Speech Quality Evaluation in VoIP using the E-model

submitted by

Lijing Ding, B. Eng.

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

Master of Applied Science

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

Carleton University
Ottawa, Ontario, Canada K1S 5B6
August 2002

© Copyright 2002, Lijing Ding

National Library
of Canada

Acquisitions and
Bibliographic Services

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale
du Canada

Acquisitions et
services bibliographiques

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-79698-1

Canada

The undersigned recommend to the Faculty of Graduate Studies and Research
acceptance of the thesis

Speech Quality Evaluation in VoIP using the E-model

submitted by

Lijing Ding, B. Eng.

in partial fulfillment of the requirements for
the degree of Master of Applied Science

Thesis Supervisor, Dr. R. A. Goubran

Chair, Department of Systems and Computer Engineering

Carleton University

August 28, 2002

Abstract

Today, the demand for voice over Internet protocol (VoIP) applications is growing rapidly. In VoIP, the speech quality suffers due to packet loss, delay and delay jitter. The E-model is a transmission planning tool that is used to predict the Mean Opinion Score (MOS). The E-model does not include the impairment from delay jitter and only provides limited impairment information about packet loss.

The thesis investigates the effects of packet loss and delay jitter on speech quality. A new parameter is proposed to represent the impairment from delay jitter. These effects are then formulated and incorporated into the present E-model. In the simulation, vocoders ITU-T Rec. G.723.1 and G.729 are used. Speech samples are degraded by random packet losses and Pareto-distributed network delay. A number of error concealment techniques and delay jitter buffer algorithms are investigated. MOS is used as an index of speech quality and is measured by the Perceptual Analysis Measurement System (PAMS) algorithm.

The formula extends the coverage of the E-model in VoIP, and demonstrates good accuracy and flexibility. It provides a useful computational tool for MOS prediction and transmission planning.

Acknowledgements

First of all, I would like to thank my supervisor, Dr. Rafik A. Goubran, for all of his guidance, patience and encouragement for my research.

I would like to thank Dr. Samy El-Hennawey of Nortel Networks and Dr. Dieter Schulz of Mitel Networks for their invaluable discussions and suggestions.

I would like to thank all of the graduate students in the DSP lab for their helpful discussions. Thanks also to Danny Lemay and the other technicians for their help in keeping the network operational.

I would like to thank Mitel Networks, Communications and Information Technology Ontario (CITO), Ontario Graduate Scholarship-in Science and Technology (OGSST) and Carleton University for their financial support.

Finally, I would like to thank my wife Ling Chen, for her constant support and encouragement during my university career.

Table of Contents

ABSTRACT.....	III
ACKNOWLEDGEMENTS.....	IV
TABLE OF CONTENTS.....	V
LIST OF TABLES.....	VIII
LIST OF FIGURES.....	IX
LIST OF ABBREVIATIONS.....	XII
CHAPTER 1 INTRODUCTION	1
1.1 OVERVIEW	1
1.1.1 Voice over IP	1
1.1.2 Speech Quality Evaluation	2
1.2 PROBLEM STATEMENT	4
1.3 OBJECTIVE	5
1.4 THESIS OUTLINE	5
CHAPTER 2 BACKGROUND REVIEW	8
2.1 SPEECH QUALITY METRICS	8
2.1.1 Subjective Measurement Method	8
2.1.2 Objective Measurement Methods	9
2.2 E-MODEL	12
2.2.1 Introduction	12
2.2.2 Algorithm Description	13
2.2.3 MOS Prediction	16
2.3 PERCEPTUAL ANALYSIS MEASUREMENT SYSTEM (PAMS)	17
2.3.1 Introduction	17

2.3.2 Algorithm Description	18
2.3.3 Robustness and Accuracy	20
2.4 DIGITAL SPEECH LEVEL ANALYZER (DSL _A)	21
2.5 VOCODERS	24
2.5.1 Speech Coding	24
2.5.2 ITU-T Rec. G.723.1	26
2.5.3 ITU-T Rec. G.729	30
2.6 NETWORK IMPAIRMENTS IN VOIP	32
2.6.1 Delay.....	33
2.6.2 Delay Jitter	35
2.6.3 Packet Loss	36
CHAPTER 3 SIMULATION SYSTEM DESIGN	38
3.1 SIMULATION STRUCTURE	38
3.2 SIMULATION DESIGN DESCRIPTION	39
3.2.1 Reference Signal Selection	39
3.2.2 Vocoder Selection	43
3.2.3 Packetization Considerations	45
3.2.4 Network Simulation	45
3.3 MOS MEASUREMENT AND DATA ANALYSIS	46
CHAPTER 4 EFFECT OF PACKET LOSS ON SPEECH QUALITY	48
4.1 EFFECTS OF PACKET LOSS FOR DIFFERENT ERROR CONCEALMENT TECHNIQUES	48
4.1.1 Error Concealment Techniques.....	48
4.1.2 Simulation Results	50
4.2 FORMULATION	53
4.2.1 Former Research	53
4.2.2 Derivation of Equipment Impairment Factor	56
4.2.3 Formulation of Equipment Impairment Factor	59
4.3 VALIDATION.....	62
4.4 SUMMARY	72

CHAPTER 5	EFFECT OF DELAY JITTER ON SPEECH QUALITY.....	74
5.1	DELAY DISTRIBUTION.....	74
5.1.1	<i>Self-similarity in Network Delay</i>	74
5.1.2	<i>Pareto Distribution</i>	75
5.2	JITTER BUFFER ALGORITHM.....	80
5.3	SIMULATION RESULTS	82
5.4	FORMULATION	85
5.4.1	<i>Derivation of Jitter Impairment Factor</i>	85
5.4.2	<i>Formulation of Jitter Impairment Factor</i>	89
5.4.3	<i>MOS Prediction under Delay Jitter</i>	90
5.5	VALIDATION.....	91
5.6	SUMMARY	95
CHAPTER 6	COMBINED EFFECTS OF PACKET LOSS AND DELAY	
	JITTER ON SPEECH QUALITY	97
6.1	SIMULATION RESULTS	97
6.2	ADDITIVITY CHECK	101
6.3	DISCUSSION.....	105
CHAPTER 7	CONCLUSIONS.....	109
REFERENCES.....		113
APPENDIX A: DETAILED ALGORITHM OF THE E-MODEL.....		119
APPENDIX B: I_e VALUES FROM ITU-T		125

List of Tables

Table 2.1 MOS in Absolute Category Rating (ACR) [4]	8
Table 2.2 Provisional planning values for the equipment impairment factor I_e [7].....	14
Table 2.3 Provisional examples for the expectation factor A [6].....	15
Table 2.4 Definition of categories of speech transmission quality [19]	15
Table 2.5 Vocoder delay contributions.....	34
Table 3.1 MOS scores for 20 speech samples under vocoder G.723.1	44
Table 4.1 I_e value pairs used in linear interpolation.....	58
Table 4.2 $C1$ and $C2$ for repetition method.....	60
Table 4.3 $C1$ and $C2$ for splicing, silence insertion methods.....	61
Table 4.4 $C1$ and $C2$ for ITU-T proposed I_e values	62
Table 4.5 I_e value pairs used in linear interpolation – speech set 2	65
Table 4.6 I_e value pairs used in linear interpolation – speech set 3	69
Table 5.1 Coefficients and time constant for prediction model.....	90
Table 6.1 Combined effects for vocoder G.723.1 B 5.3 kb/s — measured MOS	99
Table 6.2 Combined effects for vocoder G.723.1 B 6.3 kb/s — measured MOS	99
Table 6.3 Combined effects for vocoder G.729 — Measured MOS	100
Table 6.4 Combined effects for vocoder G.729 with $H = 0.75$ — calibrated MOS.....	101
Table 6.5 Prediction errors for vocoder G.723.1 B 5.3 kb/s.....	103
Table 6.6 Prediction errors for vocoder G.723.1 B 6.3 kb/s.....	103
Table 6.7 Prediction errors for vocoder G.729	104

List of Figures

Figure 1.1 Voice over IP data flow	2
Figure 1.2 Speech quality measurement by SNR in PSTN	2
Figure 1.3 A configuration of VoIP	3
Figure 2.1 Structure of signal-based measurement system.....	9
Figure 2.2 Structure of parameter-based measurement system	12
Figure 2.3 MOS as a function of R in the E-model	16
Figure 2.4 Relationships between subjective test, objective test and the E-model [17]...	17
Figure 2.5 Structure of PAMS	18
Figure 2.6 Auditory transform in PAMS [23][5].....	19
Figure 2.7 DSLA hardware.....	22
Figure 2.8 PAMS display window-1	23
Figure 2.9 PAMS display window-2	23
Figure 2.10 Speech synthesis model.....	24
Figure 2.11 Analysis-by-synthesis linear predictive coder.....	25
Figure 2.12 Block diagram of the G.723.1 speech coder [28].....	28
Figure 2.13 Block diagram of the G.723.1 speech decoder [28]	29
Figure 2.14 Block diagram of the G.729 speech coder [29].....	31
Figure 2.15 Block diagram of the G.729 speech decoder [29]	32
Figure 2.16 RTP datagram structure.....	33
Figure 3.1 Simulation block diagram.....	38
Figure 3.2 A speech sample	40
Figure 3.3 MOS vs. speech listening level	41
Figure 3.4 MOS vs. linear PCM resolution (bits).....	43
Figure 4.1 Simulation diagram for packet loss	49
Figure 4.2 MOS vs. packet loss for vocoder G.723.1 B 5.3 kb/s	51
Figure 4.3 MOS vs. packet loss for vocoder G.723.1 B 6.3 kb/s	51
Figure 4.4 MOS vs. packet loss for vocoder G.729.....	52

