Single-Ended Non-Intrusive Speech Quality Monitoring in VoIP

submitted by

Lijing Ding, B. Eng., M.A.Sc.

A thesis submitted to
the Faculty of Graduate Studies and Research
in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

Ottawa-Carleton Institute for Electrical and Computer Engineering
Faculty of Engineering and Design
Department of Systems and Computer Engineering

Carleton University
Ottawa, Ontario, Canada, K1S 5B6
April 2007

Your file Votre référence
ISBN: 978-0-494-27092-9
Our file Notre référence
ISBN: 978-0-494-27092-9

Abstract

Voice over Internet Protocol (VoIP) is a promising technology and it is expected to replace the traditional telephone networks in the next few years. However, speech quality in VoIP is not guaranteed, due to various new impairments introduced by Internet and Internet Protocol (IP) terminals. Evaluating VoIP speech quality in a non-intrusive fashion is challenging, as a reference signal is not accessible.

The thesis develops a single-ended, non-intrusive speech quality classification and assessment algorithm in VoIP, which reflects end-users' true quality of experience. A novel assessment structure is proposed, which utilizes a three-step strategy: impairment detection, individual effect modeling and an overall assessment model. The algorithm combines the merits of voice payload and IP header analysis approaches, and several major impairments in VoIP, including temporal clipping, echo, packet loss and noise, are investigated in the thesis.

To model the effects of temporal clipping on speech quality, an algorithm based on the clipping statistics is developed, with different weighting factors assigned to different clipping locations. For the effects of packet loss, a scheme is proposed to first classify the lost packet into three types: silence, unvoiced and voiced, then an algorithm is developed by using loss localization information. To reflect the effects of loss burstiness, a new codec-dependent parameter is introduced. Two prevailing VoIP codecs, ITU-T Rec. G.711 and G.729A, are examined. Echo detection is achieved by measuring its echo path delay and echo path loss. Two algorithms suitable for VoIP scenarios where echo delay is

excessive and echo path is nonlinear are developed. For the overall model, the individual models for temporal clipping and packet loss are first combined, with the noise and echo perception models in the E-model, an overall assessment model is developed. Particularly, a two-step noise power estimation method is adopted. The noise additivity assumption in the E-model is examined and a correction curve is suggested.

In the thesis, ITU-T Rec. P.862.1 is used to objectively measure the speech quality. The simulation results show the accuracy and effectiveness of the proposed algorithm. The correlation between prediction and measurement is 0.90, and standard error is 0.27 Mean Opinion Score (MOS). A subjective MOS test covering some key scenarios is also conducted; the ratings are analyzed to verify the novel concepts and to calibrate the proposed models. Moreover, the performance limitations of several leading objective measures are pointed out.

Acknowledgments

I would like to express my sincere thanks to my supervisors, Dr. Rafik A. Goubran and Dr. Samy El-Hennawey. I greatly appreciate their inspiration, valuable guidance and encouragement throughout my graduate studies at Carleton University. I truly value their patience, generous financial support and their effort on providing interactions with industries.

I gratefully acknowledge the financial support from the Ontario Graduate Scholarship in Science and Technology (2002-2004) and Nortel Networks Scholarship (2004-2006). I would also like to acknowledge the financial support of Research Assistantship and Teaching Assistantship from Carleton University.

I would like to extend my appreciation to Dr. Leigh Thorpe of Nortel for her valuable suggestions, subjective MOS test design and comments on the testing results. I am indebted to Dr. Roch Lefebvre, Martin Brousseau and Cédric Demers of Speech and Audio Processing Laboratory, University of Sherbrooke, for organizing and conducting the subjective MOS test. I also acknowledge Study Group 12 of ITU-T for the permission of reproducing a figure in Chapter 7 of the thesis, for a generic conversational model.

Thanks go to Christine McGregor for her careful proofreading of the thesis manuscript. My fellow graduate students in the Digital Signal Processing Lab, Carleton University, deserve a special thank. My thanks go to James Gordy, Zhong Lin, Joseph Gammal, Ayman Radwan and Ahmad Rami Abu-El-Quran, for their companionship and fruitful discussions. I am grateful to the secretarial and technical staff of the Department

of Systems and Computer Engineering, in particular Danny Lemay, Daren Russ and Jennifer Poll for their assistance.

I could not have succeeded in finishing this thesis without the help, support and love from my beloved wife, Ling Chen, who consistently encouraged me and stayed beside me during the course of research. Finally, my gratitude goes to my parents and parents-in-law, for their best wishes and encouragement all the time.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ACR	Absolute Category Rating
ADPCM	Adaptive Differential Pulse Code Modulation
AMR	Adaptive Multi-Rate
API	Application Programming Interface
BEC	Back End Clipping
BSD	Bark Spectral Distortion
CCI	Call Clarity Index
clp	conditional loss probability
CN	Comfort Noise
CNG	Comfort Noise Generation
CQO	Conversational Quality Objective
DPCM	Differential Pulse Code Modulation
DS	Down Sampling
DSL	Digital Subscriber Line
DSLA	Digital Speech Level Analyzer
DTD	Double Talk Detection
DTX	Discontinuous Transmission
EC	Echo Canceller
EMBSD	Enhanced Modified Bark Spectral Distortion
ENR	Echo-to-Noise Ratio
EPL	Echo Path Loss
ETSI	European Telecommunications Standards Institute
EVQEM	Embedded Voice Quality Estimation Module
FEC	Front End Clipping
FFT	Fast Fourier Transform
FIR	Finite Impulse Response

Ie	Equipment impairment factor
IETF	Internet Engineering Task Force
IIR	Infinite Impulse Response
INMD	In-service Non-intrusive Measurement Device
IP	Internet Protocol
IRS	Intermediate Reference System
ITU-T	International Telecommunication Union - Telecommunication
LE	Listening Effort
LQ	Listening Quality
LTI	Linear Time-Invariant
MCRA	Minima Controlled Recursive Averaging
MIRS	Modified Intermediate Reference System
MNB	Measuring Normalizing Block
MNRU	Modulated Noise Reference Unit
MOS	Mean Opinion Score
MOS-LQO	Mean Opinion Score - Listening Quality Objective
MOS-CQE	Mean Opinion Score - Conversational Quality Estimated
MSC	Middle Speech Clipping
NiNA	Non-intrusive Network Assessment
NiQA	Non-intrusive speech Quality Assessment
NLP	Nonlinear Processor
NS	Noise Suppression
OPINE	Overall Performance Index model for Network Evaluation
OSI	Open Systems Interconnection
P3SQM	Perceptual Single Sided Speech Quality Measure
PAMS	Perceptual Analysis/Measurement System
PCM	Pulse Code Modulation
PESQ	Perceptual Evaluation of Speech Quality
PESQ-LQ	Perceptual Evaluation of Speech Quality - Listening Quality

PESQM	Perceptual Echo and Sidetone Quality Measure
PLC	Packet Loss Concealment
Ppl	Packet loss probability
PSQM	Perceptual Speech Quality Measure
PSQM99	Perceptual Speech Quality Measure 1999 Version
PSTN	Public Switched Telephone Network
qdu	quantization distortion unit
QoS	Quality of Service
Rec.	Recommendation
RFC	Request for Comments
RMSE	Root Mean Square Error
RTCP	Real-time Transfer Control Protocol
RTP	Real-time Transport Protocol
SEAM	Single Ended Assessment Model
SEPL	Speech Echo Path Loss
SID	Silence Insertion Descriptor
SNR	Signal-to-Noise Ratio
SPL	Sound Pressure Level
S/U/V	Silence/Unvoiced/Voiced
SW	Sparse Window
TELR	Talker Echo Loudness Rating
UDP	User Datagram Protocol
ulp	unconditional loss probability
VAD	Voice Activity Detection
VoIP	Voice over Internet Protocol
VTQ	Voice Transmission Quality

CHAPTER 1 INTRODUCTION

1.1. Thesis Motivation

Today the telecommunications industry is undergoing dramatic changes. Voice over Internet Protocol (VoIP) is a promising technology that converts voice into Internet Protocol (IP) packets and then transmits them over the Internet backbone. The major deployment of VoIP services has begun in recent years, and it is expected to replace the traditional Public Switched Telephone Network (PSTN) [1], [2]. However, speech quality in VoIP is not guaranteed, due to various new impairments introduced by Internet and IP terminals, including excessive delay, delay jitter, packet loss, temporal clipping, echo, and codec nonlinear distortion [3], [4].

Many techniques have been implemented to enhance the speech quality in VoIP, leading to a large number of services offered with different levels of price and quality in the market. Assessing the VoIP speech quality is an imperative task to the service providers. They need to maintain certain levels of speech quality by monitoring VoIP call flows, and take actions whenever necessary to keep their competitiveness [5].

Speech quality is inherently subjective, as it is determined by the listener's perception. Therefore, the most reliable approach for assessing the speech quality is through subjective test. The Mean Opinion Score (MOS) test is defined in International Telecommunication Union - Telecommunication (ITU-T) Recommendation (Rec.) P.800 [6], and is widely accepted as a norm for subjective speech quality rating. In the test, listeners express their opinions on the quality of the speech materials by an integer score,

and the rating results are averaged to produce a MOS. In general, the subjective test is time-consuming and expensive. Therefore, it is quite desirable to replace the subjective test with objective methods that correlate well with it.

Several objective methods have been developed to fill the need to produce a good estimate of the subjective MOS. They can be classified into two categories [7], [8], intrusive or non-intrusive, based on whether a reference speech is needed or not.

In the intrusive methods, a reference speech is injected into the network under test and the corresponding degraded one is captured at the point of interest, then these two speeches are compared to produce a MOS. Most objective methods today are intrusive in nature. The most widely used algorithms include Perceptual Analysis/Measurement System (PAMS) [9] and Perceptual Evaluation of Speech Quality (PESQ) [10]. They can achieve a relatively good MOS estimate. However, these methods are not suitable for live call quality monitoring purposes, as in this case the reference speech is unavailable.

On the other hand, the non-intrusive methods do not need to access the reference speech; they rely on the degraded speech only, or some statistics collected from the network. More attention has been given to the non-intrusive methods in recent years, due to their effectiveness in live call quality monitoring. Some algorithms are proposed in this category however none of them works well.

1.2. Problem Statement

As introduced earlier, assessing speech quality in VoIP is an important task. The subjective test is the most accurate. However, it is time-consuming and expensive. It is

not suitable for automated, repeated test purposes. More importantly, the subjective test provides little technical information on the causes of speech quality degradation. It is of no use for service providers to tune up their networks. For the objective methods, the intrusive approaches need a reference signal to compare with, which is inaccessible for on-line monitoring. Moreover, the speech quality fluctuates as network conditions change from time to time. The service providers need to adjust the network parameters, like bandwidth, route and codec assigned, even drop a call, to meet certain service level agreements. This requires a real-time speech quality feedback by non-intrusively monitoring a large number of calls.

Currently, the E-model [11] is the leading non-intrusive algorithm. It is a computational model whose inputs are based on the experience and statistics collected from the network, such as the packet loss rate, delay, echo and noise levels. The E-model is not a measurement tool, but a planning tool. It does not guarantee accuracy because of its statistical nature.

Other non-intrusive methods are based on either voice payload analysis, like ITU-T Rec. P.563 [12], or IP header analysis, like VQMon [1] and PsyVoIP [14]. The former algorithm reconstructs a pseudo reference speech from the degraded counterpart, then the speech quality is intrusively determined by comparing the two speech signals through a psychoacoustic model. This method operates in a listening-only mode, that is, it does not take into account impairments that affect two-way conversational quality. Furthermore, it is seen to be inaccurate and very complex to implement. In the latter algorithms, only delay, delay jitter and packet loss information can be retrieved from the IP header and

Internet protocol analysis; thus it is hard to include the impairments related to speech payload like temporal clipping and echo.

In short, there is a need to design a non-intrusive algorithm that can monitor the quality of a live VoIP call by examining the degraded speech only, without any information about the reference speech.

1.3. Thesis Objective and Approach

This thesis aims at developing a non-intrusive VoIP speech quality assessment algorithm, which produces a subjective MOS estimate by utilizing the receive-end degraded speech only. It assesses the individual as well as combined effects of several major impairments in VoIP, mainly based on speech payload analysis. Also, the merits of the Internet protocol analysis approach and the current E-model are incorporated into the proposed algorithm. The algorithm can be effectively implemented as close as possible to receive-end users, like at VoIP media gateways or IP terminals, for in-service speech quality monitoring purposes.

Different types of impairments have different effects on speech quality. The thesis focuses on the following impairments in VoIP:

- Temporal clipping
- Echo
- Packet loss
- Codec distortion
- Noise

In the thesis, a novel non-intrusive VoIP speech quality assessment structure is proposed [15], as illustrated in Figure 1.1. It consists of three steps: (1) impairment detection, (2) individual effect modeling and (3) combined speech quality assessment model. Novel detection and modeling algorithms are developed for this assessment structure.

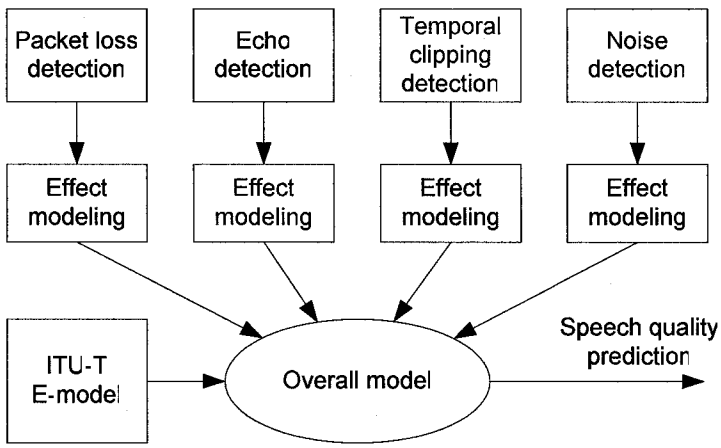


Figure 1.1 The general approach of the proposed non-intrusive algorithm

First, the impairments, including temporal clipping, echo, packet loss with codec type and noise, are detected. Detections of the temporal clipping, echo and noise are solely based on the voice payload analysis. In order to detect echo, two algorithms suitable for VoIP scenarios are proposed, where echo delay is excessive and echo path is nonlinear. The first one is based on a downsampling approach; the second one uses a sparse window to keep portions of speech samples at regular intervals. The detection of packet loss and codec types mainly relies on the analysis of IP header and Internet protocols, such as the Internet Engineering Task Force (IETF) Request for Comments (RFC) 3611, Real Time

Control Protocol – Extended Reports (RTCP-XR) [16]. Further, a scheme is proposed for Silence/Unvoiced/Voiced (S/U/V) classification of the lost packets by processing speech payload, for accurate effect modeling purposes. For the temporal clipping and noise, the detection is not our primary concern in the thesis, two algorithms developed in [17] and the algorithm developed in [18] are directly adopted, respectively.

In the second step, the effect of individual impairment is modeled separately. A multi-variable linear regression model is proposed to quantify the effect of temporal clipping on speech quality. The effects of packet loss and codec distortion are jointly investigated with packet loss S/U/V location, loss burstiness pattern and Packet Loss Concealment (PLC) technique. A nonlinear model is proposed to quantify the effect of packet loss with considerations of all above factors. Echo and noise are not new impairments to IP networks; their effects on speech quality have been studied over the years, for traditional telephone networks. Many models have been developed to quantify their impacts on speech quality, such as the E-model [11] and the complicated noise model in ITU-T Rec. P.563 [12]. It is not our objective to investigate effects of echo and noise here; the existing E-model is adopted in the thesis.

In the final step, an overall listening-only model and a conversational model are built by integrating individual models. The work here focuses on combining the effects of packet loss and temporal clipping. The effects of echo and noise are then incorporated using the E-model results, which assume that they are additive in the psychological scale [11].

The proposed structure is advantageous in that it can efficiently utilize the processing resources. For example, when some impairments are not detected, their effect modeling is unnecessary and the resources can be allocated some place else, even to speech quality enhancement if speech suffers from other impairments. Also, as the degradation of each impairment to speech quality is available as an intermediate result, it helps service providers troubleshoot their networks. The algorithm is a conversational model rather than a listening-only model. It better reflects the perceived speech quality during a live VoIP call because the call may be good in listening-only quality while poor in conversation. Finally, the structure's extensibility makes it easy to incorporate other new impairments. For example, if we have one-way delay measurement, then it can be directly modeled using the results in the E-model and combined into the overall speech quality.

The non-intrusive algorithm is developed through extensive simulation. Tens of thousands of degraded speech samples are generated for different impairment conditions. An objective algorithm is used to measure the resulting speech quality. Although the subjective MOS test is most accurate, it is not feasible to conduct subjective testing in such a large scale due to financial and time limitations. On the other hand, it is difficult to find subjective MOS databases covering all the impairments we are interested in.

In the thesis, the ITU-T Rec. P.862.1 MOS-LQO (Mean Opinion Score - Listening Quality Objective) [19] is used as the subjective MOS estimate. It is a mapping version of the current PESQ and has a bit higher correlation with the subjective MOS than the PESQ does. Meanwhile, in order to overcome the weakness of any objective algorithms,

a subjective MOS test that covers some key impairment scenarios is prepared; the MOS ratings are collected by University of Sherbrooke. The results are used to evaluate the applicability and accuracy of several objective algorithms, including P.862.1, to point out their performance limitations in general VoIP testing. The real MOS results are also used to calibrate the proposed algorithm and to verify findings based on the objective approach in the thesis.

1.4. Contributions

In the thesis, a novel non-intrusive VoIP speech quality assessment structure is proposed, and the assessment algorithm is developed. The algorithm investigates the individual as well as combined effects of several major impairments in VoIP, such as temporal clipping, echo, packet loss, codec distortion and noise.

The proposed algorithm includes an impairment detection phase and an individual effect modeling phase, and finally an overall model is developed. The simulation results show the effectiveness of the algorithm when a single impairment is presented in the degraded speech, based on the P.862.1. Furthermore, the subjective MOS test is designed and conducted. The models are calibrated and tested using the subjective MOS, and limitations of several objective measures are pointed out. The contributions made by the thesis are as follows:

- For the temporal clipping effect modeling, an algorithm is developed to map clipping statistics to speech quality by exploiting clipping locations. It identifies that front end clipping has the most severe impact on speech quality, and back end

clipping has the least impact. The algorithm shows excellent performance for different frame sizes and Comfort Noise (CN) spectrums. The result was published in [20] and [21].

- For the echo detection, two cross-correlation based algorithms are proposed. They aim at maintaining good detection accuracy while significantly reducing the computational complexity in VoIP environments. The simulation and field test show that they are robust under interferences including background noise, double talk, packet loss and codec nonlinear distortion. The computational efforts can be reduced to as low as 1.6% compared to the reference condition, as seen in subsection 5.3.4. The result was published in [22].
- For the packet loss and codec distortion, a new scheme is proposed to classify the lost packets into silence, unvoiced and voiced by exploiting correctly received neighbouring packets. For effect modeling, a novel algorithm is developed to quantify the loss S/U/V location on speech quality. Moreover, a new codec dependent factor is introduced to quantify the effect of loss burstiness on speech quality. Two prevailing codecs in VoIP today, ITU-T Rec. G.711 and G.729 are investigated in the thesis. For PLC algorithms, the repetition, silence insertion and built-in methods are examined. The algorithm extends our previous work [23] and shows excellent performance as well.
- For the overall model, a novel three-phase structure is proposed. A two-stage noise power estimation algorithm is also adopted. The impacts of packet loss and temporal clipping are first combined by consideration of their interactions. With

noise and echo perception models in the E-model, the overall assessment model is finally developed, which provides an overall speech quality estimate. It can also be used to identify the root causes of speech quality impairment in VoIP. The concept, structure and results were filed for two patents [15], [24]. The model was submitted to ITU-T Study Group 12 for competing a standard [25] by Nortel. A paper was also submitted to Speech Communications [18], and it was accepted for publication in April 2007.

- For the model verification and calibration purposes, a subjective MOS database is also developed. The speech database contains nine broad impairment categories in VoIP, with 324 cases. Each testing case has four different samples, from two male and two female talkers respectively. An English database and a French database are prepared, thus containing 2592 speech samples in total. The collected subjective MOS is analyzed to produce calibrated models; also, it is compared with several leading objective measures to point out their performance limitations. The work contributes to a large subjective MOS database in two languages, for typical VoIP scenarios. Also, some analysis result was submitted in a paper [26], and it will be published in July 2007.

1.5. Thesis Organization

The rest of the thesis is organized as follows:

Chapter 2 reviews the background materials. The VoIP technology and major speech quality impairments in VoIP are introduced. Also, speech quality assessment methods, subjective, objective, intrusive and non-intrusive approaches are presented.

Chapter 3 gives the high level description of the proposed non-intrusive algorithm. The simulation setup, reference speech database selection and pre-processing are also discussed.

Chapter 4 investigates the effects of temporal clipping on speech quality. The clipping effect is introduced by an energy-based Voice Activity Detection (VAD) algorithm. A MOS prediction model is developed based on clipping statistics and the validation results are also presented.

Chapter 5 proposes two algorithms for detection of network echo that is characterized by two parameters, echo path delay and echo path loss. The performance of the algorithms under common interferences is evaluated. The application of the two detected echo parameters in the E-model is discussed.

Chapter 6 investigates the effects of packet loss on speech quality with consideration of a pool of factors. A nonlinear model that can effectively capture the combined effects of these factors is proposed. The validation results are given finally.

Chapter 7 investigates the combined effects of all the impairments. A listening-only model is built by combining the temporal clipping model, packet loss model and noise

model. It is finally combined with the echo model to take into account the conversational speech quality.

Chapter 8 presents the design of a subjective MOS database. The collected MOS rating is then analyzed to verify and calibrate the proposed models. The performance limitations of objective MOS metrics are also evaluated.

Chapter 9 concludes the thesis and points out directions for future studies.

CHAPTER 2 BACKGROUND REVIEW

In this chapter, the concept of VoIP technology and its advantage over the traditional PSTN are discussed. The challenges facing VoIP today are reviewed from the speech quality perspective. Then, different speech quality assessment methods: subjective, objective, intrusive and non-intrusive approaches are presented. Particularly, the methods that directly link to our thesis, such as the E-model, the payload analysis model and PESQ, are given in more detail. At the end, the device toolbox used in speech quality measurement is introduced.

2.1. VoIP Technology

VoIP refers to the real-time transmission of voice signals as packetized data across networks by using Internet protocol [1]. In VoIP, speech is encoded into frames, encapsulated into packets, and then transmitted through the Internet. At the receive end, packets are buffered to compensate for delay variations and reassembled into the correct order. A PLC scheme is also implemented to recover the lost packets during the decoding process. Finally, the speech is played back.

A reference connection model for a typical IP-to-PSTN call is illustrated in Figure 2.1. At the send side, a regular telephone set is plugged into an adaptor that converts the voice signal into IP packets and vice versa. The Internet access can be provided by various means such as dial-up, wireless, Ethernet, cable or Digital Subscriber Line (DSL). At the

receive side, a media gateway interfaces between the IP network and PSTN and transcodes the voice streams.

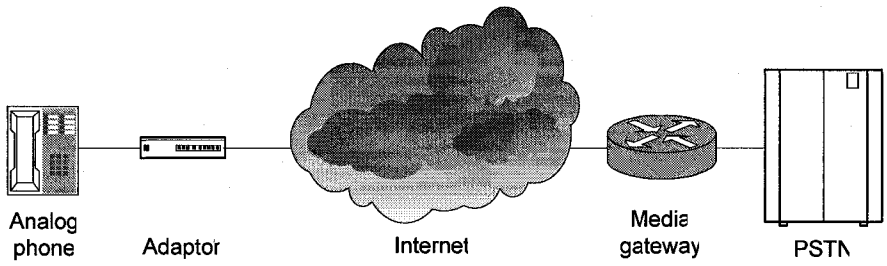


Figure 2.1 A reference connection model for an IP-to-PSTN call

Many different topologies of IP telephony exist. For example, a VoIP call may be placed between regular phone sets, IP phones, or even soft phones (like computers). The call may bypass the traditional telephone network if it is made exclusively within the IP network. In this case, the media gateway is not involved during the call at all. Instead, the adaptor or IP phone at the receive end assumes the functionalities of the media gateway, that is, buffering, concealing lost packets and decoding.

The relationships between some VoIP protocols and the Open Systems Interconnection (OSI) seven-layer model are illustrated in Figure 2.2. Real-time Transport Protocol (RTP) [27] is used for end-to-end VoIP delivery services. RTP is designed for data transmission with real-time characteristics and itself does not provide any mechanism to ensure timely delivery or provide other Quality of Service (QoS) guarantees. It relies on RTP Control Protocol (RTCP) [48] to monitor the data delivery and to provide minimum control functionality.

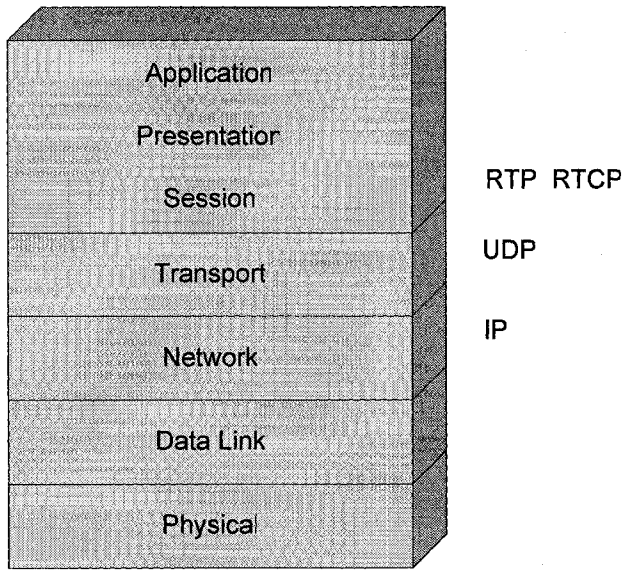


Figure 2.2 VoIP protocols and OSI seven-layer model

The structure of an IP packet is shown in Figure 2.3, which consists of a fixed 40 bytes header and *n* bytes voice payload. The RTP header includes a sequence number that can be used to detect packet loss, and a timestamp for synchronization and jitter calculation. The RTP runs on top of the User Datagram Protocol (UDP) to make use of its multiplexing and checksum functionalities [27].

IP Header 20 bytes	UDP Header 8 bytes	RTP Header 12 bytes	Voice payload *n* bytes

Figure 2.3 IP packet structure

The main advantage of VoIP is its low cost structure compared to traditional telephone services, especially for long distance calls [28]. VoIP services can efficiently utilize the bandwidth because no channel is dedicated to a call. Further, bandwidth

savings are achieved by using codecs and silence suppression techniques. Finally, as the Internet has evolved into a universal communication network that carries all kinds of traffic, including data, voice and video, only one network needs to be managed.

2.2. Impairments in VoIP

The IP network is originally designed for non-real time data communications. It offers best-effort service with no QoS guarantee. Speech quality in VoIP is mainly affected by packet loss, temporal clipping, delay, delay jitter, echo and codec distortion [3], [4]. Many approaches have been proposed to mitigate the above impairments to improve the speech quality.

2.2.1. Packet Loss

Packet loss is common in packet-switched transmission due to its nature. When the network is congested or the router buffer is full, the tail part of the queue is dropped. Link failure and channel error also result in packet loss. In addition, the jitter buffer can drop packets for two reasons, either because they arrive too late to be played out or the jitter buffer is overrun, as explained in subsection 2.2.5.

Packet loss may occur as random or burst, and this loss pattern affects the speech quality. In general, random loss results in better speech quality because burst loss leads to longer time clipping in the speech signal and its effect is more annoying. Also, the speech quality is affected by the PLC algorithm implemented and frame size, as a model shown in the appendix I to ITU-T Rec. G.113 [29].

Several PLC techniques have been used to deal with packet loss [28], [30], [31]. They can be sender-based or receiver-based. The former includes forward error correction, interleaving and multiplexing. The idea for the latter is to replace the lost packet with a similar one without sender retransmission. It can reduce but not eliminate quality impairments due to packet loss. The replacement may be silence, noise or the previous packet. Complicated interpolative-based or regenerative-based techniques can also be used, which require higher delay and computational complexity.

In silence substitution, the lost packet is filled with silence to keep the timing relationship between the surrounding packets. Studies show it achieves adequate performance only for smaller packet sizes and low packet loss rate. Substitution with background noise performs better than simple silence substitution because of the ability of the human brain to repair the received message if there is some background noise. Repeating the lost packet by the one received immediately before the loss has low computation complexity and performs reasonably well. Complicated PLC techniques, like those built-in algorithms in ITU-T Rec. G.729 and G.723.1, provide better speech quality than the repetition method at the cost of increasing computational requirement and delay.

2.2.2. Temporal Clipping

There are two types of clipping for speech signals. One is amplitude clipping, also called saturation clipping. The other is temporal clipping, also called time clipping, or syllable clipping. The thesis focuses on temporal clipping only. In VoIP, there are two sources of temporal clipping:

- Silence suppression
- Nonlinear Processor (NLP) of Echo Canceller (EC)

Silence suppression is usually achieved through a VAD, which is widely used in modern telecommunication applications, such as speech recognition, speech coding, wireless telephony, VoIP and echo cancellation [32]. In VoIP, VAD is used with speech codecs to further reduce the bandwidth consumption.

Human speech is composed of talkspurts separated by pauses or silences. In a long run, each speaker is only active no more than 50 percent during a conversation. The VAD algorithm finds the beginning and the end of the talkspurt on a frame basis. Accordingly, input speech frames are classified into active frames (talkspurt) or non-active frames (silence). If we look into talkspurts further, they consist of unvoiced sounds and voiced sounds. The former is generated by airflow passing through the vocal tract without the vocal cord vibrating, including some consonants like */th/*, */f/*, */s/*. The latter is produced as a result of excitation of the vocal tract by the periodic flow of air at the glottis, including all the vowels and some consonants like */b/*, */m/*, */v/*. The VAD algorithm attempts to find unvoiced frames and voiced frames while suppressing silence frames. A silence frame does not mean a frame of zero energy at all. In reality, the background noise is always present. Also, some fricative consonants (unvoiced sounds) have a very limited energy [33], such as */th/* in thing, */f/* in frame; they may be classified as non-active.

Generally, VAD algorithms also include a Discontinuous Transmission (DTX) and a Comfort Noise Generation (CNG) module. During a non-active frame, the background noise level and spectral information are encoded with fewer bits. This kind of frame is

known as Silence Insertion Descriptor (SID) frame. The DTX module determines if a SID frame should be sent to the speech decoder or not. At the beginning of a non-active voice segment, a SID packet is transmitted in the same RTP stream and indicated by the CN payload type [34]. Then, the following SID packets can be transmitted periodically or only intermittently when there is an appreciable change on the background noise characteristics. At the receive side, CN is generated at the CNG module by decoding SID frames.

The VAD decision is based on multiple features extracted from the speech signal, such as varying energy content, zero crossing rate, sub-band energy or cepstral features [32], [35]. The performance of VAD is very critical. It is not always perfect, when the talkspurt of a speech is incorrectly classified as non-active, this portion gets clipped and the speech quality is therefore degraded; on the other hand, when silence is classified as active, the speech quality is not impaired but the efficiency of VAD is decreased.

The other source of temporal clipping is NLP of an EC. Echo cancellation through adaptive filtering usually doesn't cancel all echoes. When a single talk that is causing echo is detected, NLP is activated by replacing the reflected echo with silence or comfort noise. In many cases, this detector fails and NLP is activated improperly. When this happens, the speech at the cancelled end gets clipped. The detailed operation of EC is introduced in the next section. In addition to the impairment resulting from the NLP temporal clipping, the EC improper or imperfect operation will lead to the echo impairment, which is one of the major degrading factors in VoIP due to the excessive delay inherent in VoIP, as will be seen in subsections 2.2.3 and 2.2.4.

2.2.3. Echo

In telephony, echo is the delayed, attenuated and possibly distorted version of the original voice signal reflected back to the source. Primarily, there are two types of echo [36], as shown in Figure 2.4:

- Network echo
- Acoustic echo

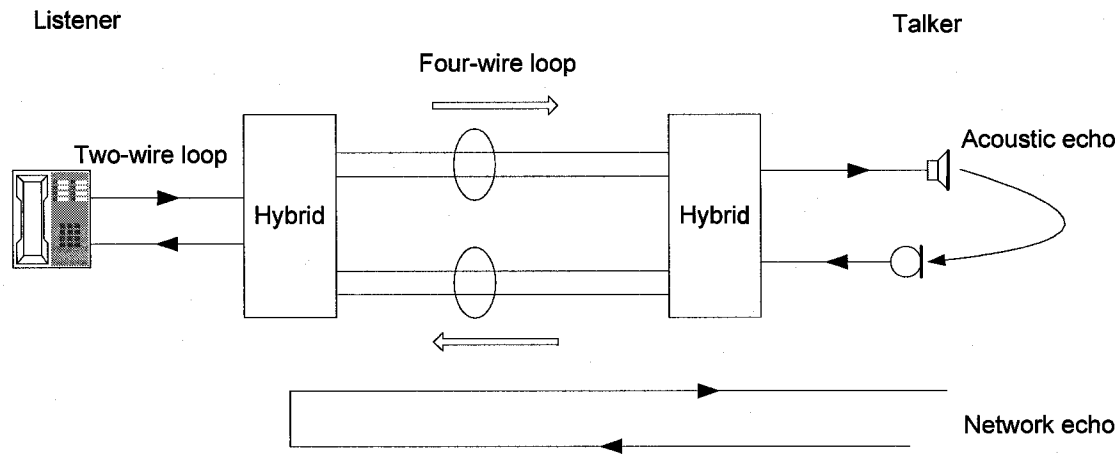


Figure 2.4 Echoes in telephony network

Network echo is sometimes called "line echo", "hybrid echo" or "electrical echo". It is a reflection of the talker's voice, caused by impedance mismatch at the hybrid of the opposite side. The hybrid is a circuit that converts between a local two-wire loop and four-wire digital trunk at the telephone exchange. The two-wire loop connects the handset to the local exchange and provides bi-directional transmission. The hybrid separates the two-wire loop into two pairs of trunks, one for send path and the other for receive path. Impedance mismatch often occurs because of variations of local loop

lengths, and wire types. As a result, the talker's voice is leaked into the receive path and returned back to the talker, causing an echo. That is, the hybrid on the opposite side of the network produces the echo heard by the talker [36], [37].

This echo is referred to as talker echo if it is only reflected once. However, if the echo is reflected twice, it is referred to as listener echo. Generally, the effect of listener echo on speech quality can be ignored, given that talker echo is well controlled [38].

Acoustic echo is caused by acoustic coupling in the open-air between a handset's microphone and earpiece. The howling which may occur when using a hands-free phone is an example of such an echo. Also, it can be caused by multiple reflections of the speech back to the microphone from surrounding objects such as walls, windows, floors and ceilings.

Human ear is sensitive to echo, the annoyance of echo on speech quality is determined by the echo level and its associated delay. Generally speaking, the greater the echo level, or the longer the delay, the worse the speech quality. However, not all echoes degrade speech quality. Particularly, sidetone is an echo of much less delay, in the order of 28-30 ms [39], [40]. It is often deliberately inserted in handset design, which assures that the talker would hear his or her own voice in the earpiece while talking. When the delay is greater than 30 ms, the echo will be heard as a hollowness of sound, and the speech quality begins to degrade.

A typical echo tolerance curve is shown in Figure 2.5. The impact of talker echo is characterized by two parameters: Talker Echo Loudness Rating (*TELR*) and one-way echo path delay *T*. *TELR* is the sum of the all losses from the talker's mouth to the

talker's ear. In the figure, the label "1% objection" means 1% probability of encountering an objectionable echo, and this corresponds to the E-model rating $R = 74$. Similarly, the label "10% objection" means 10% probability of encountering an objectionable echo, and this corresponds to the E-model rating $R = 60$ [41].

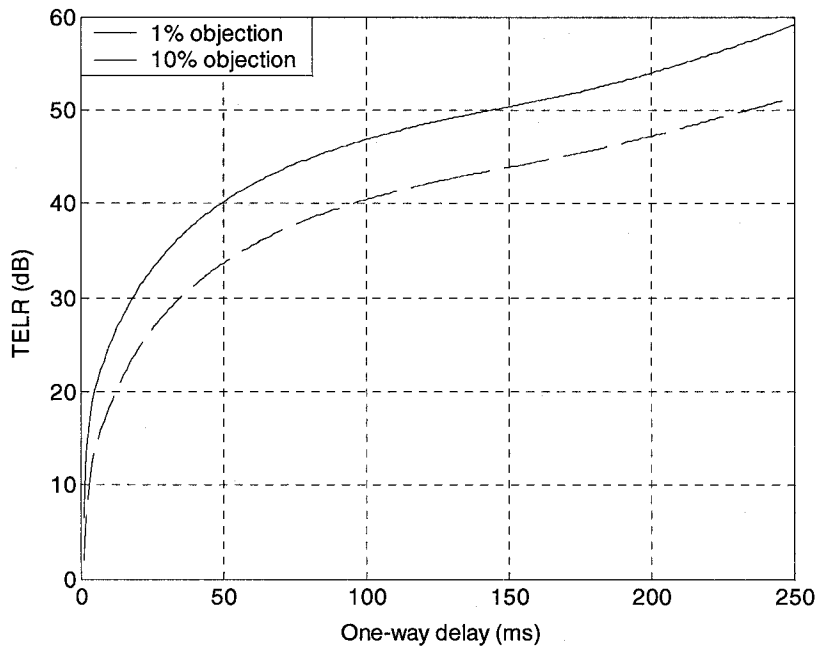


Figure 2.5 Talker echo tolerance curves

When the (round-trip) echo delay is greater than 30 ms, echo becomes annoying and has to be cancelled using an EC. EC is a voice operated device; it reduces the echo by first generating a replica of the echo through adaptive filtering techniques, and then subtracting it from the echo [42]. EC is commonly installed at the local telephone exchange, VoIP media gateway or IP terminal. It should be as close as possible to the echo source.

The structure of EC is shown inside the dashed lines in Figure 2.6. It consists of three main function blocks, an adaptive filter, a digital subtractor and a NLP. The adaptive filter attempts to estimate the impulse response of the hybrid, and then convolutes it with the voice signal on the receive path to generate a replica of the echo. Then this replica is subtracted from the reflected echo through the digital subtractor. The error signal (residual echo) is used to continuously update the filter taps. EC cannot completely remove all echoes due to reasons such as the adaptive nature of the algorithm, echo path change and finite word length of the processor. So, NLP is used to further attenuate the residual echo by replacing it with background noise.

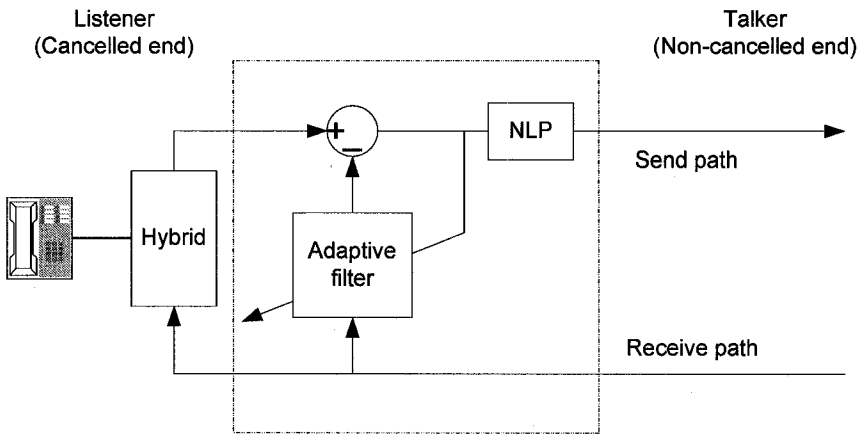


Figure 2.6 Structure of echo canceller

In a telephone conversation, typically the two parties alternate talking and listening. However, sometimes both parties speak simultaneously, and such a situation is called double talk. Double talk brings a technical challenge to the EC. Because the cancelled end speech is superimposed on the reflected echo, the updating of the adaptive filter taps must be stopped to avoid divergence. In this case, the last saved line echo model is used

to cancel the echo. The presence of double talk is detected by a Double Talk Detection (DTD) algorithm. When detected, NLP should be disabled to avoid clipping the cancelled end speech.

However, in many cases the DTD fails and NLP is erroneously activated, and temporal clipping occurs to the cancelled end speech. Typically, the NLP does not respond quickly enough to the beginning of speech. Also, it may confuse the fading of the speech level at some inter-syllable pause and at the end of a talkspurt with the echo.

In the traditional PSTN, except long distance calls via satellite links or marine cables, network echo is usually not problematic, as the relatively short distance between two parties does not introduce significant delay.

However, in VoIP networks, excessive delay is introduced by voice codecs, IP network and jitter buffers, one way end-to-end delay may be over 200 ms. This makes any minor echoes more perceivable and annoying to the listener, as seen in Figure 2.5. Furthermore, the performance of most ECs is affected by the VoIP environment due to excessive delay, packet loss, and codec distortion [43]. So, network echo is often one major issue in VoIP networks.

2.2.4. Delay

Delay is the time required for a signal traveling from a speaker's mouth to a listener's ear. The delay, as perceived by a listener, is the end-to-end delay that includes acoustic delay, filtering, digitization, compression, frame and packet formation, network delay, jitter buffer delay, decompression and audio playback [44].

Delay may result in difficulties during conversation and also increase a listener's perception of echo. According to subjective tests, these effects are unnoticeable with one-way delay below 100 ms; they are acceptable when one-way delay is below 150 ms. The limit value for most connections should be 300-350 ms with an upper limit of 400 ms. Delays above 400 ms are unacceptable for general network planning purposes [45], [46].

In a VoIP call, several sources contribute to the total one-way delay, as follows:

- Access delay: this is the time required to put the packet on the transmission line. It varies between access technologies. For example, in voice over DSL, it mainly depends on the DSL upstream bit rate and interleaving depth. It is about 2 ms for DSL fast channel, and the worst case is 20 ms [47].
- Codec delay: it includes the algorithm delay, look-ahead delay and encoding/decoding processing delay. Some codecs, like ITU-T Rec. G.711 and G.726 operate on a sample basis; the algorithm delay is the sampling interval. However, most codecs today like ITU-T Rec. G.729 and G.723.1 operate on a frame basis; the algorithm delay is the frame size. Furthermore, some codecs also look into the succeeding frame to improve the compressing efficiency; this amount of time is called look-ahead time. The processing delay is usually assumed to be equal to the frame size for efficient use of the processing power [46]. So, the typical delay introduced by an encoder/decoder pair is:

  $$2 \times \textit{sample size or frame size} + \textit{look-ahead}$$

  The one-way codec delays for several common VoIP codecs are summarized in Table 2.1 below.

Table 2.1 One-way codec delay

Codec	Sample/Frame size (ms)	Look-ahead (ms)	Total delay (ms)
G.711	0.125	0	0.25
G.729/A/AB	10	5	25
G.726	0.125	0	0.25
G.723.1	30	7.5	67.5
G.728	0.625	0	1.25

- Packetization delay: this is the time required to fill in the IP packet. For one frame per packet, this delay is zero, as the packet encapsulation can be started instantaneously after the frame is encoded. When multiple frames are encapsulated into one packet, an additional $(N\text{-}1) \times$ *frame size* amount of delay is introduced, where N is the number of frames per packet. Common VoIP packet size is about 10 - 40 ms, with the default size 20 ms [4], [48].
- Core network delay: it includes the various delays introduced by IP networks, such as queuing, routing and processing. It also includes the propagation delay, which is the time required for the signal to travel through the network at the speed of light. The core network delay may vary from several milliseconds to hundreds of milliseconds.
- Jitter buffer delay: it is based on the average time packets spend in the buffer, which is assumed to be half of the peak buffer size for network planning purposes [46]. The buffer size is usually set to be an integer multiple of the expected packet size in order to buffer an integer number of packets. Common buffer sizes range from 20 to 80 ms [49], more explanations are given in the next subsection.

- Other miscellaneous delays, such as sample capturing, digitization, echo cancellation, PSTN switching and device playback. They can be assumed to be a few milliseconds in total.

2.2.5. Delay Jitter

Delay jitter is caused by the variations of packet inter-arrival times, because packets carrying speech samples of a call may be transported through distinct routes in the network. A jitter buffer is implemented at the receive end (e.g. media gateway, analog phone adapter, IP phone) to overcome this problem. The operation of the jitter buffer is illustrated in Figure 2.7. If a packet arrives too late or the jitter buffer is overrun, then this packet will be discarded for the decoding process. In this case, the resulting degradation is mapped to packet loss. If the jitter buffer is underrun, then short pause is inserted in the decoding processes.

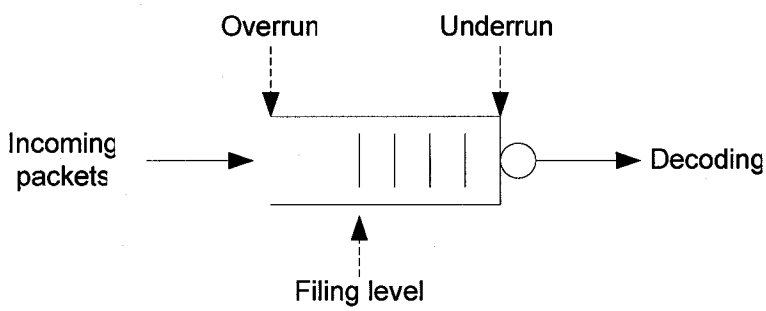


Figure 2.7 Operation of the receive jitter buffer

The jitter buffer also introduces an additional delay of its own. The size and structure of the buffer is an important design compromise. The jitter buffer size can be fixed or adaptive, while the latter approach is more efficient in minimizing packet discard rate and

introduced buffer delay. In adaptive algorithms, the buffer resizing is implemented during silences of the speech [50], [51]. Thus the silence interval between two talkspurts may be artificially elongated or shortened. This small length change in silence is unnoticeable to the listener. However, the buffer resizing may occur in the talkspurt if the adaptation algorithm is not correctly implemented, and the subjective effect is much more annoying.

