet requests for comments 1889, January 1996, http://info.internet.isi.edu:80/in-notes/rfc/_les/rfc1889.txt.

SC03 H.Schuzrinne and S.Casner."*RTP Profile for Audio and Video Conferences with Minimal Control*". Internet Draft, IETF Audio-Video Transport Group, July 2003. ftp://ftp.ietf.org/internet-drafts/draft-ietf-avt-profile-new-13.txt

SCFJ96 H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson. "*RTP: a transport protocol for real-time applications*". RFC 1889, IETF, January 1996. ftp://ftp.ietf.org/rfc/rfc1889.txt.

TEL99 Committee T1 Telecommunications. "*American National Standard for Packet Loss Concealment for Use with ITU-T Recommendation G.722*". Draft T1 Standard T1A1.7/99-012r4, Alliance for Telecommunications Industry Solutions (ATIS)/American National Standards Institute (ANSI), 1999

VA89 R.Valenzuela and C.Animalu."*A New Voice Packet Reconstruction Technique*", In Proceedings ICASSP, Glasgow, pp. 1334-1336, May 1989

VS02 U. Varshney, A. Snow, M. McGivern, and C. Howard, "*Voice over IP*". Communications of the Acm, vol. 45, pp. 89-96, January 2002.