Figure 4.5 MOS vs. packet loss under repetition error concealment.....	53
Figure 4.6 MOS curve fitting with $K = 0.50$	54
Figure 4.7 I_e values in ITU-T Rec. G.113 Appendix I [7]	55
Figure 4.8 $I_{e,mea}$ values from scale transformation.....	57
Figure 4.9 Linear interpolation line for I_e	58
Figure 4.10 Interpolated I_e values	59
Figure 4.11 I_e curve fitting for G.729.....	61
Figure 4.12 Measured MOS – speech set 2	63
Figure 4.13 Measured MOS – speech set 3	64
Figure 4.14 $I_{e,mea}$ after interpolation – speech set 2	66
Figure 4.15 Calibrated MOS – speech set 2	66
Figure 4.16 MOS prediction error for G.723.1 B 5.3 kb/s – speech set 2	67
Figure 4.17 MOS prediction error for G.723.1 B 6.3 kb/s – speech set 2	67
Figure 4.18 MOS prediction error for G.729 – speech set 2.....	68
Figure 4.19 MOS prediction error – speech set 2	68
Figure 4.20 MOS prediction error for G.723.1 B 5.3 kb/s – speech set 3	70
Figure 4.21 MOS prediction error for G.723.1 B 6.3 kb/s – speech set 3	70
Figure 4.22 MOS prediction error for G.729 – speech set 3.....	71
Figure 4.23 MOS prediction error – speech set 3	71
Figure 5.1 Complementary CDF of Pareto distribution and Poisson distribution.....	77
Figure 5.2 Pareto-distributed delay trace	78
Figure 5.3 Histogram of Pareto delay	79
Figure 5.4 Jitter PDF	79
Figure 5.5 I-policy and E-policy	81
Figure 5.6 Simulation diagram for network delay	82
Figure 5.7 MOS vs. H and buffer size for vocoder G.723.1 B 5.3 kb/s	83
Figure 5.8 MOS vs. H and buffer size for vocoder G.723.1 B 6.3 kb/s	84
Figure 5.9 MOS vs. H and buffer size for vocoder G.729.....	84
Figure 5.10 $I_{j,mea}$ values from scale transformation for vocoder G.723.1 B 5.3 kb/s...	86

Figure 5.11 $I_{j,mea}$ values from scale transformation for vocoder G.723.1 B 6.3 kb/s...	86
Figure 5.12 $I_{j,mea}$ values from scale transformation for vocoder G.729	87
Figure 5.13 I_j values for vocoder G.723.1 B 5.3 kb/s	88
Figure 5.14 I_j values for vocoder G.723.1 B 6.3 kb/s	88
Figure 5.15 I_j values for vocoder G.729	89
Figure 5.16 Predicted MOS as a function of H and T for G.729	91
Figure 5.17 MOS vs. H and buffer size for vocoder G.723.1 B 5.3 kb/s – speech set 2 ..	92
Figure 5.18 MOS vs. H and buffer size for vocoder G.723.1 B 6.3 kb/s – speech set 2 ..	92
Figure 5.19 MOS vs. H and buffer size for vocoder G.729 – speech set 2	93
Figure 5.20 MOS prediction error for vocoder G.723.1 5.3 kb/s	94
Figure 5.21 MOS prediction error for vocoder G.723.1 6.3 kb/s	94
Figure 5.22 MOS prediction error for vocoder G.729	95
Figure 6.1 Simulation block diagram for combined effects	98
Figure 6.2 Combined effects for vocoder G.729 with $H = 0.75$	100
Figure 6.3 Combined effects for vocoder G.729 with $H = 0.75$ — calibrated MOS.....	102
Figure 6.4 Prediction error for vocoder G.729 with $H = 0.75$	104
Figure 6.5 Predicted MOS vs. packet loss from I_e values in [7]	107
Figure 6.6 Additivity of packet loss.....	107

List of Abbreviations

ACELP	Algebraic-Code-Excited Linear-Prediction
ACR	Absolute Category Rating
CDF	Cumulative Distribution Function
CELP	Code Excited Linear Prediction
CS-ACELP	Conjugate-Structure Algebraic-Code-Excited Linear-Prediction
DSLA	Digital Speech Level Analyzer
FEC	Forward Error Correction
IP	Internet Protocol
ITU	International Telecommunication Union
ITU-T	ITU Telecommunication Standardization Sector
LAN	Local Area Network
LD-CELP	Low Delay Code Excited Linear Prediction
LP	Linear Prediction
LPC	Linear Prediction Coding
LSP	Line Spectral Pair
MIRS	Modified Intermediate Reference System
MOS	Mean Opinion Score
MP-MLQ	Multipulse Maximum Likelihood Quantization
MSE	Mean Square Error
PAMS	Perceptual Analysis Measurement System

PCM	Pulse Code Modulation
PDF	Probability Distribution Function
PESQ	Perceptual Evaluation of Speech Quality
PSQM	Perceptual Speech Quality Measure
PSTN	Public Switched Telephone Network
qdu	Quantization Distortion Unit
QoS	Quality of Service
RTP	Real-Time Transport Protocol
SNR	Signal-to-Noise Ratio
SPL	Speech Pressure Level
UDP	User Datagram Protocol
VAD	Voice Activity Detection
VoIP	Voice over Internet Protocol

CHAPTER 1 INTRODUCTION

1.1 Overview

Today, the telecommunications industry is undergoing dramatic changes. The traditional public switched telephone network (PSTN) and data networks are converging. Voice over Internet protocol (VoIP) has emerged as an important application and is expected to carry more and more voice traffic [1][2]. However, the present Internet only offers *best-effort* service due to its nature, rendering the speech quality neither guaranteed nor predictable. Testing speech quality on the new convergent network is imperative.

1.1.1 Voice over IP

Voice over IP (VoIP) refers to the real-time transmission of voice signals as packetized data across networks by using Internet protocol [1]. The speech samples are coded into frames by vocoders to reduce bandwidth consumption, and then encapsulated into packets before transmission. The packets arrive at the destination with varying delay, and some may even be dropped. At the receive end, packets are buffered to compensate for delay variation and reassembled into the correct order. Some error concealment schemes are also used to retrieve the lost packets. The reconstructed packets are sent into the decoder and finally played back. This VoIP data flow is shown in Figure 1.1.

Figure 1.1 Voice over IP data flow

1.1.2 Speech Quality Evaluation

The PSTN has served voice traffic over many years and evolved to provide optimal speech quality in design. The circuit is reserved in each call connection; the speech quality is reliable and predictable.

Signal-to-noise ratio (SNR), which was developed for linear time-invariant systems, is used to assess the speech quality in PSTN. As shown in Figure 1.2, the speech quality is determined by passing a single-frequency or dual-frequency tone through the channel under test [3].

Figure 1.2 Speech quality measurement by SNR in PSTN

Figure 1.3 shows a typical convergent network configuration, where the local access is provided by PSTN and long distance transmission is provided by the IP network. The non-linearity caused by low-bit rate vocoders makes such SNR metric inadequate.

Figure 1.3 A configuration of VoIP

Speech quality is determined by the listener's perception, thus it is inherently subjective. The mean opinion score (MOS) test, specified by ITU-T Rec. P.800 [4], is widely accepted as a norm for comparative rating of different systems. The subjective MOS test is time-consuming and expensive. Therefore, it is desirable to replace it with an objective method that should correlate well with the subjective MOS test.

In recent years, several objective MOS measurement methods were developed to fill a need for repeatable, cheap and fast MOS tests. They can be classified into two categories [5]. One is signal-based: it uses a reference signal and a degraded signal as the input of the measurement. This category includes some algorithms widely used today, such as Perceptual Analysis Measurement System (PAMS) and Perceptual Evaluation of Speech Quality (PESQ). The other is parameter-based: it is a computational model which combines all the impairment parameters into a total value.

The ITU-T Rec. G.107, the E-model [6], belongs to the parameter-based category. The E-model is not a measuring tool, but a planning tool. The E-model is used for end-to-end transmission planning; the output of it can be transformed into a MOS score for prediction.

1.2 Problem Statement

In the VoIP scenario, the speech quality is mainly impaired by packet loss, delay and delay jitter.

Packet loss is common in packet-switched transmission; it makes the speech sound clipped in time. The speech quality is greatly affected by factors such as packet loss rate, loss burstiness, vocoder type and error concealment method. In the E-model, the equipment impairment factor, I_e , is used to represent the impairment from packet loss. I_e values are tabulated in ITU-T Rec. G.113 Appendix I [7]. For vocoders ITU-T Rec. G.723.1 and G.729, which are widely used in VoIP applications today, these values only cover limited packet loss rates without error concealment for a certain number of frames per packet. In addition, they are provisional only, as they were determined in a single or a few tests [7]. Thus, the available I_e values lack accuracy and flexibility.

Packet-based transmission exhibits varying delay in packet delivery. This is caused by the variability of queuing delays. In addition, the packets carrying speech samples for the same stream may transmit via different routes through the Internet. Delay jitter is the variance of interarrival times at the receive side. To obtain steady packet stream, jitter must be compensated. This is achieved by using a jitter buffer at the receive side. The first received packet is held for a while before being played out. An optimum buffer size has to be found which balances removing jitter and limiting delay, packet loss to tolerable levels. In the output stream, a gap may be introduced during the silence intervals or even during the talkspurts. The implementation of jitter buffer algorithm has a vital role on speech quality. Delay jitter is potentially more disruptive for VoIP applications. However, this impairment is not included in the E-model.