2.2.6. Codec Distortion

In telecommunications, a codec is used to encode speech signals into compact digital representation for efficient transmission and storage, and then decode them back to the original form.

Generally, there are two classes of codec in VoIP, based on its coding principle [52]:

- Waveform coding
- Vocoding

Waveform coding is the method for discrete approximation of the analog waveform and it usually operates above 16 kb/s. Examples include Pulse Code Modulation (PCM), Differential PCM (DPCM) and Adaptive Differential PCM (ADPCM).

In vocoding, the speech is assumed to be produced by exciting a linear filter, the human vocal tract, by sequences of impulse. The vocal tract filter is an all-pole adaptive filter whose coefficients are derived from linear prediction analysis. The excitation sequence is determined by analysis-by-synthesis optimization. Vocoding usually operates between 4 kb/s and 16 kb/s. Common VoIP codecs, such as ITU-T Rec. G.729, G.723.1 and G.728, belong to this class.

The bandwidth reduction is at the expense of lower speech quality and additional complexity. The impairment from a specific codec to speech quality can be presented in either MOS or E-model rating *R* scale. For example, the subjective MOS for G.729 is reported to be 3.92 in an experiment [53].

In the E-model, two methods are used to quantify the effect of codec distortion [54]. In the first one, the quantization distortion unit (*qdu*) is used to represent impairment by PCM processing. For example, one *qdu* is assigned to an A- or μ-law PCM pair. In the second one, the equipment impairment factor (*Ie*) is used to characterize the impairment from non-waveform codecs by assigning a value. Note that the *Ie* method is also recommended for ADPCM codecs. The provisional planning *Ie* values for some codecs are listed in Table 2.2. For a complete list please refer to Table I.1 in appendix I to ITU-T Rec. G.113 [29]. These *Ie* values are usually determined through subjective listening tests. They can also be objectively measured through instrumentation approaches [55], [23].

Table 2.2 Codec *Ie* values under non-error conditions

Codec	Rate (kb/s)	*Ie*
G.711	64	0
G.729	8	10
G.729AB	8	11
G.723.1	6.3	15
G.723.1	5.3	19
G.728	16	7
G.728	12.8	20

When codecs operate in cascade, their impairments are simply assumed to be additive in the E-model, although an order effect may exist, that is, the impairment of codec A

followed by codec B is different from that of codec B followed by codec A [56], probably due to masking effects.

2.3. Speech Quality Assessment

The introduction of VoIP technology and liberalization of the telecommunication market has lead to a large number of voice services being offered with different levels of quality and price. Speech quality is a very important aspect in voice communication. On the one hand, service providers need to assess the quality they offer to keep their competitiveness. On the other hand, users also need ways to compare the quality and cost structure of different services and to choose the one that suits them best.

The PSTN has served voice traffic over many years and evolved to provide optimal speech quality in design. The speech quality is reliable and predictable. Signal-to-Noise Ratio (SNR), which was developed for Linear Time-Invariant (LTI) systems, is used to assess the speech quality in PSTN [57]. However, such a metric is inadequate in VoIP, mainly because the use of low bit rate codecs and delay jitter buffers invalidates the LTI assumption.

2.3.1. Subjective Methods

Subjective testing is the most reliable approach for speech quality assessment; it provides an overall score of the speech quality from the listener's subjective point of view. The widely accepted measure is the MOS test described by ITU-T Rec. P.800 [6] and ITU-T Rec. P.830 [58]. In the test, listeners express their opinions on the quality of

the speech materials using an integer opinion score. Then, the rating results are averaged to produce a MOS value. The subjective test can be carried out in listening-only or conversational mode. For practical reasons, the listening-only mode may be the only feasible method during the development of new transmission equipment when the real-time deployment is not available [6].

There are several rating methods in the subjective test. Absolute Category Rating (ACR) is the most commonly used. It is a five-point scale rating where listeners judge the "absolute" quality without comparison to a reference. The Listening Quality (LQ) and Listening Effort (LE) scale MOS in ACR is summarized in Table 2.3. The MOS in ACR LQ is often simply referred to as MOS. The highest score a call can achieve is about 4.5 in the subjective test. A rating of 4.0 or higher is considered as toll quality [59].

Table 2.3 MOS in absolute category rating

Score	LQ scale	LE scale
5	Excellent	Complete relaxation possible; no effort required
4	Good	Attention necessary; no appreciable effort required
3	Fair	Moderate effort required
2	Poor	Considerable effort required
1	Bad	No meaning understood with any feasible effort

Other rating methods, such as degradation category rating and comparison category rating, are designed for other purposes and less commonly used.

Although the subjective MOS test is the most reliable, its drawbacks are apparent. The MOS test is time-consuming and expensive. It is quite complicated; the design of the

test is strongly influenced by both human subjects and elaborate testing settings. Most importantly, the test provides little technical information on the causes of speech quality degradation. In sum, the subjective test is impractical for automated, frequent testing purposes such as routine network monitoring.

2.3.2. Objective Methods

Objective methods are those carried out by machines only, without human listener involvement. They produce a MOS estimate that should correlate well with the subjective test.

Efforts have been focused on developing objective methods to measure the speech quality, especially in VoIP. Objective methods can be classified into two categories based on whether a reference signal is needed or not [7], [8]:

- Intrusive methods
- Non-intrusive methods

A) Intrusive methods

In this category, a reference signal is injected into the network under test, and the corresponding degraded one is compared with the reference to produce a MOS. Many measures have been proposed to quantify the difference between these two speech signals, but they all share a basic structure, as shown in Figure 2.8.

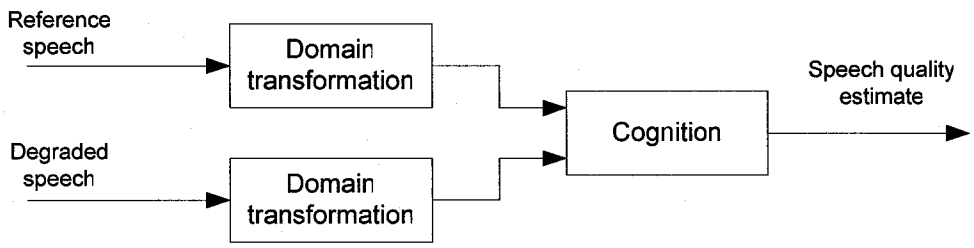


Figure 2.8 Structure of intrusive speech quality assessment methods

The two speeches are first transformed into a speech quality relevant domain. Then, the difference is compared by using various methods from simple distance measure to complicated neural network approach. Finally, it is mapped into MOS.

The measures may be time-domain based (e.g. SNR, segmental SNR), frequency-domain based (e.g. cepstral distance) or perceptual-domain based [60]. The comparative performance of these measures is usually examined by the correlation with the subjective test [61]. Among them, the perceptual-domain based measures, which transform speech signals into a psychoacoustic relevant domain such as bark spectrum or loudness, and incorporate human auditory models have the best performance, and therefore are widely used today.

The well-known perceptual measures include Bark Spectral Distortion (BSD) [62], Enhanced Modified BSD (EMBSD) [60], Perceptual Speech Quality Measure (PSQM) [63], [64], Measuring Normalizing Block (MNB) [65], [66], PAMS [9] and PESQ [10], [67], [68]. Among them, PSQM was standardized as ITU-T Rec. P.861 in 1996. MNB was added in 1998 to ITU-T Rec. P.861 as Appendix II. And, PESQ was standardized as ITU-T Rec. P.862 in 2001. These measures give good estimates of MOS scores. The comparative performances of several perceptual-based algorithms are studied in [69],

based on several categories of impairments including noise, single codec, codec in tandem, time shifting (front-end clipping), bit errors, frame errors and automatic gain control.

In this thesis, the PESQ algorithm is used in speech quality measurement, and will be introduced in detail in Section 2.6.

There are some drawbacks to the intrusive methods. They need a reference speech to compare with, which may not be feasible in many cases. Also, circuit needs to be removed from service for injecting test calls, potentially affecting revenue and thus the method cannot be used to measure a great volume of calls. As network conditions change from time to time, the intrusive methods are inadequate for instantaneous speech quality measurement.

All of the intrusive methods so far use a listening-only model. Some factors related to two-way conversation, such as delay and echo, which could adversely affect the conversational quality if too excessive, are not reflected in these methods. So, it is possible to have a high objective MOS, but the overall quality is still poor.

B) Non-intrusive methods

In this category, no reference signal is injected in the network. Instead, the non-intrusive algorithms are based on the analysis of some key parameters and statistics of the network or received degraded signals.

This non-intrusive approach is quite challenging because the original signal is unknown. The general block diagram for non-intrusive methods is illustrated in Figure 2.9. The impairments are parameterized from the degraded speech only, and finally

mapped to speech quality. Sometimes, network parameters may be used to supplement the analysis. Also, some models, like the E-model [11] and Call Clarity Index (CCI) model [70], are based on the network parameters alone. A different approach, Embedded Voice Quality Estimation Module (EVQEM) [71] is also developed, it utilizes a small portion of the bandwidth available during silences to transmit a reference signal with different payload types, and then at the receive side, the payload is decoded and an intrusive algorithm is used to measure the speech quality.

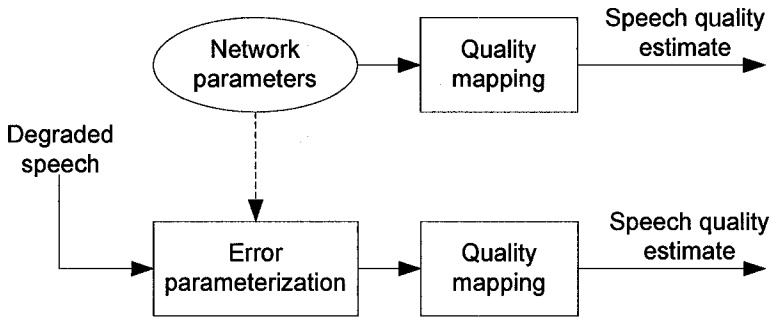


Figure 2.9 Structure of non-intrusive speech quality assessment methods

More attention has been given to the non-intrusive methods in recent years. In VoIP, they can be used for baseline network performance monitoring, troubleshooting, certifying new deployments and network optimization [72].

The non-intrusive methods can be grouped into three categories [7], [73], [74]:

- Voice payload analysis (Speech layer methods)
- IP header and Internet protocol analysis (Packet layer methods)
- Transmission rating models (Opinion methods)

The first two category methods are also called absolute estimation methods [74]. The best-known speech layer method is the ITU-T Rec. P.563 [12]. The packet layer approaches include VQMon [1] and PsyVoIP [14]. The leading opinion method is the E-model [11]. In practice, these non-intrusive methods may be implemented in combination [74].

In summary, MOS is the most widely accepted measure for speech quality. It can be measured by means of subjective or objective methods. The objective methods may be intrusive or non-intrusive depending on whether a reference speech is available or not. The non-intrusive methods are suitable for live network performance monitoring. The relationship between different MOS measurement methods is illustrated in Figure 2.10.

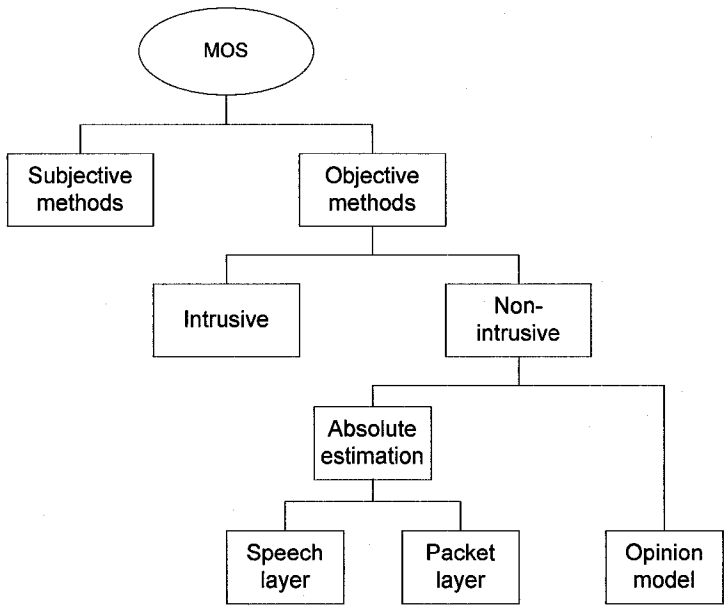


Figure 2.10 Classification of MOS measurement methods

As the non-intrusive speech quality assessment is the topic of this thesis, the existing methods are reviewed in more detail. ITU-T Rec. P.563, VQMon and PsyVoIP will be introduced in Section 2.4, and, the E-model will be introduced in Section 2.5.

2.4. Non-Intrusive Methods – Absolute Estimation Approaches

For a VoIP call, two kinds of parameters can be non-intrusively monitored. The first kind of parameters are related to the speech layer, that is, they can be derived from voice payload analysis, like speech level, echo, noise and clipping. The other kind of parameters are related to the packet layer, that is, they can be derived from the IP header or Internet protocol analysis, such as packet loss statistics, delay and delay jitter. Therefore, two distinct approaches are developed to absolutely estimate the speech quality from each layer analysis.

2.4.1. Voice Payload Analysis

The best-known voice payload analysis algorithm is the ITU-T Rec. P.563 [12]. In 2001 the ITU-T determined the need for a standard for non-intrusive voice quality measurement [75], under the working title "Single Ended Assessment Model (P.SEAM)." Three individual algorithms:

- Non-intrusive speech Quality Assessment (NiQA) [76] of Psytechnics Ltd., UK,
- Non-intrusive Network Assessment (NiNA) [77] of SwissQual Inc, Switzerland, and

- Perceptual Single Sided Speech Quality Measure (P3SQM) of Opticom GmbH, Germany,

were combined to create the best possible algorithm. After a considerable testing and verification process, the combined algorithm was standardized as ITU-T Rec. P.563 in 2004. This is the first ITU-T single ended method for predicting the subjective quality in narrow band telephone networks.

ITU-T Rec. P.563 produces the speech quality in MOS-LQO on ACR scale. It is designed for listening-only quality. Therefore, factors that degrade conversational quality, like delay and echo, are not included in the algorithm. ITU-T Rec. P.563 has shown acceptable accuracy for factors including codecs above 4 kb/s, transcoding, speech input levels to a codec, transmission channel errors (bit error, packet loss, cell loss), delay variation, environmental noise at the sending side, etc. It is not intended to measure the delay, listening level, sidetone, talker echo, listener echo, music input and codecs operating below 4 kb/s, etc. For 24 known ITU-T benchmark tests, the average correlation between ITU-T Rec. P.563 MOS and the subjective MOS is 0.88 [12].

The general block diagram of the algorithm is shown in Figure 2.11 and explained below [12], [75].

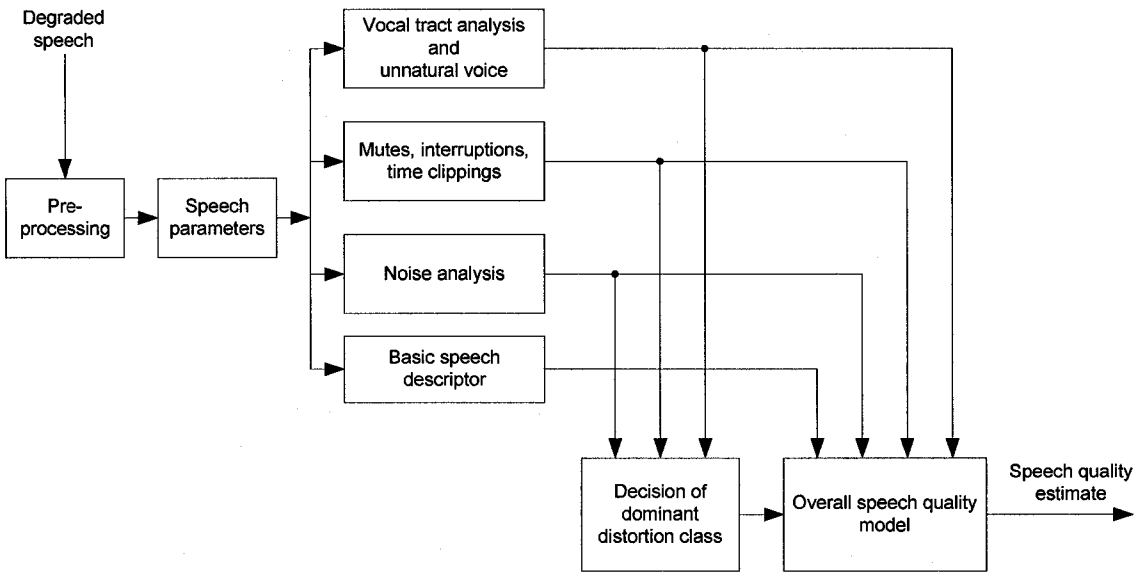


Figure 2.11 Structure of ITU-T Rec. P.563

The degraded speech needs to be pre-processed before quality judgment; this step includes the voice/unvoice separation by a VAD, Intermediate Reference System (IRS) receive filtering and speech level normalization. Then, the distortions and speech parameters are extracted by three fundamental functional blocks, which also correspond to the three main distortion classes in ITU-T Rec. P.563:

- Vocal tract analysis [78] and unnatural voice. This examines the unnaturalness of the voice: robotic voice, male voice or female voice. The basic voice quality is determined by an intrusive approach, as shown in Figure 2.12. A pseudo reference speech is reconstructed from a speech enhancer and then is compared with the degraded one using a psychoacoustic model, which is PESQ-like but simpler.

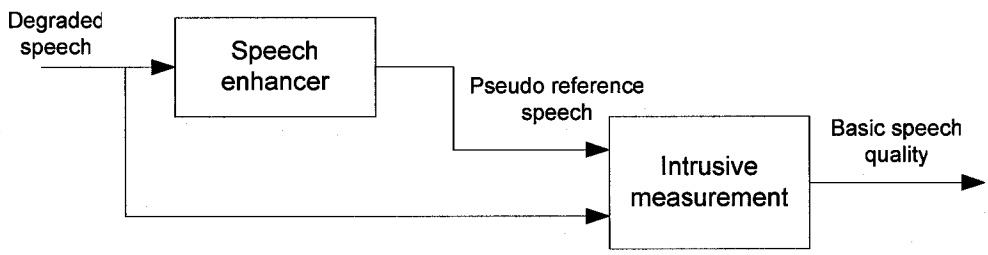


Figure 2.12 Scheme for determination of basic speech quality

- Mutes, interruptions and time clippings. These impairments may result from packet loss or VAD. Related parameters are calculated.
- Noise analysis. This characterizes the static background noises and multiplicative noises.

In addition to these three main blocks, the basic speech descriptor block contains information such as active speech level, speech activity and level variation.

After the distortion analysis by functional blocks, the dominant distortion class is determined. In the case that several distortions occur to the speech, a prioritization scheme is used. Then, the speech quality for the corresponding distortion class is calculated by using a linear combination of parameters obtained in the previous step, producing an intermediate speech quality. Finally, the overall speech quality is calculated by combining this intermediate quality with some additional speech signal features.

ITU-T Rec. P.563 involves the reconstruction of a pseudo reference, and it is complicated to implement. Moreover, it is seen to be inaccurate for those impairment conditions that it is not intended to measure, as listed in the standard [12].

There are some other methods based on general voice analysis. In [79], a small portion of the bandwidth that is available during silences is utilized to transmit a reference signal, and then an intrusive algorithm is used to measure the speech quality. In [80], [81], [82], [83], [84], a pseudo reference is used. The idea is that features of the degraded speech are extracted and compared to an artificial reference model from high quality, clean speech databases. Then, the difference is mapped to speech quality. For example, these features can be perceptual linear prediction coefficients or Mel frequency cepstrum coefficients. The reference model may be represented by continuous hidden Markov models or Gaussian mixture models.

2.4.2. Internet Protocol Analysis

Examples for this approach are VQMon [1] of Telchemy Inc., USA, and PsyVoIP [14] of Psytechnics Ltd., UK. They are currently competing for a new ITU-T standard under the working title P.VTQ, "Derivation of Voice Transmission Quality from non-intrusive Internet Protocol analysis" [74].

VQMon is a passive VoIP monitoring system. It is based on the E-model and incorporates the effects of packet loss rate, packet loss pattern and delay jitter buffer. These metrics can be obtained from RTP and RTCP protocol analysis. The packet loss is preferred to be measured after the jitter buffer. The loss pattern is divided into gaps and bursts and characterized by a four-state Markov model, then the instantaneous speech quality is modeled separately according to the gap or burst states in *Ie* domain using a linear piecewise function of loss rate. The perceived quality is estimated from the instantaneous quality by using an exponential function. The final speech quality is

obtained by averaging the perceived quality along the time scale and incorporating the recency effects [85], [86].

PsyVoIP is also a passive VoIP quality monitor [87]. First, packets belonging to a call are identified, and the relevant information for the rest of the model is extracted. Then the out-of-order packets are re-sequenced. A VAD is then used to discriminate speech and silence because packet loss and delay jitter have little effect during silence. Finally, the statistical descriptor of the call is extracted and mapped to MOS, which is further calibrated by the PESQ.

These approaches cannot take into account impairments that are related to voice payload like temporal clipping.

2.5. E-model

In the thesis, the echo perception model and some concepts for packet loss on speech quality in the E-model [11] are adopted into the proposed non-intrusive algorithm. Therefore, these relevant parts are reviewed here and the rest of the E-model is given in detail in Appendix A to this thesis.

2.5.1. Introduction

The E-model stands for the European Telecommunications Standards Institute (ETSI) computational model [88]. It was developed by an ETSI *ad hoc* group during the work on ETSI technical report 250 [89] in 1996, and then chosen by ITU-T as Rec. G.107 in 1998.

Detailed guidelines and planning examples of the E-model are given in ITU-T Rec. G.108 [45].

Essentially, the E-model is not a measurement tool, but rather a planning tool for transmission quality. It assesses the combined effects of a wide range of end-to-end transmission parameters that affect the conversation quality of narrow band (300-3,400 Hz) telephone networks [11]. The fundamental principle of the E-model is based on the following two assumptions:

- Transmission impairments can be transformed into psychological factors
- Psychological factors on the psychological scale are additive

The second assumption is based on a concept in the Overall Performance Index model for Network Evaluation (OPINE) model [90] from NTT, Japan.

The reference connection of the E-model is divided into a send side and a receive side. The E-model estimates the conversational speech quality from mouth to ear as perceived by the user at the receive side, both as listener and talker.

The E-model has 21 input parameters covering handset characteristics, noises, delays, echoes, codec distortions and so on. These parameters are available at time of planning, either from the internationally accepted standards, network experience or measurements. The primary output of the E-model is a transmission rating factor *R*, which can be transformed into other quality measures, such as MOS.

The E-model has been revised four times since its first 1998 version. In the 2000 version, formulas were updated to better take into account the effects of room noise at the

send side and quantization distortion. In the 2002 version, the impairment due to random packet loss is formulated to replace the old tabulated form. In the 2003 version, the formula for talker sidetone is updated. In the latest 2005 version, the effects of burst packet loss are formulated.

In this thesis, the 2005 version E-model is used.

2.5.2. Algorithm Description

The E-model first computes a base value from noise, and each impairment is expressed by a value and then subtracted from this base value. The transmission rating factor R is given by:

$$R = Ro - Is - Id - Ie\_eff + A \tag{2.1}$$

Ro represents the basic SNR. *Is*, simultaneous impairment factor, is a combination of all impairments which occur more or less simultaneously with the voice signal. *Id*, delay impairment factor, represents the impairments caused by delay and echo. *Ie_eff*, effective equipment impairment factor, represents the impairments caused by low bit rate codecs, including codec distortion and packet loss. Finally, advantage factor *A* allows the compensation to speech quality in situations where users can tolerate some decrease in quality.

Ro in principle represents the basic SNR. In the absence of other impairments, SNR provides a good indicator of speech quality. The default value of *Ro* is 94.8.

Is can be further divided into three terms:

$$Is = Iolr + Ist + Iq \tag{2.2}$$

where *Iolr*, *Ist* and *Iq* represent the impairments of too loud speech level, non-optimum sidetone and quantization distortion respectively. The default value is 1.4.

Id can be also divided into three factors, *Idte*, *Idle* and *Idd*:

$$Id = Idte + Idle + Idd \tag{2.3}$$

where *Idte*, *Idle* and *Idd* represent the impairments due to talker echo, listener echo and too-long absolute delay even under perfect echo cancellation, respectively.

The factor *Idte* is calculated using (2.4), (2.5) and (2.6) in turn, when other parameters that we are not concerned with are set to their default values:

$$TERV = TELR - 40\log_{10}\frac{1+\frac{T}{10}}{1+\frac{T}{150}} + 6e^{-0.3T^2} \tag{2.4}$$

where *TELR* is the talker echo loudness rating in dB and T is the one-way talker echo path delay in ms, as introduced in subsection 2.2.3.

$$\mathrm{Re} = 80 + 2.5(TERV - 14) \tag{2.5}$$

$$Idte = \left[\frac{94.7688 - \mathrm{Re}}{2} + \sqrt{\frac{(94.7688 - \mathrm{Re})^2}{4} + 100} - 1\right]\left(1 - e^{-T}\right) \tag{2.6}$$

For $T < 1$ ms, the talker echo should be considered as sidetone, i.e. $Idte = 0$.

Similarly, the factor *Idle* is calculated as follows:

$$Rle = 1228.5\,(2T+1)^{-0.25} \tag{2.7}$$

$$Idle = \frac{94.7688 - Rle}{2} + \sqrt{\frac{(94.7688 - Rle)^2}{4} + 169} \tag{2.8}$$

It can be seen that when T increases from 0 to 150 ms, *Idle* would increase from 0.1491 to 0.8408. In other words, its effect is minimal and is therefore ignored in the thesis.

The factor *Idd* is calculated as follows:

For $T < 100$ ms:

$$Idd = 0$$

For $T > 100$ ms:

$$X = \frac{\log_{10}\left[\frac{T}{100}\right]}{\log_{10} 2}. \tag{2.9}$$

$$Idd = 25\left[\left(1+X^6\right)^{\frac{1}{6}} - 3\left\{1+\left(\frac{X}{3}\right)^6\right\}^{\frac{1}{6}} + 2\right] \tag{2.10}$$

Ie represents the impairments due to low bit rate codecs under no packet loss. The *Ie* values depend on subjective MOS tests and network experience; also, they can be determined through instrumentation approaches. When packet loss occurs, the effective equipment impairment factor *Ie_eff* is derived using the codec specific *Ie* and the packet loss robustness factor *Bpl*:

$$Ie\_eff = Ie + (95 - Ie) \cdot \frac{Ppl}{\frac{Ppl}{BurstR} + Bpl} \tag{2.11}$$

where *Ppl* is the packet loss probability in percentage, *BurstR* is the burst ratio of the packet loss and is given by:

$$BurstR = \frac{Average\ length\ of\ observed\ bursts\ in\ an\ arrival\ sequence}{Average\ length\ of\ bursts\ expected\ for\ the\ network\ under\ "random"\ loss}. \tag{2.12}$$

When packet loss is random, *BurstR* = 1, and when packet loss is bursty, *BurstR* > 1.

Some recommended *Ie* and *Bpl* values are given in the Appendix I to the ITU-T Rec. G.113 [29].

The advantage factor *A* sometimes is called expectation factor; the default value is 0 in wire-line VoIP services.

As will be seen in Chapter 6, a parametric model is developed to quantify the effect of burstiness on speech quality in a different approach, the concept of *BurstR* here will be adopted.

2.5.3. Interpretations of the E-model Rating Factor *R*

The E-model rating factor *R* ranges from 0 to 100, and it can be transformed into a five-point MOS by:

$$MOS = \begin{cases} 1, & R < 0 \\ 1 + 0.035R + R(R-60)(100-R)7 \cdot 10^{-6}, & 0 < R < 100 \\ 4.5, & R > 100 \end{cases}. \qquad (2.13)$$

The relationship between *R* and MOS is illustrated in Figure 2.13, and the transmission speech quality can be interpreted in Table 2.4 as defined in ITU-T Rec. G.109 [91].

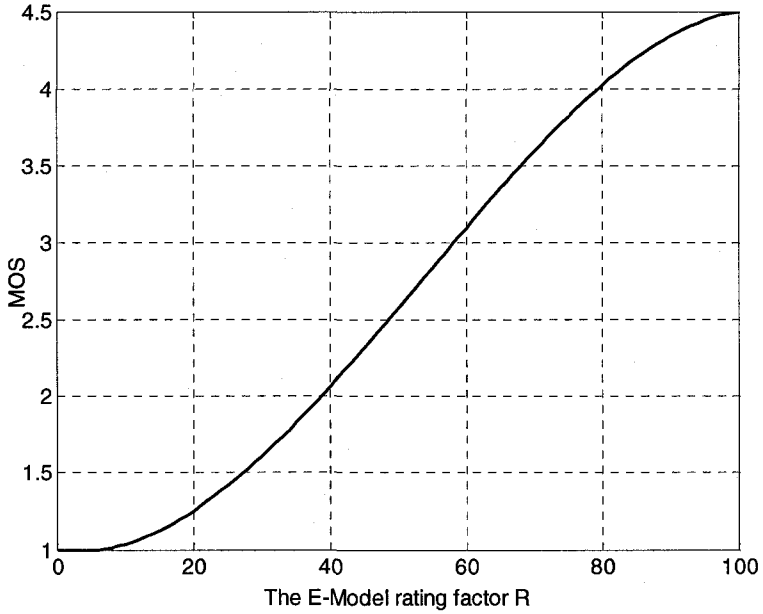


Figure 2.13 MOS as a function of rating factor *R*

Table 2.4 Relationship between *R* and user satisfaction

R value range	MOS range	Speech transmission quality category	User satisfaction
90 ... 100	4.34 ... 4.50	Best	Very satisfied
80 ... 90	4.02 ... 4.34	High	Satisfied
70 ... 80	3.60 ... 4.02	Medium	Some users dissatisfied
60 ... 70	3.10 ... 3.60	Low	Many users dissatisfied
50 ... 60	2.58 ... 3.10	Poor	Nearly all users dissatisfied

2.5.4. The E-model and INMD

The E-model provides a computational model for speech quality planning. How to non-intrusively measure these related parameters is not covered by the E-model.

The In-service Non-intrusive Measurement Device (INMD) defined in the ITU-T Rec. P.561 can be viewed as the "measurement counterpart" of the E-model [92]. INMD can

be connected at any four-wire point, but is most commonly deployed at international gateways [93]. It primarily measures the voice-grade parameters, including speech and noise level, echo level, echo path delay, packet delay variation and packet loss rate. The interfaces, measurement ranges and accuracy requirements for parameters are defined in the ITU-T Rec. P.561.

The interpretation of the speech quality from INMD measurement parameters is described in ITU-T Rec. P.562 [70]. The CCI model developed by British Telecom in 1998 is used to model the average speech quality based on speech, noise, echo levels and echo path delay. Also, these INMD measurement parameters can be mapped to some E-model parameters, as shown in Annex B to the ITU-T Rec. P. 562 [70].

In short, ITU-T Rec. 561 defines what should be measured non-intrusively in the networks for quality monitoring purposes; then these measured results can be interpreted by ITU-T Rec. 562. Note that, as explained in subsection 2.4.1, ITU-T Rec. P.563 is an objective algorithm for measuring speech quality, based on processing the voice payload of the receive-end speech signal only.

2.6. PESQ

In the thesis, the PESQ algorithm is used to objectively measure the speech quality.

2.6.1. Introduction

In 1999, five draft perceptual speech quality algorithms were reviewed by the ITU-T. Among them, PSQM99 (PSQM 1999 version, an updated version of PSQM [94]) and

PAMS performed best in the benchmark tests, with an average correlation of 0.93 and 0.92 respectively [68]. Then, it was decided that the merits of these two algorithms could be combined into a new measurement algorithm. Such a draft was jointly submitted to the ITU-T in 2000, by J. G. Beerends and A. P. Hekstra of KPN Research (developers of PSQM), and A. W. Rix and M. P. Hollier of British Telecom (developers of PAMS). The new algorithm is called PESQ and was standardized as ITU-T Rec. P.862 in February 2001. PESQ replaces PSQM.

In PESQ, the perceptual model of PSQM99 and the variable delay estimation of PAMS are kept. Also, new methods for transfer function equalization and averaging distortion over time are added [57], [68].

2.6.2. Algorithm Description

The block diagram of PESQ is shown in Figure 2.14 [95]. It includes the following three steps [10], [57], [67], [68], [95]:

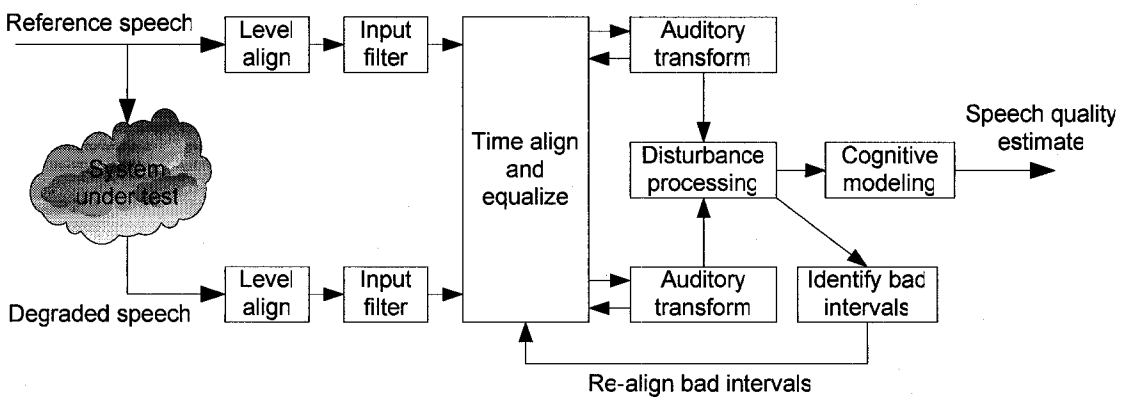


Figure 2.14 Structure of PESQ

Step 1: Signal pre-processing

- Level alignment: PESQ assumes a standard listening level of 79 dB SPL (Sound Pressure Level) at the ear reference plane. This level is judged as preferred on a loudness preference scale [96]. Both signals are compensated to this level.
- Input filtering: PESQ also assumes that the listening tests are conducted using an IRS receive or a Modified IRS (MIRS) receive characteristics in the handset. Therefore, an IRS filter is applied to both scaled signals to take into account this effect. The filtered signals are then used in the time alignment and the perceptual model.
- Time alignment: To compare the difference between two signals, the delay must be correctly identified. Several procedures are used here. First, the crude delay is calculated across both entire signals using the envelope-based correlation method. Then, the reference signal is divided into utterances, and the envelope-based and histogram-based fine delay estimations are carried out on utterances. Finally, because the delay may change during the utterance in some cases, utterance is further divided into two parts and re-aligned to search for fast delay variations.

Step 2: Perceptual modeling

This step produces the internal representation of both signals that takes into account human perception.

- Time-frequency transformation: It is performed by using a short-term Fast Fourier Transform (FFT) with a Hanning window. The window size is 32 ms with 50% overlapping.
- Frequency warping: As human ears have finer resolution in low frequency than in high frequency, the power spectrum is grouped into 42 bins, equally spaced in the Bark domain. It maps the frequency scale in Hertz to the pitch scale in Bark. The result is the pitch power density.
- Intensity warping: The reference and degraded pitch power densities are transformed to a Sone loudness scale.

Step 3: Cognitive modeling

This step calculates the difference between two internal representations and maps it to speech quality in MOS.

- Disturbance density: The difference between the reference and degraded loudness density is computed. Then, small distortions that are inaudible are masked.
- Asymmetry processing: "The asymmetry effect is caused by the fact that, when a codec distorts the input signal, it will, in general, be very difficult to introduce a new time-frequency component that integrates with the input signal, and the resulting output signal will thus be decomposed into two different components, the input signal and the distortion, leading to clearly audible distortion. When the codec leaves out a time-frequency component, the resulting output signal cannot be decomposed in the same way and the distortion is less objectionable [10]."

PESQ computes two different error averages, one with and one without an asymmetry factor, known as asymmetrical disturbance density and symmetric (normal) disturbance density, respectively. The former is calculated by multiplying an asymmetry factor to the latter. As a result, only the time-frequency cells whose degraded pitch power density exceeds the original pitch power density remain.

- Frame disturbance: The normal disturbance density and the asymmetrical disturbance density are aggregated over the frequency bands to calculate the frame disturbance.
- Time alignment re-assessed: Consecutive frames with a frame disturbance above a threshold are called bad intervals [10]. They may be caused by incorrect time alignment during pre-processing and result in large disturbance to speech quality prediction. Therefore, the time delay for bad intervals is re-assessed and the corresponding frame disturbance is also recalculated.
- Disturbance aggregation: The normal disturbance density and the asymmetrical disturbance density are then aggregated progressively from utterance split to the overall speeches.
- MOS prediction: The final PESQ score is a linear combination of the average normal disturbance value and the average asymmetrical disturbance value [95]:

$$PESQ\ MOS = 4.5 - 0.1 Disturbance_{SYM} - 0.0309 Disturbance_{ASYM}. \quad (2.14)$$

The PESQ MOS ranges from -0.5 to 4.5. But for most cases, it only lies between 1.0 and 4.5.

2.6.3. Applications and Performance

PESQ is designed to measure the end-to-end speech quality in narrow band telephone networks and speech codecs in the listening-only model. Factors such as delay and echo that are related to two-way conversation are not reflected in the PESQ score.

PESQ can effectively measure the impairments such as codecs operating above 4 kb/s, transcoding, speech input levels to a codec, transmission channel errors (bit error, packet loss and cell loss), fast delay variation, environmental noise at the sending side. It is not intended to measure the effects of delay, listening level, sidetone, talker echo, listener echo, music input and codecs operating below 4 kb/s, etc [10].

For 22 known ITU-T benchmark tests, the average correlation between PESQ after a monotonic 3rd order polynomial mapping and the subjective MOS is 0.935. For the prediction errors, 69.2% and 91.3% of the absolute residual errors are within 0.25 and 0.50 MOS respectively [10]. Detailed performance analysis under different kinds of network conditions can be found in [95].

2.6.4. Mapping Versions of PESQ

The PESQ MOS raw score ranges from -0.5 to 4.5. It is desirable to provide a mapping function from the PESQ score to an average ITU-T Rec. P.800 MOS in ACR scale. Thus, a linear comparison can be made between them.

In 2003, a single mapping function was standardized in ITU-T Rec. P.862.1 [19] as follows:

$$y = 0.999 + \frac{4}{1 + e^{-1.4945x + 4.6607}} \tag{2.15}$$

where x is the raw PESQ MOS score, y is the mapped score and is referred to as P.862.1 MOS-LQO.

This mapping function has been optimized over a large number of subjective listening quality tests in ACR scale. The test databases include different applications and conditions, such as VoIP, wireless, fixed and clean channel, in nine languages.

A small performance improvement for the P.862.1 MOS-LQO over the original raw PESQ score has been noted. For example, the overall correlation coefficient increases from 0.876 to 0.879, and the percentage for absolute residual error below 0.25 MOS increases from 36.1% to 41.92% [19].

Another mapping version of PESQ is proposed by Psytechnics [97], [98]. It is called PESQ - Listening Quality (PESQ-LQ) and is given by:

$$y = \begin{cases} 1.0, & x \le 1.7 \\ -0.157268\,x^3 + 1.386609\,x^2 - 2.504699\,x + 2.023345, & x > 1.7 \end{cases} \tag{2.16}$$

where x is the raw PESQ MOS score, and y is the mapped PESQ-LQ.

The P.862.1 MOS-LQO ranges between 1.02 and 4.55, whereas the PESQ-LQ ranges between 1.0 and 4.5. The relationships between PESQ MOS raw score and two mapped versions are illustrated in Figure 2.15. A dashed diagonal line running through 1.0 to 4.5 is also shown in the figure for characterizing the mapping functions. In both mapped versions, the PESQ raw score is lifted up in the high-end (around 3.37-3.50 or higher), and is generally pushed down when it is below 3.37-3.50. For an extreme low PESQ raw score (close to 1.0 or lower), it is clipped to 1.0-1.2.

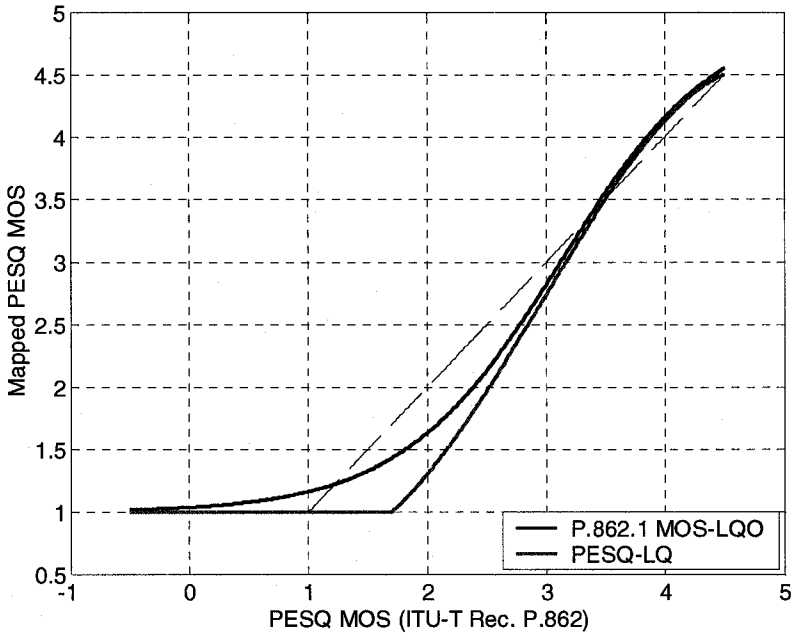


Figure 2.15 PESQ mapping functions

In the thesis, the ITU-T Rec. P.862.1 MOS-LQO is used as the subjective MOS estimate.

2.7. DSLA

Digital Speech Level Analyzer (DSLA), as shown in Figure 2.16, is a hardware tool box and is used to implement the PESQ algorithm for speech quality measurement.

DSLA is manufactured by Malden Electronics Ltd., U.K. It measures the characteristics of the speech channel and predicts the speech quality on the MOS scale for several algorithms, including PAMS, PESQ and PESQ-LQ. Also, DSLA can generate a variety of useful signals to simulate networks and network elements [99]. It can be connected to a computer via a standard serial port, or to a network through Ethernet.

Figure 2.16 DSLA measurement toolbox

For speech quality measurement, the speech signals used can be recorded by the DSLA or directly from the user's PC. DSLA supports sound files in .wav (containing 44-byte header) or raw (without header) format, mono channel, 16-bit, 8,000 or 16,000 Hz. The speech can be a real speech sample or an artificial speech-like test stimulus complying with ITU-T Rec. P.50. A common PESQ result display window is shown in Figure 2.17 for illustration purposes only. The MOS results have a precision of two decimal digits.

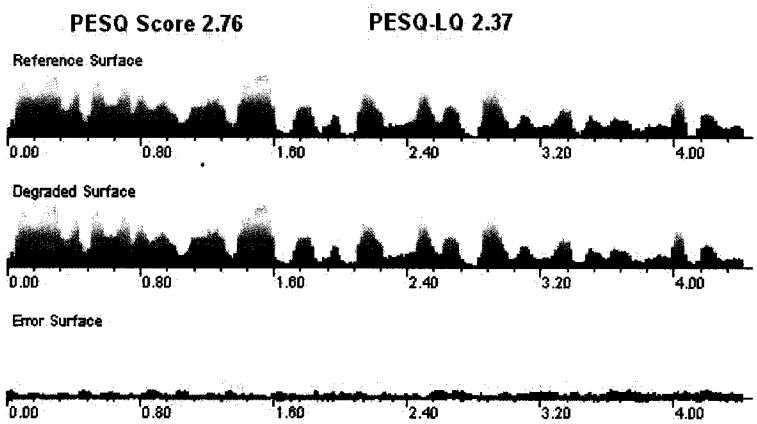


Figure 2.17 PESQ result display window

DSLA also provides the ability for batch processing; it is quite useful when a large number of files need to be processed without user intervention. In this model, the results are saved in a log file for further analysis.

The firmware version of the DSLA used in the thesis is 4.28, and the PESQ algorithm version is 1.4.2. It does not include the P.862.1 MOS-LQO, therefore we convert the measured PESQ to MOS-LQO by (2.15).

CHAPTER 3 SYSTEM DESIGN AND SIMULATION SET-UP

In this chapter, the proposed overall non-intrusive algorithm is described, followed by the simulation and measurement set-ups. The selection of the reference speech samples and the statistics used in performance evaluation are given as well.

3.1. The Overall Non-Intrusive Model

This thesis develops a novel non-intrusive VoIP speech quality assessment algorithm by assessing the individual as well as combined effects of several major speech quality impairments, including temporal clipping, echo, packet loss, codec distortion and noise.

As reviewed in Sections 2.3, 2.4 and 2.5, the voice payload analysis models are listening-only models; they do not consider factors affecting conversational quality, such as echo and delay. On the other hand, the protocol analysis models hardly cover impairments directly linked to the speech signal itself, such as temporal clipping and noise. These two approaches complement each other; it is advantageous to combine their merits in a new non-intrusive model, which is further substantiated by the E-model.