The E-model already includes the impairment caused by delay; hence we will not focus on it in the thesis.

1.3 Objective

The objectives of the thesis are as follows:

- Investigating and quantifying the effects of packet loss on speech quality.
- Investigating and quantifying the effects of delay jitter on speech quality.
- Investigating the combined effects of packet loss and delay jitter on speech quality.
- Incorporating the above parameters into the E-model and checking the additivity.

The main contribution of the thesis is extending the coverage of the E-model in evaluating speech quality in the VoIP scenario. We add a new parameter to the E-model to represent the impairment from delay jitter, and quantify the effects of packet loss and delay jitter. The formula is very helpful in MOS prediction and transmission planning.

1.4 Thesis Outline

The rest of the thesis is organized as follows:

Chapter 2 reviews the background material, including subjective and objective speech quality measurement methods, the E-model, PAMS, Digital Speech Level Analyzer (DSL), analysis-by-synthesis speech coders and network impairments related to speech quality. For MOS measurement, we introduce the subjective and objective methods. The E-model is a transmission planning tool that can be used for MOS prediction. The DSLA

is a measurement tool that includes the PAMS algorithm and is used through our MOS measurement. We also introduce two vocoders, ITU-T Rec. G.723.1 and G.729, which are based on Analysis-by-Synthesis linear predictive coding. Finally, we review the impacts of delay, delay jitter and packet loss on speech quality.

Chapter 3 discusses how the simulation system is set up. We examine the effects of speech level and bit resolution on MOS and explain how to select the reference speech signals. Also, we explain the testing conditions for vocoder, packet loss, delay, delay jitter buffer and error concealment methods.

Chapter 4 investigates the effects of packet loss on speech quality. Several receiver-based error concealment methods are tested and the repetition method is adopted for its effectiveness and easy implementation. Our simulation shows the vocoder G.729 always outperforms vocoder G.723.1 in random packet loss. We also reveal that the equipment impairment factor I_e , which is used in the E-model to represent impairment from vocoder with packet loss, increases logarithmically with increasing packet loss rate. This relationship is formulated based on our measurement. Linear interpolation is used to correct the bias of speech samples and stable I_e values are derived. Finally, the proposed formula is validated by using other speech samples, and demonstrates good accuracy.

Chapter 5 investigates the effects of delay jitter on speech quality. Pareto distribution is used to simulate the self-similar network delay. E-policy is used as the jitter buffer algorithm. We introduce a new parameter I_j into the E-model to represent the impairment from delay jitter. A formula is proposed to quantify the I_j as a function of intensity of delay self-similarity and buffer size. Linear interpolation is also used, the derived I_j values can be considered stable and used in the extended E-model for MOS

prediction. The formula is validated by using other speech samples as well, and demonstrates good accuracy.

Chapter 6 investigates the combined effects of packet loss and delay jitter on speech quality. We assume the derived I_e and I_j are additive in the psychological scale; this chapter describes the additivity check. The results show that our formula works well for low packet loss rates up to 5 percent. It gives a lower bound for MOS prediction as the prediction error is negative in general and can reach -0.60 MOS with 20 percent packet loss. The additivity is discussed for the combined effects.

Chapter 7 concludes the thesis and proposes some suggestions for further studies.

CHAPTER 2 BACKGROUND REVIEW

2.1 Speech Quality Metrics

2.1.1 Subjective Measurement Method

Speech quality is a very important aspect in voice communication. It is subjective, as it is decided by the perception of the human listener. The subjective measurement method is defined in ITU-T Rec. P.800 as the MOS test. In the test, a panel of subjects listens to the speech material and votes on its quality using an integer opinion scale in a well-designed environment. The rating results are averaged to produce a MOS score. The commonly used MOS is a five-point scale in Absolute Category Rating (ACR). Several five-point category-judgment scales may be used for different purposes; the listening-quality and listening-effort scales are summarized in Table 2.1. The former is widely used, referred to as MOS.

Table 2.1 MOS in Absolute Category Rating (ACR) [4]

MOS Score	Listening-quality Scale	Listening-effort 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

The subjective MOS test is quite complicated; the design of the test is strongly influenced by both human subjects and elaborate testing conditions. It is impractical in frequent tests such as routine network monitoring.

2.1.2 Objective Measurement Methods

Objective measurement methods can be classified into two categories [5]:

- Signal-based methods
- Parameter-based methods

Signal-based methods use two signals as the input of the measurement, namely, a reference signal and the degraded signal, which is the output of the system under test. They identify the audible distortions based on the perceptual domain representation of two signals incorporating human auditory models. The structure of signal-based methods is shown in Figure 2.1.

Figure 2.1 Structure of signal-based measurement system

The perceptual layer is relevant to lower layers of human reception in contrast to the cognitive layer. The frequency unit in the perceptual domain is called bark; frequencies at each bark are perceived as equally-space by human listeners. Because the human hearing

system has a finer frequency resolution at low frequencies than at high frequencies, the bark scale is non-linear in Hz.

The signal-based methods compare the two signals represented in bark scale. These algorithms include: Perceptual Speech Quality Measure (PSQM) [8][9], Measuring Normalizing Blocks (MNB) [10], Perceptual Analysis Measurement System (PAMS) [11][12], and Perceptual Evaluation of Speech Quality (PESQ) [13].

PSQM was developed by J.G. Beerends and J.A. Stemerdink of KPN Research in The Netherlands, and standardized by ITU-T in 1998 as Rec. P.861. PSQM was designed for assessing telephone-band vocoders. It is not able to take proper account of impairments such as transmission channel errors, environmental noise, clipping, delay, delay variation and vocoders operating below 4 kb/s [14].

MNB was proposed by S. Voran of the Institute for Telecommunication Sciences, U.S. Department of Commerce, and added in 1998 as Appendix II to Rec. P.861. MNB mainly solved the problems of PSQM in measuring transmission channel errors and vocoders operating below 4 kb/s [14].

PAMS was based on the model introduced by M.P. Hollier of the PsyTechnics group at British Telecommunications. The current version 3.0 was released in December 1999. PAMS is the first model that takes account of the end-to-end behavior. It is recommended for use in measuring the impairments from the following processes or parameters: waveform coders, vocoders, transcodings, speech input levels to a coder, talker dependencies, fast delay variation, temporal clipping, amplitude clipping and added noise. It is not intended to measure the delay, slow delay variation, overall system

gain/attenuation, analog phone filtering, environmental noise at the sending side and music as input signal [15].

PESQ was jointly developed by J.G. Beerends and A.P. Hekstra of KPN Research, and A.W. Rix and M.P. Hollier of British Telecommunications. PESQ was standardized by ITU-T in 2001 as Rec. P.862 and replaced Rec. P.861. It combines the merits of PAMS and PSQM99 (an updated version of PSQM), and adds new methods for transfer function equalization and averaging distortions over time. PESQ can effectively measure the impairments such as waveform coders, vocoders at or above 4 kb/s, other coders in mobile systems, transcoding, speech input levels to a coder, transmission channel errors (bit error, packet loss, cell loss), delay variation, environmental noise at the sending side. It is not intended to measure the delay, listening levels, sidetone, talker echo, music input and vocoders operating below 4 kb/s [16].

PESQ is the latest algorithm for objective speech quality measurement, and would be more accurate than PAMS. In our research, we use PAMS algorithm due to availability.

Speech quality testing can be conducted by listening-only or conversational methods. For practical reasons, the listening-only method is the only feasible way for subjective testing during the network deployment period, when real-time implementation is not available.

The signal-based algorithms estimate the subjective quality obtained in listening-only tests. However, some impairment 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 algorithms. It is possible to have a high objective MOS score, but the overall quality is still not acceptable.

On the one hand, given the MOS score, the reasons of degradation can't be broken down into related network impairments, such as delay, delay jitter, packet loss, echo, noise, etc. It doesn't provide much insight into how to tune the networks in engineering practice. On the other hand, during the transmission planning stage, it is not feasible to conduct such subjective or objective measurement. Therefore, parameter-based methods, i.e., computational methods, which can combine all the transmission impairments into a total value, are preferable. Figure 2.2 depicts the diagram of a parameter-based method. ITU-T Rec. G.107 defines such a computational model, the E-model.

Figure 2.2 Structure of parameter-based measurement system

2.2 E-model

2.2.1 Introduction

The E-model was defined in ITU-T Rec. G.107 in 1998, and was revised in 2000 [6]. Detailed guidelines and planning examples of this Recommendation are given in ITU-T Rec. G.108 [17].

The E-model is a computational model for use in transmission planning. It assesses the combined effects of transmission parameters that affect the conversation quality of

narrow band (300-3400Hz) telephony [6]. The parameters of the E-model cover a wide range of end-to-end transmission impairment factors.

The principle of the E-model is based on the following assumptions:

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

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

The input to the E-model includes factors such as handset characteristics, noise, delays, equipment impairments and so on. These factors 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 for prediction purposes [6][18].

2.2.2 Algorithm Description

The E-model algorithm is briefly introduced here; a detailed description is given in appendix A.

The transmission rating factor R is given by:

$$R = R_o - I_s - I_d - I_e + A \quad (2.1)$$

R_o represents the basic signal-to-noise ratio, including noise sources such as circuit noise and room noise.

I_s , simultaneous impairment factor, is a combination of all impairments which occur more or less simultaneously with the voice signal. It can be further divided into three specific impairment factors:

$$I_s = I_{olr} + I_{st} + I_q \quad (2.2)$$

I_{olr} represents the decrease caused by too-low values of overall loudness rating, I_{st} represents the decrease caused by non-optimum sidetone, and I_q represents the impairments caused by quantization distortion from a pure PCM (Pulse Code Modulation) process.

I_d , delay impairment factor, represents the impairments caused by delay and echo effects and may be divided into three factors, I_{dte} , I_{dle} and I_{dd} :

$$I_d = I_{dte} + I_{dle} + I_{dd} \quad (2.3)$$

where the factors I_{dte} , I_{dle} and I_{dd} represent the impairments due to talker echo, listener echo and too-long absolute delay respectively.

I_e , equipment impairment factor, represents the impairments caused by low bit rate vocoders. Table 2.2 lists the recommended values of I_e from ITU-T Rec. G.113.

Table 2.2 Provisional planning values for the equipment impairment factor I_e [7]

Codec type	Reference	Operating rate kbit/s	I_e Value
PCM	G.711	64	0
CS-ACELP	G.729	8	10
	G.729-A + VAD	8	11
ACELP	G.723.1	5.3	19
MP-MLQ	G.723.1	6.3	15

The advantage factor A (called *expectation factor* in the old version ITU-T Rec. G.107) can be used for some compensation of impairment factors to account for the fact that some users will tolerate a certain degradation when there are other advantages of access to the users [17]. Table 2.3 gives examples for the advantage factor A .

Table 2.3 Provisional examples for the expectation factor A [6]

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

The rating factor R ranges from 0 to 100. The conversation quality can be interpreted by Table 2.4.

Table 2.4 Definition of categories of speech transmission quality [19]

Range of E-Model rating R	Speech transmission quality category	User satisfaction
$90 \leq R < 100$	Best	Very satisfied
$80 \leq R < 90$	High	Satisfied
$70 \leq R < 80$	Medium	Some users dissatisfied
$60 \leq R < 70$	Low	Many users dissatisfied
$50 \leq R < 60$	Poor	Nearly all users dissatisfied

2.2.3 MOS Prediction

The rating factor R is transformed into a five-point MOS by the following equation:

$$MOS = \begin{cases} 1 & R < 0 \\ 1 + 0.035R + R(R - 60)(100 - R)7 \cdot 10^{-6} & 0 < R < 100 \\ 4.5 & R > 100 \end{cases} \quad (2.4)$$

The MOS is represented by a third order polynomial of rating factor R , and is clipped between 1 and 4.5. Figure 2.3 illustrates the relation between R and MOS.

Figure 2.3 MOS as a function of R in the E-model

In general, the relationships between the subjective test, objective test and E-model are summarized in Figure 2.4. In the subjective test, MOS is voted by the subjects who listen to the degraded signals, and is used to calibrate the objective models. In the objective test, for the signal-based methods, MOS is measured by comparing the

perceptual difference between the reference signals and the corresponding degraded signals; for the parameter-based method, the E-model uses a set of parameters to calculate the rating factor R and transforms it into MOS.

Figure 2.4 Relationships between subjective test, objective test and the E-model [17]

2.3 Perceptual Analysis Measurement System (PAMS)

2.3.1 Introduction

PAMS is an objective measurement algorithm designed for robust end-to-end speech quality assessment in telephone-band transmission systems. It measures the one-way speech quality by comparing the reference signal with the degraded signal. The results

returned are predictions of subjective MOS, both in five-point listening-quality and listening-effort scales, defined in ITU-T Rec. P.800 Absolute Category Rating.

2.3.2 Algorithm Description

The block diagram of PAMS is shown in Figure 2.5. The processes include four stages [20][21]:

Figure 2.5 Structure of PAMS

Stage 1: Signal pre-processing: time alignment, level adjustment and equalization.

- Time alignment: PAMS is a listening-only model and knows nothing about the delay of the system. The input signals are divided into utterances and synchronized by finding the time offset of each utterance. This process removes the bulk delay and slow delay variation, while keeping the fast delay variation for measurement.
- Level adjustment: the MOS varies with the listening level. PAMS assumes a constant optimum listening level of 79 dB SPL (Speech Pressure Level) at the ear

reference plane as recommended by ITU-T Rec. P.830 [22]. A gain is applied to both signals to bring them to this level.

- Equalization: analog interfaces, such as telephone handset and private branch exchange (PBX), always introduce some filtering; this filtering has less subjective effects on speech quality than that of vocoder distortions. PAMS auto-detects the filtering that has taken place in the network and cancels its effects by equalizing the reference signal.

Stage 2: Auditory transform

The processing in the auditory transform is shown in Figure 2.6.

Figure 2.6 Auditory transform in PAMS [23][5]

To identify the audible difference perceived by human listeners, a psychoacoustic model is used to mimic certain key properties of the human auditory system [23]. First, models of the acoustic responses of the telephone handset, outer and middle ear are applied to the signals. Then, a bank of 19 bark-spaced Infinite Impulse Response (IIR) filters that reflects the frequency selectivity of the basilar membrane is used to transform the signals into the perceptual domain. Temporal smoothing is used to estimate the temporal and simultaneous masking effects, and finally, they are mapped to a loudness scale. The output is a time-frequency representation of the perceived loudness, known as

the sensation surface. The difference by subtracting the reference sensation surface from the degraded sensation surface is called the error surface.

Stage 3: Error parameterization

Errors that represent added signal energy have positive values in the error surface. Errors that represent lost signal energy have negative values. The error surface is analyzed and several parameters that reflect different types of errors are extracted.

Stage 4: Regression

A monotonic, non-linear mapping is performed to predict the MOS score. The mapping is calibrated by a huge database of subjective tests to ensure that PAMS is applicable to a wide range of distortion types.

2.3.3 Robustness and Accuracy

Some novel mathematics methods are used in the PAMS to ensure the one-to-one relationship between the total amount of distortion and the MOS scale. If anything becomes worse, the quality must fall.

In the perceptual layer, K error parameters e_k ($1 \leq k \leq K$) are extracted from the error surface. These error parameters describe the amount and distribution of errors. Each e_k is bounded and mapped by a third-order polynomial $f_k(\cdot)$. The coefficients are calculated by gradient optimization with a cost-based constraint to keep each $f_k(\cdot)$ to be monotonic

and of correct sign within the given bounds [24][25]. The results are summed up to produce subjective speech quality prediction y as shown in Equation 2.5.

$$y = f_1(e_1) + f_2(e_2) + \dots f_K(e_K) + c, \quad \frac{df_k}{de_k} \leq 0 \quad (2.5)$$

Equation 2.5 gives $3K+1$ degrees of freedom; the optimal value of K is chosen to be 10.

As reported in [25], 14 subjective tests, containing 8,480 files in 694 conditions were used in the training of PAMS 3.0; A further 14 subjective tests, containing 13,252 files in 789 conditions were used to assess the performance of PAMS. The results of all 28 tests show that for 76.8% of conditions, the objective model was within 0.25 MOS of the subjective score and 95.4% within 0.5 MOS of the subjective score. The root mean square error is 0.236 MOS.

2.4 Digital Speech Level Analyzer (DSL A)

The DSLA is a measurement tool manufactured by Malden Electronics Ltd., U.K. The DSLA measures mean active speech level in accordance with ITU-T Rec. P.56 Method B. Also, it generates a variety of useful signals to simulate networks and network elements. The options include the PAMS algorithm [20].

The DSLA can be connected to a PC via a standard serial port, or via an Ethernet Local Area Network (LAN). The DSLA hardware is shown in Figure 2.7.

Figure 2.7 DSLA hardware

To implement the PAMS algorithm, the speech signals used can be recorded by the DSLA or directly from the user's PC. Currently, DSLA supports sound files in .wav (containing 44-byte header information) or raw (without header information) 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. Further, the speech must be at least 4 seconds in duration for PAMS to make a prediction.

The returned MOS results have a precision of two decimal digits. The PAMS display windows are shown in Figures 2.8 and 2.9. Figure 2.8 shows the reference sensation surface and degraded sensation surface of the signals as well as the MOS scores. Figure 2.9 shows MOS scores and other measures, like time offset, number of utterances and so on.

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 log files for further analysis.

In our simulation, DSLA with the PAMS algorithm is used for MOS measurement.

Figure 2.8 PAMS display window-1

Figure 2.9 PAMS display window-2

2.5 Vocoders

2.5.1 Speech Coding

Speech coding is the technique of encoding the speech signals into compact digital representations for efficient transmission and storage. Speech coding can be classified into three categories [26]: 1) waveform coding 2) source coding, and 3) hybrid coding.

Waveform coding is the method for discrete approximation of the analog waveform and it usually operates at a bit rate above 16 kb/s. The waveform coding family includes Pulse Code Modulation (PCM), Differential PCM (DPCM), Delta Modulation (DM) and Adaptive DPCM (ADPCM) [27].

Source coding makes specific assumptions about the source properties of the analog signal. Speech signals are quasi-stationary over short segments in terms of milliseconds. The speech sound can be classified into voiced, unvoiced and mixed. Voiced sound is harmonic and quasi-stationary while the unvoiced sound is random-like with broadband. Source coding is based on the linear speech source-system production model to produce perceptually intelligible speech without necessarily matching the waveform. The speech synthesis model is shown in Figure 2.10 [27].

Figure 2.10 Speech synthesis model

The speech is produced by a linear filter, the human vocal tract filter. In this model, the throat and mouth are modeled as a linear slowly time-varying filter, and it is periodically excited by a series of impulses (for voiced sound) or by random noise (for unvoiced sound). The vocal tract filter is an all-pole adaptive filter with 10 parameters, which are derived from the linear prediction (LP) analysis. The synthesized speech quality is not suitable for telephone communication because of the simple two-state excitation (impulse/noise) assumption.