The proposed non-intrusive model is a parametric model, based on the structures developed in [15], as illustrated in Figure 3.1.

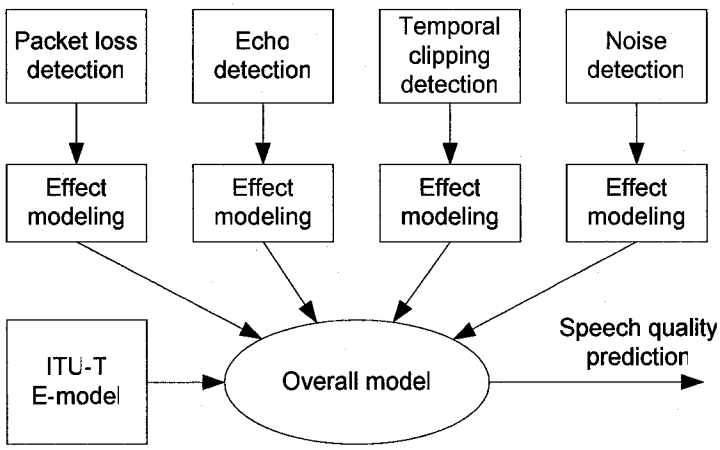


Figure 3.1 The structure of the proposed VoIP speech quality assessment model

The model consists of three steps. First, the impairments, including packet loss, echo, temporal clipping and noise, are detected. In order to detect the occurrences of packet loss, loss pattern and codec configurations, the Internet protocol analysis approach is used. Further, a scheme is adopted to classify the lost packets into three types, voiced, unvoiced or silence, through voice payload analysis. The detections of echo, temporal clipping and noise are solely based on processing the voice payload. Then, for each impairment, its effect is quantified by using the corresponding parameters detected from the first step. Finally, the overall listening-only model and the conversational model are built by integrating individual models and the E-model.

Note that echo and noise are not new impairments to IP networks; their effects on speech quality have been studied over the years, for traditional telephone networks. Many models have been developed to quantify their impacts on speech quality, such as the E-model [11], CCI model [70], and the complicated noise model in ITU-T Rec. P.563 [12]. It is not our objective to investigate effects of echo and noise here; existing models, such as the E-model, can be adopted.

The proposed non-intrusive algorithm can be used for network design, identification of root causes of speech quality degradation, and speech quality assessment purposes in VoIP. As stated earlier, it is advantageous in several aspects. First, it can efficiently utilize the processing resources by the three-step structure. Second, the model provides the overall speech quality rating as well as the contributions from individual impairments. Finally, the structure's extensibility makes it easy to incorporate other new impairments.

3.2. System Setup and Simulation Design

3.2.1. Simulation Structure

The simulation design is illustrated in Figure 3.2. The simulation system includes a reference speech database, an impairment simulator, an objective intrusive speech quality measurement block and a non-intrusive measurement block.

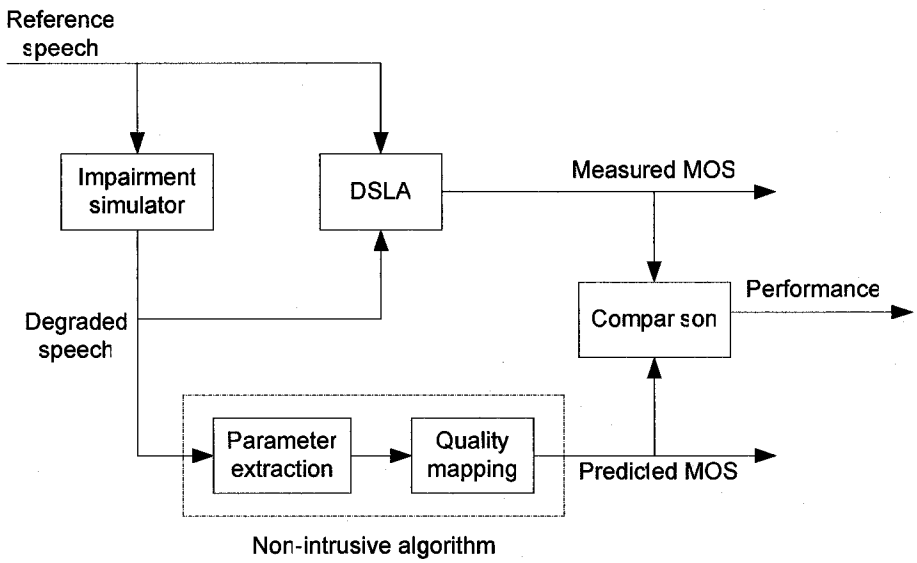


Figure 3.2 The structure of simulation system

As shown in Figure 3.2, the degraded speech is obtained by passing the reference speech through the impairment simulator. The simulator can generate the designated impairments in a fashion of either one kind of impairment alone or combined together. Then, our proposed algorithm tries to detect the impairments and maps them to speech quality in MOS from the degraded speech sample only. On the other hand, the speech quality is also intrusively measured by P.862.1 MOS-LQO using DSLA. Finally, in order to evaluate the performance of the proposed algorithm, the predicted MOS is compared to the measured MOS through statistical analysis.

The subjective MOS test is currently under preparation. When the real MOS is obtained, it can be used to calibrate the proposed non-intrusive model.

3.2.2. Speech Database

The speech samples used in the thesis are selected from the TIMIT speech corpus [100]. The TIMIT speech corpus contains 6,300 American English sentences, spoken by 630 people from 8 major dialect regions in the U.S., with 10 sentences for each speaker. Each sentence contains one talkspurt only, and is saved in 16-bit, 16,000 Hz linear PCM format. The start and end of the talkspurt are provided in a separate transcription file.

In the thesis, two sets of speech samples are used as the reference speech waveforms. The first set includes 20 samples from dialect region 1, and is used as the training set to develop the speech quality prediction algorithm. The second set is used as the testing set to evaluate the performance of the proposed algorithm. To take into account the dialect diversity, 32 well-balanced samples from all of the 8 dialect regions are used in the second set. The details of the speech sample selections are given in Table 3.1.

Table 3.1 Reference speech sample selection

Set	Dialect region	# Speakers/ Region	# Samples /speaker	Total samples
1	1	Male: 1 Female: 1	10 10	20
2	1, 2... 8	Male: 1 Female: 1	2 2	32

A pre-processing stage is applied to the selected TIMIT speech samples before they are used as the reference speeches, as shown in Figure 3.3.

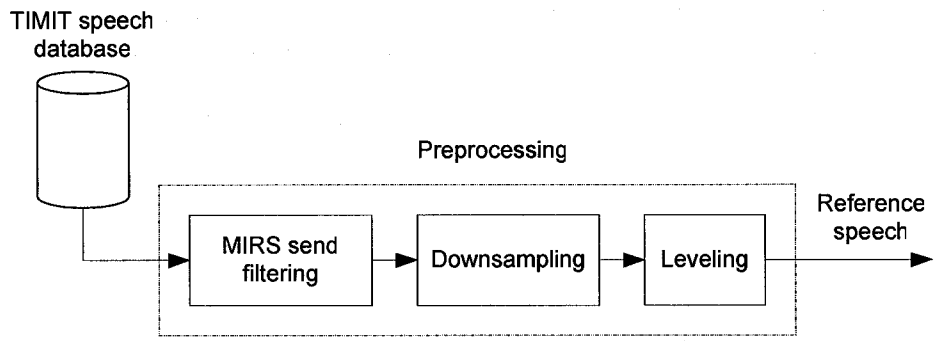


Figure 3.3 Pre-processing of the TIMIT speech samples

First, the speech sample is filtered by the MIRS send filtering characteristics defined in the ITU-T Rec. P.830 Annex D [58], which reflect the average send frequency response of telephone handsets used in modern digital networks, particularly when the low bit-rate codec is located in the handset. On the other hand, the MIRS receive characteristics are less important and do not need to be implemented to the speech database, because the PESQ algorithm already includes them (subsection 2.6.2). Also, the amplitude response of the receive characteristics is relatively flat. The magnitude

response of the MIRS send filter is illustrated in Figure 3.4, which is defined at 16 KHz sampling rate. It provides a boost of about 9 dB for higher frequencies and it sharply attenuates below 300 Hz. Above 3,400 Hz, it is seen to roll-off gradually.

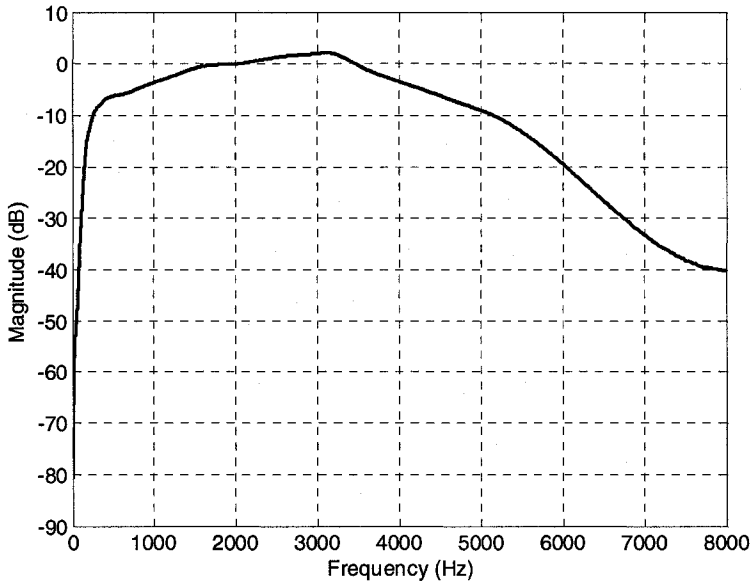


Figure 3.4 Magnitude response of the MIRS send filter

Then, the speech samples are downsampled from 16,000 Hz to 8,000 Hz in turn. Finally, the active speech level that is defined in ITU-T Rec. P.56 [102] is adjusted to -26 dBov (dB below the overload point of a digital system) to account for the standard listening level.

All the above three pre-processing steps are implemented by the C programs provided in ITU-T Software Tool Library 2000 in the ITU-T Rec. G.191 [101].

It is noticed that on average, the selected speech samples are relatively short in length (less than 4 seconds) and speech activity factor is too high (above 90%). So one second of silence is appended to the end of each speech sample. After that, the pre-processed

speech samples meet the general requirement in the ITU-T Rec. P.862 and are subsequently used as the reference speeches in simulations. For the total 52 speech samples used, the average duration is 4.396 seconds with 69.87% active.

3.2.3. Data Analysis

In order to develop the speech quality prediction algorithm, the simulations are usually run independently for 10 times and the average result is used. In the performance evaluation, on the one hand, the speech quality is predicted by the proposed algorithm. On the other hand, the speech quality is measured by the P.862.1 MOS-LQO. The predicted MOS and the measured counterpart are compared. The closeness of the prediction is characterized by the correlation coefficient ρ defined in (3.1), where x_i, y_i are the predicted and measured MOS respectively. The prediction Root Mean Square Error (RMSE) σ and distribution of the absolute prediction error e_i are also used to gauge the performance, as given by formulas (3.2) and (3.3) respectively.

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3.1}$$

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2} \tag{3.2}$$

$$e_i = |x_i - y_i| \tag{3.3}$$

As will be seen in the following chapters, when evaluating the performance of the proposed impairment detection algorithms, normally, only σ is used. In that case, x_i, y_i in

(3.2) are the measured values and true values of the impairment parameters (e.g. echo path delay), respectively.

Based on the overall structure of the proposed non-intrusive model, the detailed detection and modeling algorithms for each impairment as well as the overall model are presented in the following chapters. Particularly, the temporal clipping is presented in Chapter 4, followed by echo and packet loss in Chapters 5 and 6 respectively. And finally, the overall model, which incorporating temporal clipping, echo, packet loss and noise, is given in Chapter 7.

CHAPTER 4 TEMPORAL CLIPPING ON SPEECH QUALITY

In this chapter, the effects of temporal clipping on the perceived speech quality are investigated. The temporal clipping usually results from silence suppression, or EC's NLP, and the clipped speech portions are replaced by the CN. A non-intrusive algorithm is proposed to predict the speech quality based on the clipping statistics. The impacts of speech frame size and CN spectrum on the proposed algorithm are also investigated.

4.1. The Objective and Related Works

As introduced in subsection 2.2.2, there are two sources of temporal clipping in VoIP, silence suppression by VAD and NLP of an EC. From the speech quality point of view, the effect of NLP clipping is the same as in VAD, so every case considered here is applicable to NLP, while only mentioning VAD.

The performance of the VAD algorithm is very critical. It is not always perfect, when the talkspurt of a speech is classified as non-active, this portion gets clipped and the speech quality is degraded.

Little attention has been given to the effects of temporal clipping on speech quality so far. In [103], an intrusive method was proposed by measuring the Euclidean distance between the output of an auditory filter simulated by the reference speech and output simulated by the VAD processed speech. The mean distance over all filters provides a

measure for speech quality that suffers from VAD clipping. In [104], [105], a psychoacoustic auditory model was used. As the most significant effect of clipping consists of loss of loudness, a parameter called activity burst corruption is proposed. The parameter is defined as the ratio of the sum of the total loudness suppressed by VAD clippings to the total loudness of the activity burst, i.e., the ratio of loss in loudness scale. The speech quality is modeled as a quadratic function of this parameter. The main limitation of the both methods is that they are intrusive as the reference speech is needed in calculations.

Based on the concepts introduced in [106], [107], VAD clippings can be classified into three categories according to the clipping locations: Front End Clipping (FEC), Middle Speech Clipping (MSC) and Back End Clipping (BEC), as shown in Figure 4.1 for a speech waveform.

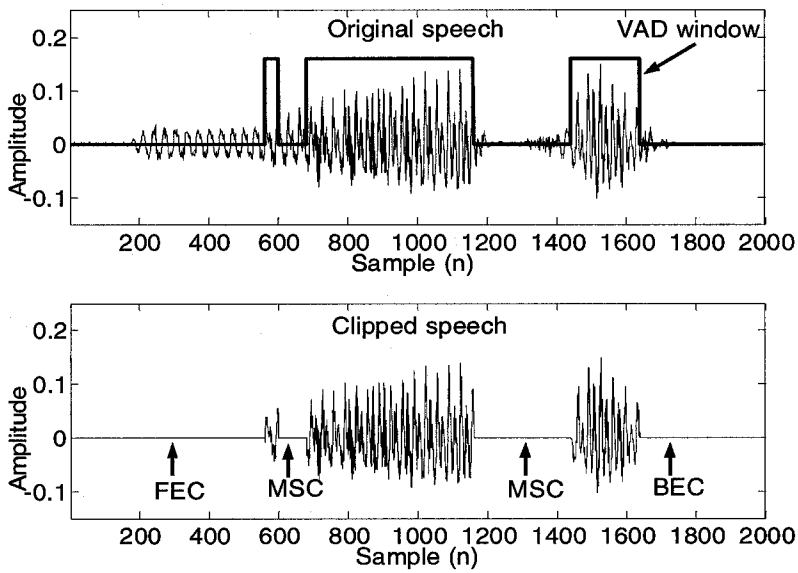


Figure 4.1 Classification of VAD clipping locations

The objective of this thesis is to develop a non-intrusive algorithm to quantify the effects of temporal clipping on speech quality. As will be seen in Section 4.4, the algorithm is based on the clipping statistics, i.e., clipping locations and clipping percentages. The thesis assumes prior knowledge of clipping statistics for the degraded speech. The detection of temporal clipping relies on the algorithms developed in [17], which will be introduced briefly in the next section.

4.2. Detection of Temporal Clipping

Two methods are proposed in [17] to detect the temporal clipping non-intrusively. The first method applies a threshold to the received speech signal and gradually increases this threshold. If the current threshold is lower than the one used to introduce temporal clipping, the resulting signal will be the same as the received signal. When the threshold is increased to a level which is a little bit higher than the VAD threshold, the resulting signal will differ from the received signal. So, the threshold used in the VAD can be identified. The second method is based on the statistical distribution of the talkspurt and silence intervals of human speeches. Detection threshold and the clipping statistics can be derived from either method; please refer to [17] for further information.

4.3. Simulation and Measurement Design

The effects of temporal clipping on speech quality are analyzed quantitatively. The overall simulation and measurement block diagram for temporal clipping is given in Figure 4.2.

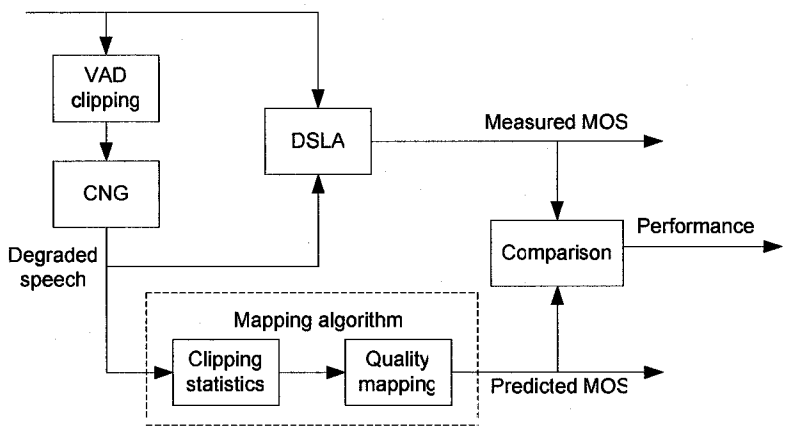


Figure 4.2 Simulation and measurement diagram

The temporal clipping is introduced by a VAD algorithm. To keep the naturalness, the clipped portions that also include some speeches are replaced by comfort noise, using a CNG. The corresponding degraded speech quality is measured by P.862.1 MOS-LQO. Meanwhile, the clipping statistics are collected. The new prediction algorithm, which maps the clipping statistics to speech quality, is proposed here. Finally, the performance of the proposed algorithm is examined by comparing the measured speech quality to the predicted one.

A) Speech Database

Speech sets 1 and 2, introduced in Chapter 3 are used here as the reference speeches. In short, the first set includes 20 samples from two speakers, and is used to develop the speech quality prediction algorithm. The second set includes 32 samples from sixteen speakers, and is used to evaluate the performance of the proposed algorithm.

B) Temporal Clipping Simulation and Statistics

To introduce the temporal clipping effect, an energy-based VAD algorithm is simulated. The speech is segmented into frames, and when the energy of a given frame exceeds the energy detection threshold, the frame is marked as active; otherwise, it is marked as non-active.

The speech frame sizes are selected to be 5, 10, 20 and 30 ms. The VAD detection thresholds are varied from 6 to 30 dB below the average energy (-26 dBov) with an interval of 3 dB. Thus, for each frame size, 180 and 288 degraded speech samples are generated from set 1 and set 2 samples respectively.

In the simulation, 30 dB pink noise (narrow-band telephone 300-3,400 Hz limited) is used to replace the clipped frames. This is because the power spectral density of pink noise is proportional to the reciprocal of the frequency, which is closer to human speech than that of white noise. Later on, white noise is also used as the comfort noise during the performance evaluation phase to test the effect of noise spectrum on the proposed algorithm.

We aim to model the speech quality as a function of clipping locations and percentages. For each speech sample, the talkspurt length is provided by the associated

transcription file. The FEC, MSC and BEC percentages are calculated as the ratio of length of the corresponding clippings to the length of the talkspurt.

C) Mapping Algorithm

Based on the clipping statistics from set 1 speech samples and measured MOS, the mapping algorithm is developed and will be given in the next section.

4.4. Speech Quality Modeling

This section introduces the proposed non-intrusive algorithm and presents the performance validation results.

4.4.1. Proposed Algorithm

In general, speech quality falls with increasing clippings. Moreover, different clipping locations have different weights on speech quality degradation. A multi-variable linear regression model is proposed to predict the speech quality as follows:

$$MOS = MOS_O - DMOS_{TC} \tag{4.1}$$

$$DMOS_{TC} = C1 \cdot FEC\% + C2 \cdot MSC\% + C3 \cdot BEC\% \tag{4.2}$$

where MOS_O = 4.55 and represents the optimum MOS without any impairment, $DMOS_{TC}$ is the MOS drop caused by temporal clipping, and is modeled as a linear combination of *FEC%*, *MSC%* and *BEC%*.

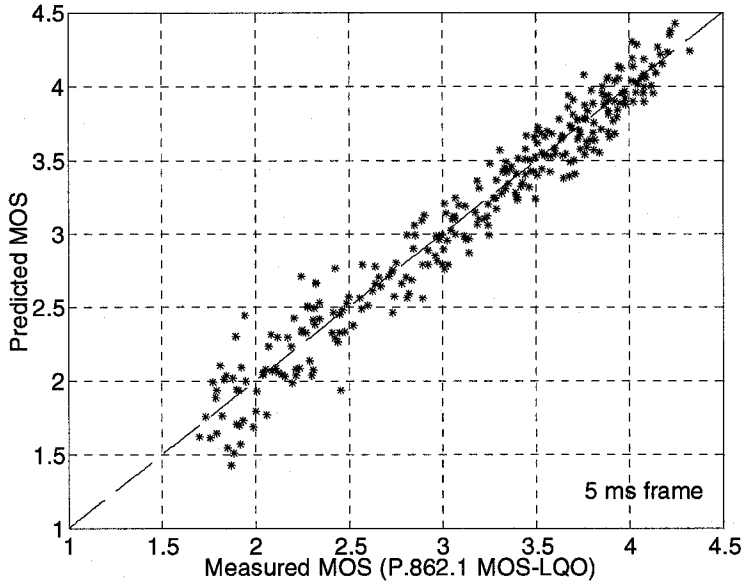
Based on the degraded samples from set 1, the coefficients in (4.2) are determined by using the linear regression, and summarized in Table 4.1. The valid ranges for *FEC%*, *MSC%* and *BEC%* (in percentage) are from 0 to 4, 0 to 50 and 0 to 10 respectively. These ranges are derived from the amount of speeches clipped under different VAD thresholds in the simulations. The overall MOS is limited between 1.02 and 4.55.

Table 4.1 Prediction model coefficients for temporal clipping

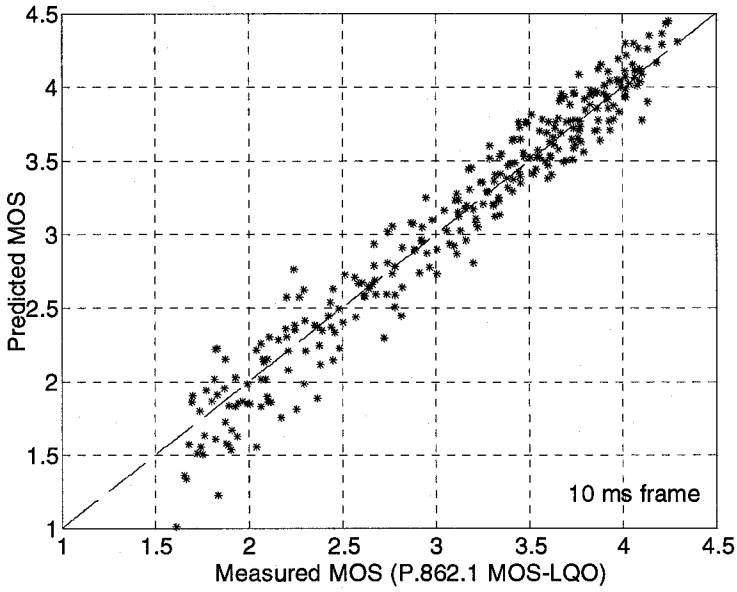
Frame size (ms)	$C1$	$C2$	$C3$
5	0.2418	0.0441	0.0340
10	0.2011	0.0503	0.0323
20	0.1955	0.0551	0.0333
30	0.2075	0.0594	0.0342

4.4.2. Performance Analysis

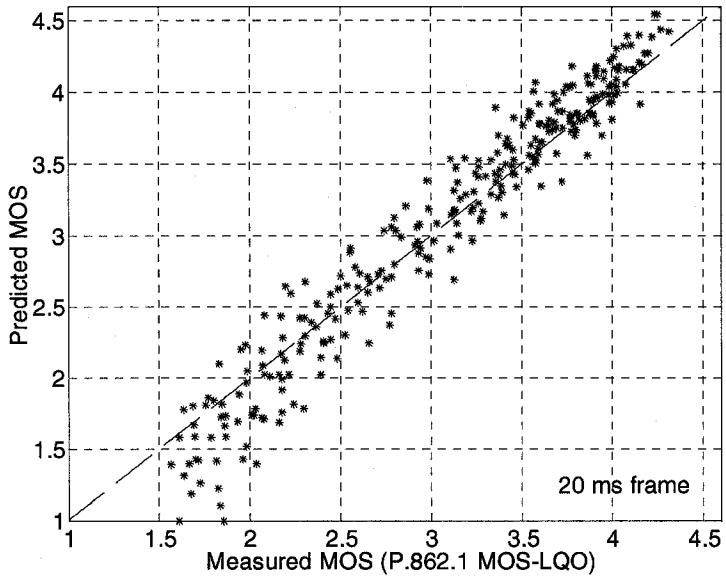
Degraded speech samples from set 2 are used to evaluate the performance of the proposed algorithm. The predicted MOS obtained using formula (4.2) and the measured MOS are illustrated in Figures 4.3 (a)-(d) for frame sizes of 5, 10, 20 and 30 ms respectively. The results show that the prediction model works very well as the majority of points closely lie on the dashed diagonal line. Also, MOS covers a wide range from the low-end (near 1.02) to the high-end (near 4.55), therefore the performance over the entire speech quality range is assured.



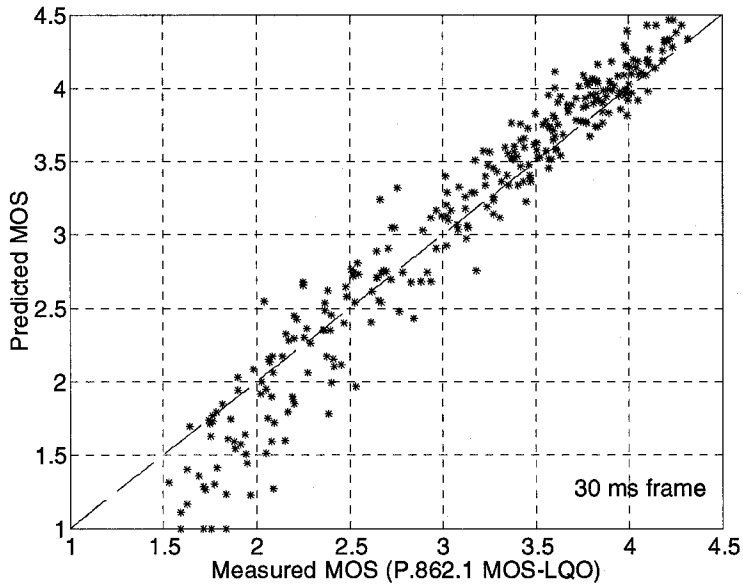
(a) 5 ms frame



(b) 10 ms frame



(c) 20 ms frame



(d) 30 ms frame

Figure 4.3 Prediction performances of the proposed algorithm

The Pearson correlation coefficient ρ, the RMSE σ and the absolute prediction error distributions are used as criteria to evaluate the performance of the algorithm. For 5, 10, 20 and 30 ms frame, ρ is 0.977, 0.975, 0.973 and 0.973; σ is 0.16, 0.18, 0.22 and 0.24 MOS respectively. The absolute prediction errors are binned with a width of 0.10 MOS, and percentages in each bin are illustrated in Figure 4.4. The results show that, on average, about 40%, 70% and 87% of the prediction is within 0.10, 0.20 and 0.30 MOS of the measurement respectively. Also, less than 2.3% of the absolute prediction error is greater than 0.50 MOS. Furthermore, the algorithm performance degrades when frame size becomes larger.

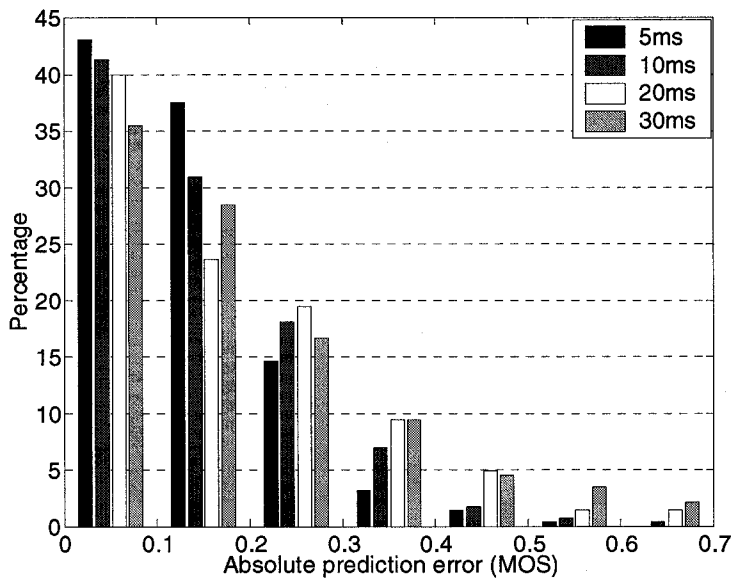


Figure 4.4 Absolute prediction error distributions under pink CN

4.4.3. Effects of Comfort Noise Spectrum

All the results so far are obtained by using pink noise as the CN. The actual noise spectrum may be different as the CNG algorithm varies. We also examine the performance under another common noise type, that is, white noise. Note that we still use the coefficients in Table 4.1, which is derived from the pink noise case.

In the same way, we introduce the temporal clipping to set 2 speech samples, but the clipped frames are replaced by the 30 dB white noise (narrow-band telephone 300-3,400 Hz limited also) instead of pink noise. The degraded speech quality is measured.

Under two different noise spectrums, the measurement shows that speech quality may appreciably change, especially when the amount of clipping is quite high. However, on average, the quality change resulting from the spectrum difference is rather limited, which is seen in the order of 0.03-0.06 MOS. This suggests that the proposed algorithm can also be applicable to other types of CN.

To evaluate the effects of noise spectrum on the algorithm performance, we still use the coefficients in Table 4.1 in the prediction model, although the actual noise is white. Note under white noise, the clipping statistics are unchanged for each degraded speech. For 5, 10, 20 and 30 ms frame, ρ is 0.972, 0.972, 0.970 and 0.971; σ is 0.19, 0.21, 0.26 and 0.28 MOS respectively. Similarly, the absolute prediction errors are binned with a width of 0.10 MOS and percentages within each bin are given in Figure 4.5. The results show about 36%, 67% and 82% of the prediction is within 0.10, 0.20 and 0.30 MOS of the measurement respectively, and less than 4.9% of the absolute prediction error is greater than 0.50 MOS. The algorithm performance slightly decreases but still works well.

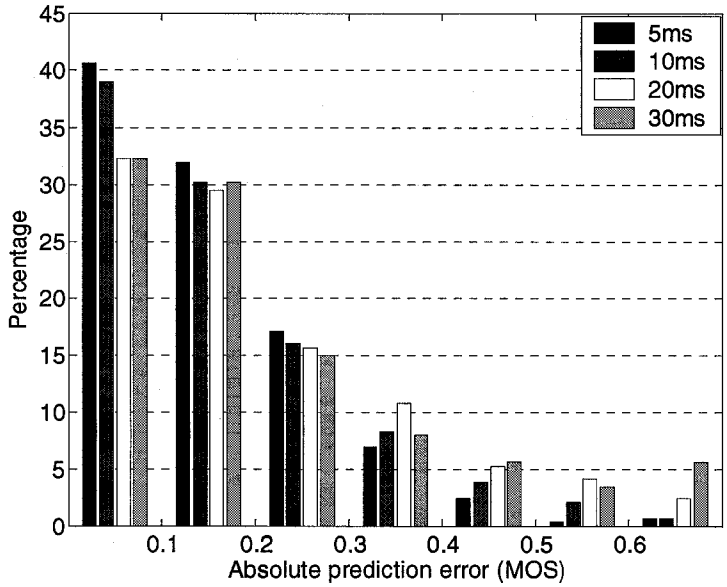


Figure 4.5 Absolute prediction error distributions under white CN

4.5. Discussion and Summary

The temporal clipping statistics provide useful information on predicting the degraded speech quality. Especially, clipping locations (FEC, MSC and BEC) have an important role here. The proposed algorithm, which differentiates between clipping locations by assigning different weight factors, shows excellent performance. We also tried other models that do not consider the clipping location factor, and they are shown to be inferior to the proposed algorithm.

Among the three coefficients, $C1$ is significantly larger than $C2$ and $C3$. The result suggests that, for the same amount of clipping percentage, front end clipping has the largest impact on speech quality. Their effect differences could be as large as 6 to 7 times in magnitude. This can be explained by the fact that the speech energy difference is

critical for all the perceptual speech quality assessment algorithms, including the PESQ. Usually, FEC happens at the beginning of the voiced part, whereas MSC and BEC happen at the unvoiced part. So, FEC would result in the largest energy loss to the degraded speech, and it significantly increases the perceptual distance. In turn, it has the largest impact on speech quality.

For given percentages of clippings to a speech, they should result in identical speech quality irrespective of the frame size. However, different frame sizes may cause different MSC distributions. For example, a 20 ms frame that is classified as active may be classified into one active frame and one non-active frame if frame size is changed to 10 ms. The distribution of the MSC also affects the speech quality, and therefore Ci (i = 1, 2 and 3) is almost the same for different frame sizes but not identical.

The VAD algorithm simulated in this thesis is energy based, while the actual ones could be quite complicated. For example, in ITU-T Rec. G.729B, the VAD decision is based on the differencing of the following four parameters: low-band (0-1 KHz) energy, full band energy, zero crossing rate and linear prediction spectrum [108]. This VAD algorithm is also used in ITU-T Rec. G.711 [109]. In ITU-T Rec. G.723.1, the VAD is basically an energy detector, and a hangover of 6 frames (180 ms) is added only in the case of speech bursts larger than or equal to 2 frames [110]. However, all kinds of VAD algorithms introduce temporal clippings. As the proposed model is built on the clipping locations and percentages, it can be applied to other VAD algorithms. Moreover, it is validated under some common CN spectrums, including pink noise and white noise.

In all cases, ρ is above 0.970 and shows excellent linear relationship between prediction and measurement. σ ranges from 0.16 to 0.28 MOS, however, the Figures 4.3 (a)-(d) show that the majority of huge prediction errors occur at the low-end MOS. Generally speaking, such a high amount of temporal clipping in addition to other impairments would cause the speech totally to be unintelligible, and should be avoided in any phone calls. In most cases, the temporal clipping is not so severe, and σ would be smaller.

As the real MOS is hard to obtain, MOS-LQO, which is a simple mapping version of the PESQ algorithm, is used to estimate the subjective speech quality in this chapter. It is the best available objective tool so far. When the real MOS is available, the modeling procedure will be slightly changed, by only providing better coefficients of the polynomial curve fitting.

The algorithm is suitable for general speech quality assessment schemes, such as VoIP, as the frame size of 10-30 ms considered in this thesis falls in the common sizes of the VoIP packet. Also, the model suggests that the correct detection of the onset of the talkspurt to avoid FEC is critical in the VAD design. However, the hangover time, which is used to reduce MSC and BEC by prolonging the opening of the VAD window, is not so important.

Summing up, this chapter has proposed a non-intrusive algorithm to estimate the effects of temporal clipping on speech quality in the MOS scale. The algorithm is based on the statistics of clipping, and shows how the clipping locations affect the speech quality in a quantitative way. The results show that the proposed algorithm can efficiently

predict the speech quality that suffers from temporal clipping. The correlation coefficient between the prediction and the measurement is about 0.975, and the RMSE for the prediction is about 0.16-0.28 MOS. The algorithm can be used as an integral part of the overall VoIP speech quality assessment algorithm.

CHAPTER 5 ECHO ON SPEECH QUALITY

In this chapter, the detection and effect modeling of network echo on speech quality are investigated. The echo detection is achieved by measuring its two parameters, echo path delay *D* and Echo Path Loss (*EPL*). Two methods are proposed to measure this delay in VoIP environments, where the echo suffers from excessive delay and nonlinear distortion. The methods aim at greatly reducing the computational requirements while maintaining good accuracy. *EPL* is also measured by using the obtained delay information. The performance under codec distortion, packet loss, noise and double talk conditions is examined through simulations and real field measurements. To model the effects of echo on speech quality, we rely on the current E-model.

As introduced in Chapter 2, there are two kinds of network echoes in telephony, talker echo and listener echo. We only focus on the former, as the effect of the latter is usually ignored, given that talker echo is well controlled.

5.1. The Objective and Related Works

Echo is problematic in VoIP networks. As stated in subsections 2.2.3 and 2.2.4, the additional delay introduced by IP networks makes any minor echoes more perceptible and annoying to listeners. Furthermore, the performance of most ECs is affected by the VoIP environment, due to excessive delay, packet loss, and codec distortion [43]. The subjective effects of echo on speech quality are characterized by its two associated parameters, *D* and *EPL*.

Note that the objective of this thesis is not to control the echo, but rather to provide efficient algorithms to measure the echo at the receive-end for the general speech quality assessor [15].

When the echo path delay exceeds 30 ms, the echo has to be cancelled using an EC. Normally, the EC is deployed in a media gateway, to interface a VoIP system to a circuit-switched network, such as the PSTN. In such cases, the corresponding PSTN parameters for *D* and *EPL* are known at the EC. However, those parameters greatly change when measured at the end IP terminal, due to the excessive delay added and any volume control and distortion involved in the IP network. Even when the voice quality is assessed at the media gateway, it is unlikely to easily access those parameters from a third party EC, or from the RTCP-XR protocol [16], in many cases.

As will be seen in the following section, the echo measurement algorithm is implemented at the nearest possible point to the end user. The delay estimation at the measurement point differs from that at the EC, due to the relative excessive delay associated with the former that would require significant computing resources. Also, the algorithm shares the processing power with other VoIP impairment measurements and analyses, which mandates a much simpler approach for efficient deployment. On the other hand, small measurement errors are acceptable since they have little impact on the speech quality estimate, as suggested by the E-model. Therefore, *D* and *EPL* still need to be measured but in a computationally efficient manner.

In traditional PSTN networks, except for satellite or marine cable calls, the echo delay is relatively small and the echo path is usually linear. The cross correlation method

and adaptive filtering method [93] can be used for echo delay measurement. The former explores the interdependence of the waveforms between the non-cancelled end speech and the returned echo. The latter assumes the LTI properties of the echo path, which is not valid in VoIP environments, due to the presence of VoIP codecs and the receiver jitter buffer. The background noise and double talk also affect the measurement performance. In addition, as VoIP systems introduce excessive delay to the echo, great computational efforts are required for calculating the cross correlation between two long speech sequences, or updating the lengthy adaptive filter taps.

Other methods, such as the combination of cross correlation and adaptive filtering techniques [111], frequency-domain block-processing algorithms [112], are aimed at the accurate measurement of echo parameters in the linear echo path, meeting the resolution and mean accuracy requirements specified in [93]. Subband [113], filter bank [114], or wavelet packet decomposition algorithms [115], are used for efficiently identifying echo path impulse response for echo cancellation purposes. They are beyond our considerations, since they rely on the linearity of the echo path; and most importantly, there is no need to estimate the impulse response for our speech quality monitoring.

In addition, we are interested in the performance of the echo measurement algorithm under some impairments particular to VoIP, such as codec nonlinear distortion and packet loss. Also, the performance under background noise and double talk situations needs to be examined.

For modeling the effects of talker echo on speech quality, current widely used objective algorithms, such as PAMS and PESQ, are not suitable. Essentially, they are

listening-only algorithms. However, the echo is simply a filtered version of the talker's voice. If it is used as the degraded signal together with the original speech as the reference to such measurement algorithms, the effects of delay and echo level will be cancelled out due to the pre-processing. Therefore, these objective algorithms cannot correctly predict the effect of echo on speech quality.

The most reliable way to evaluate the effect of echo is still through the subjective test, as described by the parameter *Idte* in the E-model; see subsection 2.5.2 for detailed formulas. Therefore, we here rely on the existing E-model for effect modeling.

Several other subjective models [116], [90], [70], and an objective algorithm, called Perceptual Echo and Sidetone Quality Measure (PESQM) [117], are also proposed. For example, in [117], the degraded signal is selected by summing up the reflected echo and the original speech, and the original speech is also used as the reference. Then, the psychoacoustic model in PSQM is used as a starting point, and several modifications are made to compare the difference between two signals, and finally map to MOS. However, those models have not gained much popularity.

5.2. Echo Path Model and Measurement Principles

5.2.1. Echo Path Model and Measurement Block

A reference connection model for an IP-to-PSTN call is shown in Figure 5.1. The network echo is generated at the hybrid of the local switch, and returned back to the talker. The EC installed at the media gateway performs the task of echo cancellation. The

residual echo parameters will be measured by our proposed algorithms that are implemented close to the talker.

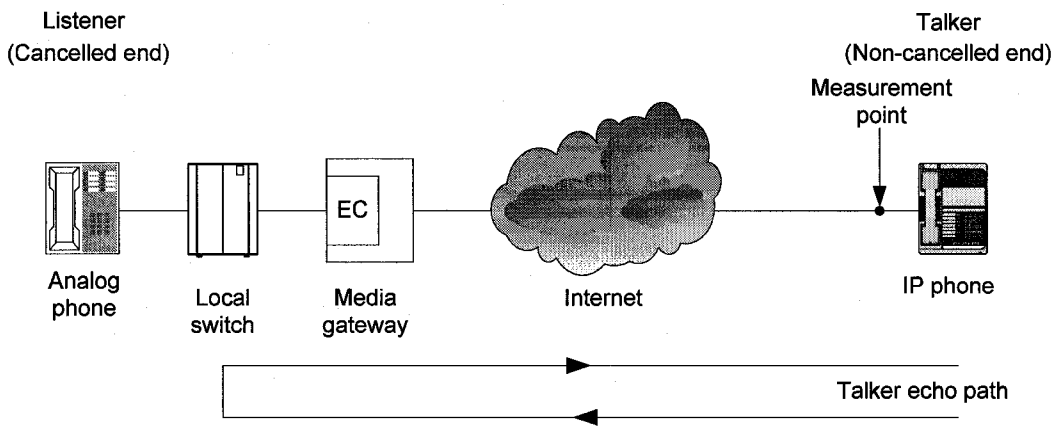


Figure 5.1 The reference connection model for an IP-to-PSTN call

The block diagram for the echo path model and measurement block is shown in Figure 5.2. Signal *x(n)* is the non-cancelled end speech, and *y(n)* is the residual echo. At the measurement point, these two signals will be monitored to measure the echo path parameters. The hybrid is modeled by a FIR filter *h(n)*. Signals *w(n)* and *v(n)* are the background noise and possible cancelled end speech, respectively. *D* is the round trip echo path delay, including all the delays from codecs, networks, receiver jitter buffer, hybrid, etc. *EPL* includes the attenuations not only from the hybrid, but also from the EC, codecs and transmission circuits.

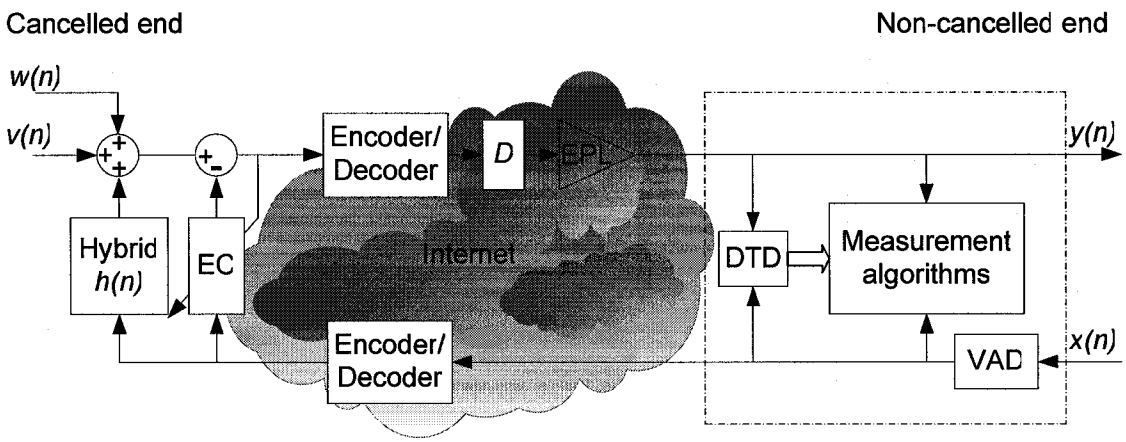


Figure 5.2 Echo path model and the measurement block diagram

The echo measurement block is shown inside the dashed lines in Figure 5.2. Besides the measurement algorithm block, it also includes a VAD and a DTD for measurement controlling purposes.

- *VAD block:* When the non-cancelled end speaker does not talk, the measurement is not needed. The VAD block serves as a detector for the onset of the non-cancelled end speech. For example, an algorithm based on the frame energy and LPC distance [118] can be used here.
- *DTD block:* Double talk incidentally occurs during a conversation. Although the cancelled end speech usually does not correlate with the non-cancelled end speech, it acts as high level noise to the measurement and may cause erroneous results. The measurement should be suspended when double talk is declared.

  Double talk detection algorithms may be energy-based or correlation-based [119]. The Geigel algorithm [120], a well-known energy-based algorithm used in practice, can be used here. The double talk status is declared when:

$$|y(n)| \geq \tau \cdot \max\{|x(n-1)|, |x(n-2)|, \cdots, |x(n-D)|\}. \quad (5.1)$$

The detection threshold τ is determined by the *EPL*. It is necessary to compare $y(n)$ to the most recent D samples of $x(n)$ because of the possible round trip delay. A hangover time is also specified in the algorithm; the double talk status is held for this amount of time beyond the last detection.