Hybrid coding combines the features of the waveform coding and source-system coding. The parameters of vocal tract filter are still determined by linear prediction, but the excitation sequence is determined by analysis-by-synthesis optimization. The encoder searches through the parameter space and synthesizes a different representation of the input signal in each search loop. The synthesized signal is compared with the input signal and the parameters which minimize the weighted difference in a mean square error (MSE) sense are selected. The block diagram is shown in Figure 2.11.

Figure 2.11 Analysis-by-synthesis linear predictive coder

The excitation models for analysis-by-synthesis coders include Multi-Pulse Excitation, Regular Pulse Excitation and Code Excited Linear Prediction (CELP). Among them, CELP coders are very popular for VoIP applications [27]. It can operate at 4.8 kb/s to 8 kb/s and provide good speech quality.

CELP uses a fixed codebook for the unvoiced speech and an adaptive codebook for the voiced speech. The excitation is determined by the index to an appropriate vector in the codebook; the index is found by an exhaustive search over the codebook and the vector minimizes the MSE between the original and synthesis signal is selected. The CELP family includes ACELP (Algebraic-Code-Excited Linear- Prediction), MP-MLQ (Multipulse Maximum Likelihood Quantization), LD-CELP (Low Delay CELP) and CS-ACELP (Conjugate-Structure ACELP) [26].

The term *vocoder*, which is the acronym of voice coder, is generally used to represent coders using source coding or hybrid coding. To avoid confusion, vocoder only refers to hybrid coding in our thesis.

Two vocoders, ITU-T Rec. G.723.1 and G.729, which belong to the CELP family, are used in our simulation.

2.5.2 ITU-T Rec. G.723.1

The ITU-T Rec. G.723.1 specifies a coded representation that can be used for compressing the speech or other audio signal component of multimedia services at a very low bit rate as part of the overall H.324 family of standards [28]. The typical applications of this speech coder are in telephony over packet-switched networks, like VoIP, as well as in coding the speech component in video conferencing. This coder has two bit rates

associated with it, 5.3 and 6.3 kbit/s. The coder and decoder support both rates and can switch between the two rates at any frame boundary. An option for variable rate operation using discontinuous transmission and noise fill during non-speech intervals is also possible.

The G.723.1 coder encodes speech in frames by linear predictive analysis-by-synthesis coding. The 6.3 kbit/s coder uses MP-MLQ for the excitation signal and the 5.3 kbit/s coder uses ACELP. The higher bit rate has greater quality. The lower bit rate gives good quality and provides system designers with additional flexibility [28]. The frame size is 30 ms and look-ahead is 7.5 ms, resulting in a total algorithm delay of 37.5 ms. For the high rate, each frame is 24 bytes, for the low rate, each frame is 20 bytes.

This coder operates with a digital signal in 16-bit linear PCM format and sampled at 8000 Hz. The digital signal is obtained by telephone band filtering of the analog input, or converted from other formats, such as 64 kbit/s PCM defined in ITU-T Rec. G.711.

The encoder operates on blocks (frames) of 240 samples each. Each block is first high pass filtered to remove the DC component and then divided into four 60-sample subframes. A 10th order linear prediction coding (LPC) filter is computed to model each of the subframes. For every two subframes, the open loop pitch period is computed using the weighted speech signal. Using the computed pitch period, a harmonic noise shaping filter is constructed. The combination of the LPC synthesis filter, the formant perceptual weighting filter, and the harmonic noise-shaping filter is used to create an impulse response. A 5th order closed loop pitch predictor is computed by using the pitch period estimation, and the impulse response. The contribution of the pitch predictor is then

subtracted from the initial target vector. Both the pitch period and the differential value are transmitted to the decoder. The block diagram of the encoder is shown in Figure 2.12.

Figure 2.12 Block diagram of the G.723.1 speech coder [28]

The decoder also operates on a frame-by-frame basis. The quantized LPC indices, adaptive codebook excitation and fixed codebook excitation are decoded. The LPC synthesis filter is constructed and applies to the excitation signal. The adaptive postfilter

consists of a formant and a pitch postfilter. The excitation signal passes through the pitch postfilter, the synthesis filter and the formant postfilter in turn. A gain scaling unit is applied to keep the energy at the input level of the formant postfilter. The block diagram of the decoder is shown in Figure 2.13.

Figure 2.13 Block diagram of the G.723.1 speech decoder [28]

G.723.1 is a fixed-point version. The annex A, silence compression scheme, uses a Voice Activity Detection (VAD) algorithm to reduce the transmitted bit rate during the silent speech intervals. The annex B introduces an alternative specification based on floating-point arithmetic.

2.5.3 ITU-T Rec. G.729

The ITU-T Rec. G.729, Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP) coder, is designed for cellular systems, digital satellite systems and network applications. It is particularly useful in applications with stringent delay requirements. This coder also operates with speech signals in 16-bit linear PCM, 8000 Hz format.

The CS-ACELP coder is based on the CELP coding model [29]. The coder operates on speech frames of 10 ms and uses a 5 ms look-ahead for linear prediction analysis, resulting in a total algorithm delay of 15 ms. Each frame is encoded in 10 bytes; the rate of the coder is 8.0 kb/s.

For each 10 ms frame, linear prediction analysis is performed to compute the LP filter coefficients. The coefficients are then transformed into Line Spectral Pair (LSP) coefficients. The excitation signal is modeled using a combined fixed and adaptive codebook method. Each frame is divided into 2 subframes of 5 ms to determine the excitation parameters. The adaptive codebook represents the periodic component in the excitation signal, the parameters of this codebook are the gain and pitch lag [30]. The fixed codebook consists of 4 nonzero pulses, with amplitude of either plus one or minus one and a fixed set of positions. The parameters of the encoded signal — LPC parameters, pitch lag, pitch gain, fixed codebook pulse positions, fixed codebook pulse signs and gains — are quantized and transmitted to the decoder. The block diagram of the encoder is shown in Figure 2.14, where the G_c and G_p represent the fixed-codebook gain and adaptive codebook gain respectively.

Figure 2.14 Block diagram of the G.729 speech coder [29]

At the decoder, the LSP coefficients are interpolated and converted to LP filter coefficients for each subframe. For each subframe, the excitation is constructed by adding the adaptive and fixed codebook vectors scaled by their respective gains; the speech is reconstructed by filtering the excitation through the LP synthesis filter and further enhanced by post-processing, which includes an adaptive postfilter based on the long-

term and short-term synthesis filters, followed by a high-pass filter and scaling operation. The block diagram of the decoder is shown in Figure 2.15.

Figure 2.15 Block diagram of the G.729 speech decoder [29]

Vocoder G.729 is a fixed-point version. G.729 A is a reduced-complexity version of the G.729 with a simplified the search of adaptive and fixed codebook. It is bit stream interoperable with the full version but the performance may not be as good as the full implementation of Recommendation G.729 in certain circumstances [31].

2.6 Network Impairments in VoIP

The IP network was originally designed for non-real time data communications. It offers *best-effort* service with no Quality of Service (QoS) guarantee. The main impairments related to voice over IP are delay, delay jitter and packet loss.

Real-time Transport Protocol (RTP) [32] is used for end-to-end delivery services for data with real-time characteristics, such as interactive audio and video over Internet. RTP is designed to send data in one direction without acknowledgement and itself does not

provide any mechanism to ensure timely delivery or provide other QoS guarantees. However, it relies on RTP Control Protocol (RTCP) to monitor the data delivery. The RTP header includes a sequence number that can be used to detect packet loss, a timestamp for synchronization and jitter calculation. The RTP runs on top of the User Datagram Protocol (UDP), and hence all the services of UDP are inherited. The common RTP datagram structure is shown in Figure 2.16. The RTP datagram includes a fixed 40 bytes header and n bytes payload.

Figure 2.16 RTP datagram structure

2.6.1 Delay

End-to-end delay is the time required for a signal traveling from a speaker's mouth to a listener's ear. The analog speech are sampled, coded, encapsulated into packets, transmitted over the Internet, received in the jitter buffer, decoded and finally converted back into the original analog form. A stringent delay budget must be ensured to keep a certain speech quality.

Delay may result in difficulties during conversation. According to subjective tests, this effect is unnoticeable with one-way delay below 100 ms, and it is acceptable below 150 ms. The limit value for most connections should be in the range of 300 to 350 ms with an upper limit of 400 ms. Above 400 ms, it is unacceptable for general network planning purposes [33][17].

The end-to-end delay consists of the following parts [34][35]:

1. Vocoder delay, including the algorithm delay, look-ahead delay and processing delay.

The vocoder operates on a frame basis; the algorithm delay is the frame size. In addition, the vocoder also looks into the next frame to provide guidance in compressing the current frame, causing a look-ahead delay. The time required to process an input frame is assumed to be equal to the frame size since efficient use of processor resources will be accomplished when an encoder/decoder pair fully uses the available processing power [17]. Thus, the vocoder delay is given by Equation 2.6 and summarized in Table 2.5:

$$\text{vocoder delay} = 2 \times \text{frame size} + \text{look-ahead} \quad (2.6)$$

Table 2.5 Vocoder delay contributions

Vocoder	Algorithm Delay (ms)	Look-ahead Delay (ms)	Processing Delay (ms)	Total Delay (ms)
G.723.1	30	7.5	30	67.5
G.729	10	5	10	25

2. Network delay, including packetization delay and queuing delay. At the send side, if more than one frame is filled into a single packet, further delay is added. Queuing delay occurs at the routers and gateways of the network. It depends on the routing algorithm, traffic load and packet priority. It also changes significantly because of the statistical nature of the arrival process.