5.2.2. Principles of the Cross Correlation Method

When the echo path is linear, the echo can be expressed as:

$$y(n) = x(n) \otimes h(n) + w(n) + v(n) \quad (5.2)$$

where $\otimes$ denotes convolution. If there is no double talk, equation (5.2) is simplified to:

$$y(n) = x(n) \otimes h(n) + w(n) \quad (5.3)$$

For a limited observation length N, the cross correlation function between the echo and non-cancelled end speech can be estimated by:

$$C_{xy}(m) = \sum_{n=0}^{N-m-1} x(n+m)y^*(n) \quad (5.4)$$

where * denotes the complex conjugate. The normalized cross correlation is then calculated by:

$$\rho_{xy}(m) = \frac{C_{xy}(m)}{\sqrt{\sum_{n=0}^{N-m-1} x(n+m)x^*(n+m)}\sqrt{\sum_{n=0}^{N-m-1} y(n)y^*(n)}}. \quad (5.5)$$

ρ_{xy} ranges between -1.0 and 1.0. The lag m maximizing the normalized cross correlation is regarded as the delay:

$$\hat{D} = \arg \max_{m} \{ | \rho_{xy}(m) | \}. \quad (5.6)$$

The cross correlation method is based on the waveform similarities. Its performance degrades under noise, double talk, codec distortion and packet loss.

5.3. Proposed Echo Measurement Methods

To quantify the effects of echo on speech quality, *D* and *EPL* need to be measured. The new algorithms suggested in this thesis are described below. Two methods are proposed to measure *D*. They are cross correlation based, but the sequences used for computation are greatly reduced in terms of lengths or information contained. Then, *EPL* is measured by directly using the obtained echo delay information.

5.3.1. Downsampling (DS) Method

Signals *x(n)* and *y(n)* are both downsampled by a factor F_1 before calculating the cross correlation, as shown in Figure 5.3.

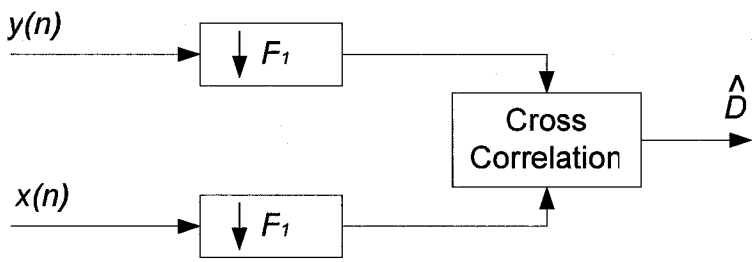


Figure 5.3 Structure of the *DS method.*

To avoid aliasing, both signals are band limited before being downsampled. Since linear phase property is not required here, an Infinite Impulse Response (IIR) bandpass

filter is designed, with central frequency ranges from 1,125 to 1,500 Hz, depending on the different F_1 values. The ranges are selected for the following two reasons. One is that human auditory system does not perceive all audible frequencies with equal sensitivity; signals at about 1 KHz appear to be loudest, as described by the equal-loudness-level contours [121]. The other reason is that the MIRS filtering boosts the higher frequencies as shown in Figure 3.4.

The filter design specifications are given in Table 5.1. The order of such elliptic IIR filters is about 10.

Table 5.1 Design parameters for the IIR bandpass filters

F_1	Fstop1 (Hz)	Fpass1 (Hz)	Fpass2 (Hz)	Fstop2 (Hz)	Astop (dB)	Apass (dB)	Fc (Hz)	BW (Hz)	Fs (Hz)	Order
4	1,000	1,050	1,950	2,000	40	1	1,500	1,000	2,000	12
8	1,000	1,050	1,450	1,500	40	1	1,250	500	1,000	10
16	1,000	1,050	1,200	1,250	40	1	1,125	250	500	8

Fstop1, Fpass1, Fpass2 and Fstop2 are the lower stopband, lower passband, upper passband and upper stopband frequencies respectively. Astop is the stopband attenuation. Apass is the passband ripple. Fc is the passband central frequency, BW is the bandwidth, Fs the sampling frequency and Order is the order of the IIR filter.

The frequencies in Table 5.1 satisfy:

$$Fc + \frac{BW}{2} = k \cdot \frac{Fs}{2} \tag{5.7}$$

where k is an integer [122]. After passing the IIR bandpass filter, every F_1^{th} sample, starting with the first sample, is kept. This method is referred to as the *IIR-DS method.*

A brutal downsampling method, which merely keeps every F_1^{th} sample, starting with the first sample without band limiting the signals, is also proposed. The aliasing happens

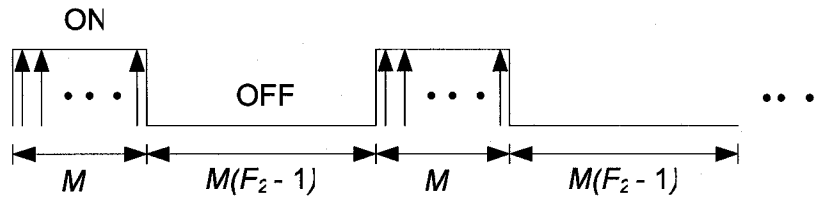
to both signals in this case. Similarly, this method is referred to as the *NONE-DS method.* It further reduces the computational requirements by eliminating the bandpass filtering.

Although not suggested, a FIR filter with the same design parameters as those of the IIR filter is also examined. The FIR filter has the linear phase property, but the order is around 230 by the Parks-McClellan algorithm [122], which requires more computational resources. This method is called the *FIR-DS method.* Note that the *FIR-DS method* is only examined here for performance comparison purposes; it is not suggested in this thesis unless it has significant performance improvement over the *IIR-DS* and *NONE-DS methods.* However, as will be seen later, the analysis and simulation results show it does not.

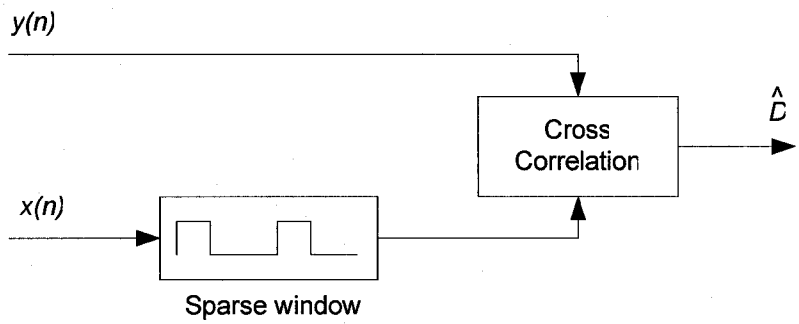
5.3.2. Sparse Window (SW) Method

A window with a certain repeated ON/OFF pattern is designed, as shown in Figure 5.4 (a). We call it the sparse window. The block diagram of this method is shown in Figure 5.4 (b).

In the sparse window, samples within ON portions are kept, while samples within OFF portions are set to zero. The durations for each ON and OFF portion are M and $M(F_2 - 1)$ samples respectively. This ON/OFF pattern repeats, so only one F_2^{th} of the total samples are kept, the rest are set to zero, reducing the information contained. A special case is $M = 1$, which is similar to the *NONE-DS method.*



(a) The sparse window



(b) Block diagram of the *SW method*

Figure 5.4 Structure of the *DS method*

The window is applied to either $x(n)$ or $y(n)$, but not to both. Then, the cross correlation is calculated by using one windowed signal and the other signal in full.

5.3.3. Property Analysis of the Algorithms

To simplify the analysis, we assume that the echo path impulse response $h(n)$ can be represented by a delta function:

$$h(n) = \alpha\delta(n - \hat{D}) = \begin{cases} \alpha, & n = \hat{D} \\ 0, & n \neq \hat{D} \end{cases} \tag{5.8}$$

with $\alpha = 10^{-EPL/20}$. When there is no double talk, equation (5.3) can be simplified to:

$$y(n) = \alpha x(n - \hat{D}) + w(n) \tag{5.9}$$

Further, we assume that $x(n)$ and $w(n)$ are stationary with zero mean, flat spectrum with the same bandwidth, the correlation durations of the signals are very small relative to N,

and $x(n)$ and $w(n)$ are uncorrelated with each other. It can be shown that $\hat{D}$ is an unbiased estimator of the true D, and the mean square error $E[(\hat{D}-D)^2]$ is proportional to [123]:

$$E[(\hat{D}-D)^2] \propto 1/(NB_s^3 ENR) \tag{5.10}$$

where B_s is the signal bandwidth, and ENR is the Echo-to-Noise Ratio.

This thesis is not intended to derive the explicit error bounds under nonlinear distortions. Rather, under the above assumptions, we show the performance and limitations of the proposed algorithms using (5.10), which quantifies the relationships between error variance and several contributing factors. Also, the implications of using speech signals to the error variance are analyzed.

For the *DS method*, its effects on the delay error variance are two-fold. First, when F_1 increases, the error variance increases too, because N used for the calculation is smaller. Second, for the *IIR-DS* and *FIR-DS methods*, B_s would decrease when F_1 increases, resulting in higher variances. Under the flat spectrum assumption, there is no change for ENR between the downsampled signals. However, the spectrum of the speech signal $x(n)$ is not really flat; it often has more energy content around 1,000 Hz. Therefore, the *IIR-DS* and *FIR-DS methods* would achieve a higher ENR ratio, as the bandpass filter is designed for this frequency range. The filtering operation increases the ENR while reducing the signal bandwidth at the same time. On the other hand, the *NONE-DS method* has no such frequency selection advantage, but the whole bandwidth is kept.

For the *SW method*, N and B_s in (5.10) wouldn't change, but many samples of $x(n)$ are set to zero, as shown in Figure 5.4 (b); this can be considered as a special case in which

ENR is reduced. Under the stationary assumption of $x(n)$, the resulting ENR of the *SW method* is given by:

$$ENR_{SW} = 10\log_{10}\frac{\sum_i x_{SW}{}^2(n)}{\sum_i w^2(n)} = 10\log_{10}\frac{\frac{1}{F_2}\sum_i x^2(n)}{\sum_i w^2(n)} \qquad (5.11)$$
$$= ENR - 10\log_{10} F_2$$

with the summation index i from 1 to N. When F_2 increases, the resulting variance would be higher as the ENR decreases. As (5.11) suggests, the window size MF_2 has no effect on variance. However, the speech signal $x(n)$ can only be considered quasi-stationary over segments of 20-40 ms [52]. Therefore, equation (5.11) only holds for very small windows. When the window size tends to be larger, the energy of the ON portion samples may deviate significantly from the assumed $1/F_2$ of the energy of the whole window, because of the speech variations. This uncertainty would increase the variation of the ENR under the *SW method*, so using a smaller sparse window is preferable.

5.3.4. Computation Complexity for Delay Measurement

These two methods provide simple and efficient ways to reduce the samples used in calculations. To compare the computation reduction achieved by our methods, the cross correlation calculated using full $x(n)$ and $y(n)$ by (5.4) is used as the reference. We denote the lengths of $x(n)$ and $y(n)$ as L_X and L_Y respectively, and the maximum searchable delay D_{MAX} = min(L_X , L_Y). If we assume $L_X = L_Y = L$, then $D_{MAX} = L$. In the reference, L multiplications and L–1 additions are required for computing one lag. Thus, to compute the D_{MAX} delays, L^2 multiplications and $L^2 - L$ additions are required.

Under the *NONE-DS method*, both $x(n)$ and $y(n)$ are downsampled by F_1 times. Only L/F_1 multiplications and L/F_1-1 additions are required for computing one lag, and only a total of D_{MAX}/F_1 lags need to be computed for D_{MAX} delays. The total multiplications and additions will be roughly $1/F_1^2$ of those in the reference.

Under the *SW method,* the portions of $x(n)$ set to zero will not be used in computation, but D_{MAX} lags still need to be computed. So, the total multiplications and additions will be roughly $1/F_2$ of those in the reference.

These comparisons are summarized in Table 5.2. The *NONE-DS method* is more efficient than the *SW method*, but its resolution is also reduced correspondingly. As an example, when $F_1 = 8$ and $F_2 = 8$, compared with the reference, only 1.6% and 12.5% computations are required for the respective methods. For both methods, FFT can be implemented as an efficient way to calculate the cross correlation [122].

Table 5.2 Comparison of delay computation requirements

Method	Multiplication	Addition
Reference	L^2	$L^2 - L$
NONE-DS	L^2/F_1^2	$L^2/F_1^2 - L/F_1$
SW	L^2/F_2	$L^2/F_2 - L$

5.3.5. EPL Measurement

EPL measured here is the Speech Echo Path Loss (SEPL) [93], which is the ratio of the root mean square values of the non-cancelled end speech to the echo, when the echo path delay is removed.

The measurement utilizes the delay information obtained from either of the proposed methods. All the measurements are expressed in dB. Before D, only background noise exists; the noise power P_N is measured. After D, the sum of the noise and echo power P_{N+E} is measured. Then, EPL is calculated by (5.12) and (5.13) in turn:

$$P_E = 10\log_{10}(10^{\frac{P_{N+E}}{10}} - 10^{\frac{P_N}{10}}) \tag{5.12}$$

$$EPL = P_X - P_E \tag{5.13}$$

where P_X and P_E are the powers of $x(n)$ and echo, respectively.

5.4. Simulation Setup and Measurement Design

5.4.1. Speeches

The 32 samples from the speech set 2 are used here as both non-cancelled end and cancelled end speech waveforms. When simulating the double talk, $v(n)$ is randomly selected from them, provided that $x(n)$ and $v(n)$ come from different speakers.

5.4.2. Echo Path Elements

A) D: First, ten close round-trip delays, from 302 ms to 320 ms with increments of 2 ms, are used, as explained below. The one-way delay of 150 ms is suggested as a limit for most interactive applications in ITU-T Rec. G.114, so the base delay is selected to be around 300 ms. On the other hand, delay resolution is decreased in the *DS method.* For example, we cannot distinguish a delay of 300.125 ms from another delay of 300.250 ms if F_1 is 4, which implies only a 0.5 ms resolution at the 8,000 Hz sampling rate. The 2 ms

increment is selected to examine if our methods can achieve consistent accuracy at delays with a 2 ms resolution, which is our interest.

Then, two other delays of a different order, 80 ms and 400 ms, are also selected to examine the performance of the proposed algorithms over a wider delay range.

The codecs may introduce additional but fixed delay. This delay is adjusted accordingly when the desired delay is added.

B) Encoder/Decoder: Two popular VoIP codecs, ITU-T Rec. G.711 and G.729A are used, the latter introducing nonlinear distortion. The quantization effect of G.711 is very small and is ignored here.

To take into account the effect of packet loss, different packet loss rates, 0 (no loss), 1, 2, 5 and 10%, are simulated in both paths. The default 20 ms VoIP packet size [4], [48] is used, the loss is only introduced during speeches (no loss during silences), and the ITU-T standard, built-in packet loss concealment algorithm for each codec is used to recover the lost packets.

C) Hybrid: Because the 2-wire loop is balanced by a lumped network, the echo is not just an attenuated, but a filtered version of the input speech [37]. The echo path model 5 in [42], which represents a realistic digital echo path measured from North America, is used to model $h(n)$. It has a single reflection and the length is 12 ms, the dispersion time is about 6 ms. This kind of echo path occurs most often in the measurements [42]. The impulse response and frequency response of $h(n)$ are shown in Figure 5.5.

D) EPL: 20 dB are used in the simulation. In the echo path, the hybrid can be viewed as a bandpass filter, as shown in Figure 5.5 for its frequency response. The overall energy

of the signal will be attenuated when it convolves with the hybrid. The EC, codecs and packet loss contribute to the total energy loss, too. In order to introduce a certain *EPL*, as the exact delay between *x(n)* and *y(n)* is known in the simulation, the energy of the echo is adjusted to the desired *EPL* relative to *x(n)* when this delay is removed.

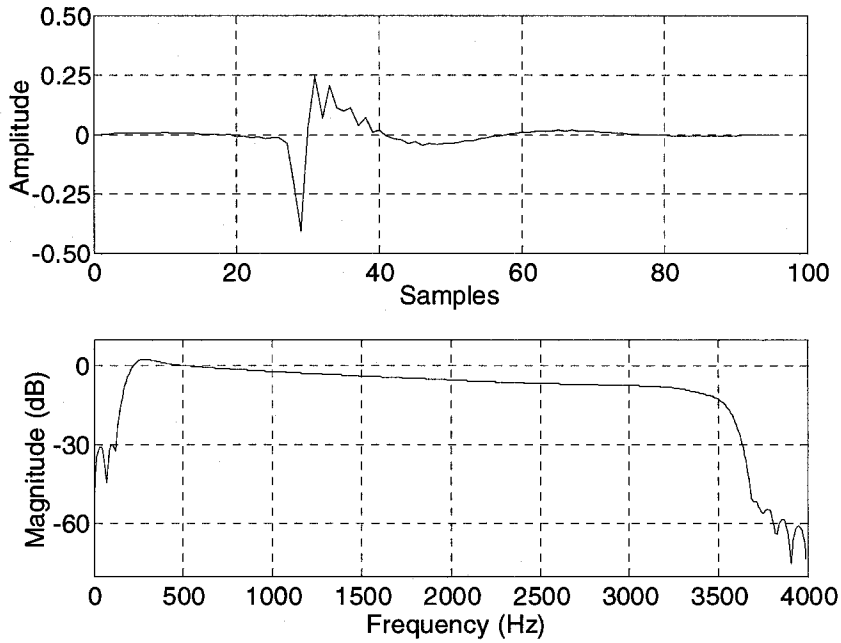


Figure 5.5 The impulse response and frequency response of the hybrid model

As shown in subsection 5.3.3, the performance of the delay measurement is affected by the ENR. Although the real *EPL* values may be different from 20 dB in the simulation, this effect is evaluated through using different ENR.

E) Noise: w(n) is generated by passing white Gaussian noise through a 300-3,400 Hz bandpass filter. ENR is set to: ∞ (no noise), 10, 6 and 0 dB.

5.4.3. Echo Measurement Block Elements

A) VAD: In the TIMIT database, a marker file is provided for each speech. It contains

the locations of the beginning and end of the talkspurt. This information is directly used as the detecting decision of the VAD instead of really implementing one, because it is desired to exclude the effects of VAD misdetection on measurement accuracy.

B) DTD: The Geigel algorithm is used. The initial value of τ is set to 0.5012, corresponding to the assumption of at least 6 dB attenuation. When *EPL* is measured later, τ can be updated with some safe margin. In the simulation, τ is set to 0.1259, corresponding to 18 dB, i.e., with a 2 dB margin. The hangover time is 60 ms.

C) DS method: The factor F_1 is set to 4, 8 and 16 for both the *IIR-DS* and *NONE-DS methods*. Also, these F_1 values are used for the *FIR-DS method* for comparison.

D) SW method: The factor F_2 is set to 4, 8 and 16, and M is set to 32, 64 and 128 samples. The full combinations of F_2 and M are investigated.

For various conditions examined, all 32 speech samples are used as the non-cancelled end speeches for each case. D and *EPL* are measured by the proposed methods in Section 5.3. RMSE is used here as the indicator of algorithm performance. The ideal case, no noise, codec distortion, packet loss and double talk, and using the cross correlation without any computation simplification, is used as the reference condition, where RMSE of the measurement is zero.

5.5. Results

In this section, the simulation results are presented. Also, field measurements are conducted to verify the performance of the proposed methods under real VoIP networks. The measurement setup and results are given as well.

For the simulations, RMSE at different delays, 80 ms, 302 to 320 ms and 400 ms are analyzed. It is shown that RMSEs do not depend on the delay; this is expected. Also, the algorithms can achieve consistent accuracy at 2 ms delay resolution. The performance of the *SW method* with $M = 32$, $F_2 = 8$, ENR = 0 dB is shown in Figure 5.6 as an example. RMSEs for G.729A are always higher than those of G.711, due to the nonlinearity introduced by the former.

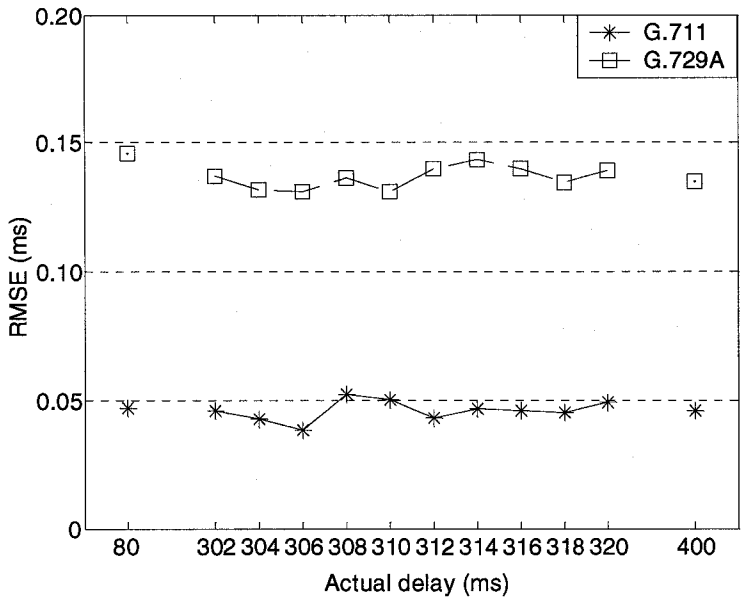


Figure 5.6 RMSE of delay measurement across different delays

In order to reduce the dimensions of the results presented in this thesis, all the results hereafter are the average of RMSEs for delays from 302 to 320 ms, unless otherwise specified.

5.5.1. Performance of the DS Method

As a starting point, the performance of the *DS method* under different F_1 and filtering methods is illustrated in Figure 5.7, for simple cases, by using G.729A, no packet loss, no

double talk. It shows that RMSE increases under bigger F_1 or lower ENR. For the three filtering methods, the *FIR-DS method* has slight improvement over the *IIR-DS method*, suggesting that there is no need to use a FIR filter to maintain the linear phase. Interestingly, the *NONE-DS method* outperforms the other two under $F_1 = 4$ and 8. This could be explained as follows. In the *IIR-DS* and *FIR-DS methods*, the frequency selection improves the ENR, however the signal bandwidth is reduced; perhaps the role of the latter is more dominant here, although not deterministic. When $F_1 = 16$, RMSE is seen to dramatically increase to 2-6 ms (not shown in Figure 5.7 for clearness), which is unacceptable for our purposes.

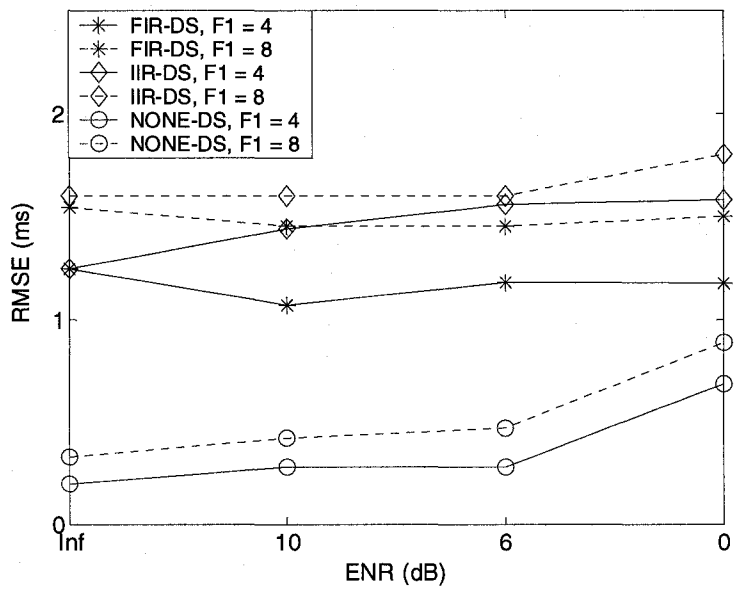


Figure 5.7 RMSE of the three variations of the *DS method* for G.729A

Next, comparison results for codec distortion and double talk are given in Figure 5.8, for the *NONE-DS method* as an example. Under single talk, codec nonlinearity caused by G.729A increases the RMSE several fold relative to G.711; RMSE for G.729A is seen to

be about 0.4 ms except for the extreme low ENR of 0 dB case. However, double talk dramatically increases the RMSEs for both codecs.

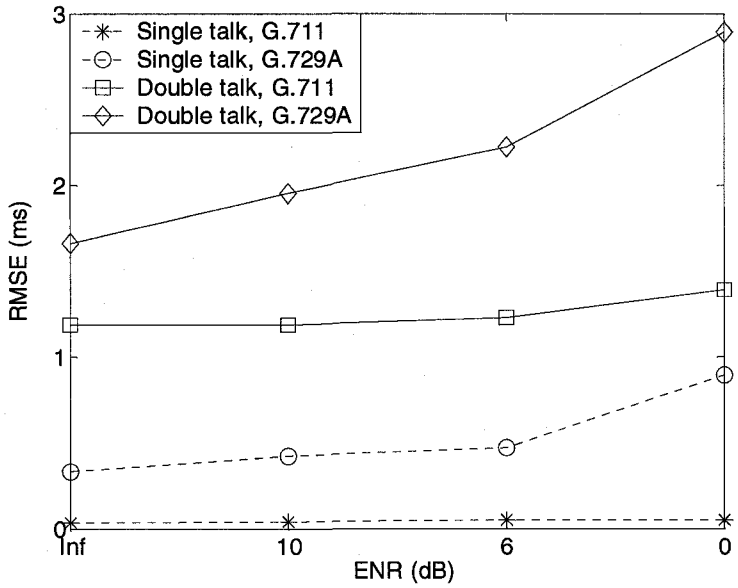


Figure 5.8 RMSE of the *NONE-DS method* under $F_1 = 8$ and double talk

Finally, the impacts of packet loss are summarized in Table 5.3 for the *NONE-DS method* using $F_1 = 8$. RMSE becomes larger when packet loss rate increases.

Table 5.3 RMSE of the *NONE-DS method* with $F_1 = 8$, under packet loss (unit: ms)

Codec	Loss rate	ENR (dB)			
		∞	10	6	0
G.711	1%	0.0370	0.0420	0.0487	0.0526
	2%	0.0442	0.0443	0.0615	0.0692
	5%	0.0806	0.0824	0.0898	0.0938
	10%	0.9548	0.9724	0.9729	0.9769
G.729A	1%	0.3339	0.4338	0.4842	0.8963
	2%	0.3418	0.4681	0.5418	0.9346
	5%	0.4045	0.5013	0.6382	1.1387
	10%	1.3316	1.3963	1.6311	1.7632

In short, double talk dominates the delay measurement performance; background

noise, codec distortion and moderate packet loss (e.g., loss rate < 5%) have limited impacts here. As we are not concerned about reproducing the speech signals, the *NONE-DS method* with $F_1 = 8$ is preferred in both accuracy and computational complexity.

5.5.2. Performance of the SW Method

First of all, the acceptable ranges of M and F_2 are determined. M is selected from 1, 2, 3, … , 200; F_2 is selected from 4, 8 and 16. As a trial, simulations are run on several pairs of *x(n)* and *y(n)* with actual delay of 302 ms for G.711. The results are shown in Figure 5.9, large measurement errors manifest themselves by spikes. The performance of $F_2 = 4$ is quite consistent through different values of M. However, when M becomes larger, big measurement errors begin to occur for $F_2 = 8$ and 16. That is, a smaller M is preferred to make sure the speech within the sparse window is as stationary as possible, as we mentioned in subsection 5.3.3.

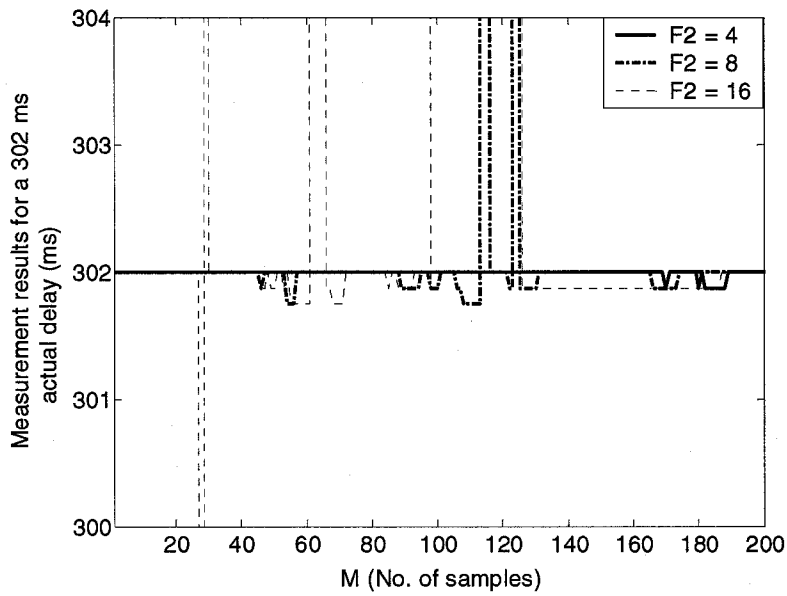
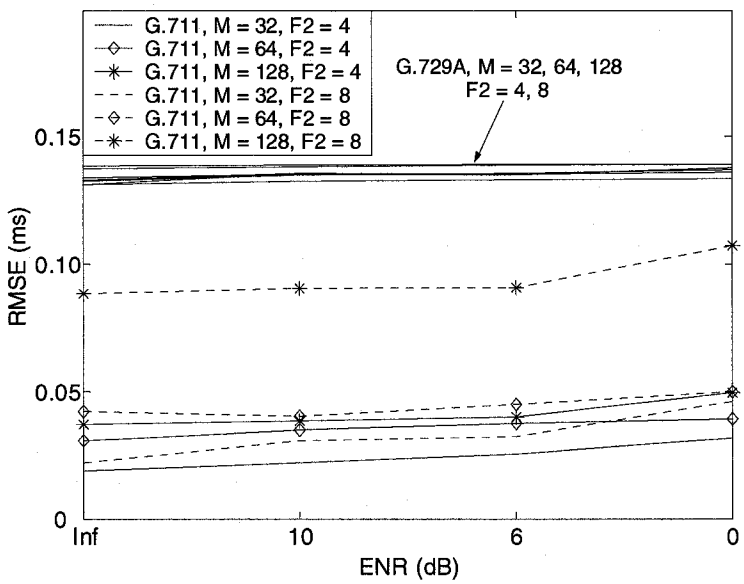
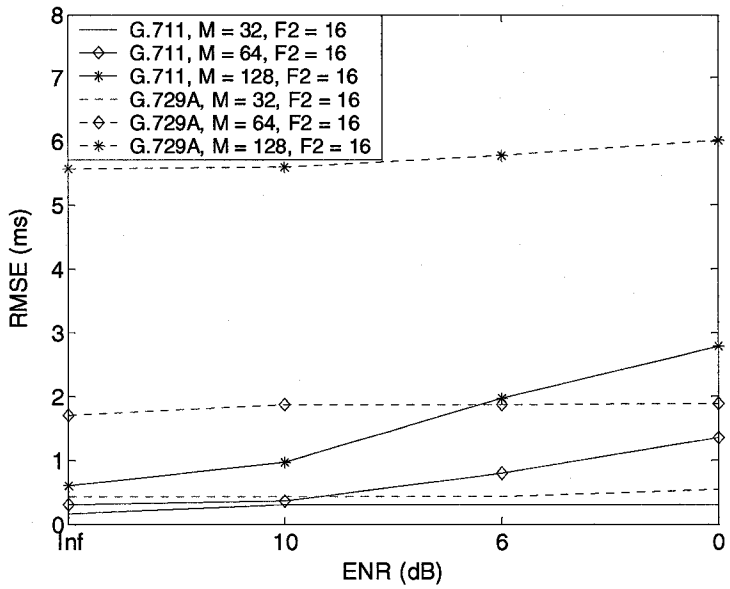


Figure 5.9 The effect of selections of M and F_2 on the performance of the *SW method*

In the simulations, we therefore focus on the performance of the *SW method* under M = 32, 64 and 128, and F_2 = 4, 8 and 16. Similarly, as a starting point, the performance under simple cases, no packet loss and double talk, is illustrated in Figure 5.10 (a) for F_2 = 4 and 8, and in Figure 5.10 (b) for F_2 = 16. In general, RMSE increases when M or F_2 are bigger or ENR is lower. This effect is not obvious for G.729A, under F_2 = 4 and 8, as shown in Figure 5.10 (a), the six curves almost mingle with each other. This may be explained by that the nonlinearity introduced by G.729A dominates the performance here. In addition, under the same M and F_2, all the RMSEs from G.729A are bigger than those of G.711, demonstrating the effects of nonlinearity on the measurement. As seen in Figure 5.10 (b), under F_2 = 16, RMSE rises sharply when M increases. When other impairments such as double talk occur, the performance is expected to be even worse. Therefore, F_2 = 16 is not recommended.



(a) F_2 = 4 and 8



(b) $F_2 = 16$

Figure 5.10 RMSE of the *SW method* under different F_2

Under double talk, the performance for $M = 32$ and $F_2 = 8$ is illustrated in Figure 5.11, as an example, for both codecs. The results show that double talk is still more dominant to measurement than codec distortion.

Finally, the RMSEs under packet loss are summarized in Table 5.4, for the *SW method* with $M = 32$ and $F_2 = 8$. Even under 10% packet loss, RMSEs are about 0.69 and 1.46 ms, under ENR = 0 dB, for G.711 and G.729A, respectively.

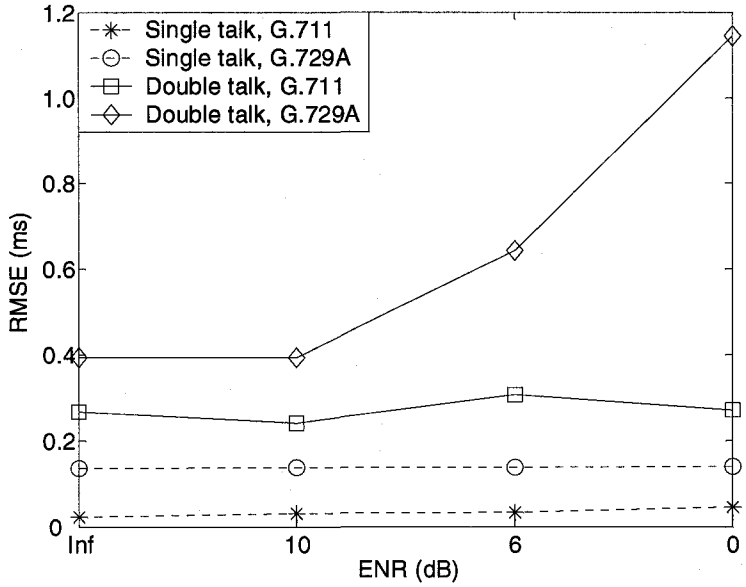


Figure 5.11 RMSE of the *SW method* with $M = 32$ and $F_2 = 8$, under double talk

Table 5.4 RMSE of the *SW method* with $M = 32$, $F_2 = 8$, under packet loss (unit: ms)

Codec	Loss rate	ENR (dB)			
		∞	10	6	0
G.711	1%	0.0235	0.0311	0.0327	0.0476
	2%	0.0383	0.0387	0.0504	0.0494
	5%	0.0494	0.0674	0.0778	0.0813
	10%	0.5519	0.6274	0.6215	0.6889
G.729A	1%	0.1455	0.1488	0.1501	0.1511
	2%	0.2370	0.2480	0.2418	0.2519
	5%	0.2548	0.2783	0.2812	0.2955
	10%	1.4363	1.4365	1.4375	1.4602

In summary, the *SW method* with $M = 32$ and $F_2 = 8$ has the best tradeoff between computation and accuracy.

5.5.3. Performance of EPL Measurement

As the measured delay information is used to determine the beginning of the echo, its accuracy will affect the measurement of *EPL*. The RMSE results for *EPL* measurement

using the *NONE-DS method* with $F_1 = 8$, and the *SW method* with $M = 32$, $F_2 = 8$ are presented in Table 5.5, with unit in dB. For the packet loss scenarios, only the results for 5% loss are listed, for brevity. Under single talk, the two methods have similar performance; codec distortion and packet loss have little effect. This is because, when introducing the echo path loss in simulations, the attenuation from the codec and packet loss has been considered and incorporated into the total 20 dB loss. Under double talk, the performance of both methods degrades, but is still acceptable, as the RMSE is seen to be about 0.36 dB even under high noise. Similar performance is observed for other packet loss rates. Double talk dominates the delay measurement and therefore has the biggest impact on *EPL* measurement accuracy.

Table 5.5 RMSE of *EPL* measurement (unit: dB)

Method	Scenarios	ENR (dB)			
		∞	10	6	0
Single talk:					
NONE-DS	G.711	0.0048	0.0220	0.0531	0.1519
	G.729A	0.0076	0.0236	0.0624	0.1723
	G.711 + 5% loss	0.0162	0.0319	0.0533	0.2159
	G.729A + 5% loss	0.0195	0.0312	0.0912	0.2555
SW	G.711	0.0033	0.0198	0.0347	0.1346
	G.729A	0.0051	0.0223	0.0344	0.1541
	G.711 + 5% loss	0.0107	0.0284	0.0398	0.1640
	G.729A + 5% loss	0.0146	0.0221	0.0511	0.2194
Double Talk:					
NONE-DS	G.711	0.0576	0.0945	0.1034	0.1673
	G.729A	0.0456	0.1077	0.0835	0.2518
	G.711 + 5% loss	0.1044	0.1399	0.1533	0.2943
	G.729A + 5% loss	0.1780	0.1445	0.2219	0.3599
SW	G.711	0.0783	0.1231	0.1195	0.3348
	G.729A	0.1131	0.2086	0.1414	0.2712
	G.711 + 5% loss	0.0788	0.1237	0.1445	0.2832
	G.729A + 5% loss	0.1346	0.1930	0.2311	0.3289

5.5.4. Real Field Measurements

VoIP to PSTN calls are made between several locations in North America with real VoIP deployment. At the IP phone side (non-cancelled end), the speeches on both channels are recorded. Sometimes, the echoes can be clearly perceived. Sometimes, however, they are only transitory or do not exist at all. Also, double talk occurs from time to time during the calls.

Note that the value of $\max\{|\rho_{xy}(m)|\}$ in (5.6) depends on the degree of the correlation between $x(n)$ and $y(n)$. It would be very big (close to 1.0) if there is a perfect echo, or very small (close to 0) when there is no echo. In the latter case, the maximum search of the normalized cross correlation would still return a delay estimate, but this result is erroneous. To avoid this, a minimum threshold ρ_0 is used here, that is, the echo is only assumed to exist when $\max\{|\rho_{xy}(m)|\} > \rho_0$.

For our measurements, ρ_0 is empirically determined to be 0.35 and 0.20, for the *DS method* and the *SW method*, respectively. Using a ρ_0 close to 1 (e.g., 0.80) would lead to most of the echoes being undetected. The results for ρ_0 demonstrate the challenges of echo measurement in real VoIP networks, under which echo suffers from many impairments, and its correlation with the far-end speech decreases significantly.

A total of 29 pairs of far-end speeches and returned echoes are identified from three VoIP calls we made, with 26, 1, 2 pairs from each call respectively. The reference *D* and *EPL* are measured using the full cross correlation without any computational reduction. These results are verified using the off-line utility software Adobe Audition [124], which can visualize the speeches and echoes and measure the two parameters as well.

Then, our proposed methods are applied to these 29 pairs of samples for echo measurement. The performance of the *NONE-DS method* ($F_1 = 8$) and *SW method* ($M = 32$ and $F_2 = 8$) is summarized in Table 5.6, with the VAD and DTD algorithms specified in Section 5.2. For example, for the first call, 26 pairs of echo and non-cancelled end speech are detected. Using the full speeches as the reference, D is found to be 402.125 ms. One *EPL* is 21.34 dB; the rests range from 31.58 to 41.97 dB. Then, the proposed algorithms are applied to measure these two parameters, the RMSE of the measurement are given in columns 5 and 6 for D and *EPL*, respectively.

Table 5.6 Real Field results for the proposed echo measurement methods

Call No.	Reference D and *EPL* (ms; dB)	Pairs of samples	Method	RMSE of D (ms)	RMSE of *EPL* (dB)	Speech quality (R; MOS)
1	(402.125; 21.34)	1	*NONE-DS*	0	0	(21.9; 1.31)
			SW	0	0	
	(402.125; 31.58 ... 41.97)	25	*NONE-DS*	2.1325	0.1193	(46.6 ... 70.0; 2.40... 3.60)
			SW	0.2558	0.0352	
2	(332.125; 38.15)	1	*NONE-DS*	1.1250	0.0030	(67.4; 3.47)
			SW	2.2500	0.0038	
3	(91. 250; 32.33)	1	*NONE-DS*	0.2500	0.0011	(79.7; 4.01)
			SW	0	0	
	(176.500; 33.13)	1	*NONE-DS*	0.5000	0.0045	(68.6; 3.53)
			SW	0	0	

The measured D and *EPL* are used in the E-model [11] to compute the effect of talker echo for our speech quality monitoring purposes, as will be seen in details in the next section. For example, by assuming that all the other inputs to the E-model are set to their default values, the resulting E-model rating R and expected speech quality in MOS are calculated and presented in Table 5.6, as well. It can be seen that, for these VoIP calls,

the real D ranges from 91 to 402 ms, and EPL ranges from 21 to 42 dB. The results confirm that the two methods work well under real networks. For the speech quality, MOS is about 3.50 for most cases; however, echo sometimes could severely degrade the speech quality as MOS is only 1.31 for one case.

5.6. Talker Echo Modeling by the E-model

In the E-model, the effect of talker echo is represented by parameter *Idte*, calculated using (2.4)-(2.6), with other parameters that we are not interested in set to their default values. Two basic parameters, $TELR$ and T, are used in the formulas.

The parameter $TELR$ is the sum of all the losses from the talker's mouth to the talker's ear, and is given by:

$$TELR = SLR + SEPL + RLR \tag{5.14}$$

where SLR is the Send Loudness Rating, and RLR is the Receive Loudness Rating of the talker's handset, and $SEPL$ is the Speech Echo Path Loss we measured. The default value for SLR is 8 dB, and for RLR is 2 dB for North American phone sets [11].

The parameter T is the one-way echo path delay, and is assumed to be half of the D we measured.

$SEPL$ and D are continuously measured by our proposed algorithms, and serve as the inputs to the E-model.

5.7. Discussion and Summary

In this chapter, we develop two methods suitable for measuring the echo parameters in VoIP networks. The measurements are then used as inputs to the E-model to characterize the effects of echo on speech quality.

The proposed algorithms aim to reduce the computational complexity while maintaining good accuracy. The simulations show that, compared with double talk, background noise, codec distortion and moderate packet loss have limited impacts on measurement. On the other hand, double talk, which acts as a high power noise, dominates the measurement performance.

In the two methods, when F_1 is equal to F_2, by comparing Figures 5.7 and 5.10 (a), Figures 5.8 and 5.11, Tables 5.3 and 5.4, it is clear that the RMSEs for the *SW method* are always smaller than those of the *DS method*, under the same conditions. This is because computation reduction is only applied to one of the signals in the *SW method*. Even under the dominant double talk and low ENR of 0 dB, the RMSE for D is about 1.2 ms for the *SW method* with $M = 32$ and $F_2 = 8$; it is about 3 ms for the *NONE-DS method* with $F_1 = 8$. For the majority of cases when echo is noticeably perceived, the ENR value is large and RMSE becomes significantly lower.

The delay measurement here assumes zero delay jitter in a talkspurt in VoIP. For most of the adaptive jitter buffer algorithms today, jitter buffer resizing only occurs during speech silences. Also, the amount of the resizing is usually designed in step sizes of 10-20 ms, in order to accommodate an integer number of voice packets. As shown in the simulations, the RMSE is only about 3 ms for the worst case, which has sufficient

accuracy to detect any delay changes in the order of 10-20 ms.

The computational complexity of the delay measurement depends on the setting of the maximum searchable delay, which should cover the possible delays for most practical systems. Under this maximum delay setting, different real delays require the same computational efforts.

The VAD and DTD are used to facilitate the delay measurement. The performance requirement of the VAD is not very strict. A less aggressive VAD would include some silences in the talkspurts; a more aggressive one would result in some temporal clippings to the talkspurts. The impact is very limited, provided that the speech samples used in the measurement are still long enough compared to the delay; commonly available VAD algorithms can serve this purpose well. However, the role of the DTD is critical; a more robust DTD may be used.

For the multi-path echoes, e.g., two media gateways inter-connected via an IP network, the performance of the algorithms relies on the relative echo path losses and delays for the respective echoes. When two delays are very close, or one of the echo path losses is very large, the two normalized cross correlation functions may overlap each other, and only one echo can be distinguished. With that one exception, the algorithms still work for identifying multi-path echoes.

The performance of the proposed algorithms is also justified through the real field measurements. The measured *D* and *EPL* are effective for our speech quality monitoring purposes, utilizing the ITU-T E-model. Note that for the *NONE-DS method*, the delay resolution is reduced correspondingly. A local maximum search may be performed

around the measured delay with an interval of 0.125 ms, should other applications need a finer resolution.

CHAPTER 6 PACKET LOSS ON SPEECH QUALITY

In this chapter, the detection of packet loss and modeling its effect on speech quality are investigated. A novel strategy is proposed by first classifying the degraded speech frames into three categories: silence, unvoiced and voiced, as they are not equally important to perceptual quality. Then, a non-intrusive algorithm is developed to predict the resulting speech quality by packet loss patterns on the above three categories. To derive such an algorithm, the random packet loss scenario is first investigated as the basis case, then the effects of loss burstiness are modeled on top of the random loss case by introducing a new codec-dependent parameter called the codec burstiness index. The proposed algorithm covers several common VoIP codecs and PLC algorithms.

6.1. The Objective and Related Works

Packet loss is common in VoIP networks. Its effects on speech quality depend on many factors [23], such as loss rate, loss pattern, codec type, packet size, PLC algorithm, etc. Human speech consists of voiced, unvoiced sounds and silences. During a VoIP call, however, not all packets are equally important to the perceptual quality [17], [20], [125], [126], [127]. Therefore, for a given packet loss rate, loss at different locations may result in big fluctuations in speech quality. Generally speaking, voiced packets are more relevant to speech quality than unvoiced packets are. Also, there is no perceived speech quality degradation when loss occurs during silences. Utilizing the loss S/U/V location information will therefore improve the prediction accuracy. Furthermore, packet loss may

be random or bursty, and this loss pattern also affects the speech quality differently. Bursty loss in general leads to longer time clipping of the speech signal and its effect is more annoying than that of random loss.

Our objective here is two-fold: (1) detecting and classifying the lost packets, (2) developing a non-intrusive algorithm to predict the effects of packet loss by further exploring the loss S/U/V location and loss burstiness.

The proposed non-intrusive algorithm consists of a loss pattern extraction module and a quality mapping module. The pattern extraction module is implemented right after the receiver jitter buffer and is introduced in Section 6.3. The quality mapping module produces the MOS estimate based on the packet loss profile obtained; it is introduced in Sections 6.4 and 6.5.

There are many studies on the effects of packet loss on speech quality, either in the MOS domain or the E-model *Ie* domain. Particularly, several speech quality prediction models are proposed [5], [11], [17], [23], [85], [128]. The common approach is first to measure the speech quality under some specific loss conditions and then to develop a model by curve fitting. A logarithmic function was developed to represent the nonlinear relationships between *Ie* (or MOS) and packet loss rate [5], [23], [128]. In the E-model [11], the impairment of packet loss is represented by a factor *Ie*, and was formulated in 2002 instead of using a table form. Also, the effect of burstiness is incorporated by using a single parameter called *BurstR* in the latest 2005 revision E-model. In [85], the packet loss pattern is divided into gaps and bursts; the perceived quality is estimated from the instantaneous quality of two states. The final speech quality is obtained by averaging the

perceived quality along the time scale and incorporating the recency effects. However, these above models do not account for the impacts of loss S/U/V locations on speech quality. In [17], such impacts are first investigated.

6.2. Packet Loss Pattern

The perceived speech quality strongly depends on the packet loss pattern for a given packet loss rate [11], [85]. Packet loss can be random or bursty. The random loss can be characterized by the Bernoulli or independent model where loss probability for each packet is equal and is independent of the loss status of other packets. However, empirical studies show that Internet packet losses are often correlated. The loss is also bursty, that is, the next packet has a higher chance of being lost given the current packet is lost [129], [130], [131], [132]. The packet loss process can be modeled by low order Markov chains. Particularly, the first-order Markov chain, known as the Gilbert model [133], is widely used to capture the bursty nature of packet loss [129].

As shown in Figure 6.1, the Gilbert model has two states, where state X = 0 represents a packet is received and state X = 1 represents the packet is lost. This model can be specified in two ways with the following definitions:

- unconditional loss probability (*ulp*) and conditional loss probability (*clp*), or
- state transition probability p and q.

where *ulp* = Pr(packet lost), *clp* = Pr(X = 1| X = 1), *p* = Pr(X = 1| X = 0) and *q* = Pr(X = 0| X = 1). Note here *ulp* is simply the *Ppl* in the E-model equation (2.11), and is referred to as the packet loss rate in the thesis.

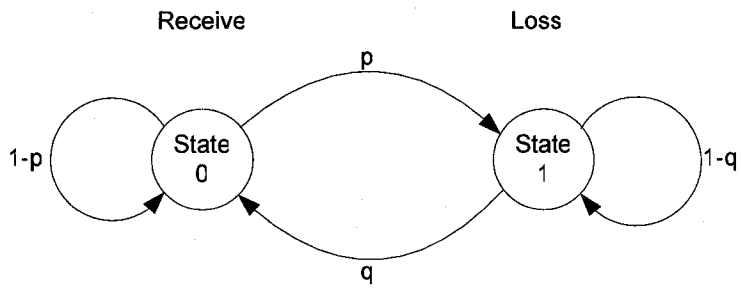


Figure 6.1 The two-state Gilbert model

If we represent the steady-state probability at states 0 and 1 with π_0 and π_1 respectively, it can be shown that:

$$\pi_0 = \frac{q}{p+q} \text{ and } \pi_1 = \frac{p}{p+q}. \tag{6.1}$$

ulp and *clp* can be expressed by *p* and *q* as follows:

$$ulp = \pi_1 = \frac{p}{p+q} \text{ and } clp = 1-q. \tag{6.2}$$

In the E-model, the packet loss burst is characterized by a parameter called *BurstR*, as defined in (2.12). When the loss process is modeled by the Gilbert model, it can be shown that:

$$BurstR = \frac{1}{p+q} = \frac{ulp}{p} = \frac{1-ulp}{q} = \frac{1-ulp}{1-clp} \tag{6.3}$$

Under random loss, *BurstR* is simply 1; when loss is bursty, $ulp < clp$ [129], *BurstR* will be greater than 1.

6.3. Packet Loss Detection and Classification

The proposed packet loss detection and S/U/V classification module is illustrated in Figure 6.2. A packet loss profile is obtained by pooling loss information including codec type, packet size, rate, burstiness, location S/U/V category and the PLC algorithm implemented.

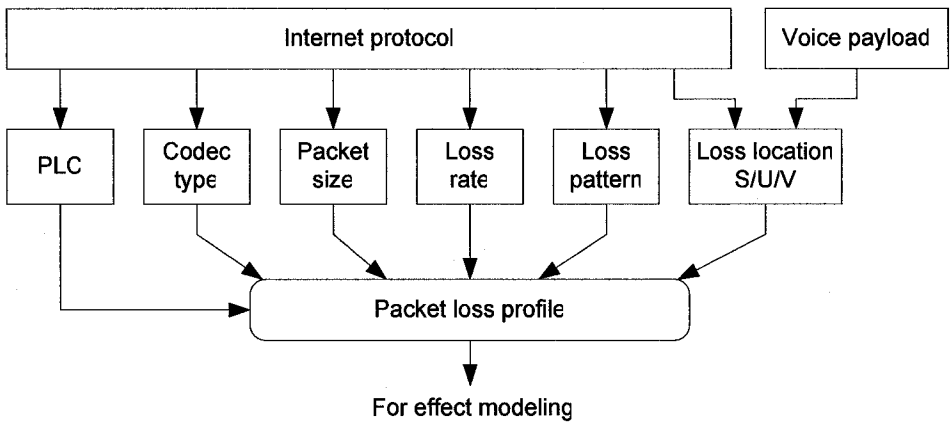


Figure 6.2 Diagram of packet loss detection and S/U/V classification

The packet loss profile is effectively extracted from the RTP header [27] and RTCP-XR [16], avoiding extensive payload analysis. Specifically, codec type is determined from the payload type (PT) field of the RTP header (e.g., 0 for G.711 μ-law, 18 for G.729 [48]); packet size is derived from the RTP payload length (e.g., for a 10 ms frame, G.711 μ-law 80 bytes; G.729 10 bytes, SID 2 bytes). The packet loss event is determined from the discontinuity in the sequence number of the RTP header. When measured after the receiver jitter buffer, the receive/loss status of a VoIP call packet stream may be

represented using a sequence like 010001011100100, where 0 and 1 represent that a packet is received or lost respectively. For such a sequence, the maximum likelihood estimators for p and q in (6.3) are given by [134]:

$$\hat{p} = n_{01} / n_0 \text{ and } \hat{q} = n_{10} / n_1 \tag{6.4}$$

where n_{01} is the number of times 0 is followed by 1 and n_{10} is the number of times 1 is followed by 0, n_0 is the number of 0s and n_1 is the number of 1s in the sequence. The loss rate *ulp* is calculated as $n_1/(n_0+n_1)$.

The PLC algorithm is read from the receiver configuration byte of the VoIP metrics report block of RTCP-XR (e.g., 01 for silence insertion, 10 for built-in concealment and 11 for repeating the last good packet [16]). In case such report is unavailable, analysis of the decoded voice payload has to be performed to determine the PLC algorithm [17]. During a talkspurt, the energy for each packet is calculated. The silence insertion PLC is characterized by a sudden energy drop and is therefore detected by peaks of inverse of energy. To detect the repetition PLC, the cross correlation between two successive packets is calculated. If it is greater than a threshold, then the repetition PLC is assumed; the threshold is selected to be 0.80 empirically. The detection of the PLC algorithm by means of the RTCP-XR report analysis is more efficient and accurate; the two above methods are only used as alternatives, in case that such report is unavailable.

Packet losses can occur during voiced, unvoiced and silence packets (SID packets if VAD is enabled). To determine the type of a lost packet, we propose a scheme by first recovering its features and then feeding the features into a S/U/V classifier. Note our objective here is not to conceal the lost packets, but rather to recover some of their features from the surrounding received packets.

Analyzing the concealed packets is not a good approach here, because common PLCs, such as the built-in and repetition algorithms, only use the information from the previously received good packets, therefore they do not work faithfully during the transitional packets. For example, if the one right before the loss is an unvoiced packet, then a lost, but voiced, packet is highly likely to be reconstructed as an unvoiced packet. To solve this problem, we instead estimate those features from adjacent packets, two immediately before and two immediately after the loss, through interpolation. A method that is shown to be quite effective in speech recognition, nonlinear interpolation using cubic Hermite polynomials [135], is adopted here. For a burst of length β, the feature vector of the n^{th} packet within the burst is estimated by [135]:

$$\hat{x}_{b+n} = x_b(1-3t^2+2t^3) + x_{b+\beta+1}(3t^2-2t^3) + x'_b(t-2t^2+t^3) + x'_{b+\beta+1}(t^3-t^2) \qquad (6.5)$$

where $1 \le n \le \beta$, x_b and $x_{b+\beta+1}$ are the feature vectors from the packets immediately before and after the loss, respectively, $t = n/(\beta+1)$, $x'_b = \beta(x_b - x_{b-1})$ and $x'_{b+\beta+1} = \beta(x_{b+\beta+2} - x_{b+\beta+1})$.

The estimated feature vector of a lost packet, which consists of the frame energy and 8-order linear prediction coefficients, is then sent into the S/U/V classifier proposed by Rabiner [118] for location type determination.

6.4. Setup Design

This section describes system design for our proposed algorithm, including the simulation and measurement setups, approaches, reference speech samples, as well as testing conditions.

The overall simulation and measurement block diagram for packet loss effect modeling is depicted in Figure 6.3. At the sender side, the reference speech samples are first encoded and then passed through a packet loss simulator, where a specific loss pattern is introduced. A S/U/V classifier is used to categorize the packets into silence, unvoiced and voiced, and this information is provided to the loss simulator to determine the appropriate loss locations. At the receiver side, a PLC is implemented and then the reconstructed VoIP packet streams are decoded back to the original waveform.

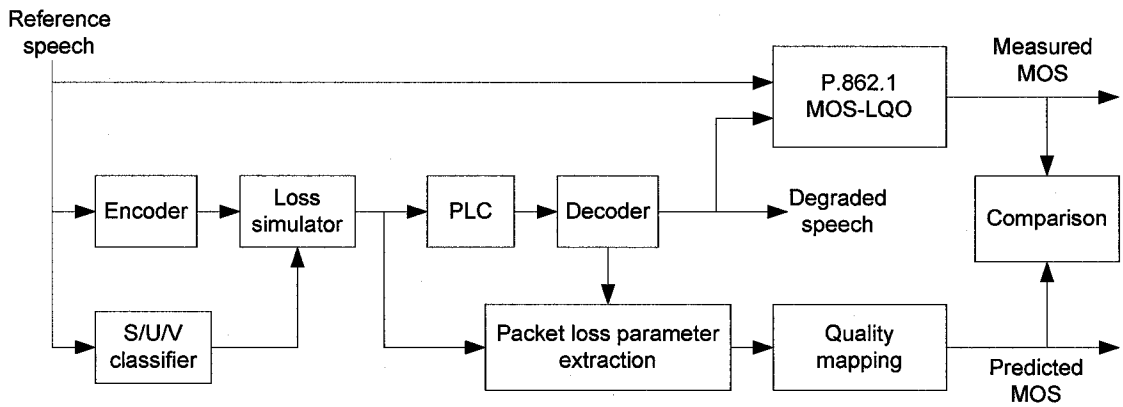


Figure 6.3 Packet loss effect modeling simulation diagram

In this thesis, a non-intrusive model is developed to predict the speech quality based on the extracted packet loss parameters, including the loss rate, loss location category, loss pattern, codec type, packet size, and PLC algorithm. Meanwhile, the speech quality

in P.862.1 MOS-LQO is also intrusively measured by the DSLA. The measured P.862.1 MOS-LQO score and our predicted MOS score are compared to evaluate the performance of the proposed model.

Common VoIP codecs such as G.711, G.729 and G.723.1 usually include a VAD module. When a speech frame is classified as active, it is encoded and transmitted as usual. When the speech frame is classified as non-active, it would be encoded with fewer bits and transmitted periodically or only intermittently. The implementation of a VAD algorithm is application specific. If VAD is disabled, there is no differentiation between silence frames and voiced/unvoiced frames. Packet loss could occur at the silence frames as well, which however has little effect on perceived speech quality in this case. This thesis only investigates the effects of packet loss on the voiced or/and unvoiced frames. The S/U/V classifier is used to serve the purpose of frame type classification.

To investigate the effects of packet loss location and loss burstiness on speech quality, the following steps are adopted:

- Mark the reference speech into silence, unvoiced and voiced. The loss will only be introduced during unvoiced and voiced periods.
- Introduce random loss only during unvoiced periods, then only during voiced periods. A model is developed for these two extreme cases. Although the Internet packet loss is bursty, random loss is first simulated to build a base model for speech quality.

- Introduce random loss, but gradually change the loss weights of unvoiced and voiced periods. By using the model for the two extreme cases as anchors, a general model taking into account the loss weight is developed.
- Introduce bursty loss. The prediction model is finally developed on top of the general random loss model.

Two common VoIP codecs, ITU-T Rec. G.729A and G.711(μ-law) are selected in the simulation. The former belongs to the code excited linear prediction family; the latter belongs to the PCM family of waveform coding. Speeches in set 1 are used to develop the prediction model, the testing conditions of which are summarized in Table 6.1. Common VoIP packet sizes range from 10 to 40 ms, and for our simulations we select the default packet size of 20 ms [4], [48].

Table 6.1 Testing conditions

Parameter	Value
Codec	G.729A, G.711
ulp	0, 1, 2, 3, 5, 7.5, 10, 12.5, 15
PLC	Built-in, repetition, silence
Packet size	20 ms
Runs	10 times, independently
Random loss:	
- Loss weighting (*Wu/Wv*)	0/100, 10/90, 20/80 ... 90/10, 100/0
Burst loss:	
- *BurstR*	1.25, 1.50, 1.75, 2.00

In Table 6.1, *Wu* and *Wv* represent the weight percentages for loss occurring at the unvoiced and voiced packets respectively. These apply when simulating the random loss. The computation of packet loss rate and weights are illustrated as follows: suppose a speech contains 200 packets, 15 packets are lost. Among them, 3, 4 and 8 are silence, unvoiced and voiced packets respectively. Then *ulp* is (4+8)/200 = 6%, and *Wu* is 4/(4+8) = 33.33% and *Wv* is 66.67%.

In order to determine the proper ranges for *BurstR*, several empirical measurement results for *ulp* and *clp* are analyzed, as shown in Table 6.2. It can be seen that under moderate packet loss, the reasonable range of *BurstR* is between 1 and 2, although (6.3) suggests that it could be very large if *clp* approaches 1. As a result, we select *BurstR* to be 1.25, 1.50, 1.75 and 2.00.

Table 6.2 Evidence of *ulp* and *clp* ranges from network measurements

Source	*ulp*	*clp*
Bolot [129]	0.10-0.23	0.18-0.60
Borella [130]	0.20	0.32-0.35
Yajnik [131]	0.0005-0.0735	0.0208-0.2085
Paxson [132]	0.028-0.053	0.20-0.50

When simulating the bursty loss, for given *BurstR* and *ulp*, transition probabilities p and q are derived from (6.3) and are used to generate the burst loss pattern. Here we do not have control over loss location weights in the bursty loss scenario.

For the S/U/V classification, the algorithm in [118] is used in the simulation. The S/U/V decision is made by proximity of energy and LPC distances between the input

frame and the class template. The decision error rate is found to be 5-10%, then we manually correct these errors based on the transcription files.

The simulation is run independently 10 times for each condition; the measurement results are then averaged to develop the prediction model.

6.5. Effects Modeling

The effect of packet loss on speech quality is directly modeled in the MOS domain. Note that it can be intermediately modeled in the *Ie* domain using (2.11) or the logarithmical formula [23], and then transformed back into the MOS domain, yet such a transformation is highly nonlinear and the model expression would be significantly more complicated. Both random and bursty cases are provided next.

6.5.1. Random Packet Loss

First we present the speech quality under random packet loss, as shown in Figures 6.4 and 6.5 for G.711 and G.729A using repetition PLC, respectively. The figures only include results for a few *Wu/Wv* pairs for clarity. It is obvious that MOS drops at a decreasing rate when packet loss rate increases, and for the same loss rate, speech quality gets worse when more voiced packets are lost. For example, under 5% random loss, the MOS-LQO for G.711 when $Wv = 0\%$ and $Wv = 100\%$ are 3.58 and 3.24 respectively, with a difference as large as 0.34 MOS. This suggests that loss location type has an impact on speech quality. In addition, adjacent curves are approximately of equal distance. The same trends are also observed for G.711 or G.729A under different PLCs.

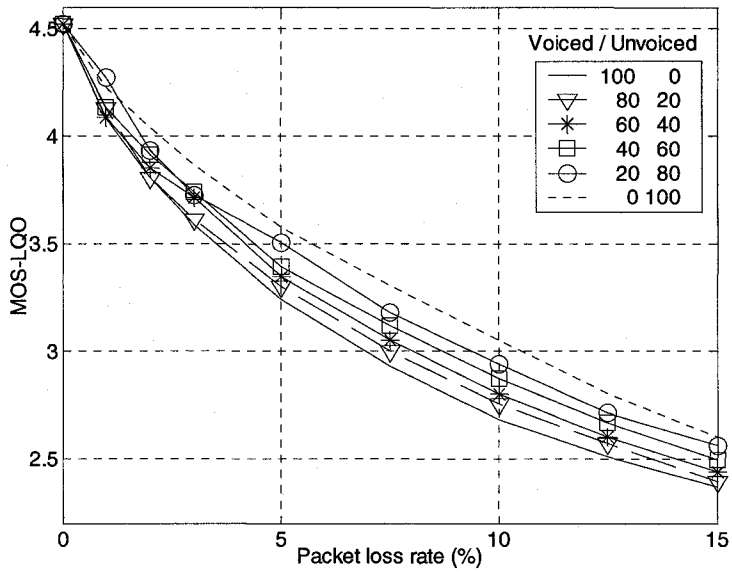


Figure 6.4 Speech quality for G.711 using repetition concealment

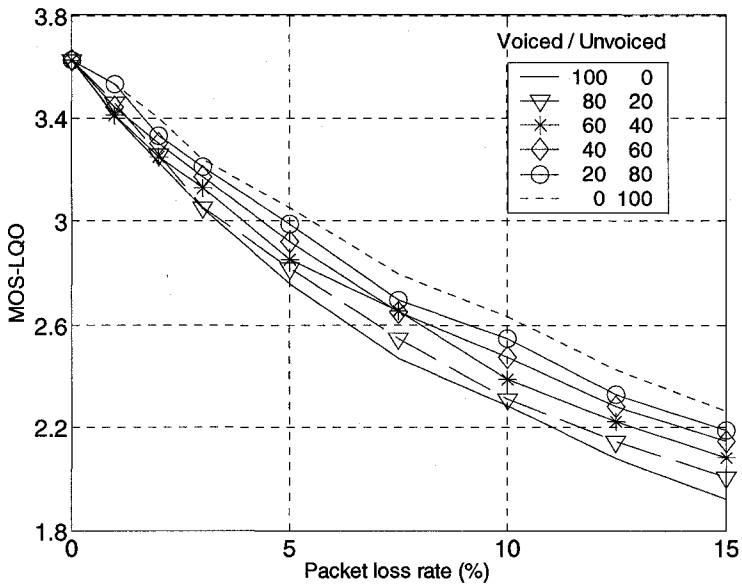


Figure 6.5 Speech quality for G.729A using repetition concealment

The two extreme loss cases, $Wv = 0\%$ and $Wv = 100\%$, provide the upper and lower boundaries for the curves whose weights lie between 0% and 100%. Based on the above

observations, a third-order polynomial is chosen to model the relationship between packet loss rate and speech quality, but only for the two extreme loss cases. Then, the curves with *Wv* between 0% and 100% are modeled by using linear interpolation between two extreme cases. Although using higher order polynomials can reduce fitting error, it also reduces the degree of freedom for the measured data and may cause overfitting.

The detailed prediction model is given by:

$$MOS = MOS_O - DMOS_{PL} \tag{6.6}$$

where MOS_O = 4.55, the same as in (4.1); $DMOS_{PL}$ is the MOS drop caused by packet loss including codec distortion. $DMOS_{PL}$ is determined using (6.7) and (6.8). The linear interpolation scheme is used:

$$DMOS_{PL} = W_u \cdot DMOS_{PLu} + W_v \cdot DMOS_{PLv} \tag{6.7}$$

where $DMOS_{PLu}$ and $DMOS_{PLv}$ are the drops in MOS if the lost packets are all unvoiced or all voiced, with loss weights *Wu* and *Wv*, respectively, and *Wu*+*Wv* = 1. $DMOS_{PLu}$ and $DMOS_{PLv}$ are determined by the third-order polynomial curve fitting:

$$DMOS_i = C_0 + C_{1i} \cdot ulp + C_{2i} \cdot ulp^2 + C_{3i} \cdot ulp^3 \tag{6.8}$$

where subscript *i* can be *PLu* or *PLv*, and *ulp* is packet loss rate expressed in percentage. When there is no packet loss, $DMOS_{PL}$ simply equals C_0 in (6.8), which represents the degradation purely caused by a codec alone. The coefficients of the curve fitting are given in Table 6.3.

Table 6.3 Coefficients for packet loss model

Codec	PLC		Voiced			Unvoiced		
		C_0	C_1	C_2	C_3	C_1	C_2	C_3
G.711	Built-in	0.0277	0.2992	-0.0201	0.00061	0.2657	-0.0160	0.00046
	Repetition	0.0277	0.3927	-0.0302	0.00091	0.2635	-0.0171	0.00054
	Silence	0.0277	0.4857	-0.0330	0.00091	0.2712	-0.0067	2.83E-05
G.729A	Built-in	0.9237	0.1970	-0.0095	0.00023	0.1441	-0.0049	8.71E-05
	Repetition	0.9237	0.2252	-0.0120	0.00031	0.1301	-0.0033	4.54E-05
	Silence	0.9237	0.5921	-0.0543	0.0017	0.3175	-0.0208	0.00054

6.5.2. Bursty Packet Loss

The speech quality gets worse when packet loss occurs in bursts rather than randomly; this effect is more pronounced when loss rate is high (e.g., *ulp* > 5). Figure 6.6 illustrates such effects for G.711, using the built-in PLC as an example.

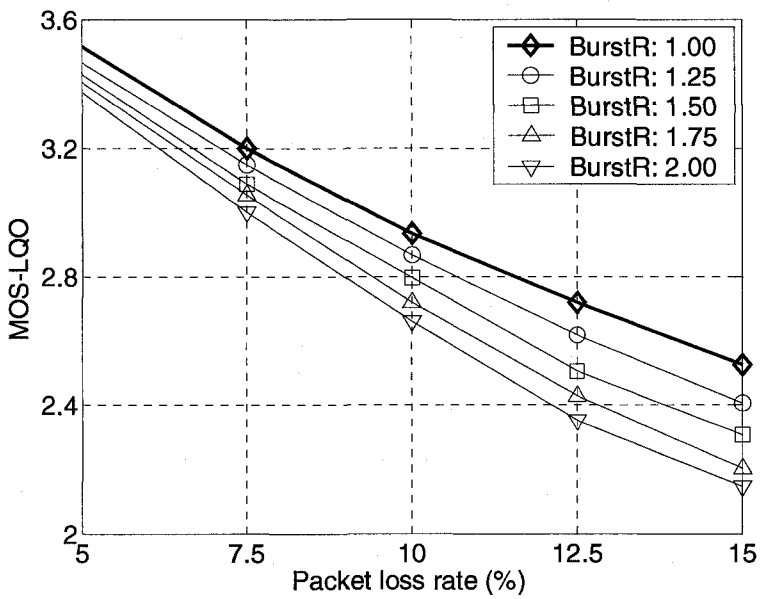


Figure 6.6 Speech quality for G.711 under bursty loss, with the built-in PLC

For a given loss rate, when *BurstR* increases from 1 to 2, the distance between adjacent curves is found to be closer by numeric calculation. It suggests that the speech

quality drops at a decreasing rate with the increasing *BurstR*, when other conditions are unchanged. The same trends are also observed for G.729A or G.711 under different PLCs.

To incorporate the effects of burstiness on top of the proposed random loss model, a new parameter *a* called the codec burstiness index is introduced. The *ulp* in (6.8) is subsequently updated to:

$$ulp_{eqv} = ulp \cdot BurstR^{a} \tag{6.9}$$

where ulp_{eqv} represents the equivalent packet loss percentage under bursty loss as if it were lost randomly. The exponent *a* depends on the codec and PLC algorithm and lies between 0 and 1. Its value is empirically determined through nonlinear optimization using Matlab Optimization Toolbox, as given in Table 6.4.

Table 6.4 Codec burstiness index for the bursty packet loss model

	Built-in	Repetition	Silence
G.711	0.3099	0.0729	0.2904
G.729A	0.1115	0.1155	0.1155

The proposed codec bursty index *a* reflects the tendency to speech quality change for a codec under bursty loss. Note that under random loss, *BurstR* = 1, ulp_{eqv} is identical to *ulp*. When loss is bursty, the equivalent *ulp* is magnified by a factor $BurstR^{a}$. For example, with *BurstR* = 1.75 and G.711 using the built-in PLC, the effect for 10% packet loss on speech quality would be equivalent to that of 11.89% random loss; it would contribute an extra drop of 0.15 MOS, as the model suggested.

6.6. Performance Evaluation

The set 2 speech samples are used to evaluate the performance of the proposed algorithm. To generate the degraded speeches, the same nine packet loss rates, two codecs and three PLC methods are used as before. *BurstR* is selected to be 1.00, 1.25, 1.50, 1.75 and 2.00, to cover the random loss and different degrees of loss burstiness.

For each degraded speech file, *Wv* and *Wu* are calculated from the loss pattern generated by the loss simulator. Together with other parameters, including *ulp*, *BurstR*, codec type and PLC method, *Wv* and *Wu* are fed into the proposed model to compute a MOS value non-intrusively. Also, the speech quality is intrusively measured by the DSLA and compared to the predicted one through statistical analysis.

The Pearson correlation coefficient ρ between the measured and predicted MOS, and the root mean square error σ are calculated. For G.711, ρ is 0.91 and σ is 0.26 MOS. The results for G.729A are slightly worse, because the linear predictive coding further introduces variability to speech quality. In particular, ρ is 0.88 and σ is 0.28 MOS. The distribution of absolute prediction errors is examined by binning them with a width of 0.10 MOS, and the percentages in each bin are given in Figure 6.7. The results show that, on average, about 27%, 52% and 72% of the prediction is within 0.10, 0.20 and 0.30 MOS of the measurement, respectively. Less than 7.3%, 2.2% and 0.2% of the absolute prediction error is greater than 0.50, 0.75 and 1.00 MOS, respectively.

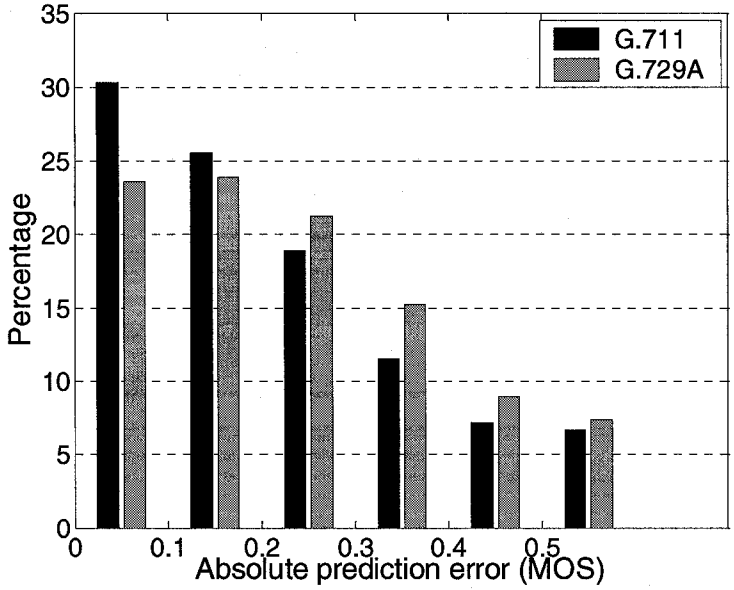


Figure 6.7 Distributions of absolute prediction error

6.7. Discussion and Summary

In this chapter, an algorithm is developed to model the effects of packet loss S/U/V locations and loss burstiness on VoIP speech quality. Also, a scheme is proposed to classify the type of the lost packet to work with the effect modeling algorithm.

In the proposed algorithm, the base speech quality is modeled as a third-order polynomial function of the packet loss rate (*ulp*). Then, a linear interpolation approach is used to represent the effects of loss location weights of voiced and unvoiced packets. In order to accommodate the bursty loss, *ulp* is updated by multiplying a codec dependent factor. The algorithm demonstrates good accuracy through the validation tests.

As seen in Table 6.3, the coefficient C_0 represents the average speech quality one can expect without any packet loss. In general, C_1 dominates the speech quality under different PLCs for moderate packet loss. Here, we mainly discuss the performance

differences between the built-in and repetition PLCs, as both of them are far superior to the silence PLC with regard to speech quality.

G.711 is sample based and memoryless. The built-in PLC uses the last 390 good samples stored in the history buffer to calculate the pitch period, and to synthesize the lost packet. Also, the overlap add method is used to smooth the bad/good packet transition [137]. The repetition PLC simply repeats the last good packet (160 samples at 20 ms packet size) with no transition smoothing. As shown in Table 6.3, the former PLC inherently outperforms the latter. When lost packets are totally unvoiced, the two PLCs almost have the same performance, as C_{1u} is about the same. However, when more and more voiced packets are lost, the built-in PLC obviously performs better, as the model suggests MOS won't change much with different packet loss location weights. For the repetition PLC, the impact of loss location is significant. Voice dominated loss results in worse speech quality, compared to unvoice dominated loss. For the silence PLC, its performance is also location sensitive.

G.729A is based on linear prediction coding. In the built-in PLC, the decoder interpolates parameters of a lost frame from those of the last good frame. Particularly, the linear prediction coefficients (in the form of line spectral pairs) are repeated, and the adaptive and fixed codebook gains are also taken from the last good frame, but are attenuated to gradually reduce their impact [138]. This built-in PLC also outperforms the simple repetition PLC. Compared to G.711, similar results are suggested for G.729A by the model.

Human speech usually contains more voiced sounds than unvoiced sounds, because the former are produced by periodic air flow exciting the vocal tract and last longer than the latter. For example, for the 20 speech samples used in set 1, the weight pair *Wu/Wv* is 33.2%/66.8%. As a result, the loss location weights over the long run would tend to fluctuate around a fixed value determined by the voiced/unvoiced speech structure. Nevertheless, the full range of weights (i.e., *Wv* from 0% to 100%) is still reasonable for the following reasons. First, for low packet loss rates, the loss would occur sporadically. In this case, it is possible that loss of voiced or unvoiced packets would dominate, perhaps even only one kind of packet. Next, bursty loss results in consecutive losses and greatly affects the loss weights. Last but not least, if a packet priority marking scheme is used in VoIP [139], [140], some perceptually more important packets are prioritized for transmission and are less likely to be lost. Many of these prioritized packets are voiced; these schemes therefore affect the loss weights as well.

CHAPTER 7 OVERALL ASSESSMENT MODEL

In this chapter, an overall assessment model is developed. The methodologies and steps for combining various impairments, including packet loss, temporal clipping, background noise and echo are described. The additivity assumption of the E-model on the noise effect is also examined. A listening-only model and a conversational assessment model are developed in this chapter. The performance of the listening-only model is presented. In addition, the relation of the model to the most recent ITU-T standardization activities is discussed.

7.1. Combining Effects Strategy

The individual models proposed in Chapters 4, 5 and 6 quantify the speech quality degradation contributed by temporal clipping, echo and packet loss respectively. The overall assessment model provides a global picture of speech quality that a subject would expect. The overall listening-only model and conversational model are developed in this thesis.

Besides the impairments discussed in the previous chapters, noise has to be considered in the overall model. In fact, noise is not a new impairment to IP networks; its effect on speech quality has been studied over the years, for traditional telephone networks. Several models have been developed to quantify its impact, such as in the E-model [11], CCI model [70], and ITU-T Rec. P.563 [12]. Furthermore, for the echo, its effect cannot be measured using objective methods, such as P.862.1 MOS-LQO. It is not

our objective to model the effects of echo and noise here; existing models, such as those in the E-model, are examined and adopted in the thesis.

Rather, we are interested in those new impairments to VoIP networks, such as packet loss and temporal clipping. The work here focuses on modeling the combined effects of packet loss and temporal clipping. Then, the E-model is used to quantify the effects of noise and echo, and further to develop the overall assessment model. The additivity assumption in the psychological scale for the noise is also examined in this chapter.

The block diagram for developing the overall assessment model is illustrated in Figure 7.1, where the noise module will be explained in Section 7.3. The packet loss and temporal clipping modules are first combined. Then, with the noise module, the non-intrusive listening-only model is developed, which yields a MOS-LQO. The echo module is added to reflect the conversational quality; the final output is a MOS-CQE (Mean Opinion Score – Conversational Quality Estimated) and an E-model rating *R*. Note that in addition to the echo, there are other factors, such as delay and double talk, which affect the conversational quality. This thesis is only concerned with echo.

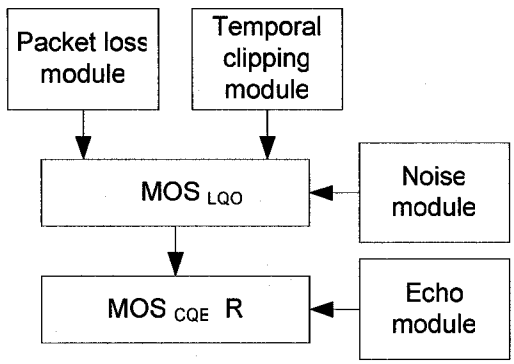


Figure 7.1 The block diagram for integrations of individual modules

7.2. Combining Packet Loss and Temporal Clipping

To investigate the combined effects of packet loss and temporal clipping, they are jointly introduced. The following testing conditions are used, as summarized in Table 7.1. The temporal clipping is introduced at 5 different levels, that is, 6, 12, ... , 30 dB below the active speech level. To introduce the effect of loss burstiness, the Gilbert model is used as before.

Table 7.1 Testing conditions for combined effects

Parameter	Value
Codec	G.729A, G.711
ulp	0, 1, 2, 3, 5, 7.5, 10, 12.5, 15
PLC	Built-in, repetition, silence
Packet size	20 ms
BurstR	1.00, 1.25, 1.50, 1.75, 2.00
Clipping threshold	6, 12, ... , 30 dB below active speech level
Runs	5 times, independently

The loss pattern is analyzed to calculate *Wu* and *Wv*, for quantifying the effects of packet loss using the findings in Chapter 6. The clipping statistics are collected to quantify the effect of temporal clipping, using the formula (4.2) in Chapter 4. Meanwhile, the resulting speech quality that suffers from both impairments is measured. The combined degradation of packet loss and temporal clipping $DMOS_C$ is computed by subtracting the measured quality from the optimum value 4.55. The relationship between $DMOS_C$ and the two individual contributors, $DMOS_{PL}$ and $DMOS_{TC}$, needs to be determined, as the strategies illustrated below.

$DMOS_{PL}$ and $DMOS_{TC}$ are integrated in the MOS domain. One simple way is to sum them up. The rationale behind the summation is that each variable is dimensionless, and if they are independent to speech quality, their total contributions could be added in a certain domain, e.g., the MOS domain. The errors of this assumption are examined, by adding $DMOS_{PL}$ and $DMOS_{TC}$, and then subtracting $DMOS_C$. If the additivity holds, the errors would have a zero mean. Figure 7.2 illustrates the error distributions, under G.711 using the built-in PLC. For brevity, only a few combined cases are given.

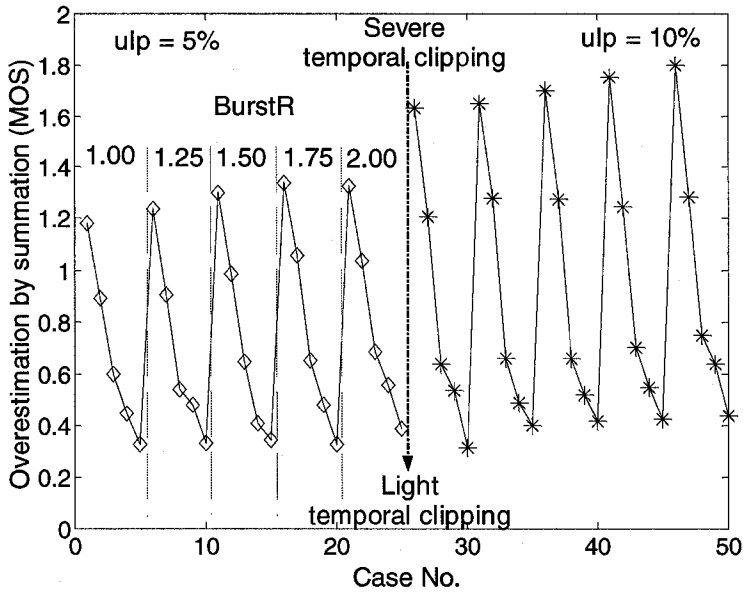


Figure 7.2 The overestimation of combined effects by simple summation

In the figure, the left 25 and right 25 cases are for 5% and 10% packet loss, respectively. Each two of the 25 cases consists of 5 different *BurstR* scenarios, ranging from 1.00 to 2.00 with an increment of 0.25, from left to right. The effects of temporal clipping are reflected vertically, for the severe clipping to light clipping, from top to

bottom, with the corresponding clipping threshold from 6 to 30 dB below the active speech level, with an interval of 6 dB.

It is clear that simple addition would overestimate the combined effects. This overestimation tends to be larger when packet loss is burstier, loss rate is higher or temporal clipping is more severe. That is, the overestimation is positively correlated with $DMOS_{PL}$ and $DMOS_{TC}$. Similar trends are also observed across different codecs, concealments, loss rates, burstiness and temporal clippings. The reason is that, the effects of packet loss and temporal clipping are not independent. The latter resembles the former in a specific way where "losses" mainly occur during packets with low energy contents. Therefore, some measures for their correlation could be used to correct this bias. The model is empirically developed by considering the interactions between the two with a correlation factor γ:

$$DMOS_C = DMOS_{PL} + DMOS_{TC} + \gamma DMOS_{PL} DMOS_{TC} \tag{7.1}$$

The heuristics is justified by the observations in Figure 7.2. It will be seen in the following performance analysis subsection that formula (7.1) works very well. The value of γ is determined from nonlinear optimization (using Matlab Optimization Toolbox) and summarized in Table 7.2, which ranges from -0.30 to -0.40.

Table 7.2 Correlation factor γ for packet loss and temporal clipping

	PLC		
	Built-in	Repetition	Silence
G.711	-0.3743	-0.3838	-0.3312
G.729A	-0.3511	-0.3643	-0.3172

Normally, it is expected that the combined effects are more annoying than any of the

two individual effects. Although possible, it very rarely happens that $DMOS_C$ calculated from (7.1) is smaller than the bigger of $DMOS_{PL}$ and $DMOS_{TC}$. As a precautionary measure, the final $DMOS_C$ is set to: $DMOS_C = \max(DMOS_C, DMOS_{TC}, DMOS_{PL})$.

7.3. Noise Detection and Effects Modeling

Noise is a common impairment to speech signals, which is normally introduced by background of microphones or communication channels. Noise reduces the speech clarity, and one of the ways to model its effects on speech quality is based on the E-model, which requires total noise power as its input parameter [11]. SNR needs to be calculated in this case; one approach is to estimate the original speech and noise respectively, from the degraded speech only.

There are many speech quality enhancement algorithms that can be used to estimate the original speech spectrum, such as short-time spectral estimation based method [141], high order statistics based method [142], etc. The noise spectrum can be measured during speech silences, with the help of a VAD. However, for low SNR and non-stationary noise environments, this method can easily fail. More accurate algorithms, such as the Minima Controlled Recursive Averaging (MCRA) [143], can be used for such cases. MCRA uses a subband spectral tracking method to estimate the noise spectrum recursively while skipping the speech spectrum. It utilizes not only speech silence periods, but also talkspurt periods to estimate the non-stationary noise spectrum. One problem of MCRA is that it is easy to overestimate noise when SNR is very high.

The proposed noise detection and effect modeling algorithms are presented below.

Note that the noise detection part is developed by Zhong Lin [18] and is adopted in the thesis. As an integral part of the whole non-intrusive assessment model, the algorithm and its performance are summarized in subsection 7.3.1.

7.3.1. Summary of Noise Detection Algorithm

The block diagram of the noise module is shown in Figure 7.3. The noise effect modeling is based on the E-model, so the noise power has to be determined. In our approach, both noise and speech spectrums are estimated from the degraded speech to calculate the SNR, to facilitate accurate noise power estimation. In order to deal with both high SNR and low SNR situations, a two-step method is considered, rough SNR estimation followed by SNR adjustment. Finally, the noise power is determined from the active speech level and adjusted SNR, and it is used in the E-model for effect modeling.

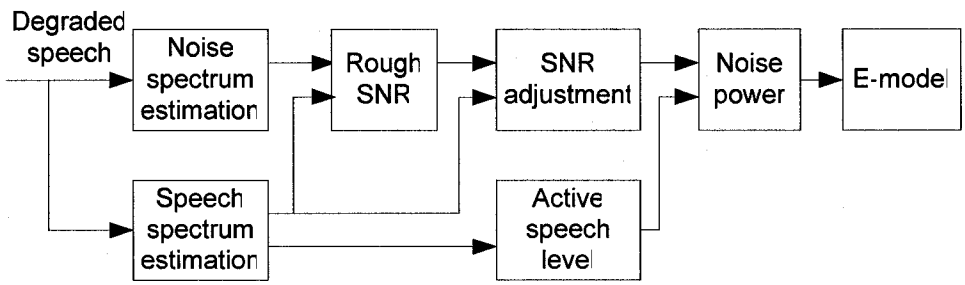


Figure 7.3 Block diagram of noise detection and effect modeling

A) Rough SNR Estimation

The speech spectrum estimation algorithm is selected from [144], which uses a method that combines a minimal mean-square error short-time spectral amplitude

estimator and a two-dimensional spectrum enhancement filter to estimate the reference speech signal. The algorithm also produces the speech in time domain.

For the noise spectrum, it is first estimated by the MCRA [143] algorithm. Then, a rough SNR is calculated as:

$$\hat{SNR}_r = \sum_l \sum_k \left|\hat{X}(l,k)\right|^2 \Big/ \sum_l \sum_k \left|\hat{N}(l,k)\right|^2 \tag{7.2}$$

where $\left|\hat{X}(l,k)\right|^2$ and $\left|\hat{N}(l,k)\right|^2$ are the estimated speech and noise spectrums. l is the index of all the signal frames that contain speech, and k is the frequency bin index.

B) SNR Adjustment

As we mentioned, MCRA can easily overestimate noise when SNR is very high, especially when SNR is greater than 20 dB. Therefore, for $\hat{SNR}_r \geq 20$ dB, an adjustment is performed as follows. First, an energy threshold is decided:

$$TH = 1/10^{(1+\hat{SNR}_r/10)}. \tag{7.3}$$

Then, a simple energy VAD is used to detect the speech silence periods. A frame whose energy is smaller than TH multiplying the maximum energy of all the frames is regarded as a silence frame. Finally, the adjusted SNR is given by:

$$SNR_{adj} = \begin{cases} \hat{SNR}_r & if\ \hat{SNR}_r < 20dB \\ \sum_l \sum_k \left|\hat{X}(l,k)\right|^2 \Big/ \sum_m \sum_k \left|\hat{N}(m,k)\right|^2 \cdot M/L & if\ \hat{SNR}_r \geq 20dB \end{cases} \tag{7.4}$$

where m is the detected silence frame index, M and L represent the number of speech frames and detected silence frames, respectively.

The active speech level is measured from the estimated time domain speech signal, by

using the algorithm defined in ITU-T Rec. P.56 [102]. Finally, the noise power P_N is given by subtracting SNR_{adj} from the active speech level.