3. Jitter buffer delay. A jitter buffer is used at the receive side to compensate for the varying delay. The receiver holds the first packet in the jitter buffer for a while before playing out, and introduces more delay.
4. Transmission delay. Transmission delay is the time required to put the packet on the transmission line and depends on the packet size and transmission bit rate. Usually, it is in the order of microseconds and can be ignored.
5. Propagation delay. Propagation delay is the time required for the signal to travel from one point to the end. The travel speed is light speed or a bit lower than it, as determined by the media. The delay depends on the distance between two sides and can be significant if the signal travels through a long distance, such as in satellite communication.

In our research, we are concerned with the delays introduced by the vocoder, IP network and jitter buffer. The transmission delay and propagation delay are supposed to be relatively small and thus neglected.

2.6.2 Delay Jitter

The nature of packet-based transmission causes varying packet interarrival time. The jitter is defined as the difference D between the inter-departure and interarrival times of packets [36]. If S_i is the RTP timestamp from packet i , and R_i is the time of arrival in the RTP timestamp units for packet i , then for two consecutive arrival packets i and $i+1$, the D is given by:

$$D(i, i+1) = (R_{i+1} - R_i) - (S_{i+1} - S_i) = (R_{i+1} - S_{i+1}) - (R_i - S_i) \quad (2.7)$$

A sequence of negative jitters (clustering) can result in buffer overrun and consecutive packet loss; a sequence of positive jitters (dispersion) can result in consecutive packets experiencing excessive delay [36]. Jitter smoothing is achieved by adding artificial delay to the received packets at the jitter buffer. Real-time applications require stringent end-to-end delay. Too large a jitter buffer will make the total delay unacceptable; too small a jitter buffer is not enough to compensate for the delay variation.

The jitter buffer size can be fixed or adaptively changed according to network variability. A jitter buffer adaptation algorithm updates the playout delay on a talkspurt-by-talkspurt basis. Thus the silence interval between two talkspurts at the receiver may be artificially elongated or shortened [37][38]. This small amount of length change in silence is not noticeable in the played out speech [39][40]. However, jitter may occur in the talkspurt if the adaptation algorithm is not correctly implemented, and the subjective effect is much more annoying than the jitter in a silence period.

2.6.3 Packet Loss

Packet loss is common in packet-switched transmission due to its nature. When the network is congested, the router buffer is full and the tail part of the queue is dropped. Link failure and channel error also result in packet loss. In addition, the jitter buffer can cause packet loss if some packets arrive too late to be played out or upon overrun when packets have to be dropped or discarded; they are considered effectively lost.

Packet loss affects the speech quality. For example, for vocoder G.729, speech quality will drop to 2.98 MOS in the case of 5% random frame erasure [30].

Packet loss may be random or burst. In [7], the results show that, for ITU-T Rec. G.711 with packet loss concealment, when the loss rate is low (no more than 3%), the random loss and burst loss cause the same distortion in speech quality; when the loss rate is increasing (more than 5%), the random loss gives higher speech quality because the burst loss leads to longer time clipping and further degrades speech quality.

Several techniques have been used to deal with packet loss; they can be sender-based or receiver-based [41]. The sender-based repair needs the participation of the sender to recover the lost packet and includes: interleaving and Forward Error Correction (FEC).

1. Interleaving. Packets are rearranged before transmission so that previously consecutive packets are separated by a guaranteed distance, and they are returned to their original order at the receiver. It disperses the effects of the burst packet loss that may last for several packets, so one long gap becomes multiple short gaps, thus the impairment is less perceptible. Interleaving doesn't increase the bandwidth requirement but it increases the delay. It is an attractive method if the delay budget is satisfied.
2. Forward Error Correction (FEC). This technique relies on the addition of repair data to the stream. A lost packet can be recovered by using the redundancy information. Using FEC results in overhead and more repairing data during transmission. It could increase network congestion and hence packet loss, worsening the problem it wants to solve.

The receiver-based method, also called error concealment, is a method for producing a replacement similar to the lost packet without sender intervention. In the thesis, a lost packet is concealed by this method, which will be given in detail in Chapter 4.

CHAPTER 3 SIMULATION SYSTEM DESIGN

3.1 Simulation Structure

In the thesis, we carry out simulations to investigate the effects of packet loss and delay jitter on speech quality. The simulation block diagram is shown in Figure 3.1.

Figure 3.1 Simulation block diagram

The reference signal is encoded by vocoders and filled into packets. Then, packet loss and variable delay are introduced by simulation during packet transmission. At the receive side, packets are buffered and an error concealment technique is applied to recover the lost packets before decoding. Finally, the reference signal and the corresponding degraded signal are sent into the DSLA for MOS measurement.

3.2 Simulation Design Description

3.2.1 Reference Signal Selection

To select the reference signal, the following factors should be considered according to ITU-T Rec. P.830 [22], P.800 [4] and the DSLA user manual [20]:

- Real or artificial voices
- Temporal structure, duration
- Filtering
- Listening level
- Resolution

(1) Real or artificial voices

In our simulation, a total of 20 clean speech samples from 2 male and 2 female speakers are used.

Both real voices and artificial voices may be used as the reference signal in objective measurement. If real voices are used, it is recommended that a minimum of two male talkers and two female talkers should be used for each testing condition. Also, for each talker, a minimum of two and a maximum of five speech samples are recommended. These speech samples should be clean.

(2) Temporal structure and duration

Our 20 speech samples are roughly 8 seconds in length with 50% of active speech.

Testing signals should include speech bursts separated by silent periods. As a guide, 1-3 seconds is a typical duration of a speech burst. Certain types of voice activity detectors

are sensitive only to silent periods that are longer than 200 ms. Speech should be active for between 40% and 80% of the time. PAMS is optimized for samples of 8s in duration containing at least 4s of active speech. As a guide, the minimum length for a measurement to give a representative PAMS score is about 6s, containing at least 3s of active speech.

A typical speech sample selected in our simulation is shown in Figure 3.2.

Figure 3.2 A speech sample

(3) Filtering

No analog filtering is applied to our speech sample.

Analog connections introduce some degree of filtering. A Modified Intermediate Reference System (MIRS) sending frequency characteristic is often used to simulate the sending frequency characteristics of a telephone handset. It is generally accepted this

makes little difference to speech quality. Furthermore, PAMS identifies any filtering that has taken place in the network and cancels its effect automatically. So, our speech signals are not pre-processed by a MIRS sending filter.

(4) Listening level

The preferred listening level is 79 dB SPL at the ear reference plane.

In a digital speech file, the preferred listening level is -26 dB relative to the overload point of the system, which is equivalent to 79 dB SPL. The MOS changes with the listening level. In the test, we use several input signals and average the results. The input signals are multiplied by 0.3 to 2.0 with increment of 0.1 in turn, and then degraded by vocoder G.729. The relationship between MOS and listening level is shown in Figure 3.3.

Figure 3.3 MOS vs. speech listening level

Too weak speech may not be distinguished clearly, resulting in a low MOS score estimate; too loud speech may be quite annoying, which also results in a low MOS score estimate.

PAMS assumes a constant listening level of approximately 79 dB SPL, by applying a gain to both the reference and degraded signals to bring them to this level.

(5) Signal resolution

In our simulation, we use 8000 Hz, 16-bit linear PCM signals.

In PSTN, analog speech is sampled at 8000 Hz and quantized with 8 bits. A-law or μ -law companding is used to increase the dynamic range of the 8 bit samples. Typically 13-bit or 14-bit data in a linear scale is companded to 8-bit logarithmic scale for transmission over the 64 kb/s channel.

However, PAMS operates on 16-bit linear PCM speech samples. The speech will be distorted if it is represented with fewer bits, leading to a lower MOS score. In the test, we also use several input signals and average the results. The input signals, originally stored in 16-bit resolution, are represented by 8 to 13 bits with increment of 1 bit in turn, and then degraded by vocoder G.729. The relationship between MOS and signal resolution is shown in Figure 3.4.

It can be seen that there is no perceivable quality degradation when using 12 bits or more linear PCM to represent the sample. So, A-law or μ -law companding has no noticeable subjective effect on speech quality. Moreover, the effects of quantization distortion are included by parameter *qdu* (Quantization Distortion Unit) in the E-model. It

is unnecessary to compand the speech sample by A-law or μ -law before tests. Therefore, in our simulation, we use the 8000 Hz, 16-bit linear PCM signals.

Figure 3.4 MOS vs. linear PCM resolution (bits)

In conclusion, we use 20 speech samples from two male and two female speakers. The samples are clean, in 16-bit, 8000 Hz linear PCM format, roughly 8 seconds in duration with 50% of active speech. They are level adjusted to 79 dB SPL without analog filtering.

3.2.2 Vocoder Selection

Vocoders G.723.1 and G.729 are used in our simulation. The C codes simulating the coders/decoders are from ITU-T.

The G.723.1 is a fixed-point dual-rate vocoder with the high rate 6.3 kb/s and low rate 5.3 kb/s. The Annex B specifies a floating-point algorithm for fast implementation. The simulation is conducted on the 20 speech samples by implementing fixed-point and floating-point algorithms at both rates. Their MOS scores are given in Table 3.1.

Table 3.1 MOS scores for 20 speech samples under vocoder G.723.1

Speech samples	G.723.1 Fixed-point 5.3 kb/s	G.723.1 Floating-point 5.3 kb/s	G.723.1 Fixed-point 6.3 kb/s	G.723.1 Floating-point 6.3 kb/s
or001.wav	4.01	4.00	4.16	4.13
or002.wav	4.08	4.06	4.08	4.09
or003.wav	4.11	4.07	4.16	4.17
or004.wav	4.13	4.12	4.29	4.27
or005.wav	4.21	4.24	4.26	4.28
or006.wav	4.17	4.16	4.20	4.23
or007.wav	4.05	4.00	4.07	4.10
or008.wav	4.12	4.14	4.21	4.18
or009.wav	4.15	4.11	4.21	4.21
or010.wav	4.13	4.14	4.29	4.23
or011.wav	4.12	4.16	4.24	4.22
or012.wav	4.18	4.11	4.24	4.22
or013.wav	4.11	4.06	4.12	4.24
or014.wav	4.24	4.20	4.27	4.23
or015.wav	4.20	4.14	4.27	4.27
or016.wav	4.13	4.17	4.24	4.30
or017.wav	4.14	4.12	4.26	4.34
or018.wav	4.22	4.17	4.25	4.25
or019.wav	4.15	4.14	4.20	4.22
or020.wav	4.17	4.22	4.20	4.19
Average	4.1410	4.1265	4.2110	4.2185

The results show that the high rate vocoder has higher speech quality. At each rate, there is no obvious difference in speech quality between the fixed-point and the floating-point algorithm. However, the simulation code for the floating-point algorithm executes faster than that of the fixed-point algorithm.

In our simulation, for the vocoder G.723.1, we adopt the floating-point algorithm at both rates for fast implementation; for the vocoder G.729, we use the fixed-point algorithm upon availability.

3.2.3 Packetization Considerations

We assume one frame per packet.

Each packet consists of a fixed header and a payload. For efficient utilization of network resources, it is desirable to encapsulate several frames into one packet or to compress the 40-byte header into several bytes. Putting several frames into one packet will lead to extra delay. We simply assume one frame per packet.

3.2.4 Network Simulation

In simulating network conditions, the following factors are considered: delay, delay jitter, packet loss, error concealment and jitter buffer algorithms. The simulation codes are written in Matlab.

(1) Delay

As we discussed in section 2.6, transmission delay and propagation delay are ignored in our simulation. The vocoder delay is fixed. We only simulate the network delay and the jitter buffer delay. It will be described in detail in Chapter 5.

(2) Packet loss

Packet loss may be random or burst. We assume random packet loss.

(3) Error concealment

A number of error concealment methods are examined. Among them, the repetition method, which replaces the lost packet by the previous packet, is used in our simulation. More details are given in Chapter 4.

(4) Jitter buffer algorithm

Jitter buffer adaptation is possible only if the encoder doesn't produce any frames during the silent interval so the buffer size can be adjusted on a talkspurt-by-talkspurt basis. It means the VAD must be enabled. In all of our simulations, the VAD is turned off, and the packets are generated at equal time interval with no differentiation of active speech and silence. We use the fixed buffer size to simulate the jitter buffer. More details are given in Chapter 5.

3.3 MOS Measurement and Data Analysis

The MOS scores are measured by the DSLA implementing the PAMS algorithm. A DSLA batch processor is used to automatically process a large number of reference and degraded file pairs.

It is noticed that DSLA sometimes gives questionable results which are much lower than the average, especially in severe degradation conditions. The questionable data should be discarded when the average is calculated. To deal with this problem, a simple statistical method is used: for each speech sample and each testing condition, the mean and the standard deviation of data from several repeated measurements are calculated first, the data which deviate more than 3 times the standard deviation from the mean are

treated as questionable data, and therefore discarded. The data discarding happens infrequently, with a probability of several in 1000. Even if not properly treated, it will not affect our results in general.

The simulation is usually repeated 10 times under each testing condition. The MOS scores in the thesis are given by averaging these repeated tests, after discarding a few questionable data.

CHAPTER 4 EFFECT OF PACKET LOSS ON SPEECH QUALITY

In this chapter, the effects of packet loss on speech quality are studied. To address the problem of packet loss, several receiver-based error concealment techniques are examined. One of them, the repetition method, which replaces the lost packet by the previous one, is adopted because of its effectiveness and simple implementation. We propose a formula to quantify the relationships between MOS and packet loss rate. The formula is validated and shows good accuracy.

4.1 Effects of Packet Loss for Different Error Concealment Techniques

4.1.1 Error Concealment Techniques

The idea of error concealment is to regenerate a replacement for the lost packet which is similar to the original one [41][42]. The receiver-based error concealment can be further divided into three classes: insertion-based, interpolation-based and regeneration-based. We aim at the insertion-based concealment method.

The insertion-based concealment method replaces the lost packet by inserting a simple fill-in, which can be:

- Nothing, just splicing packets together
- Silence
- Noise
- Previous packet, repetition

In splicing, the timing of the stream is disrupted and this method performed poorly. 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 for the lost packet with white (background) noise has been shown to perform better than silence substitution because of the ability of the human brain to repair the received message if there is some background noise (known as *phonemic restoration*) [34]. Repetition replaces the lost packet by the one arriving immediately before the loss. It has low computation complexity and performs reasonably well.

The effects of insertion-based concealment methods — splicing, silence insertion and repetition methods — are examined. The repetition method is expected to give the highest speech quality.

The simulation diagram is shown in Figure 4.1. We only consider the effects of packet loss for certain error concealment methods.

Figure 4.1 Simulation diagram for packet loss

The simulation is repeated 10 times. In all cases, the standard deviation of the MOS is within 0.13 MOS.

4.1.2 Simulation Results

The simulation results for vocoder G.723.1 B and G.729 are illustrated in Figures 4.2, 4.3 and 4.4 respectively. The packet is lost randomly up to 10 percent.

We see that in all cases, the repetition method performs best because of the short-term self-similarity in human speech; it successfully replaces the lost packet with a similar one to some extent. The splicing method performs worst, since it further distorts the timing relationship of the surrounding packets. The silence insertion method performs in between the others, but it is just a bit better than the splicing method. Although the timing of the voice stream is kept, the lost meaningful speech samples can't be reconstructed in this way.

The splicing and silence insertion methods are only applicable in very low packet loss rate scenarios. The repetition method works reasonably well when the packet loss increases.

The repetition method will be adopted as a standard error concealment method in all of our following simulations.

Figure 4.2 MOS vs. packet loss for vocoder G.723.1 B 5.3 kb/s

Figure 4.3 MOS vs. packet loss for vocoder G.723.1 B 6.3 kb/s

Figure 4.4 MOS vs. packet loss for vocoder G.729

To quantify the effects of packet loss on speech quality, we run the simulation with a total of 25 random packet loss rates: 0, 0.5, 1, 1.5, 2, 2.5, 3, 4, 5, 6, 7, 7.5, 8, 9, 10, 12.5, 15, 17.5, 20, 25, 30, 35, 40, 45 and 50 percent.

The results are illustrated in Figure 4.5. It shows that: the MOS scores drop more slowly when the packet loss rates increase; vocoder G.729 always performs better than G.723.1 because of the higher bit rate and smaller frame length (10 ms); and, the speech quality for G.729 degrades more slowly than that of G.723.1, especially when the packet loss rates increase.

Figure 4.5 MOS vs. packet loss under repetition error concealment

4.2 Formulation

4.2.1 Former Research

Some researches have been carried out to examine the numerical relationships between packet loss rate and speech quality in MOS scale. A mathematical model is suggested in [43]:

$$\text{predicted MOS} = 4.0 - 0.7 \ln(\text{loss}) - 0.1 \ln(\text{size}) \quad (4.1)$$

where 4.0 represents the PSTN quality, the *loss* is the data loss ratio in percent from 0.5% to 50% and *size* is the frame size from 3 ms to 96 ms. The equation is derived from using 16-bit linear PCM and replacing the lost frames with silence.

In [44], Equation 4.1 is slightly changed to fit the testing results by:

$$predicted\ MOS = MOS_opt - C \ln(1 + loss) \quad (4.2)$$

where MOS_opt is the optimal MOS score without any impairment for a certain vocoder, $loss$ is packet loss rate from 0 to 10 percent, and C is a constant factor. Under the condition of no packet loss concealment, the C is found to be about 0.25 for vocoders that have a smaller degradation of quality (ITU-T Rec. G.711, G.723.1, G.729), or 0.4 for vocoders that are more sensitive to packet loss (ITU-T Rec. G.726-7, G.728).

In general, these researches model that MOS scores decline linearly at a constant rate with the packet loss rate in a logarithmic scale. C differs with vocoder types and error concealment techniques and is calculated by linear regression for the best fitting.

In our repetition method, the factor C in Equation 4.2 is calculated to be around 0.50-0.51. However, the prediction does not fit our measured MOS curve well, as depicted in Figure 4.6. Equations 4.1 or 4.2 can't give a good prediction of MOS.

Figure 4.6 MOS curve fitting with $K = 0.50$

The idea of the E-model is that the transmission quality can be predicted by using relevant parameters rather than doing the real measurement. The impairment from packet loss is represented by the equipment impairment factor, I_e . The predicted MOS thus would be independent of the speech samples used.

However, in the objective measurements, different speech samples give similar but different MOS scores, although they all meet the selection requirements specified in section 3.2.1. If we directly examine the relationships in the MOS domain, we will get different results.

Such relationships should be examined in the equipment impairment factor domain (*packet loss rate vs. I_e*), where the bias from the speech sample can be eliminated.

The up-to-date I_e values are tabulated in ITU-T Rec. G.113 Appendix I and illustrated in Figure 4.7. The detailed table is given in Appendix B to the thesis.