7.3.2. Noise Detection Algorithm Performance

To examine the performance of the developed noise detection algorithm, both stationary noise (Hoth) and a recorded car noise are used. Before being added to the clean speech, noise is psophometric weighted; and 11 SNR values are selected, from 0 to 30 dB with an interval of 3 dB. The noise detection algorithm uses a 32-ms frame size with 75% overlapping.

The performance of the noise detection algorithm is depicted in Figure 7.4, which shows that it can accurately measure the background noise power; the measurement standard deviation is about 0.88 dB on average.

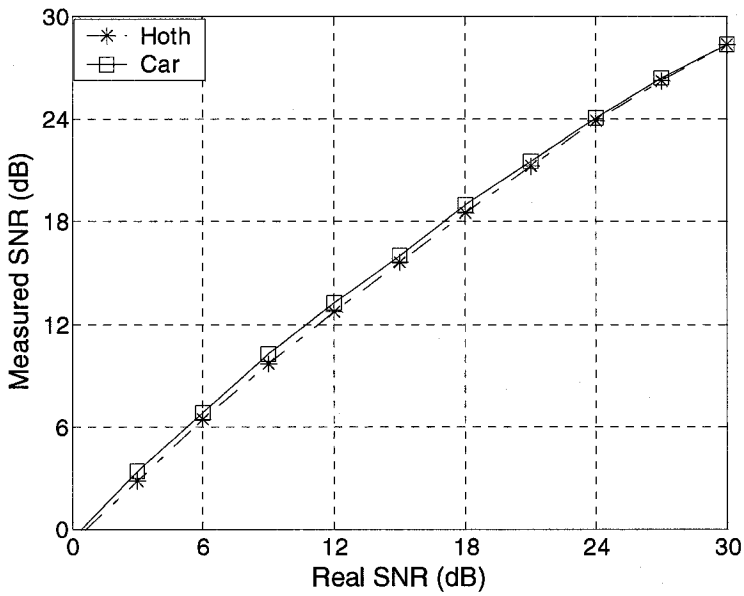


Figure 7.4 Performance of the noise detection algorithm

7.3.3. Noise Effect Modeling

To quantify the effect of noise on speech quality, the parameter *Ro* [11] in the E-model is adopted. The measured P_N can be regarded as noise power summation of circuit noise, equivalent circuit noises caused by room noise at the send side and receive side. *Ro* is calculated by using (7.5) and (7.6) in turn:

$$No = 10\log_{10}(10^{\frac{P_N}{10}} + 10^{\frac{Nfo}{10}}) \tag{7.5}$$

$$Ro = 15 - 1.5(SLR + No) \tag{7.6}$$

where *Nfo* is the noise floor at the receive side. For SLR and *Nfo*, their default values can be used, with 8 dB and -62 dBmp, respectively.

7.4. Overall Assessment Model

7.4.1. Performance

By following the strategies in Section 7.1, the effects of packet loss and temporal clipping are first combined using formula (7.1), in turn, noise effect is added for the listening-only model. Finally, echo effect is added for the conversational model. The detailed steps are given in Table 7.3.

Table 7.3 Steps for calculating overall speech quality

Step 1: Calculate the MOS after packet loss and clipping by: $MOS = MOS_O - DMOS_C$.
Step 2: Convert MOS to R by taking inverse of formula (2.13).
Step 3: Calculate R_O and *Idte* using the E-model to reflect the impacts of noise and talker echo, respectively.
Step 4: Update R from step 2 using R_O only for the listening-only model, or using both R_O and *Idte* for the conversational model.
Step 5: Convert R back to MOS by (2.13), the result is a MOS-CQE or MOS-LQO depending on whether *Idte* is included or not.

The performance of the listening-only model is evaluated. The testing conditions are similar to those in Table 7.1, also Hoth noise with SNR = 10, 20, 30 and 40 dB is added. The simulation is run only once.

It is found that, the noise additivity does not hold well for high noise cases, under which the measured MOS is often not as bad as the predicted counterpart. As SNR is the only parameter used in the model, it is proposed that the measured SNR in subsection 7.3.1, when greater than 20 dB, can be compensated a bit in effect modeling, using the curve shown in Figure 7.5. The curve is derived from the data at SNR = 20, 30 dB, using nonlinear optimization using Matlab Optimization Toolbox. For example, if the measured SNR is 20 dB, its effect is equivalent to about 25 dB, when using E-model formulas (7.5) and (7.6). The curve also shows that a higher SNR requires smaller compensation. When SNR is around 40 dB or higher, the speech can be regarded as clean and no compensation is needed. Note that when SNR is below 20 dB, using the E-model additivity assumption shows reduced correlation between measurement and prediction and increased prediction

error. Further work is needed for a proper listening-only model under such high noise scenarios. Also note that E-model does not take into account the effect of noise spectrum and nonstationarity on speech quality.

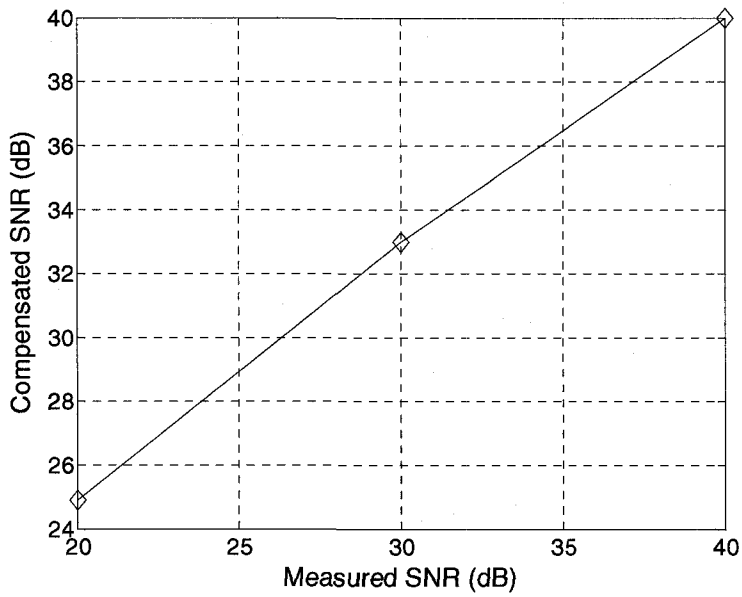


Figure 7.5 SNR compensation curve used in the listening-only quality modeling.

The performance of the listening-only model (with noise SNR $\geq$ 20 dB) is summarized in Table 7.4. It gives the correlation between prediction and measurement, standard error and absolute error distribution.

Table 7.4 Performance of the listening-only model

			Absolute error distribution percentage in each MOS bin								
	ρ	σ	< 0.10	< 0.20	< 0.30	< 0.40	< 0.50	< 0.60	< 0.75	< 1.00	< 1.25
Listening-only model											
G.711	0.91	0.27	30.7	56.0	73.9	85.9	93.1	96.7	98.8	99.9	100
G.729A	0.90	0.27	30.9	55.8	73.0	85.2	93.0	97.3	99.5	100	

An example for the predicted MOS vs. the measured MOS is also shown in Figure 7.6. It is seen that the prediction model works very well as the majority of points closely lie on the dashed diagonal line. In addition, MOS values cover a wide range from the low-end (near 1) to the high-end (near 4.55); therefore the performance over the entire speech quality range is ensured.

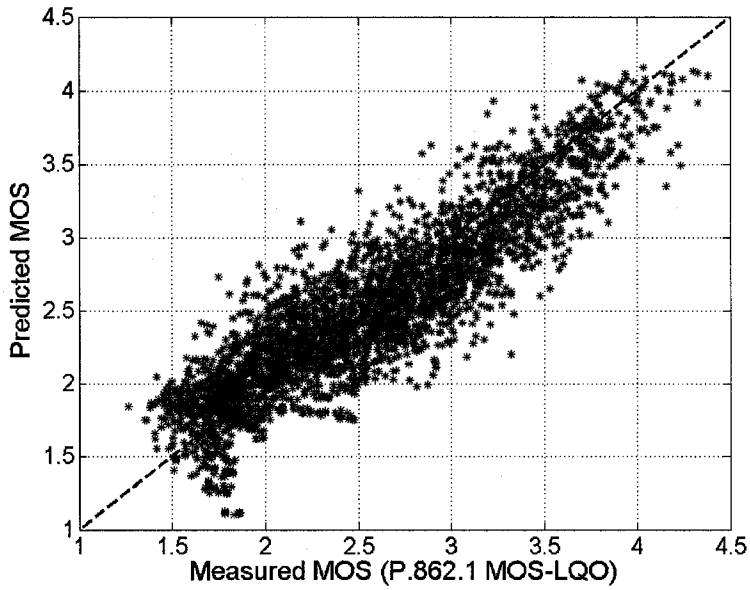


Figure 7.6 The predicted MOS and the measured MOS.

In short, the predicted MOS is highly correlated with the P.862.1 MOS-LQO, and the prediction error is reasonably small. The algorithm we proposed is effective to predict the listening-only speech quality suffering from packet loss, temporal clipping and noise.

7.4.2. Computational Complexity

The developed non-intrusive speech quality assessment algorithm is designed to operate at the receive-end media gateway or IP terminal. It interacts with a speech

Application Programming Interface (API) and a network API. The former provides the access to the two speeches (a cancelled end speech and a non-cancelled end speech); the latter provides the network information, such as packet loss pattern, codec type, and etc. The output of the algorithm is a set of parameters for each detected impairment and its effects. Also, the algorithm gives speech quality in MOS and R scales for both listening-only and conversational quality.

In typical application scenarios, the algorithm reports speech quality every five seconds for a call. In order to detect echo, both non-cancelled end and cancelled end speeches need to be buffered for analysis. At 8,000 Hz sampling rate and 16-bit resolution, this requires 160 KB data memory. The program memory is relative small and 10 KB is deemed to be sufficient.

For the computational complexity, the echo detection module is analyzed first. It consists of a VAD, a DTD and a measurement block. When the maximum searchable echo path delay is set to 500 ms and DTD is set to every 80 samples (10 ms), the total multiplication is 0.36 million and addition is 0.36 million per second for the DTD. For the VAD, the 5 seconds non-cancelled end speech has five hundred 10-ms frames. With 33% overlapping, each frame has 120 samples. For each frame, windowing, autocorrelation matrix, Levinson-Durbin iteration and S/U/V template match need to be performed. It is calculated that 0.21 million multiplications and 0.192 million additions are required in one second. For the echo measurement block, when using the *DS method* with F_1 = 8, the computation of normalized cross correlation is 1.5 million multiplications and 1.5 million additions in one second. Summing these up, the echo

detection module needs 2.07 million multiplications and 2.05 million additions per second.

For the packet loss detection module, the same VAD algorithm in the echo detection module is used for S/U/V classification, for the cancelled end speech. Therefore, 0.21 million multiplications and 0.192 million additions are required per second. For the temporal clipping detection module, 0.5 million additions are required per second. Finally, the noise detection algorithm operates on 32-ms frame with 50% overlapping. It is calculated that 0.4325 million multiplications and 0.5925 million additions are needed per second. The effect modeling only involves in some indexing and simple calculation, the workload is relatively small and can be ignored.

When considering the implementation strategy, the different modules are not real-time critical. Instead, they can be implemented in a sequential manner within the 5 seconds reporting interval. Therefore, the dominant computational requirement which is for the echo module would be the overall requirement. That is, the proposed algorithm requires 2.07 million multiplications and 2.05 million additions per second.

7.5. Discussion and Summary

In this chapter, the overall assessment model is developed, which combines the effects of packet loss, temporal clipping, noise and echo. The performance of the proposed model is evaluated. For the listening-only model, when noise SNR is no less than 20 dB, the results show that the correlation between prediction and measurement is about 0.90, and RMSE is 0.27 MOS.

The combined effects of packet loss and temporal clipping are investigated. It is shown that they are dependent and thus cannot be simply added in the MOS domain. Instead, the correlation factor γ between them is investigated, and a new term is added to correct the bias of summation. The proposed formula is justified through simulations.

The effect of background noise is considered. A noise measurement method is adopted and how to connect the measured noise SNR with the relevant E-model parameters is presented. In the listening-only model, when noise is added, it is shown that the noise additivity assumption in the E-model does not hold well. For SNR $\geq$ 20 dB, a compensation curve is suggested. When SNR $<$ 20 dB, using the E-model is unsatisfactory. Furthermore, the E-model is simplistic to some extent. For the noise, factors such as stationarity and noise spectrum also affect speech quality. For example, car noise, whose strong low frequency content will be heavily attenuated by a telephone handset, may exhibit different effects on speech quality relative to Hoth noise. These are the limitations of the E-model and it needs further validation.

For the overall conversational model, the effects of echo cannot be assessed by using objective approaches. We rely on the E-model for echo effect modeling.

The next generation IP networks exhibit time-varying performance. Interactive applications, such as voice conversations, are quite time-sensitive. For example, the increased delay in these networks will increase interaction difficulty between the talkers. Also, it will increase the perception of echo and chance of double talk. Therefore, listening-only quality may not fully reflect the experience of the end user, there is an imperative demand for real-time conversational speech quality assessment methods.

This need is addressed in Question 20 of ITU-T Study Group 12 (Q.20/12). Question 20 is concerned with working on the standardization of P.CQO (Conversational Quality Objective). Because of its multi-dimensional complexity of the variables and models involved, P.CQO has several versions, including -L (Lightweight), -E (Extended), -W (Wideband) and -F (Footprint). In Draft Recommendation P.CQO-L [145], it includes the following three aspects of a telephone call: talking quality, interaction quality and listening quality, as shown in Figure 7.7*.

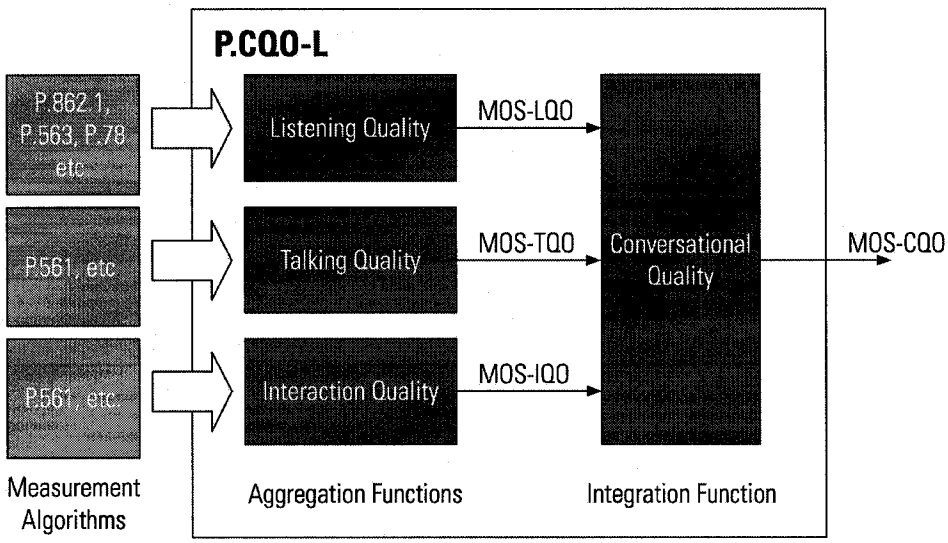


Figure 7.7 Generic conversational model (Courtesy of ITU-T SG12).

The quality evaluation is based on existing objective voice payload or protocol measurements in a real-time manner, which may use ITU-T Rec. P.862, P.561, P.562, P.563, speech level, noise level, echo, delay, etc. The listening quality reflects how the user perceives his/her partner's voice; this can be evaluated by using ITU-T Rec. P.862,

* The author wishes to acknowledge SG12 of ITU-T for the permission of reproducing this figure in the thesis.

P.563 or P.564 (P.VTQ). The talking quality means how the user perceives his/her own voice – the echo. It can be assessed by P.561, by measuring EPL and echo path delay. Interaction quality assesses how a user perceives the interaction with his/her partner during a conversation. This can be done by using P.561 too, where delay and double talk percentage are the major factors to interaction quality. The three aspects of the models are combined to yield a single quality score in MOS-CQO scale.

P.CQO model is very promising. However, it doesn't provide a decomposition of the source of degradations that is vital for network operators to improve the service quality. Moreover, using P.561 as an approach to measure delay and echo is impractical for embedded implementations in VoIP, as it requires excessive memory and CPU usage. Finally, when determining the listening quality, the results from different metrics, such as P.862 and P.563, are often inconsistent.

The developed non-intrusive algorithms in the thesis are backed by Nortel. In SG 12 meeting, it is proposed to provide a set of diagnostic abilities and enhance the current P.CQO.

CHAPTER 8 SUBJECTIVE QUALITY VERIFICATION

A subjective MOS database is developed as a joint effort from several parties, including Nortel, Carleton University and University of Sherbrooke. The database aims to evaluate the proposed non-intrusive algorithm, and to verify some ideas suggested in the thesis. First, a brief overview of the database is given. Then, the subjective MOS results are presented and compared to the objective MOS results. The applicability of PESQ-based algorithms is investigated and their performance limitations are suggested. Finally, calibrated non-intrusive assessment models are developed, based on the subjective MOS.

8.1. Subjective MOS Database Overview

A database containing impaired speech samples and corresponding subjective MOS ratings is prepared to evaluate the performance of proposed non-intrusive, single-ended VoIP speech quality monitoring and assessment algorithms. The speech database may also be used for verifying some ideas suggested in the algorithms and for speech quality enhancement purposes.

The development of the database is a team job. Nortel provides the raw speech recordings and determines the impairment cases investigated. The speech processing and impairment introduction are conducted by the Digital Signal Processing Lab at Carleton University. The subjective MOS test is conducted by the Speech and Audio Processing Lab at University of Sherbrooke. The subjective test consists of two phases, with English

samples and French samples used in the phase 1 and phase 2, which were completed in April and July 2006, respectively.

The database includes a total of 324 cases, including reference and degradation cases. For each case and for each language, four samples are used, one from each of four different talkers (two male and two female talkers). Each sample consists of two different sentences in cascade, with around 650 ms silence (background noise) in between. Talkers are native speakers of Canadian English and French from the Ottawa area. The cases investigated are summarized in Table 8.1. As a result, a total of 2,592 (324 cases × 4 samples/case × 2 languages) degraded samples are included.

Table 8.1 Cases used in the speech database

No	Case category	No. of cases
1	Reference	16
2	Random packet loss	54
3	Constraint packet loss	22
4	Bursty packet loss	54
5	Bursty constraint packet loss	22
6	Temporal clipping	21
7	Noise alone	33
8	Noise with packet loss	54
9	Noise suppression	48
	Total	**324**

8.2. Development of Database

Each raw speech recording provided by Nortel contains a sentence from a native speaker, and is stored in 16,000 Hz, 16-bit linear PCM format. It is pre-processed by ITU-T routines before being introduced with any degradation as explained next. Similar

to those steps in subsection 3.2.2, at first, it is filtered by the MIRS send characteristics defined by ITU-T Rec. P.830 Annex D. Then, it is 2:1 down-sampled to 8,000 Hz and level adjusted to -26 dBov. Finally, two pre-processed sentences are concatenated into one speech sample. Some sentences are used twice provided that none of the contents of two speech samples are identical in the database. For the temporal structure, the interval between two talkspurts is set to about 650 ms, which is determined by using the VAD algorithm in ITU-T G.729B. Background noise is inserted in between to avoid pure silence.

Based on the cases investigated, 324 different degradation cases are introduced, using batch programs written in Matlab scripts and some C routines. For each case and language, 4 different speech samples from 2 male and 2 female speakers are used. Normally, the contents of the 4 samples are different unless otherwise specified. The degraded samples are further filtered using the MIRS receive characteristics. After those steps, 1,296 English samples and 1,296 French samples are created.

To prepare the subjective test, for each language, the 1,296 samples are randomly divided into 4 experiments, subject to the constraints that each experiment contains 324 samples, covering all the cases, and each speaker has exactly 81 samples. Also, it is determined to have 60 people to participate in the listening test to minimize the variability of the rating. Considering that it is hard to gather all the people together at one time, using a small group (e.g. 2-5 people) is more reasonable during the listening test. Therefore, groups of presentation orders are used. That is, in each experiment, the 324 samples are played in a randomized order, which is also subject to the constraints that

successively played samples should come from different speakers. A total of 120 groups of presentation orders are created. After those preparations, the samples are ready for subjective listening test.

The subjective listening test is conducted by University of Sherbrooke. The testing settings and environment are summarized below [146].

The ambient noise level in the listening room is measured daily at the head position of each listener (in the absence of a listener). It is consistently situated at around 27-28 dBA. The headphones used are Beyerdynamic DT 770. Since the tests are monaural, the sound level at the ear position is only taken on the active side of the headset (chosen by the listener at the beginning of the test). This is also measured every day. A calibrating tone which is a 1004 Hz sinusoid with level at -20 dBm0 is used for this measurement. The measurement results show that for all listening systems in the listening room, the sound level at the ear position varies between 77.7 dB to 79.0 dB, satisfying the requirement set by Nortel that is 78 dB.

In phase 1, 32 male and 28 female listeners are used; and in phase 2, 38 male and 22 female listeners are used. For each phase, all the listeners participate in all the four experiments.

8.3. Subjective MOS Results and Analysis

The subjective MOS results are reported in the thesis, for each case category given in Table 8.1, followed by comments. For each case, the average result for the four MOS ratings is illustrated.

Note that based on the analysis with several objective measures, as will be seen in Section 8.4, those objective measures have better correlation to the French database than to the English database. Therefore, the phase 2 sample results are presented in this section, including some complimentary results from Adaptive Multi-Rate (AMR) codec.

8.3.1. Reference Case

The reference category contains unfiltered, HS filtering (MIRS send), Modulated Noise Reference Unit (MNRU), waveform and several CELP codecs under a clean channel, serving as an anchor for the database. The MNRU curve is depicted in Figure 8.1. It shows that MOS ratings increase monotonically with increases in dBQ. Diminishing returns for improvements above 25 dBQ are exactly as expected.

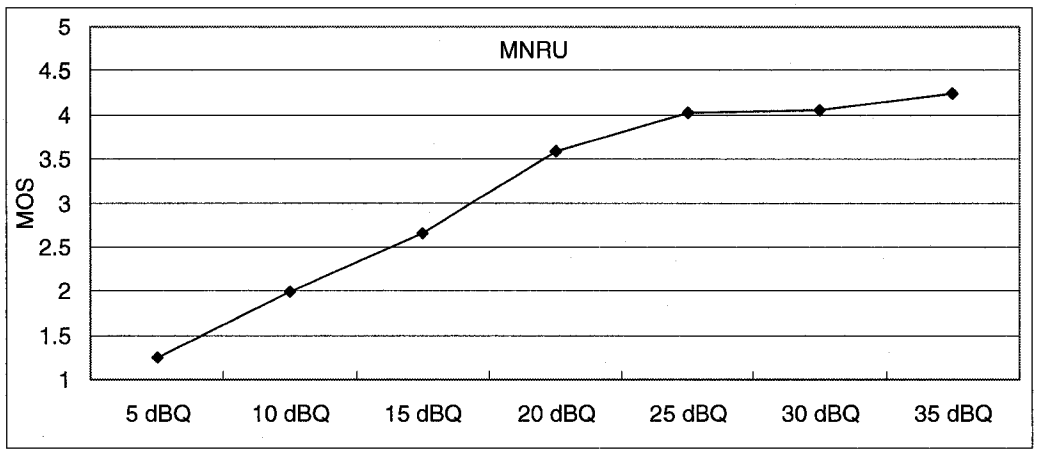


Figure 8.1 MNRU in the reference cases

Also for the reference category, the rating for the clean codec is presented in Figure 8.2. The direct case (no filtering) and the case with HS filtering both use linear quantization. G.711 introduces non-linear quantization steps, and the remaining cases introduce compression with various codecs as indicated. All the ratings but AMR (mode

4) are expected: the direct one has the highest MOS, followed by the standard HS filtered, G.711 and various codecs. The result is consistent with expectation, given the range of quality in the study as a whole.

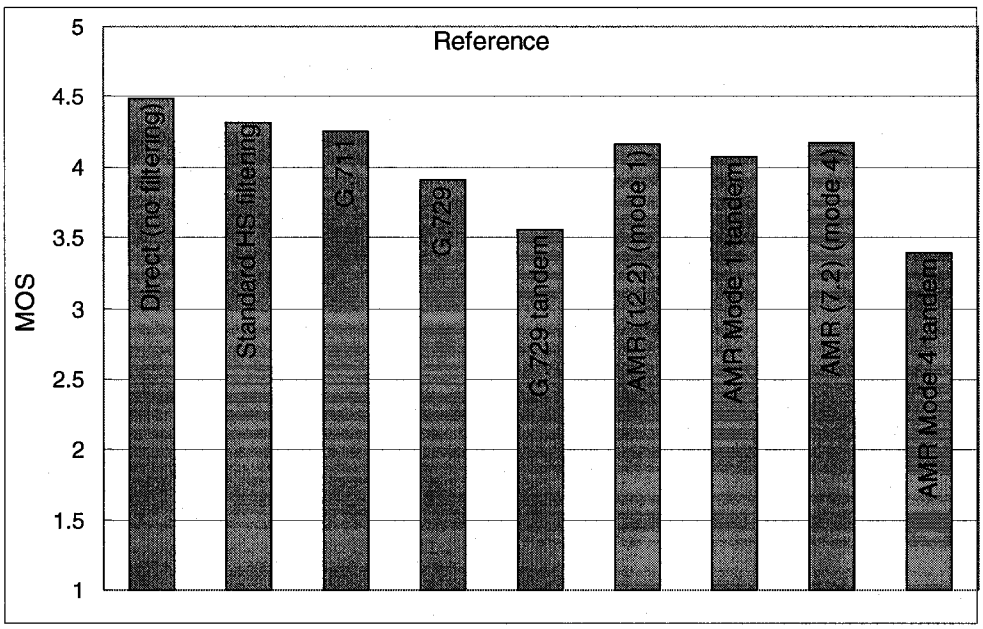
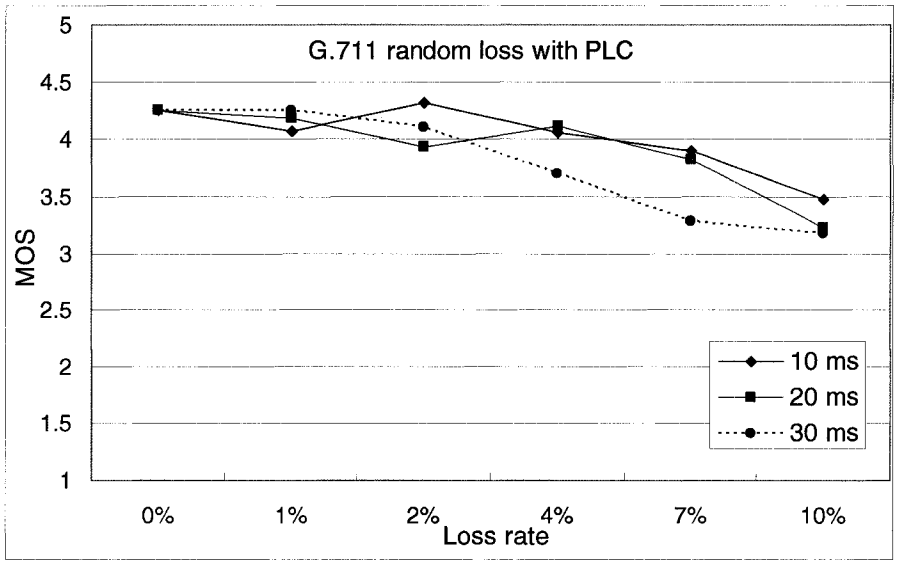


Figure 8.2 Clean codec in the reference cases

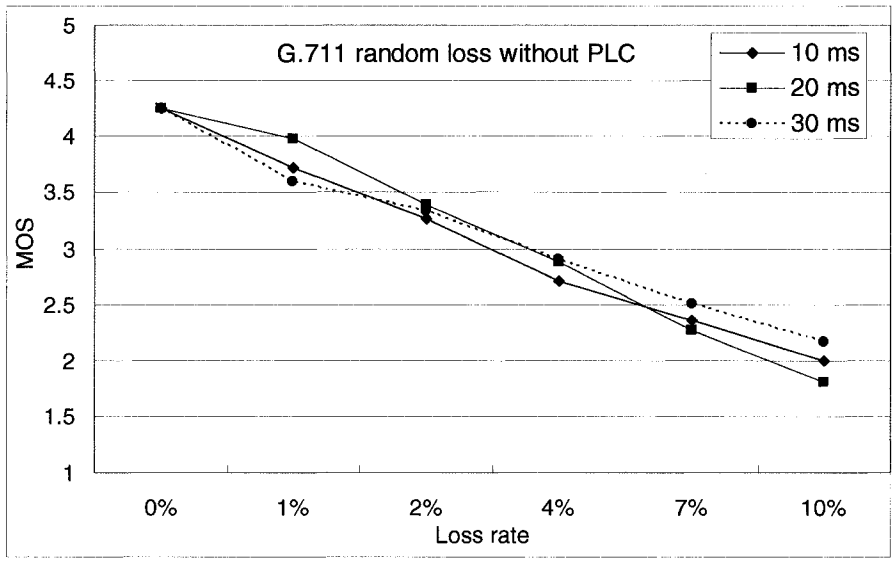
8.3.2. Random Loss Case

For the random loss case category, the MOS ratings for G.711 and G.729 are given in Figures 8.3 and 8.4 respectively. For all codecs, MOS is seen to fall off monotonically (with three exceptions) with increasing packet loss rate. A 10% packet loss without PLC is expected to be quite poor for all packet sizes. For G.711, as expected, speech quality with PLC (G.711 Appendix I) is rated higher than that without PLC (replacing lost frames by silence). Also, when PLC is used, the ratings fall off more slowly as seen in Figures 8.3 and 8.4. In addition, rating reversals are seen in a few cases where packet loss

is applied, especially for the lower loss rates. This is because of statistical variation in the importance of the acoustic feature the missing data falls on.



(a) With PLC



(b) Without PLC (silence insertion)

Figure 8.3 G.711 random loss with and without PLC

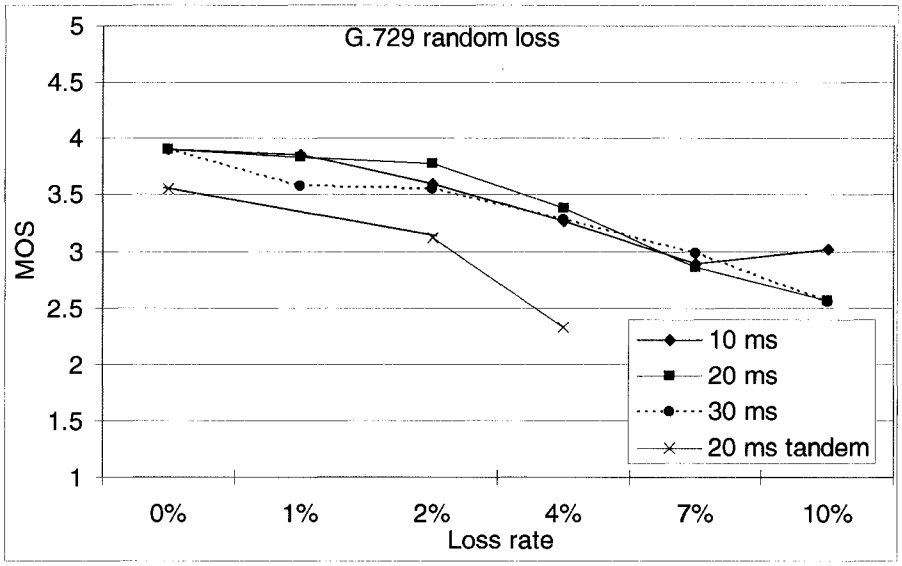
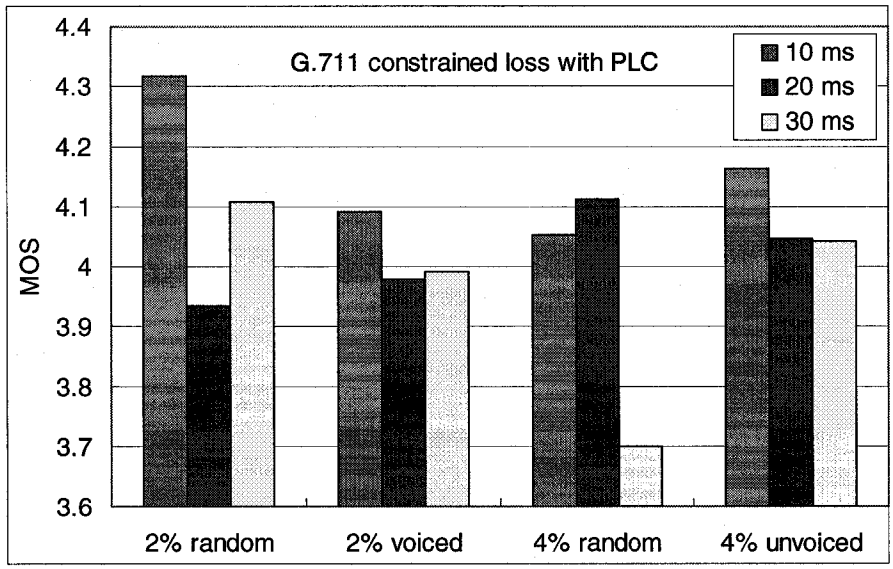


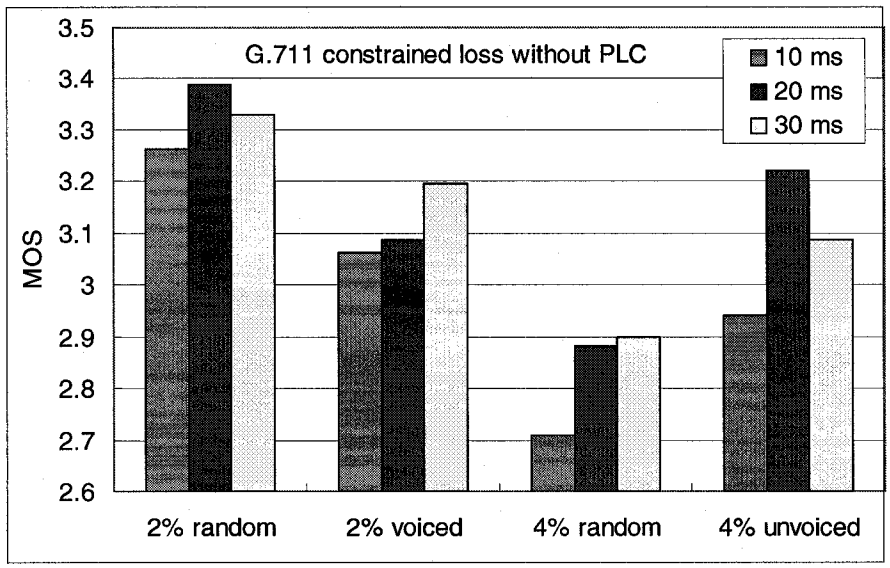
Figure 8.4 G.729 random loss with PLC

8.3.3. Constraint Loss Case

For the constraint random loss case category, only 2% and 4% packet loss rate are explored. To examine the idea that packet loss landing on a voiced speech segment degrades the quality more than that landing on an unvoiced segment, in the database, 2% loss pattern is introduced at voiced packets, whereas 4% loss is introduced at unvoiced packets. The results are depicted at Figures 8.5 (a), (b), and 8.6 for G.711 with PLC, without PLC and G.729 with PLC respectively. By comparing to their random loss counterparts (2% random loss vs. 2% voiced loss, and 4% random loss vs. 4% unvoiced loss), it is clear that voiced loss has more impact to speech quality than unvoiced loss.



(a) With PLC



(b) Without PLC

Figure 8.5 G.711 constrained random loss

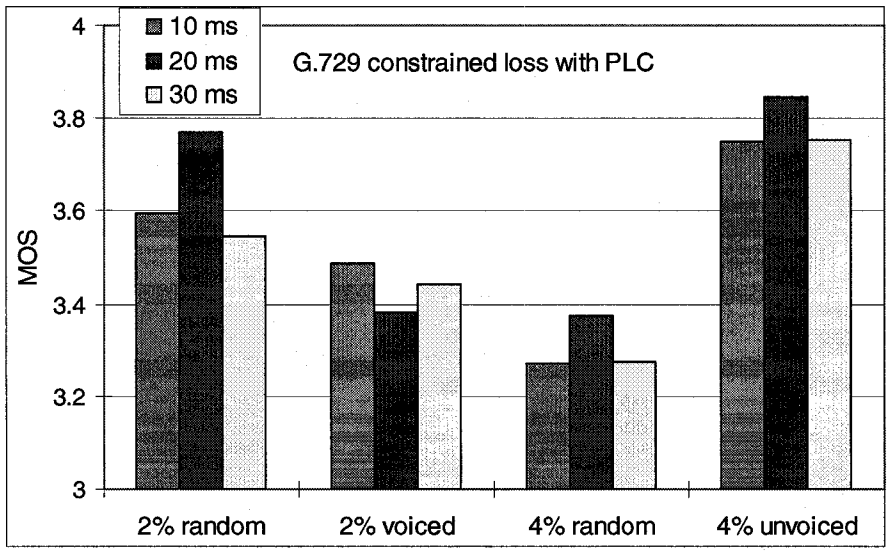
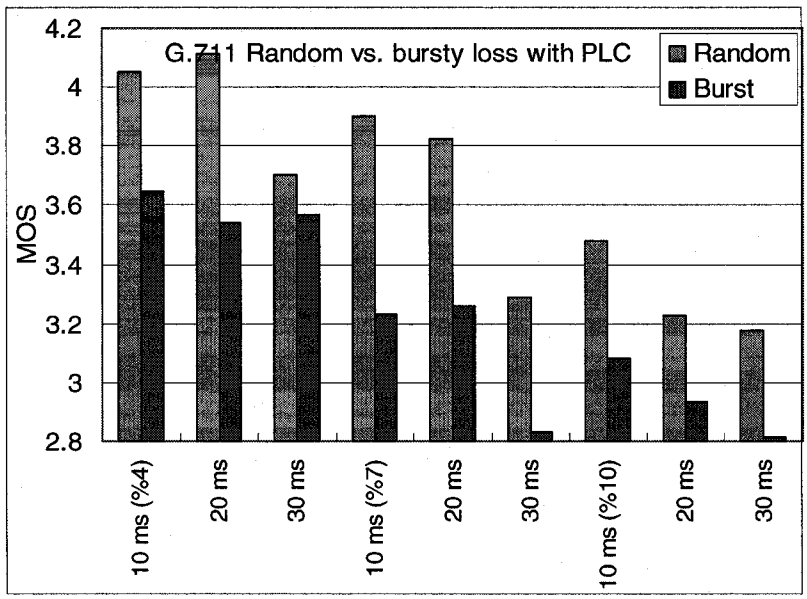


Figure 8.6 G.729 constrained random loss with PLC

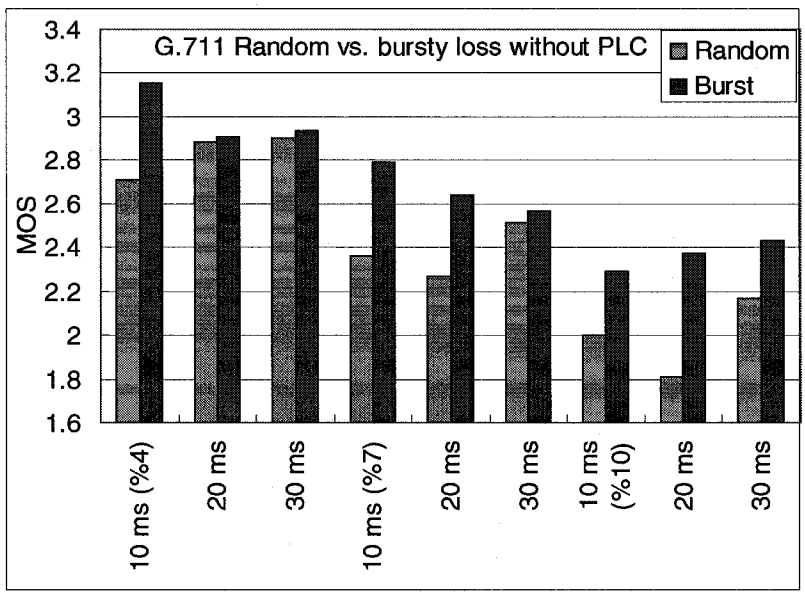
8.3.4. Bursty Loss Case

For the bursty loss case category, when the loss rate is low (e.g. 1% or 2%), the effects of loss burstiness are not so pronounced due to the small number of speech samples used in the subjective MOS test. Therefore, results for loss rate starting from 4% are depicted in Figures 8.7 (a), (b), and 8.8 for G.711 with PLC, without PLC and G.729 respectively. Each figure compares the MOS difference between random loss and bursty loss. For G.711 with PLC, it is clear that bursty loss is rated lower than random loss. When PLC is not used for G.711, packet loss results in clicks in speeches, which are more frequent for smaller packets. On the one hand, bursty loss results in longer speech mute. On the other hand, it also reduces the frequency of clicks. Interestingly, the subjective test shows that bursty loss has better speech quality. That is, the effects of reduced clicks compensate more than that of losing multiple consecutive frames. For G.729, as the built-in PLC could gracefully alleviate the effects of packet loss, it is

expected that random loss would be rated higher than bursty loss. However, Figure 8.8 shows several unexpected features.



(a) With PLC



(b) Without PLC

Figure 8.7 Random loss vs. bursty loss for G.711

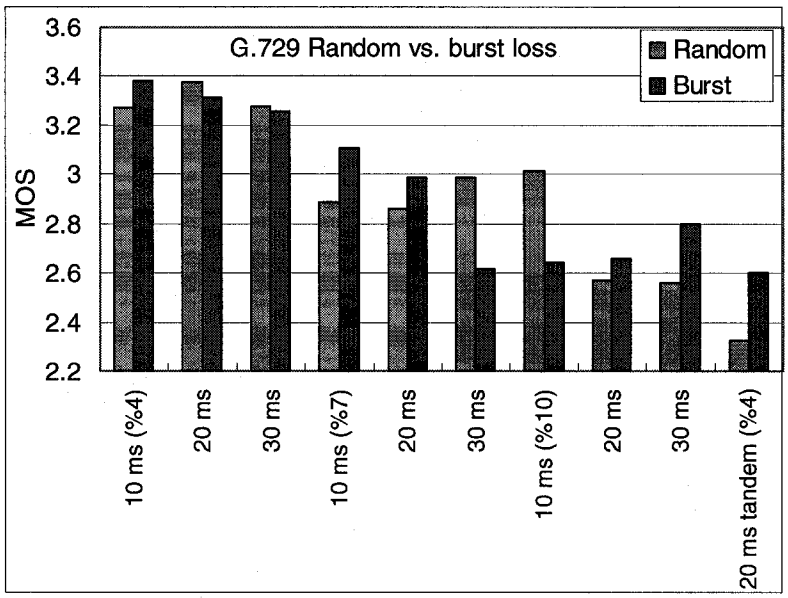


Figure 8.8 Random loss vs. bursty loss for G.729

8.3.5. Temporal Clipping Case

For the temporal clipping case category, the effects of FEC (15, 35 and 60 ms) and MSC (20, 40 and 120 ms) clipping are examined, as shown in Figures 8.9 and 8.10 respectively. In Chapter 4, we discovered that speech quality falls linearly with the amount of clipping. However, subjective test doesn't support this. This may be because the amount of clipping introduced is very small; the subjective effect of replacing the small amount of low energy contents by noise is not discernible. More speech samples with various degrees of clipping need to be examined in the future.

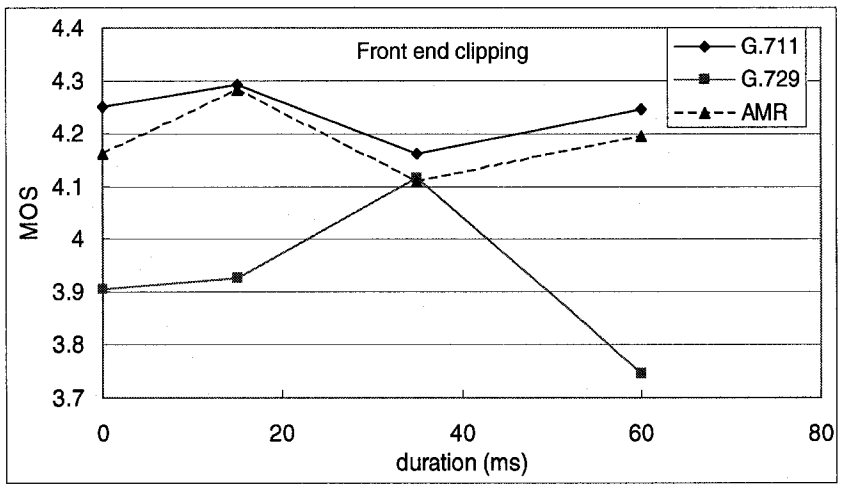


Figure 8.9 Front end clipping

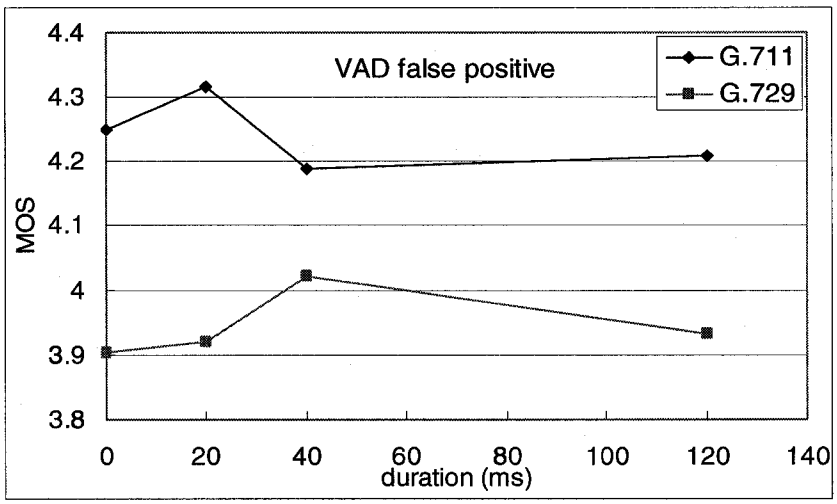


Figure 8.10 Middle speech clipping

8.3.6. Noise Alone Case

For the noise alone case category, several kinds of noises, including Hoth, car, street and babble noise, are investigated. SNR is set to be 0, 10, 20 dB; the noise floor is assumed to be 40 dB (clean case). The results are shown in Figures 8.11 and 8.12 for G.711 and G.729 respectively. It is seen that higher SNR results in higher speech quality. In all cases, SNR = 0 dB is not acceptable for listeners (MOS is around 1). Also, for all

codecs, under the same SNR, the speeches corrupted by car noise have the best speech quality; and those corrupted by Hoth noise have the worst quality. The effects of street and babble noise are appreciable under low SNR of 10 dB case.

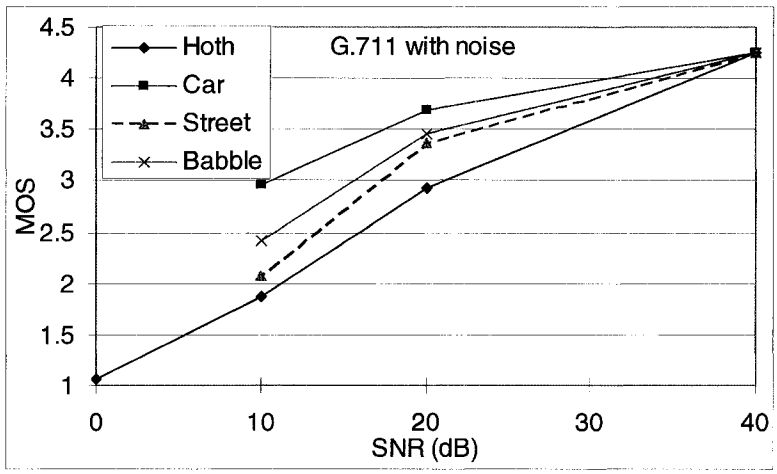


Figure 8.11 Noise alone for G.711

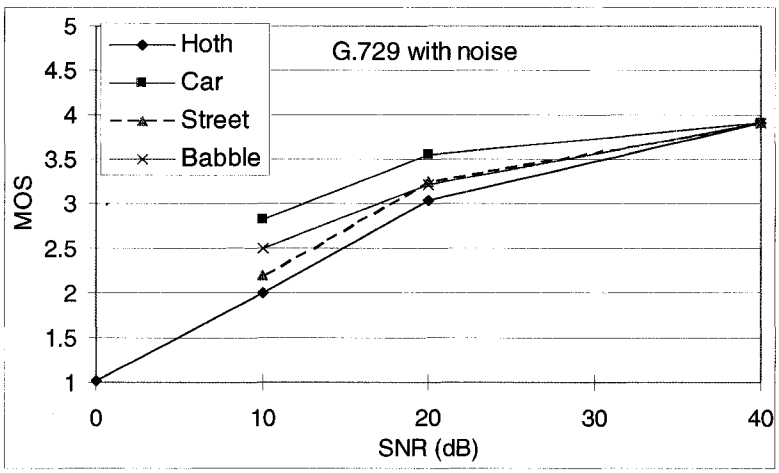
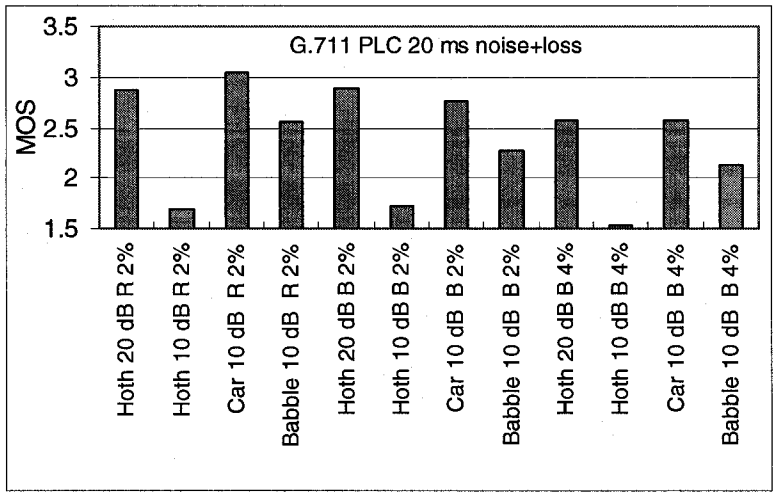


Figure 8.12 Noise alone for G.729

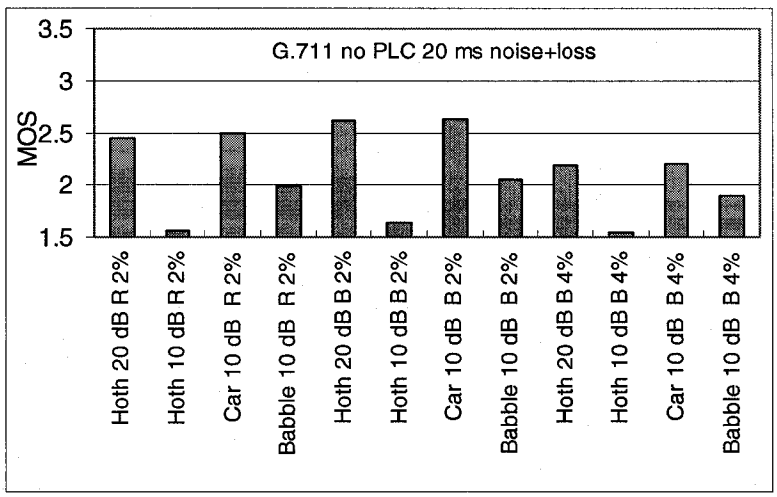
8.3.7. Noise with Packet Loss Case

For the noise with packet loss case category, both random loss and bursty loss under noisy scenarios are examined. The loss rate is set to 2% and 4%. Similarly, the results for

G.711 and G.729 are shown in Figures 8.13 and 8.14 respectively. Clearly, with other conditions the same, G.711 speech quality with PLC is better than that without PLC. Also, similar to noise alone case, under the same SNR, car noise results in the best speech quality; and Hoth noise has the worst quality.



(a) with PLC



(b) no PLC

Figure 8.13 Noise with packet loss for G.711

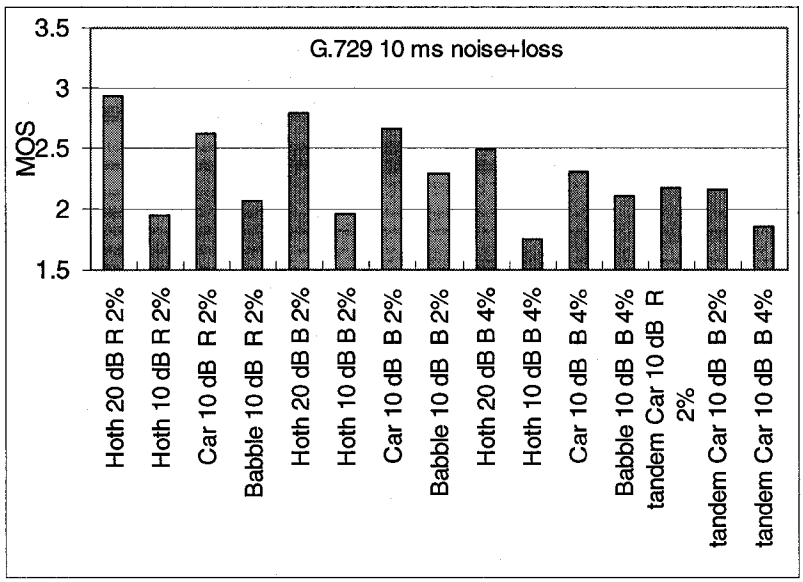


Figure 8.14 Noise with packet loss for G.729

8.3.8. Noise Suppression Case

Finally, the performance of two Noise Suppression (NS) algorithms is compared, for the preparation of selecting an appropriate algorithm to enhance the speech quality in the future. NS1 is chosen as a good NS technique; and NS2 is chosen as one showing obvious artifacts. In the database, NS1 is the noise suppression algorithm A in Selectable Mode Vocoder (SMV) and NS2 is the noise reduction algorithm in Adobe Audition 1.0 software with reduction level set to 75. The comparative performance of two NS techniques is illustrated in Figure 8.15. It is expected that NS1 to be rated higher in most cases than NS2, and as well that the rating difference might be substantial. Consistent with this expectation, NS1 is almost always rated higher than NS2. In only five cases, NS2 is rated higher than NS1. The size of the average difference between the two methods is smaller than expected. This may be due to the range of sample quality included in the study.

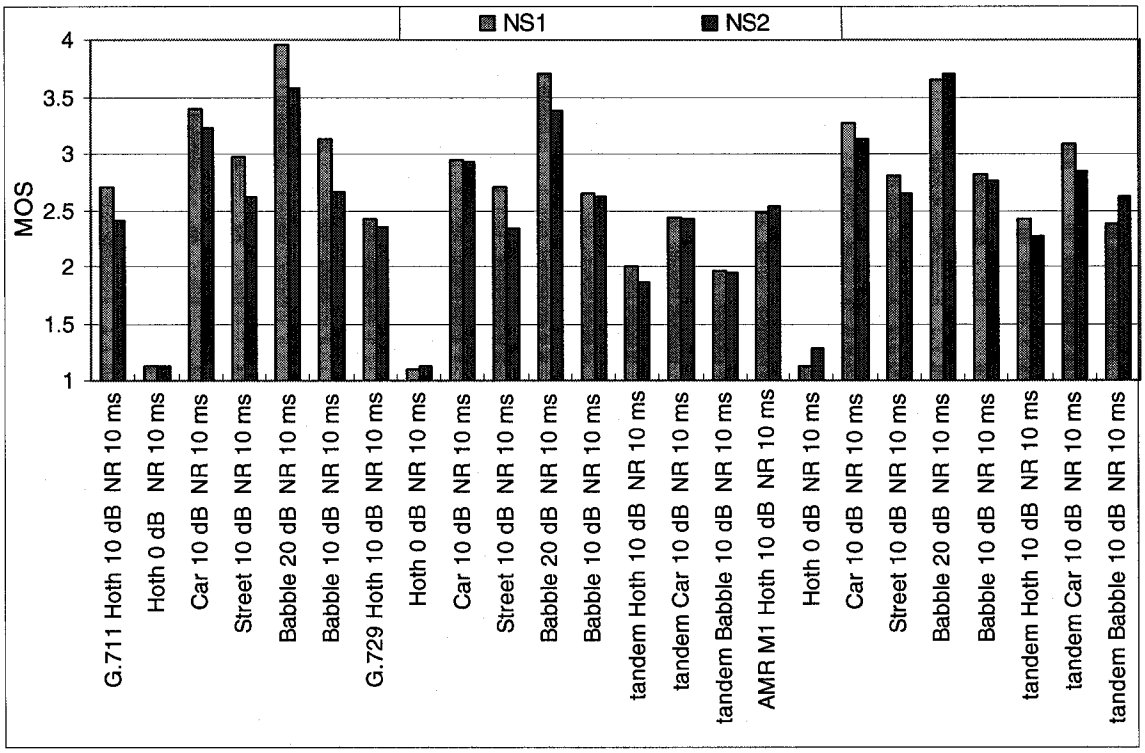


Figure 8.15 Noise suppression algorithms comparison

8.4. Comparisons with Objective MOS

8.4.1. Correlation Analysis

As we know, due to the time-consuming, expensive and unrepeatable nature of the subjective test, the developed MOS database only covers a portion of impairments we investigated. In order to develop the VoIP speech quality assessment algorithms that cover a broad range of impairments and their combinations, using some leading objective measures is one of the feasible approaches.

Then, it is important to evaluate the performance of those objective measures, using the developed subjective MOS database. The purpose is to get insights how the objective MOS is close to the subjective one and how reliable these objective measures are under

different testing conditions. Furthermore, their performance limitations will be revealed. The corresponding objective voice quality is measured by MOS-LQO. For comparison purposes, PESQ and P.563 MOS are also measured and discussed.

The MOS cloud for the whole database is illustrated in Figure 8.16. For the French database, the MOS dots are seen to be fairly distributed around the diagonal line. For the English database, the MOS cloud is biased downward. The subjective MOS is usually higher than the objective one. That is, a bit to our relief, the speech quality normally is not as bad as the objective measure indicates.

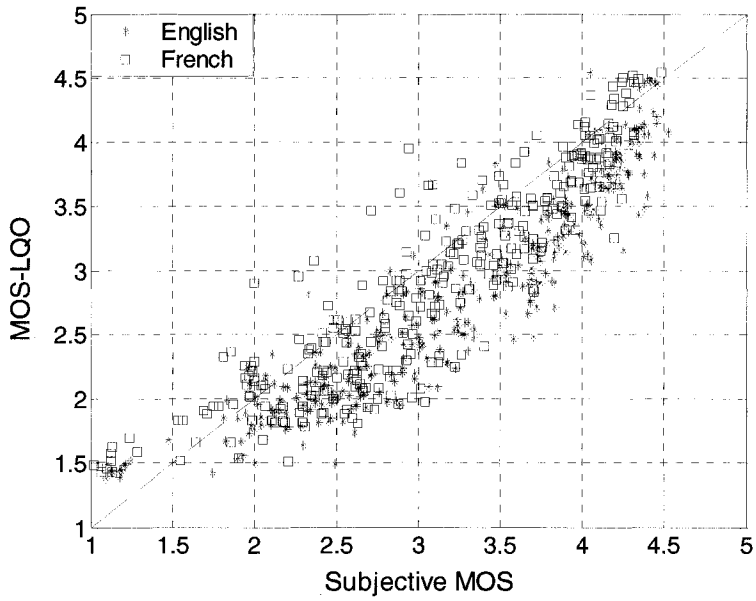


Figure 8.16 Comparison of subjective MOS and MOS-LQO

The testing categories in Table 8.1 can be divided into two groups: categories that the objective methods generally intend to measure (the target group), and those that the methods are not designed to evaluate or need further validation (the non-target group). The former includes case categories no. 1, 2, 4, 7 and 8, totalling 211 out of 324 testing

cases; the rest 113 cases are within the non-target group. The performance of these objective algorithms is characterized by the Pearson correlation coefficient ρ, RMSE σ and the absolute prediction error *e*. The same formulas (3.1) through (3.3) are used here, where x_i, y_i are the subjective MOS and the measured objective MOS, respectively. The performance is examined based on the database as a whole, the target group, non-target group and individual categories.

A) Correlation Per Sample and Per Case

As a starting point, the correlation per speech sample is analyzed. The results are given in Table 8.2, where F1, F2, M1, M2 represent female talker 1, 2, male talker 1 and 2, respectively.

Table 8.2 Correlation with objective measures per sample

Taker	MOS-LQO			PESQ			P.563		
	All	Target	Non-target	All	Target	Non-target	All	Target	Non-target
English									
F1	0.88	0.88	0.89	0.88	0.88	0.89	0.77	0.76	0.84
F2	0.89	0.90	0.86	0.89	0.90	0.88	0.78	0.77	0.85
M1	0.87	0.88	0.87	0.87	0.88	0.86	0.66	0.66	0.69
M2	0.89	0.89	0.87	0.89	0.90	0.88	0.78	0.80	0.76
All	0.88	0.89	0.87	0.88	0.89	0.87	0.72	0.72	0.75
French									
F1	0.86	0.86	0.85	0.86	0.86	0.86	0.69	0.66	0.71
F2	0.86	0.85	0.85	0.86	0.86	0.86	0.71	0.70	0.73
M1	0.84	0.83	0.86	0.84	0.82	0.87	0.68	0.66	0.74
M2	0.84	0.84	0.85	0.84	0.84	0.85	0.68	0.63	0.78
All	0.85	0.84	0.85	0.85	0.84	0.86	0.67	0.65	0.70

It is seen that, the above two intrusive objective measures (MOS-LQO and PESQ) have almost the same correlation with the subjective MOS, and their performance is much better than that of the non-intrusive algorithm P.563. For the two groups, surprisingly, they have similar correlation as well, suggesting that some categories in the non-target group could be reliably measured by PESQ and MOS-LQO algorithms. This will be seen clearly when the correlation with each category is analyzed later. Also, the correlation of the English database is a bit better relative to the French database.

For the talker wise, there is no difference across four talkers, except English talker M1 when measured by P.563. Therefore, the rest of the analysis is based on the correlation per case, that is, the four MOS ratings for a case are averaged before further analysis. The correlation results per case are shown in Table 8.3. Compared to those in Table 8.2, substantial improvement in correlations is observed, which is about 0.04-0.05 for the two PESQ-based methods. The improvement for the P.563 is even much higher. This is because both acoustic variations of individual talkers and impairment variations are averaged out, further demonstrating the importance of using several talkers in a subjective MOS database. Furthermore, σ is also given in Table 8.3.

Table 8.3 Correlation with objective measures per case

	MOS-LQO			PESQ			P.563		
	All	Target	Non-target	All	Target	Non-target	All	Target	Non-target
English									
ρ	0.93	0.93	0.91	0.93	0.93	0.91	0.85	0.84	0.88
σ	0.51	0.50	0.54	0.43	0.39	0.49	0.47	0.43	0.52
French									
ρ	0.90	0.90	0.90	0.90	0.90	0.90	0.78	0.76	0.83
σ	0.39	0.38	0.42	0.36	0.33	0.41	0.50	0.51	0.49

The detailed performance of objective measures, including correlation and RMSE for each case category is presented in Table 8.4, where the non-target case categories are shown in the shaded cells. Moreover, the absolute error distribution based on per case is given in Table 8.5.

Table 8.4 Correlation and RMSE of objective measures for each case category

English		**All**	Target	Non-target	Reference	Random loss	Constraint loss
MOS-LQO	ρ	**0.93**	0.93	0.91	0.95	0.86	0.69
	σ	**0.51**	0.50	0.54	0.31	0.47	0.42
PESQ	ρ	**0.93**	0.93	0.91	0.92	0.86	0.69
	σ	**0.43**	0.39	0.49	0.38	0.42	0.41
P.563	ρ	**0.85**	0.84	0.88	0.90	0.82	0.74
	σ	**0.47**	0.43	0.52	0.61	0.34	0.32
		Bursty loss	Constraint bursty loss	Temporal clipping	Noise	Noise with loss	Noise suppression
MOS-LQO	ρ	0.94	0.88	0.74	0.90	0.69	0.89
	σ	0.59	0.88	0.42	0.45	0.50	0.42
PESQ	ρ	0.94	0.87	0.71	0.93	0.70	0.91
	σ	0.49	0.76	0.48	0.32	0.26	0.34
P.563	ρ	0.62	0.75	0.50	0.79	0.47	0.68
	σ	0.41	0.52	0.63	0.42	0.48	0.55

French		**All**	Target	Non-target	Reference	Random loss	Constraint loss
MOS-LQO	ρ	**0.90**	0.90	0.90	0.97	0.83	0.57
	σ	**0.39**	0.38	0.42	0.23	0.34	0.39
PESQ	ρ	**0.90**	0.90	0.90	0.95	0.83	0.56
	σ	**0.36**	0.33	0.41	0.32	0.36	0.39
P.563	ρ	**0.78**	0.76	0.83	0.87	0.67	0.36
	σ	**0.50**	0.51	0.49	0.67	0.47	0.43
		Bursty loss	Constraint bursty loss	Temporal clipping	Noise	Noise with loss	Noise suppression
MOS-LQO	ρ	0.95	0.79	0.64	0.85	0.62	0.85
	σ	0.31	0.31	0.38	0.46	0.44	0.49
PESQ	ρ	0.94	0.79	0.64	0.87	0.63	0.87
	σ	0.21	0.32	0.41	0.42	0.35	0.45
P.563	ρ	0.32	-0.02	0.25	0.84	0.59	0.79
	σ	0.54	0.41	0.63	0.41	0.50	0.48

Table 8.5 Absolute error distribution with percentage in each MOS bin

	< 0.25	< 0.50	< 0.75	< 1.00	< 1.25	< 1.50
English						
MOS-LQO	25.31	60.19	86.42	97.53	100	
PESQ	37.04	70.37	95.37	99.69	100	
P.563	32.10	69.14	91.05	98.15	99.69	100
French						
MOS-LQO	48.15	79.32	93.21	99.38	100	
PESQ	50.31	82.41	96.60	99.69	100	
P.563	27.78	61.73	89.20	98.46	99.69	100

For the database as a whole, both subjective and objective MOS values are observed within range from the low-end 1.0 to the high-end 4.55. ρ is as high as 0.90-0.93 for MOS-LQO and PESQ; for P.563, only a moderate 0.78-0.85 is seen. Also, σ for PESQ and MOS-LQO is in general smaller than that of P.563. As expected, the intrusive methods are more effective than the non-intrusive one, because the former have access to the reference.

From Tables 8.4 and 8.5 and Figure 8.16, it is seen that the performance of the French database is slightly better than that of the English database, as indicated by σ and error distribution, although the correlation with English database is a bit higher. Finally, the error statistics in Table 8.5 are comparable to those listed in Table 3 of ITU-T Rec. 862.1 [19], for studying the performance improvement of MOS-LQO over PESQ. Note that, using MOS-LQO doesn't show improved performance over PESQ. Perhaps this depends on the database used.

B) Performance over Each Category

The analysis for each case category below mainly based on the MOS-LQO, because all our models are built on it.

For the two groups, σ for the target group is smaller than that of the non-target group. Surprisingly, ρ between two groups is similar around 0.90, suggesting that although not intended to be, some categories in the non-target group could be measured by the PESQ or MOS-LQO with cautions. For each case category, as its MOS may not span over a wide range, ρ often decreases when it is examined around a small region. In this case, σ is a better indicator than ρ for evaluating the objective methods [69].

For the target categories, results are reasonable in general. An exception is the bursty loss category using the English database. MOS-LQO is designed to measure such kind of degradation. Although ρ is the highest of 0.94, σ is bigger than expected. A closer look for this reveals that subjective MOS and MOS-LQO have excellent linear relationship, but the latter is lower than the former, as shown in Figure 8.17.

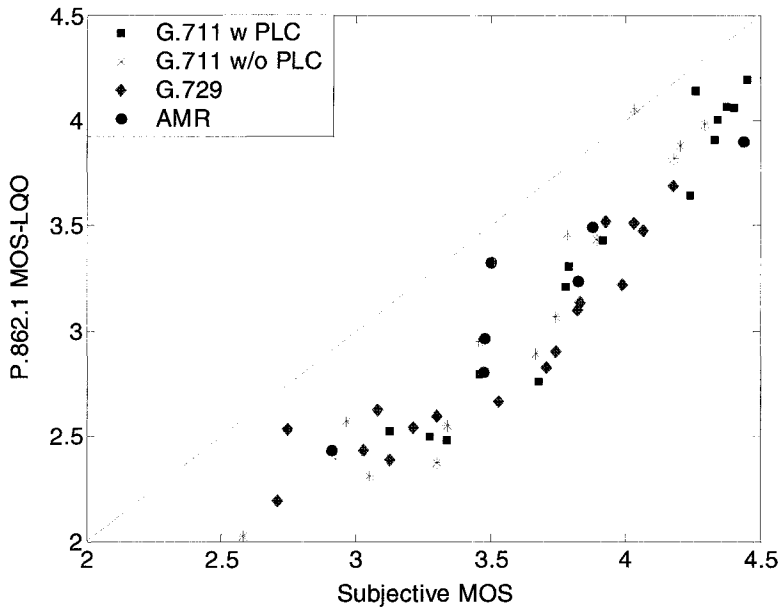


Figure 8.17 Subjective MOS vs. MOS-LQO – bursty loss (English)

Moreover, the noise and noise with packet loss categories have relative larger σ, this is investigated next in subsection 8.4.2. The noise with packet loss category shows a low correlation 0.62, because MOS is clustered around the low-end.

Among the non-target categories, σ for the constraint loss is comparable to those of the target group. A further examination of their MOS cloud shows that the distribution of MOS dots is reasonable. For the constraint bursty loss case, identical sentences are used across different talkers and codecs. At first, a certain bursty loss pattern is applied to a speech sample. Then, the packet loss is manually adjusted to the rest of the three samples to make sure that loss hits the same phonemes. This needs extensive manual operations as the speed and acoustic features of four talkers are different. Only 88 samples are produced and such case is not investigated by our previous chapters. The results are commented here. For the French database, the results are reasonable. However, for the

English database, σ is around 0.88. The applicability of MOS-LQO in this case seems to depend on the database and needs further validation. For the temporal clipping and noise suppression categories, the MOS cloud is skewed from the diagonal line. The two objective algorithms may not correctly indicate the true MOS and their measurement results must be used or interpreted with cautions. For example, an improvement in objective MOS for a noise suppression algorithm doesn't necessarily mean it improves the voice quality subjectively. The performance for the noise suppression category is illustrated in Figure 8.18 for the French database.

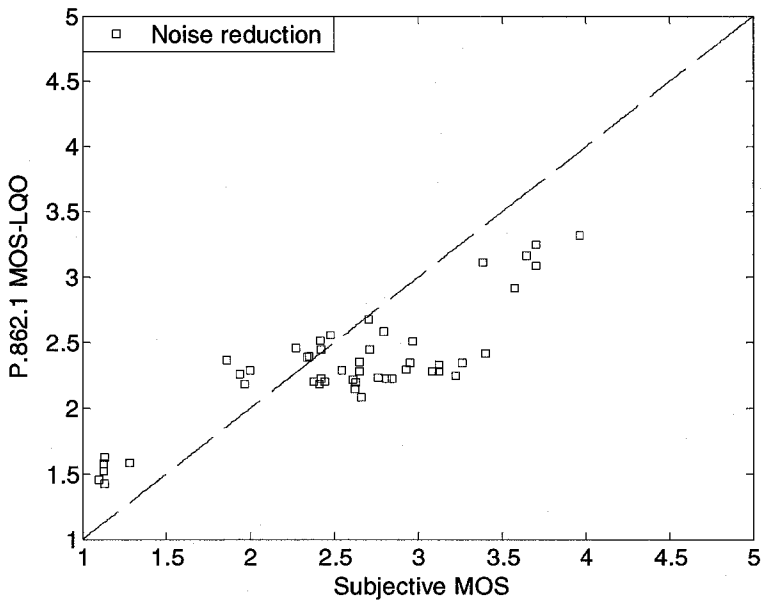


Figure 8.18 Subjective MOS vs. MOS-LQO – noise suppression (French)

In summary, MOS-LQO and PESQ are generally applicable in VoIP speech quality testing for the target impairments, and they have similar performance. They can be reliably used to evaluate the effects of loss localization (constraint loss) on speech quality.

For the constraint bursty loss, temporal clipping and noise suppression cases, they need further validation. As a non-intrusive measure, P.563 only shows moderate performance; it's better to be used when necessary.

8.4.2. Discrepancy Analysis

The subjective MOS reveals several features that are not captured by the objective measures. First of all, for G.711 with bursty loss, if no PLC is implemented, the subjective test suggests that bursty loss leads to better speech quality relative to random loss. This doesn't conform to our findings in subsection 6.5.2. Secondly, for temporal clipping, subjective MOS shows that the amount of clipping is not related to speech quality, this also conflicts with our findings in subsection 4.4.1. Finally, subjective test shows that noise spectrum is important when evaluating the effects of noise. However, such a fact is not covered by the E-model.

To model the impacts of those impairments on VoIP speech quality, therefore, it is expected that there would be some differences if subjective MOS rating is used instead. Now, it is vital to understand where the difference comes from. Is it because of the speech database used? That is, if we use the speech samples from the subjective database and measure them by MOS-LQO, will the MOS-LQO results be very close to the subjective MOS? Or, is the difference caused by the MOS measurement algorithms? That is, will the MOS-LQO results of the database be very close to those findings in Chapters 4 and 6? To serve this need, MOS-LQO and subjective MOS are also analyzed for above discrepancies and results are highlighted below.

For the first problem, MOS-LQO for G.711 without PLC is depicted in Figure 8.19 for the French database. By comparing results under random loss to those under bursty loss, it is clear that the former have higher quality ratings (except one case), supporting the findings in subsection 6.5.2.

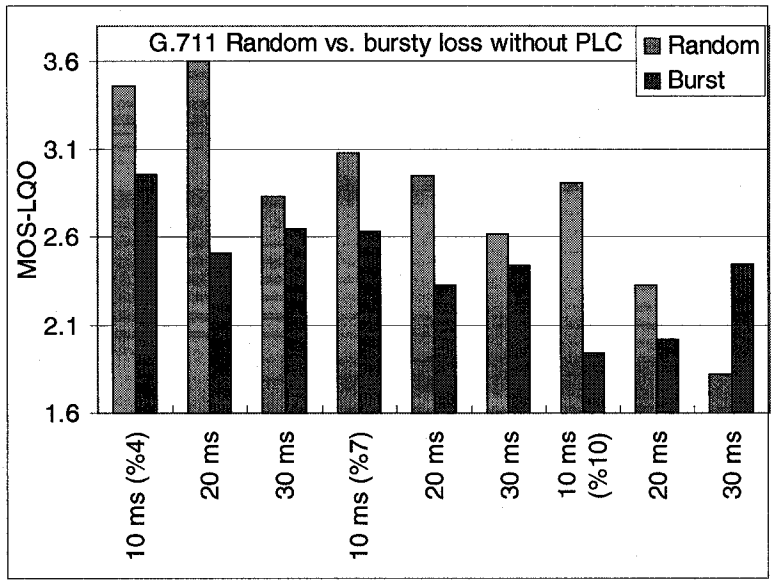
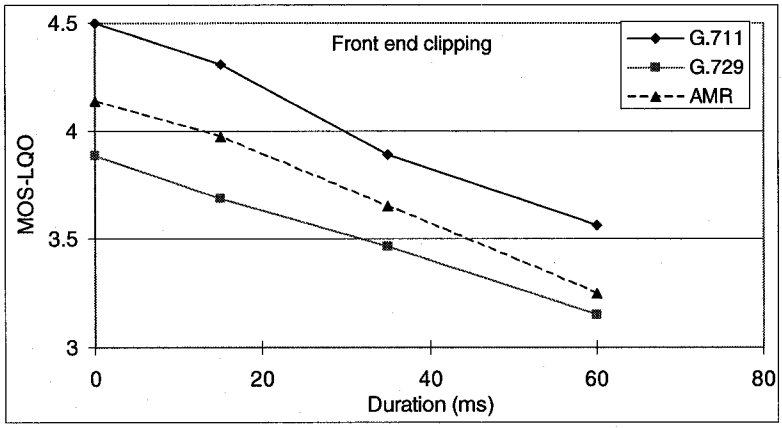
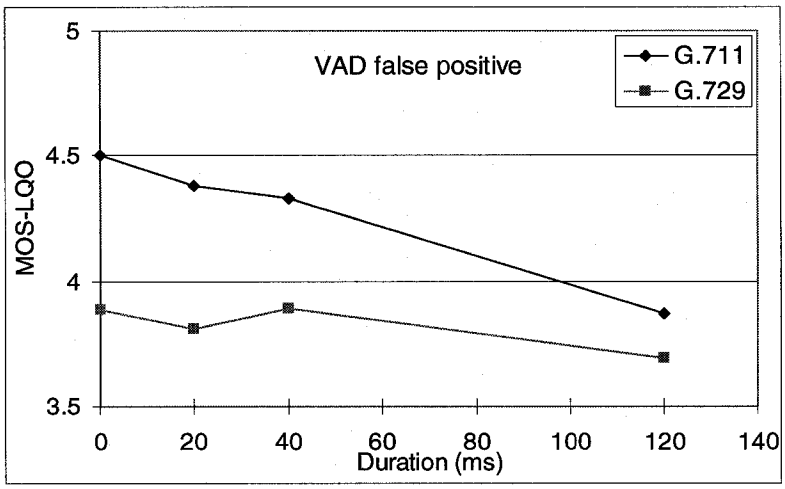


Figure 8.19 Packet loss for G.711 without PLC by using MOS-LQO

For the second problem, MOS-LQO for 15, 35 and 60 ms FEC is shown in Figure 8.20(a), and for 20, 40 and 120 ms MSC is shown in Figure 8.20(b). Both figures support the finding in subsection 4.4.1 that speech quality drops linearly with the amount of FEC or MSC. And, FEC impairs more than MSC in speech quality as shown by the slope of the curves.



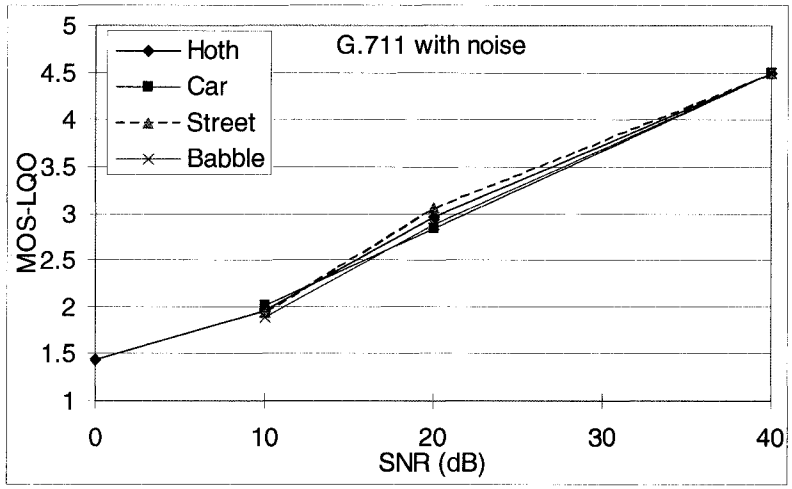
(a) FEC



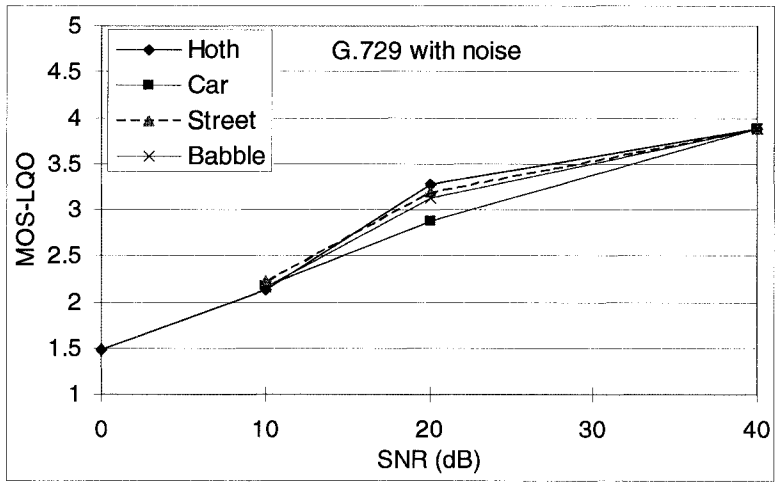
(b) MSC

Figure 8.20 Effects of temporal clipping by using MOS-LQO

For the final problem, MOS-LQO for G.711 and G.729 with different types of noise is shown in Figures 8.21(a) and (b) respectively. Difference for the impacts of noise spectrum is only sometimes observable (e.g., car noise at SNR = 10 dB for G.729). It is much smaller than that reflected in Figures 8.11 and 8.12 for subjective MOS. That is, MOS-LQO is unable to accurately capture the effects of noise spectrum.



(a) G.711



(b) G.729

Figure 8.21 Noise alone by using MOS-LQO

The quality of the rest speech samples in MOS-LQO is also analyzed together with the subjective MOS to determine the origin of the difference. And similar results are found to support the models we developed before, in which MOS-LQO metric is used.

In conclusion, it is the measurement algorithms not the speech databases that contribute to the quality rating difference.

Given that currently there are no other widely accepted objective measures and subjective MOS test is time-consuming and expensive, using MOS-LQO or PESQ is probably the best approach to measure the quality of a huge number of speech samples. Although PESQ-based algorithms are not verified to evaluate the effects of constraint loss, they may be used with cautions. Moreover, the analysis also shows that PESQ-based algorithms may not accurately predict the speech quality under the following situations, where they also should be used with caution:

- Bursty loss for G.711 without PLC. The effects of click frequency may outweigh the effects of losing consecutive packets in subjective test. This may not be reflected in the objective MOS.
- Temporal clipping. The subjective MOS does not show a clear relationship between amount of clipping and speech quality. Perhaps clipping to some syllabus may not be perceived by a listener in ACR MOS test. But in the objective measures, the difference between two signals is detected and mapped to a quality score.
- Different noise spectrums. The subjective MOS reveals big difference in effect for several noises. However, the objective MOS seems to narrow this difference down.

8.5. Model Calibration

Based on the subjective MOS database, parts of the proposed speech quality assessment models are calibrated, and the results are reported in this section.

For the database, samples in experiments 1, 2 and 3 are used as the training set (75% of the total samples), and samples in experiment 4 are used as the testing set (25% of the total samples). Note that samples in the four experiments are randomized from talkers, as stated in Section 8.2. Similarly, the training set is used to calibrate the models and the testing set is then used to evaluate the performance of the calibrated models.

The models using subjective MOS would be somewhat different from those using the MOS-LQO, as expected. For the effects of packet loss, random loss and bursty loss models adopt the same forms of the formulas (6.8) and (6.9). That is, using a third-order polynomial to represent the effect of packet loss rate; and using an exponential function to represent the effect of loss burstiness. For the effect of loss localization (i.e., voiced, unvoiced), only 2% loss all on voiced packets and 4% loss all on unvoiced packets are studied in the subjective test. Although the idea that loss landing on the voiced packets is more detrimental to speech quality is justified, there is not enough data to calibrate the model. For the effects of temporal clipping, the subjective results in general do not support the relationships between amount of clipping and speech quality, the subjective model is therefore not developed here. Further subjective tests are needed for the effect of temporal clipping. For the effects of noise with packet loss, the same steps in Table 7.3 are used. The calibrated model for packet loss is given in Table 8.6. And, for the different

noise conditions, the equivalent SNR (dB) in the E-model for calculation is given in Table 8.7, following the same methods in subsection 7.4.1.

Table 8.6 Parameters of the calibrated packet loss models

			C_0	C_1	C_2	C_3	a
English							
G.711	PLC	10 ms	0.1386	0.0734	-0.01506	0.00167	0.0521
		20 ms	0.1386	0.1558	-0.02830	0.00237	0.0184
		30 ms	0.1386	-0.0140	0.02997	-0.00185	0.1108
	No PLC	10 ms	0.1386	0.3316	-0.03128	0.00198	-0.1586
		20 ms	0.1386	0.3629	-0.01422	-0.0004	-0.3394
		30 ms	0.1386	0.4014	-0.05063	0.00294	-0.3066
G.729	PLC	10 ms	0.4111	0.0591	0.01616	-0.00102	-0.0125
		20 ms	0.4111	0.1296	0.01257	-0.00136	-0.1273
		30 ms	0.4111	0.0668	0.02689	-0.00206	-0.0746
French							
G.711	PLC	10 ms	0.3098	0.008728	0.004258	0.000247	0.1653
		20 ms	0.3098	0.2096	-0.05612	0.004653	0.0304
		30 ms	0.3098	-0.02975	0.04922	-0.00359	-0.0142
	No PLC	10 ms	0.3098	0.5750	-0.06057	0.00257	-0.3570
		20 ms	0.3098	0.4245	-0.02711	0.000929	-0.2547
		30 ms	0.3098	0.5913	-0.07559	0.003659	-0.1293
G.729	PLC	10 ms	0.5932	0.1012	0.03825	-0.00377	-0.2242
		20 ms	0.5932	0.03253	0.04233	-0.00338	-0.0496
		30 ms	0.5932	0.2669	-0.02464	0.001228	-0.0672

Table 8.7 Equivalent SNR in the E-model

	Hoth 20 dB	Hoth 10 dB	Hoth 0 dB	Car 20 dB	Car 10 dB	Street 20 dB	Street 10 dB	Babble 20 dB	Babble 10 dB
English	25	15	-5	29	26	24	12	23	19
French	30	17	-4	36	29	31	18	32	24

As seen from the two tables above, there are also some differences between the models developed from the two languages, as the ratings from two databases are different. Moreover, as a of G.729 with PLC is slightly negative, it means that bursty loss results in

better speech quality relative to random loss. The effects are not pronounced and are hardly observed from Figure 8.8. For the noise effect, it is seen that using the E-model additivity assumption will underestimate the speech quality except for Hoth 0 dB case. Finally, the performance of the calibrated models is tested for the packet loss, bursty loss, noise and noise with packet loss cases; they are a subset of the whole database. Note that codec tandem cases are excluded from the evaluation. The results are reasonable, as given in Table 8.8.

Table 8.8 Performance of the calibrated model

	Random loss (50 cases)	Bursty loss (50 cases)	Noise alone (27 cases)	Noise with loss (48 cases)	All (175 cases)
English					
ρ	0.8558	0.8686	0.9053	0.7961	0.9331
σ	0.2873	0.2560	0.3730	0.2570	0.2861
French					
ρ	0.8824	0.7827	0.9467	0.8111	0.9096
σ	0.3240	0.3656	0.2864	0.2983	0.3244

8.6. Discussion and Summary

In this chapter, a subjective MOS database is developed as a joint project from several parties. The database contains 9 categories of typical VoIP conditions, and includes 2592 speech samples, half in English and half in French. It is used to evaluate the proposed non-intrusive algorithm, and to verify some ideas suggested in the thesis. The subjective MOS is collected under elaborate lab settings. For the speech database as a whole, the subjective MOS results meet our expectation.

The MOS rating is analyzed against the corresponding objective MOS measured by several metrics, including MOS-LQO, PESQ and P.563. For the talk wise, the analysis

shows that there is little difference among four talkers. Therefore, per case analysis is conducted to examine the correlation between subjective MOS and objective ones. Similar performance is observed for the two PESQ-based metrics, MOS-LQO and PESQ. The results from P.563 are moderate as P.563 has no knowledge of the reference signal. The database is further divided into two groups, that is, a target group and a non-target group. As examined by cross correlation, standard deviation and error distributions, it is seen that PESQ-based algorithms work well for the target group. Also, they can be used to measure the impact of packet loss localization. The idea that loss landed on voiced packet is more detrimental to speech quality than that landed on unvoiced packet is verified. For the effects of bursty constraint loss, temporal clipping and noise suppression algorithms, their performance needs further validation. Furthermore, the performance limitation of the PESQ-based algorithms is also pointed out for bursty loss for G.711 without PLC and different noise spectrums.

Finally, based on the subjective MOS database, calibrated models are developed and tested. The models mainly cover the effects of packet loss and noise scenarios, and show reasonable performance. For the effects of temporal clipping, further validation work is needed.

CHAPTER 9 CONCLUSIONS AND FUTURE WORK

9.1. Conclusions

This thesis has developed a non-intrusive, single-ended speech quality classification and assessment algorithm in VoIP environment. Several major impairments, including packet loss, temporal clipping, noise and echo are investigated in the thesis. The primary application of the research is in-service, non-intrusive VoIP speech quality monitoring and testing.

With rapid deployment of VoIP service, there is a need to assess its voice quality as it is not guaranteed. Evaluating speech quality in a non-intrusive fashion is challenging in that only a degraded signal is available for analysis, which is the case for most of the on-line, live network monitoring. The subjective MOS test is not suitable for this purpose, due to its time-consuming and expensive nature. Objective intrusive algorithms do not work either, because they require a reference speech signal to compare with. Some existing non-intrusive algorithms operate on voice payload, and they generally work in a listening-only mode and require extensive computations. Other non-intrusive algorithms utilize IP protocol analysis approach and cannot take into account the impairments from voice payload itself.

In the thesis, a novel non-intrusive VoIP speech quality assessment structure is proposed. The algorithm investigates the individual as well as combined effects of several major impairments in VoIP. It adopts a three-step strategy, namely, impairment

detection, individual effect modeling and an overall model. The algorithm combines the merits of the voice payload analysis and Internet protocol analysis approaches, and further incorporates the noise and echo perception models in the ITU-T E-model to build the overall assessment model. The proposed model does not require extra core network resources and the computational complexity is quite low. An assessor using this algorithm can be implemented at the receive-end media gateway or IP terminal for on-line network performance monitoring and troubleshooting purposes.

For the temporal clipping, an algorithm is developed to map clipping statistics to speech quality by exploiting clipping locations. By using MOS-LQO as a subjective MOS estimate, the research identifies that FEC has the most severe impact on speech quality. The algorithm shows excellent performance for different frame sizes and CN spectrums, the correlation coefficient ρ between the measurement and prediction is about 0.97.

For the echo detection, two cross-correlation based algorithms are proposed, which is realized by measuring two echo parameters, echo path delay and EPL. The echo detection in VoIP is challenging in that echo suffers from excessive delay and nonlinear distortion. The first algorithm is based on a downsampling approach; the second one uses a sparse window to keep portions of speech samples at regular intervals. They both successfully reduce the computational requirements while maintaining good accuracy. When compared to the reference condition, up to 98% computational efforts is saved. The performance under codec distortion, packet loss, noise and double talk conditions is examined through simulations and real field measurements. The results show that,

compared with double talk, background noise, codec distortion and moderate packet loss have limited impacts on measurement. On the other hand, double talk, which acts as a high power noise, dominates the measurement performance. The algorithms also have sufficient resolution and accuracy to deal with delay jitter and multi-path echo.

For the packet loss and codec distortion, a new scheme is proposed to classify the lost packets into silence, unvoiced and voiced by exploiting correctly received neighbouring packets. Based on a packet loss profile, which consists of codec type, packet size, loss rate, loss distribution and PLC, a novel quality model is developed. The base speech quality is modeled as a third-order polynomial function of loss rate. Then, a linear interpolation approach is used to represent the effects of loss location weights of voiced and unvoiced packets. In order to accommodate the bursty loss, a codec dependent factor is also proposed. The research examines two prevalent VoIP codecs, G.711 and G.729A. The results show excellent performance as well. For G.711, ρ is 0.91 and σ is 0.26 MOS; for G.729A, ρ is 0.88 and σ is 0.28 MOS. Also, it is seen that voice dominated loss results in worse speech quality, compared to unvoice dominated loss.

For the overall model, the combined effects of packet loss and temporal clipping are investigated. It is shown that they are dependent and thus cannot be simply added in the MOS domain. Instead, the correlation between them is examined and a new term is added to correct the bias of summation. Then, the effect of background noise is considered for the listening-only model. A noise measurement method is adopted that uses a two-step approach, rough SNR estimation, followed by SNR adjustment, to take into account different noise levels. The measured SNR is then used for noise power calculation and

effect modeling in the E-model. In the listening-only model, the research shows that the noise additivity assumption in the E-model does not hold well. For SNR $\geq$ 20 dB, a compensation curve is suggested. When SNR < 20 dB, using the E-model is unsatisfactory. The performance of the listening-only model is evaluated. When SNR $\geq$ 20 dB, the results show that ρ is about 0.90, and σ is 0.27 MOS.

In the thesis, the non-intrusive algorithm is developed through extensive simulation. Tens of thousands of degraded speech samples are generated for different impairment conditions. MOS-LQO is used to measure the resulting speech quality as it is not feasible to conduct subjective testing in such a large scale. Meanwhile, a subjective MOS test that covers small but key impairment scenarios is prepared. The subjective MOS rating is analyzed and compared with the corresponding MOS-LQO. On the one hand, the results show that MOS-LQO is generally acceptable in VoIP speech quality testing for those impairments it is designed to measure. Also, the idea that loss landed on voiced part degrades the speech quality more than those landed on unvoiced part is justified. On the other hand, the analysis shows that MOS-LQO is limited in scenarios including bursty loss for G.711 without PLC, temporal clipping and different noise spectrums. Finally, the real MOS results are used to calibrate the proposed algorithm in a small scale.

Modern telecommunication networks are becoming more and more complex nowadays. From a speech quality point of view, the effects of many impairments are correlated, such as the packet loss and temporal clipping we examined in the thesis. The proposed non-intrusive algorithm mainly utilizes parameters extracted from voice payload and Internet protocols, that is, it is rather a parametric model than an acoustic

processing model. One of the simplifications in the parametric models is that they do not take into account the acoustic variations of individual talkers. Under the same degradation conditions, speeches from different talkers are assumed to be of the same speech quality. This result can be interpreted as an expected average speech quality, in the statistical sense. On the other hand, the parametric models have merit in that they are simple to implement; also, they reveal the contributions from different impairments and their interactions, which is quite useful for tuning up the networks.

The speech database used for this thesis is generated from simulations, and only a small subjective MOS database is available. Although these simulated degradations occur in real VoIP networks, the model needs further validation with speech samples collected under real network conditions and all using subjective MOS results. Also, the model requires further validation for a wider range of packet loss, clipping and noise conditions, including combinations of these effects.

Noise and echo are not specific to VoIP networks. The E-model is used to quantify their effects and further to integrate them into the overall assessment model. The assumed additivity in the psychological scale needs further verification. Moreover, the E-model is simplistic to some extent. For the noise, factors such as stationarity and noise spectrum also affect speech quality. For example, car noise, whose strong low frequency content will be heavily attenuated by a telephone handset, exhibits different effects on speech quality relative to Hoth noise as justified by the subjective test. These are the limitations of the E-model and it needs further validation. For the conversational aspects of the model, subjective MOS tests are needed for model validation purposes.

9.2. Future Work

Future research work will be carried out in the following several directions. First of all, as mentioned, the model requires further validation for a wider range of packet loss, clipping and noise conditions, including combinations of these effects. The model needs further validation with speech samples collected under real network conditions and all using subjective MOS results.

The temporal clipping investigated in the thesis is introduced by using energy based VAD. As mentioned in Section 4.5, the actual VAD could be complicated. Under the same percentage of clippings, the MSC distribution could be different. It is beneficial to verify the performance of the proposed temporal clipping effect model under other VAD algorithms.

For the overall assessment, it is suggested that the noise additivity assumption doesn't work well under high noise scenarios in Section 7.4. Further work can be done to enhance the performance of the E-model or our developed overall assessment model. One approach could be first determining a dominant impairment and using its effect as an anchor, then assigning different weights to other impairments for an overall model.

In addition to echo studied in the thesis, there are other factors that affect conversational quality, such as end-to-end delay and double-talk. The round-trip delay information may be obtained from RTCP-XR if it is accessible, or it can be estimated by the echo path delay obtained from the developed echo measurement module. By assuming symmetric echo path, end-to-end delay is half of the round-trip delay and its effects on conversation quality can be quantified by the E-model. Even though, recent

research [148] suggested that the E-Model is somewhat pessimistic in predicting user opinion of voice quality for one-way echo-free pure delay exceeding 150 ms. The delay measuring method and subjective effect of one-way delay deserve further research. How double talk affects conversation also needs further investigation.

More research is now being carried out on developing efficient, accurate non-intrusive assessment algorithms for speech quality, particularly in VoIP. Finally, future work will focus on extension of the model to cover other impairment types, such as delay jitter, amplitude clipping, and other prevalent VoIP codecs, using the same structure. In addition to its diagnostic functionalities, specific speech quality enhancement routines, such as adaptive noise cancellation, call flow routing and codec configuration selection, could be added to the assessment algorithm to enhance the VoIP speech quality.

REFERENCES

[1] U. Black, *Voice over IP*, Upper Saddle River, NJ: Prentice Hall, 2000.

[2] D. Minoli and E. Minoli, *Delivering Voice over IP Networks*, 2nd ed., New York, NY: John Wiley & Sons, 2002.

[3] A. P. Markopoulou, F. A. Tobagi and M. J. Karam, "Assessing the quality of voice communications over Internet backbones," *IEEE/ACM Trans. Networking*, vol. 11, no. 5, pp. 747-760, October 2003.

[4] B. Goode, "Voice over Internet protocol (VoIP)," *Proc. IEEE*, vol. 90, no. 9, pp. 1495-1517, September 2002.

[5] R. G. Cole and J. H. Rosenbluth, "Voice over IP performance monitoring," *J. ACM SIGCOMM Comput. Commun. Rev.*, vol. 31, no. 2, pp. 9-24, April 2001.

[6] ITU-T Rec. P.800, *Methods for subjective determination of transmission quality*, August 1996.

[7] A. W. Rix, "Perceptual speech quality assessment – a review," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing,* May 2004, vol. 3, pp. 1056-1059.

[8] S. Möller and A. Raake, "Telephone speech quality prediction: Towards network planning and monitoring models for modern network scenarios," *Speech Commun.*, vol. 38, no. 1-2, pp. 47-75, September 2002.

[9] A. W. Rix and M. P. Hollier, "The perceptual analysis measurement system for robust end-to-end speech quality assessment," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing,* June 2000, vol. 3, pp. 1515-1518.

[10] ITU-T Rec. P.862, *Perceptual Evaluation of Speech Quality (PESQ): An objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs*, February 2001.

[11] ITU-T Rec. G.107, *The E-Model, a computational model for use in transmission planning*, March 2005.

[12] ITU-T Rec. P.563, *Single-ended method for objective speech quality assessment in narrow-band telephony applications,* May 2004.

[13] A. D. Clark, Telchemy Inc. "Description of VQMon algorithm," ITU-T Study Group 12, Delayed Contribution, COM12-D105, January 2003.

[14] S. Broom, Psytechnics Ltd., "High level description of Psytechnics ITU-T P.VTQ candidate," ITU-T Study Group 12, Delayed Contribution, COM12-D175, September 2003.

[15] M. S. El-Hennawey, R. A. Goubran, A. Radwan and L. Ding, "Method and apparatus for non-intrusive single-ended voice quality assessment in VoIP," International patent publication no. WO2006/035269, April 6, 2006.

[16] IETF RFC 3611, RTP Control Protocol Extended Reports (RTCP-XR), November 2003.

[17] A. Radwan, "Non-intrusive speech quality assessment in VoIP," M.A.Sc. thesis, Carleton University, Ottawa, Canada, August 2003.

[18] L. Ding, Z. Lin, A. Radwan. M. S. El-Hennawey and R. A. Goubran, "Non-intrusive single-ended speech quality assessment in VoIP," *Speech Comm.*, accepted for publication in April 2007.

[19] ITU-T Rec. P.862.1, *Mapping function for transforming P.862 raw result scores to MOS-LQO*, November 2003.

[20] L. Ding, M. S. El-Hennawey and R. A. Goubran, "Measurement of the effects of temporal clipping on speech quality," in *Proc. IEEE Instrumentation and Measurement Technology Conf.*, May 2005, vol. 2, pp. 1135-1138.

[21] L. Ding, A. Radwan, M. S. El-Hennawey and R. A. Goubran, "Measurement of the effects of temporal clipping on speech quality," *Special Issue of IEEE Trans. Instrum. Meas.*, vol. 55, no. 4, pp. 1197-1203, August 2006.

[22] L. Ding, M. S. El-Hennawey and R. A. Goubran, "Non-intrusive measurement of echo-path parameters in VoIP environments," *IEEE Trans. Instrum. Meas.*, vol. 55, no. 6, pp. 2062-2071, December 2006.

[23] L. Ding and R. A. Goubran, "Speech quality prediction in VoIP using the extended E-Model," in *Proc. IEEE Global Telecommunications Conf.*, December 2003, vol. 7, pp. 3974-3978.

[24] M. S. El-Hennawey, R. A. Goubran, A. Radwan and L. Ding, "Method and apparatus for non-intrusive single-ended voice quality assessment in VoIP," International patent publication no. WO2006/136900, December 28, 2006.

[25] Nortel Networks, "An algorithm to enhance non-intrusive speech quality assessment testing capabilities," ITU-T Study Group 12, Delayed Contribution, COM12-D161, May 2006.

[26] L. Ding, A. Radwan, M. S. El-Hennawey and R. A. Goubran, "Performance study of objective voice quality measures in VoIP," in *Proc. IEEE Sym. Computers and Communications*, to be published in July 2007.

[27] IETF RFC 3550, RTP: A transport protocol for real-time applications, July 2003.

[28] M. Hassan, A. Nayandoro and M. Atiquzzaman, "Internet telephony: services, technical challenges, and products," *IEEE Commun. Mag.*, vol. 38, no. 4, pp. 96-103, April 2000.

[29] ITU-T Rec. G.113 Appendix I, *Provisional planning values for the equipment impairment factor Ie and packet-loss robustness factor Bpl*, May 2002.

[30] B. W. Wah, X. Su and D. Lin, "A survey of error-concealment schemes for real-time audio and video transmissions over the Internet," in *Proc. IEEE Int. Sym. Multimedia Software Engineering*, December 2000, pp. 17-24.

[31] C. Perkins, O. Hodson and V. Hardman, "A survey of packet loss recovery techniques for streaming audio," *IEEE Network*, vol. 12, no. 5, pp. 40-48, September-October 1998.

[32] S. G. Tanyer and H. Özer, "Voice activity detection in nonstationary noise," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 4, pp. 478-482, July 2000.

[33] C. Rowden, *Speech Processing*, New York, NY: McGraw-Hill, 1992.

[34] IETF RFC 3389, Real-time Transport Protocol (RTP) payload for Comfort Noise (CN), September 2002.

[35] R. V. Prasad, A. Sangwan, H. S. Jamadagni, M. C. Chiranth, R. Sah and V. Gaurav, "Comparison of voice activity detection algorithms for VoIP," in *Proc. IEEE 7th Int. Sym. Computers and Communications*, July 2002, pp. 530-535.

[36] Empirix Inc., "Hammer VoIP test system echo detection and analysis," 2001. Available: http://wireless.feld.cvut.cz/mesaqin2002/Echo_Detection.pdf

[37] W. P. Ng, J. M. H. Elmirghani and S. Broom, "Parallel DAF measurement device (PDMD) for non-intrusive whitening of speech communications," in *Proc. IEEE Int. Conf. Communications,* June 2001, vol. 4, pp. 1052-1056.

[38] ITU-T Rec. G.126, *Listener echo in telephone networks,* March 1993.

[39] R. J. B. Reynolds and A. W. Rix, "Quality VoIP – an engineering challenge," *BT Technol. J.,* vol. 19, no. 2, pp. 23-32, April 2001.

[40] S. Pracht and D. Hardman, "Voice quality in converging telephony and IP networks," Agilent Technologies, May 2001. Available: http://literature.agilent.com/litweb/pdf/5980-0989E.pdf

[41] ITU-T Rec. G.131, *Talker echo and its control,* November 2003.

[42] ITU-T Rec. G.168, *Digital network echo cancellers,* August 2004.

[43] Y. Huang and R. A. Goubran, "Effects of vocoder distortion on network echo cancellation," in *Proc. IEEE Int. Conf. Multimedia and Expo*, July-August 2000, vol. 1, pp. 437-439.

[44] T. Yensen, J. P. Lariviere, I. Lambadaris and R. A. Goubran, "HMM delay prediction technique for VoIP", *IEEE Trans. Multimedia*, vol. 5, no. 3, pp. 444-457, September 2003.

[45] ITU-T Rec. G.108, *Application of the E-model: A planning guide*, September 1999.

[46] ITU-T Rec. G.114, *One-way transmission time*, May 2003.

[47] Alcatel, "Voice over DSL quality study," 2001. Available: http://www.alcatel.com/doctypes/articlepaperlibrary/pdf/VODSLqos.pdf

[48] IETF RFC 3551, RTP profile for audio and video conferences with minimal control, July 2003.

[49] R. V. Cox, B. G. Haskell, Y. Lecun, B. Shahraray and L. Rabiner, "On the applications of multimedia processing to communications," *Proc. IEEE*, vol. 86, no. 5, pp. 755-824, May 1998.

[50] R. Ramjee, J. Kurose, D. Towsley and H. Schulzrinne, "Adaptive playout mechanisms for packetized audio applications in wide-area networks," in *Proc. IEEE INFOCOM*, June 1994, vol. 2, pp. 680-688.

[51] J. D. Rosenberg, L. Qiu and H. Schulzrinne, "Integrating packet FEC into adaptive voice playout buffer algorithms on the Internet," in *Proc. IEEE INFOCOM*, March 2000, vol. 3, pp. 1705-1714.

[52] A. S. Spanias, "Speech coding: a tutorial review," *Proc. IEEE*, vol. 82, no. 10, pp. 1541-1582, October 1994.

[53] M. E. Perkins, K. Evans, D. Pascal and L. A. Thorpe, "Characterizing the subjective performance of the ITU-T 8 kb/s speech coding algorithm – ITU-T G.729," *IEEE Commun. Mag.*, vol. 35, no. 9, pp. 74-81, September 1997.

[54] ITU-T Rec. G.113, *Transmission impairments due to speech processing*, February 2001.

[55] ITU-T Rec. P.834, *Methodology for the derivation of equipment impairment factors from instrumental models*, July 2002.

[56] S. Moller and J. Berger, "Describing telephone speech codec quality degradations by means of impairment factors," *J. Audio Eng. Soc.* vol. 50, no. 9, pp. 667-680, September 2002.

[57] J. Anderson, "Methods for measuring perceptual speech quality," Agilent Technologies, October 2001. Available: *http://literature.agilent.com/litweb/pdf/5988-2352EN.pdf*

[58] ITU-T Rec. P.830, *Subjective performance assessment of telephone-band and wideband digital codecs,* February 1996.

[59] P. Denisowski, "How does it sound?" *IEEE Spectr.*, vol. 38, no. 2, pp. 60-64, February 2001.

[60] W. Yang, "Enhanced Modified Bark Spectral Distortion (EMBSD): an objective speech quality measure based on audible distortion and cognition model," Ph.D. Dissertation, Temple University, Philadelphia, USA, May 1999.

[61] K. H. Lam, O. C. Au, C. C. Chan, K. F. Hui and S. F. Lau, "Objective speech quality measure for cellular phone," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, May 1996, vol. 1, pp. 487-490.

[62] S. Wang, A. Sekey and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE J. Select. Areas Commun.*, vol. 10, no. 5, pp. 819-829, June 1992.

[63] J. G. Beerends and J. A. Stemerdink, "A perceptual speech-quality measure based on psychoacoustic sound presentation," *J. Audio Eng. Soc.*, vol. 42, no. 3, pp. 115-123, March 1994.

[64] ITU-T Rec. P.861, *Objective quality measurement of telephone-band (300-3400 Hz) speech codecs*, February 1998.

[65] S. Voran, "Objective estimation of perceived speech quality – Part I: Development of the measuring normalizing block technique," *IEEE Trans. Speech Audio Processing*, vol. 7, no. 4, pp. 371-382, July 1999.

[66] S. Voran, "Objective estimation of perceived speech quality – Part II: Evaluation of the measuring normalizing block technique," *IEEE Trans. Speech Audio Processing*, vol. 7, no. 4, pp. 383-390, July 1999.

[67] A. W. Rix, M. P. Hollier, A. P. Hekstra and J. G. Beerends, "Perceptual Evaluation of Speech Quality (PESQ): The new ITU standard for end-to-end speech quality assessment Part I – time-delay compensation," *J. Audio Eng. Soc.*, vol. 50, no. 10, pp. 755-764, October 2002.

[68] J. G. Beerends, A. P. Hekstra, A. W. Rix and M. P. Hollier, "Perceptual Evaluation of Speech Quality (PESQ): The new ITU standard for end-to-end speech quality assessment Part II – psychoacoustic model," *J. Audio Eng. Soc.*, vol. 50, no. 10, pp. 765-778, October 2002.

[69] L. A. Thorpe and W. Yang, "Performance of current perceptual objective speech quality measures," in *Proc. IEEE Workshop Speech Coding,* June 1999, pp. 144-146.

[70] ITU-T Rec. P.562, *Analysis and interpretation of INMD voice-service measurements,* May 2004.

[71] M. S. El-Hennawey and D. Lee, "Embedded real-time voice quality analysis system," in *GSPx: The International Embedded Solutions Event*, Sept. 2004.

[72] J. Anderson, "Addressing VoIP speech quality with non-intrusive measurements," Agilent Technologies, September 2002. Available: http://literature.agilent.com/litweb/pdf/5988-7878EN.pdf

[73] A. Takahashi, H. Yoshino and N. Kitawaki, "Perceptual QoS assessment technologies for VoIP," *IEEE Commun. Mag.*, vol. 42, no. 7, pp. 28-34, July 2004.

[74] ETSI Guide, EG 201 377-3 V1.1.1, *Speech processing, Transmission and Quality aspects (STQ); Specification and measurement of speech transmission quality; Part 3: Non-intrusive objective measurement methods applicable to networks and links with classes of services,* June 2003.

[75] Opticom GmbH, "3SQM™ advanced non-intrusive voice quality testing," June 2004. Available: www.opticom.de/download/3SQM-WP-290604.pdf

[76] Psytechnics Ltd., UK, "NiQA – Non-intrusive speech quality assessment," ITU-T Study Group 12, Delayed Contribution, COM12-D48, October 2001.

[77] SwissQual Inc., Switzerland, "Non-intrusive speech quality measurement," ITU-T Study Group 12, Contribution, COM12-27, July 2001.

[78] P. Gray, M. P. Hollier and R. E. Massara, "Non-intrusive speech-quality assessment using vocal-tract models," *IEE Proc. Vision, Image, Signal Processing*, vol. 147, no. 6, pp. 493-501, December 2000.

[79] M. S. El-Hennawey and D. Lee, "Embedded real-time voice quality analysis system," in *GSPx: The International Embedded Solutions Event*, September 2004.

[80] J. Liang and R. F. Kubichek, "Output-based objective speech quality," in *Proc. IEEE 44th Vehicular Technology Conf.*, June 1994, vol. 3, pp. 1719-1723.

[81] C. Jin and R. F. Kubichek, "Vector quantization techniques for output-based objective speech quality," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, May 1996, vol. 1, pp. 491-494.

[82] W. Li and R. F. Kubichek, "Output-based objective speech quality measurement using continuous hidden Markov models," in *Proc. IEEE 7th Int. Sym. Signal Processing and Its Applications,* July 2003, vol. 1, pp. 389-392.

[83] T. H. Falk, Q. Xu and W.-Y. Chan, "Non-intrusive GMM-based speech quality measurement," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing,* March 2005, vol. 1, pp. 125-128.

[84] T. H. Falk and W.-Y. Chan, "Nonintrusive speech quality estimation using Gaussian mixture models," in *IEEE Signal Processing Lett.,* vol. 13, no. 2, pp. 108-111, February 2006.

[85] A. D. Clark, "Modeling the effects of burst packet loss and recency on subjective voice quality," in *Proc. 2nd Internet Telephony Workshop*, April 2001, pp. 123-127.

[86] ETSI Technical Specification, TS 101 329-5 V1.1.2, *Telecommunications and Internet Protocol Harmonization Over Networks (TIPHON) Release 3; End-to-end quality of service in TIPHON systems; Part 5: Quality of Service (QoS) measurement methodologies,* January 2002.

[87] S. Broom and M. P. Hollier, "Speech quality measurement tools for dynamic network management," in *Proc. Measurement Speech Audio Quality in Networks Workshop*, June 2003.

[88] N. O. Johannesson, "The ETSI computation model: a tool for transmission planning of telephone networks," *IEEE Commun. Mag.*, vol. 35, no. 1, pp. 70-79, January 1997.

[89] ETSI Technical Report, ETR 250, *Transmission and Multiplexing (TM); Speech communication quality from mouth to ear for 3.1kHz handset telephony across networks*, July 1996.

[90] N. Osaka, K. Kakehi, S. Iai and N. Kitawaki, "A model for evaluating talker echo and sidetone in a telephone transmission network," *IEEE Trans. Commun.*, vol. 40, no. 11, pp. 1684-1692, November 1992.

[91] ITU-T Rec. G.109, *Definition of categories of speech transmission quality*, September 1999.

[92] ETSI Guide, EG 201 377-1 V1.2.1, *Speech Processing, Transmission and Quality Aspects (STQ); Specification and measurement of speech transmission quality; Part 1: Introduction to objective comparison measurement methods for one-way speech quality across networks,* December 2002.

[93] ITU-T Rec. P.561, *In-service, non-intrusive measurement device – Voice service measurements,* July 2002.

[94] KPN, Netherlands, "Improvement of the P.861 perceptual speech quality measure," ITU-T Study Group 12, Contribution, COM12-20, December 1997.

[95] A. W. Rix, J. G. Beerends, M. P. Hollier and A. P. Hekstra, "PESQ – the new ITU standard for end-to-end speech quality assessment," in *Proc. 109th Audio Eng. Soc. Convention*, September 2000, pre-print no. 5260.

[96] ITU-T Rec. P. 10, *Vocabulary of terms on telephone transmission quality and telephone sets*, December 1998.

[97] A. W. Rix, "Comparison between subjective listening quality and P.862 PESQ score," in *Proc. Measurement Speech Audio Quality in Networks Workshop*, June 2003.

[98] A. W. Rix, Psytechnics Ltd., UK, "A new PESQ-LQ scale to assist comparison between P.862 PESQ score and subjective MOS," ITU-T Study Group 12, Delayed Contribution, COM12-D86, May 2002.

[99] *Digital Speech Level Analyzer, User Guide*, Revision 4.0, Malden Electronics Ltd., UK, 2003.

[100] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett and N. L. Dahlgren, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM," National Institute of Standards and Technology, October 1990.

[101] ITU-T Rec. G.191, *Software tools for speech and audio coding standardization,* November 2000.

[102] ITU-T Rec. P.56, *Objective measurement of active speech level*, March 1993.

[103] P. Murrin, D. M. Howard, A. M. Tyrrell and P. Barrett, "Objective measure of performance of voice activity detectors," *IEE Electron. Lett.*, vol. 35, no. 22, pp. 1922-1923, October 1999.

[104] F. Beritelli, S. Casale, and G. Ruggeri, "A psychoacoustic auditory model to evaluate the performance of a voice activity detector", *J. Signal Processing*, vol. 80, no. 7, pp. 1393-1397, July 2000.

[105] F. Beritelli, S. Casale, G. Ruggeri and S. Serrano, "Performance evaluation and comparison of G.729/AMR/fuzzy voice activity detectors," *IEEE Signal Processing Lett.*, vol. 9, no. 3, pp. 85-88, March 2002.

[106] F. Beritelli, S. Casale and A. Cavallaero, "A robust voice activity detector for wireless communications using soft computing," *IEEE J. Select. Areas Commun.*, vol. 16, no. 9, pp. 1818-1829, December 1998.

[107] D. K. Freeman, G. Cosier, C. B. Southcott and I. Boyd, "The voice activity detector for the Pan-European digital cellular mobile telephone service," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing,* May 1989, vol. 1, pp. 369-372.

[108] ITU-T Rec. G.729 Annex B, *A silence compression scheme for G.729 optimized for terminals conforming to Recommendation V.70,* October 1996.

[109] ITU-T Rec. G.711 Appendix II, *A comfort noise payload definition for ITU-T G.711 use in packet-based multimedia communication systems,* February 2000.

[110] ITU-T Rec. G.723.1 Annex A, *Silence compression scheme,* November 1996.

[111] M. Bertocco and P. Paglierani, "In-service nonintrusive measurement of echo parameters in telephone-type networks," *IEEE Trans. Instrum. Meas.*, vol. 47, no. 5, pp. 1322-1325, October 1998.

[112] T. Gansler and G. Salomonsson, "Nonintrusive measurements of the telephone channel," *IEEE Trans. Commun.*, vol. 47, no. 1, pp. 158-167, January 1999.

[113] H. Yasukawa and K. Watanabe, "Subband adaptive digital filter for impulse response estimation of huge tap system," in *Proc. IEEE Instrumentation and Measurement Technology Conf.*, April 1995, pp. 472-475.

[114] Q. Jin, Z. Q. Luo, and K. M. Wong, "Optimum filter banks for signal decomposition and its application in adaptive echo cancellation," *IEEE Trans. Signal Processing*, vol. 44, no. 7, pp. 1669-1680, July 1996.

[115] J. Liu, "Efficient and robust cancellation of echoes with long echo path delay," *IEEE Trans. Commun.*, vol. 52, no. 8, pp. 1288-1291, August 2004.

[116] J. R. Cavanaugh, R. W. Hatch and J. L. Sullivan, "Models for the subjective effects of loss, noise and talker echo on telephone connections," *Bell Syst. Tech. J.*, vol. 55, no. 9, pp. 1319-1371, November 1976.

[117] R. Appel and J. G. Beerends, "On the quality of hearing one's own voice," *J. Audio Eng. Soc.*, vol. 50, no. 4, pp. 237-248, April 2002.

[118] L. R. Rabiner and M. R. Sambur, "Application of an LPC distance measure to the voiced-unvoiced-silence detection problem," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. 25, no. 4, pp. 338-343, August 1977.

[119] T.-A. Vu, H. Ding and M. Bouchard, "A Survey of double-talk detection schemes for echo cancellation applications", *J. Canadian Acoustical Association,* vol. 32, no. 3, pp. 144-145, October 2004.

[120] D. L. Duttweiler, "A twelve-channel digital echo canceler," *IEEE Trans. Commun.*, vol. 26, no. 5, pp. 647-653, May 1978.

[121] Y. Suzuki and H. Takeshima, "Equal-loudness-level contours for pure tones," *J. Acoust. Soc. Am.*, vol. 116, no. 2, pp. 918-933, August 2004.

[122] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing: Principles, Algorithms, and Applications*, Upper Saddle River, NJ: Prentice-Hall, 1996.

[123] M. Azaria and D. Hertz, "Time delay estimation by generalized cross correlation methods," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. 32, no. 2, pp. 280-285, Apr. 1984.

[124] Adobe Audition, http://www.adobe.com/products/audition/overview.html.

[125] L. Sun, G. Wade, B. M. Lines and E. C. Ifeachor, "Impact of packet loss location on perceived speech quality," in *Proc. 2nd Internet Telephony Workshop*, April 2001, pp. 114-122.

[126] H. Sanneck, N. T. L. Le and A. Wolisz, "Intra-flow loss recovery and control for VoIP," in *Proc. ACM Int. Multimedia Conf.*, September 2001, vol. 4, pp. 441-454.

[127] C. Hoene and E. Dulamsuren-Lalla, "Predicting performance of PESQ in case of single frame losses," in *Proc. Measurement Speech Audio Quality in Networks Workshop*, June 2004.

[128] B. Duysburgh, S. Vanhastel, B. De Vreese, C. Petrisor and P. Demeester, "On the influence of best-effort network conditions on the perceived speech quality of VoIP connections," in *Proc. 10th IEEE Int. Conf. Computer Communications and Networks*, October 2001, pp. 334-339.

[129] J. C. Bolot, "End-to-end packet delay and loss behavior in the Internet," *J. ACM SIGCOMM Comput. Commun. Rev.*, vol. 23, no. 4, pp. 289-298, October 1993.

[130] M. S. Borella, D. Swider, S. Uludag and G. B. Brewster, "Internet packet loss: measurement and implications for end-to-end QoS," in *Proc. IEEE ICPP Workshop Architectural and OS Support for Multimedia Applications/Flexible Communication Systems/Wireless Networks and Mobile Computing*, August 1998, pp. 3-12.

[131] M. Yajnik, S. Moon, J. Kurose and D. Towsley, "Measurement and modeling of the temporal dependence in packet loss," in *Proc. IEEE INFOCOM*, March 1999, vol. 1, pp. 345-352.

[132] V. Paxson, "End-to-end Internet packet dynamics," *IEEE/ACM Trans. Networking*, vol. 7, no. 3, pp. 277-292, June 1999.

[133] E. N. Gilbert, "Capacity of a burst-noise channel," *Bell Syst. Tech. J.*, vol. 39, no. 9, pp. 1253-1265, September 1960.

[134] P. Billingsley, *Statistical Inference for Markov Processes*, Chicago, IL: The University of Chicago Press, 1961.

[135] A. B. James and B. P. Milner, "An analysis of interleavers for robust speech recognition in burst-like packet loss," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, May 2004, vol. 1, pp. 853-856.

[136] L. Sun and E. C. Ifeachor, "Subjective and objective speech quality evaluation under bursty losses," in *Proc. Measurement Speech Audio Quality in Networks Workshop*, June 2002.

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

[138] ITU-T Rec. G.729, *Coding of speech at 8 kbit/s using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP)*, March 1996.

[139] W. Feng; D. D. Kandlur, D. Saha and K. G. Shin, "Maintaining end-to-end throughput in a differentiated-services Internet," *IEEE/ACM Trans. Networking*, vol. 7, no. 5, pp. 685-697, October 1999.

[140] J. C. De Martin, "Source-driven packet marking for speech transmission over differentiated-services networks," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing,* May 2001, vol. 2, pp. 753-756.

[141] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 32, no. 6, pp. 1109-1121, Dec. 1984.

[142] E. Nemer, R. A. Goubran and S. A. Mahmoud, "Speech enhancement using fourth-order cumulants and optimum filters in the subband domain", *Speech Communication*, vol. 36, no. 3, pp. 219-246, Mar. 2002.

[143] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Lett.,* vol. 9, no. 1, pp. 12-15, Jan. 2002.

[144] Z. Lin and R. A. Goubran, "Musical noise reduction in speech using two-dimensional spectrogram enhancement," in *Proc. 2nd IEEE Int. Workshop on Haptic, Audio and Visual Environments and Their Applications*, Sept. 2003, pp. 61-64.

[145] Initial Draft Recommendation for P.CQO-L, ITU-T Study Group 12, Temporary Document 49, Revision 1, January 2007.

[146] R. Lefebvre and C. Demers, "Subjective tests for the development of automatic QoS control in voice communication networks," Final Report to Nortel, July 2006.

[147] L. Sun and E. Ifeachor, "New models for perceived voice quality prediction and their applications in playout buffer optimization for VoIP networks," in *Proc. IEEE Int. Conf. Communications,* June 2004, vol. 3, pp. 1478-1483.

[148] British Telecom, "Comparison of E-Model and subjective test data for pure-delay conditions," ITU-T Study Group 12, Contribution, COM12-30, January 2007.

APPENDIX A: DETAILED ALGORITHMS OF THE E-MODEL

The reference connection of the E-model is divided into a send side and a receive side, as shown in Figure A.1. The E-model estimates the conversational speech quality from mouth to ear as perceived by the user at the receive side, both as listener and talker.

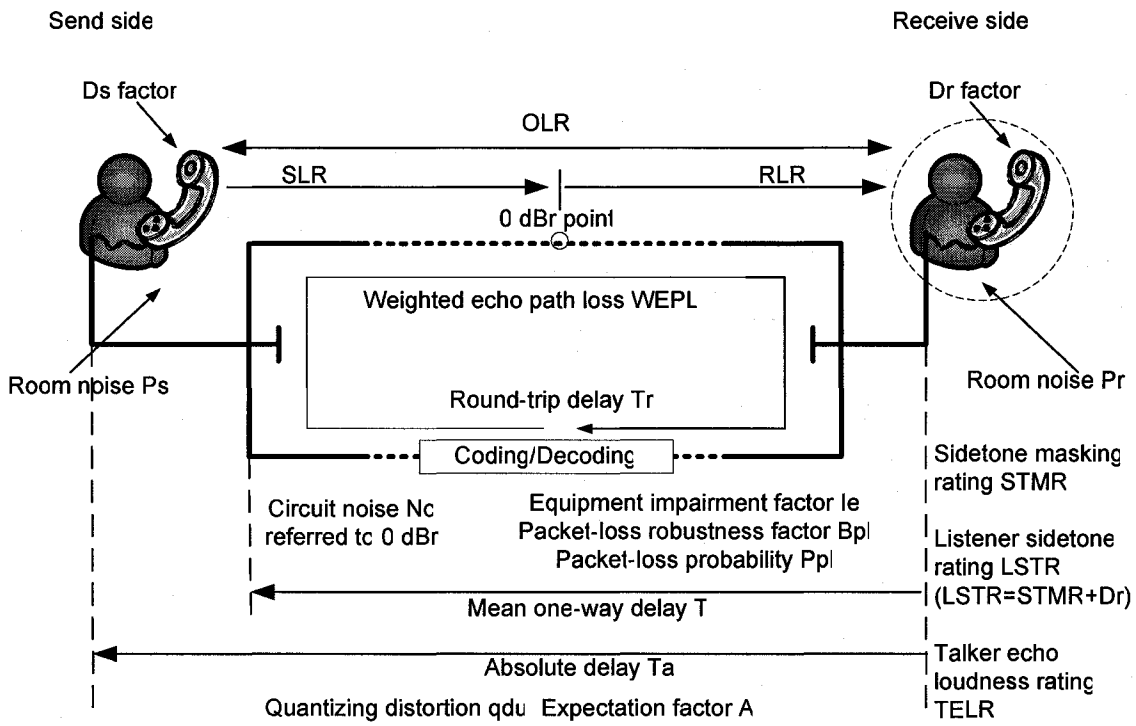


Figure A.1 Reference connection of the E-model from ITU-T Rec. G.107

For all 21 input parameters used in the E-model, their default values and permitted ranges are listed in Table A.1 from ITU-T Rec. G.107 and G.108. Only those parameters in bold letters are usually subject to planning, other parameters can be set to their default values in calculation.

Table A.1 Default values and permitted ranges for the E-model parameters

Parameter	Abbr.	Unit	Default value	Permitted range
Send Loudness Rating	**SLR**	dB	+8	0...+18
Receive Loudness Rating	**RLR**	dB	+2	–5...+14
Sidetone Masking Rating	STMR	dB	15	10...20
Listener Sidetone Rating	LSTR	dB	18	13...23
D-Value of Telephone, Send Side	Ds	–	3	–3...+3
D-Value of Telephone, Receive Side	Dr	–	3	–3...+3
Talker Echo Loudness Rating	**TELR**	dB	65	5...65
Weighted Echo Path Loss	WEPL	dB	110	5...110
Mean one-way Delay of the Echo Path	**T**	ms	0	0...500
Round-Trip Delay in a 4-wire Loop	Tr	ms	0	0...1000
Absolute Delay in echo-free Connections	**Ta**	ms	0	0...500
Number of quantization distortion units	**qdu**	–	1	1...14
Equipment Impairment Factor	**Ie**	–	0	0...40
Packet-loss Robustness Factor	**Bpl**	–	1	1...40
Random Packet-loss Probability	**Ppl**	%	0	0...20
Burst Ratio	**BurstR**	–	1	1...2
Circuit Noise referred to 0 dBr-point	Nc	dBm0p	–70	–80...–40
Noise Floor at the Receive Side	Nfor	dBmp	–64	–
Room Noise at the Send Side	Ps	dB(A)	35	35...85
Room Noise at the Receive Side	Pr	dB(A)	35	35...85
Advantage Factor	**A**	–	0	0...20

Fixed relation: LSTR = STMR+D

The transmission rating factor *R* is given by:

$$R = Ro - Is - Id - Ie\_eff + A \tag{A.1}$$

where *Ro* represents in principle the basic SNR, including circuit noise and room noise. The factor *Is* represents all impairments occurring more or less simultaneously with the voice signal. The factor *Id* represents the impairments caused by delay and echo. The

factor *Ie_eff* represents effective equipment impairment caused by low bit rate codecs. The advantage factor *A* is used for compensation when there are other advantages of access to users.

A.1. Basic Signal-to-Noise Ratio, *Ro*

Ro is given by:

$$Ro = 15 - 1.5(SLR + No) \tag{A.2}$$

The term *No* [in dBm0p] is the power addition of different noise sources:

$$No = 10\log_{10}\left[10^{\frac{Nc}{10}} + 10^{\frac{Nos}{10}} + 10^{\frac{Nor}{10}} + 10^{\frac{Nfo}{10}}\right] \tag{A.3}$$

Nc [in dBm0p] is the sum of all circuit noise powers, all referred to the 0 dBr point.

Nos [in dBm0p] is the equivalent circuit noise at the 0 dBr point, caused by the room noise *Ps* at the send side:

$$Nos = Ps - SLR - Ds - 100 + 0.004(Ps - OLR - Ds - 14)^2 \tag{A.4}$$

where $OLR = SLR + RLR$.

Nor [in dBm0p] is the equivalent circuit noise at the 0 dBr point, caused by the room noise at the receive side:

$$Nor = RLE - 121 + Pre + 0.008(Pre - 35)^2 \tag{A.5}$$

Pre [in dBm0p] is the "effective room noise" caused by the enhancement of *Pr* by the listener's sidetone path:

$$Pre = Pr + 10log_{10}\left[1 + 10^{\frac{(10 - LSTR)}{10}}\right] \tag{A.6}$$

Nfo [in dBm0p] represents the "noise floor" at the receive side,

$$Nfo = Nfor + RLR\,. \tag{A.7}$$

A.2. Simultaneous Impairment Factor, *Is*

The factor *Is* can be divided into three specific impairment factors:

$$Is = Iolr + Ist + Iq \tag{A.8}$$

Iolr represents the decrease in quality caused by too-low values of *OLR* and is given by:

$$Iolr = 20\left[\left\{1+\left(\frac{Xolr}{8}\right)^{8}\right\}^{\frac{1}{8}} - \frac{Xolr}{8}\right] \tag{A.9}$$

where:

$$Xolr = OLR + 0.2(64 + No - RLR) \tag{A.10}$$

Ist represents the decrease caused by non-optimum sidetone:

$$Ist = 12\left[1+\left(\frac{STMRo-13}{6}\right)^{8}\right]^{\frac{1}{8}} - 28\left[1+\left(\frac{STMRo+1}{19.4}\right)^{35}\right]^{\frac{1}{35}} - 13\left[1+\left(\frac{STMRo-3}{33}\right)^{13}\right]^{\frac{1}{13}} + 29 \tag{A.11}$$

where:

$$STMRo = -10\log_{10}\left[10^{-\frac{STMR}{10}} + e^{-\frac{T}{4}}\,10^{-\frac{TELR}{10}}\right] \tag{A.12}$$

Iq represents the impairments caused by quantization distortion from a pure PCM process:

$$Iq = 15\log_{10}\left[1 + 10^{Y} + 10^{Z}\right] \tag{A.13}$$

where:

$$Y = \frac{Ro - 100}{15} + \frac{46}{8.4} - \frac{G}{9} \tag{A.14}$$

$$Z = \frac{46}{30} - \frac{G}{40} \tag{A.15}$$

and:

$$G = 1.07 + 0.258Q + 0.0602Q^2 \tag{A.16}$$

$$Q = 37 - 15\log_{10}(qdu). \tag{A.17}$$

A.3. Delay Impairment Factor, *Id*

The factor *Id* is further subdivided into the three factors *Idte*, *Idle* and *Idd*:

$$Id = Idte + Idle + Idd \tag{A.18}$$

The factor *Idte* represents the impairment of talker echo:

$$Idte = \left[\frac{Roe - \mathrm{Re}}{2} + \sqrt{\frac{(Roe - \mathrm{Re})^2}{4} + 100} - 1 \right] \left(1 - \mathrm{e}^{-T}\right) \tag{A.19}$$

where:

$$Roe = -1.5(No - RLR) \tag{A.20}$$

$$\mathrm{Re} = 80 + 2.5(TERV - 14) \tag{A.21}$$

$$TERV = TELR - 40\log_{10}\frac{1 + \frac{T}{10}}{1 + \frac{T}{150}} + 6\mathrm{e}^{-0.3T^2} \tag{A.22}$$

T is the one-way talker echo path delay in ms. For $T < 1$ ms, the talker echo should be considered as sidetone, i.e. *Idte* = 0. The computation algorithm furthermore combines the influence of *STMR* to talker echo. Taking into account that low values of *STMR* may

have some masking effects on the talker echo and for very high values of *STMR* the talker echo may become more noticeable, the terms *TERV* and *Idte* are adjusted as follows:

For *STMR* < 9 dB:

TERV in Equation (A.21) is replaced by *TERVs*, where:

$$TERVs = TERV + \frac{Ist}{2} \tag{A.23}$$

For 9 dB ≤ *STMR* ≤ 20 dB:

the above Equations (A.19) to (A.22) apply.

For *STMR* > 20 dB:

Idte in Equation (A.18) is replaced by *Idtes*, where:

$$Idtes = \sqrt{Idte^2 + Ist^2} \tag{A.24}$$

The factor *Idle* represents impairments due to listener echo:

$$Idle = \frac{Ro - Rle}{2} + \sqrt{\frac{(Ro - Rle)^2}{4} + 169} \tag{A.25}$$

where:

$$Rle = 10.5(WEPL + 7)(Tr + 1)^{-0.25} \tag{A.26}$$

The factor *Idd* represents the impairment caused by too-long absolute delay *Ta*, even with perfect echo cancellation.

For *Ta* < 100 ms:

$$Idd = 0$$

For *Ta* > 100 ms:

$$Idd = 25\left[\left(1+X^6\right)^{\frac{1}{6}} - 3\left\{1+\left(\frac{X}{3}\right)^6\right\}^{\frac{1}{6}} + 2\right] \tag{A.27}$$

with:

$$X = \frac{\log_{10}\left[\frac{Ta}{100}\right]}{\log_{10} 2}. \tag{A.28}$$

A.4. Equipment Impairment Factor, *Ie*

Ie represents the impairments due to low bit rate codecs. The *Ie* values depend on subjective MOS tests and network experience; also, they can be determined through instrumentation approaches. The recommended *Ie* values are given in Table I.1 of the Appendix I to the ITU-T Rec. G.113.

When packet loss occurs, the effective equipment impairment factor *Ie_eff* is derived using the codec specific value for the equipment impairment factor at zero packet loss *Ie* and the packet loss robustness factor *Bpl*:

$$Ie\_eff = Ie + (95 - Ie)\cdot\frac{Ppl}{\frac{Ppl}{BurstR} + Bpl} \tag{A.29}$$

where *Ppl* is the packet loss probability in percentage, *BurstR* is the burst ratio of the packet loss and is given by:

$$BurstR = \frac{Average\ length\ of\ observed\ bursts\ in\ an\ arrival\ sequence}{Average\ length\ of\ bursts\ expected\ for\ the\ network\ under\ "random"\ loss} \tag{A.30}$$

when packet loss is random, *BurstR* = 1 and when packet loss is bursty, *BurstR* > 1.

Some *Bpl* values are given in Tables I.3, I.4 and I.5 of the Appendix I to the ITU-T Rec. G.113.

A.5. Advantage factor, *A*

Advantage factor *A* allows the compensation to speech quality in situations where users can tolerate some decrease in quality. Provisional values are given in Table A.2 below.

Table A.2 Provisional examples for the advantage factor *A*

Communication system example	**Maximum value of A**
Conventional (wirebound)	0
Mobility by cellular networks in a building	5
Mobility in a geographical area or moving in a vehicle	10
Access to hard-to-reach locations, e.g. via multi-hop satellite connections	20