Figure 4.7 I_e values in ITU-T Rec. G.113 Appendix I [7]

These values only cover limited conditions and lack flexibility. For example, what are the I_e values for 10 percent packet loss, or using some error concealment methods? Furthermore, they are provisional only as they were determined in single or a few experiments [7]. So far, I_e values provided by ITU-T don't suffice.

4.2.2 Derivation of Equipment Impairment Factor

Our goal is to be able to predict the speech quality for other packet loss rates, error concealment methods or frame sizes. This is conducted in the equipment impairment factor domain by deriving I_e values used in the E-model.

To get stable I_e values, the measured MOS scores must be calibrated first. These steps are shown as follows [45][46]:

Step 1: Scale Transformation

This step transforms the measured MOS into the scale of equipment impairment factor by Equations 2.4 and 2.1 in turn with all the other parameters set to their default values. These I_e values are denoted by $I_{e,mea}$, reflecting the results from the specific speech samples, and are not necessarily equal to those defined by the ITU-T standard. The results of step 1 are illustrated in Figure 4.8.

Figure 4.8 $I_{e, mea}$ values from scale transformation

Step 2: Linear Interpolation

In this step, a linear interpolation is applied to the $I_{e, mea}$ values to correct them to the ITU-T proposed values, which is denoted by $I_{e, std}$.

According to the reference conditions of [45], several pairs of equipment impairment factor $I_{e, std}$ and $I_{e, mea}$ are available. These pairs can be represented by a straight line given by equation:

$$I_{e, std} = a \cdot I_{e, mea} + b \quad (4.3)$$

The coefficients a and b are determined by minimizing the MSE. Due to the availability of vocoders, five pairs of I_e values are used in linear interpolation, as tabulated in Table 4.1. We can see the $I_{e, mea}$ values without packet loss are lower than the ITU-T proposed values.

Table 4.1 I_e value pairs used in linear interpolation

Vocoder	MOS	$I_{e,mea}$	$I_{e,std}$
G.729	4.3435	3.0081	10
G.729 x2	3.9300	15.5966	20
G.729 x3	3.4995	25.2417	30
G.723.1 B 5.3 kb/s	4.1265	10.3579	19
G.723.1 B 6.3 kb/s	4.2185	7.5406	15
Note: x2, x3 designates double and triple tandems for the same vocoder respectively.			

The linear interpolation is shown in Figure 4.9, and the coefficients are found to be: $a = 0.8532$ and $b = 8.2640$. Then, Equation 4.3 is applied to all the $I_{e,mea}$ values from step 1. Thus, a stable equipment impairment factor, which is consistent with the ITU-T standard, is obtained. The interpolated I_e values are given in Figure 4.10.

Figure 4.9 Linear interpolation line for I_e

Figure 4.10 Interpolated I_e values

4.2.3 Formulation of Equipment Impairment Factor

As an observation from Figure 4.10, the I_e values increase logarithmically with the packet loss. We proposed a new formula to predict I_e values as a function of packet loss rate:

$$I_e = I_{e_opt} + C1 \cdot \ln(1 + C2 \cdot loss_rate) \quad (4.4)$$

where I_{e_opt} is the optimum (without packet loss) I_e values from the ITU-T standard, $loss_rate$ is the packet loss rate in percent from 0 to 50, and factors $C1$ and $C2$ are constants used to adjust the shape of the curve. $C1$ and $C2$ depend on the vocoder types, error concealment methods and number of frames per packet.

The fitness of our formula can be examined by root mean square error σ and correlation coefficient ρ :

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - y_i)^2} \quad (4.5)$$

$$\rho = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 (y_i - \bar{y})^2}} \quad (4.6)$$

where x_i is the interpolated I_e based on measurement for condition i , and $\bar{x}$ is the average of x_i ; y_i is the predicted I_e by the proposed Equation 4.4 for condition i , and $\bar{y}$ is the average of y_i . For the repetition error concealment method, the results are summarized in Table 4.2.

Table 4.2 $C1$ and $C2$ for repetition method

Vocoder	I_{e_opt}	$C1$	$C2$	ρ	σ
G.723.1 B 5.3 kb/s	19	37.40	0.05	0.9989	1.3080
G.723.1 B 6.3 kb/s	15	36.59	0.06	0.9986	0.9286
G.729 8.0 kb/s	10	25.05	0.13	0.9987	0.9548

Figure 4.11 graphically shows how our prediction fits the interpolated I_e for vocoder G.729. The solid line is for $C1 = 25.05$ and $C2 = 0.13$ by formula; the stars are the interpolated I_e . It can be seen that nearly perfect match is achieved.

Figure 4.11 I_e curve fitting for G.729

Under other testing conditions, the I_e values can be predicted by Equation 4.4 as well, with different sets of $C1$ and $C2$. For the error concealment by splicing or silence insertion we examined in Section 4.1, the results are summarized in Table 4.3, where the factors $C1$ and $C2$ are valid for packet loss rate up to 10 percent, as both methods only works under low packet loss scenarios.

Table 4.3 $C1$ and $C2$ for splicing, silence insertion methods

Vocoder	Method	I_{e_opt}	$C1$	$C2$	ρ	σ
G.723.1 B 5.3 kb/s	Silence	19	71.38	0.06	0.9988	1.6581
	Splicing	19	75.55	0.06	0.9992	1.5297
G.723.1 B 6.3 kb/s	Silence	15	90.00	0.05	0.9983	0.9164
	Splicing	15	94.59	0.05	0.9987	0.7487
G.729 8.0 kb/s	Silence	10	47.82	0.18	0.9997	0.3682
	Splicing	10	69.84	0.12	0.9980	1.1682

For the ITU-T proposed I_e values, such $C1$ and $C2$ can also be found and summarized in Table 4.4. Thus, for these cases, a I_e value for any packet loss rate up to 16 percent can be calculated directly.

Table 4.4 $C1$ and $C2$ for ITU-T proposed I_e values

Vocoder	I_{e_opt}	$C1$	$C2$	ρ	σ
G.723.1.A+VAD 6.3 kb/s	15	30.50	0.17	0.9991	0.5396
G.729A+VAD, 20 ms packet	11	30.00	0.16	0.9998	0.2945

The proposed formula 4.4 is very useful for MOS prediction. For example, if we want a fair speech quality ($MOS > 3$) in listening-only model, the tolerable packet loss rate is 10.8, 12.2 and 13.3 percent for vocoders G.723.1 B 5.3 kb/s, G.723.1 B 6.3 kb/s and G.729 respectively when lost packets are concealed by the repetition method.

4.3 Validation

In this section, we examine whether our formula holds if we use any arbitrary speech samples and how much the prediction error will be. We focus the validation test on the repetition error concealment method.

Two speech sample sets, namely, speech set 2 and speech set 3, are used in validation. Each set contains five speech samples which are totally different from the 20 speech samples we used before (called set 1).

Set 2 samples are good speech samples and meet the selection requirements. They usually yield similar speech quality as those in set 1. However, set 3 samples are not as good as set 1 samples in terms of speech quality, and thus not necessarily meet all the

selection requirements. For example, they may be clipped a little in magnitude or almost no silence intervals. Set 3 samples are designed to check the applicability of our formula to the worse speech sample scenarios.

The simulation is carried out under the same testing conditions and the measured MOS scores are shown in Figures 4.12 and 4.13.

Figure 4.12 Measured MOS – speech set 2

Figure 4.13 Measured MOS – speech set 3

The MOS curves in Figures 4.5 (set 1), 4.12 (set 2) and 4.13 (set 3) are similar but not identical for each vocoder as speech samples differ. Moreover, these measured MOS scores are also expected to deviate from the E-model prediction. For example, the MOS scores for G.729 without packet loss are 4.3435, 4.2560, 4.0700 and 4.1388 for speech set 1, set 2, set 3 and the E-model formula respectively.

Directly comparing the difference between the predicted MOS by our formula and the measured MOS from set 2 or set 3 doesn't mean too much, because the measured MOS is biased if not calibrated, while the MOS prediction formula is derived by calibrating the measurement.

So, a calibration is applied to speech set 2 and set 3 individually. The measured MOS scores will be changed a little after calibration, and then used for validating.

The procedures for speech set 2 and set 3 are same, and similar to those in section 4.2.2, here we just explain them briefly. First, the scale transformation is used to get $I_{e,mea}$ values from the measured MOS; then a linear interpolation line is found from the five pairs of I_e values of ITU-T standard and measurement. Finally, the measured I_e is interpolated and substituted into the E-model to derive the calibrated MOS scores.

The prediction error is given by Equation 4.7, where MOS_{pre} is the MOS by our prediction formula, and MOS_{cal} is the measured MOS after calibration.

$$MOS_{err} = MOS_{pre} - MOS_{cal} \quad (4.7)$$

For speech set 2, these procedures are given in Table 4.5, Figures 4.14 and 4.15. The coefficients of linear interpolation are found to be: $a = 0.8726$ and $b = 5.8340$. The MOS prediction errors are illustrated in Figures 4.16, 4.17 and 4.18 respectively for three vocoders and summarized in Figure 4.19.

Table 4.5 I_e value pairs used in linear interpolation – speech set 2

Vocoders	MOS	$I_{e,mea}$	$I_{e,std}$
G.729	4.2560	6.2880	10
G.729 x2	3.8300	18.0030	20
G.729 x3	3.3860	27.5618	30
G.723.1 B 5.3 kb/s	4.0420	12.7132	19
G.723.1 B 6.3 kb/s	4.1480	9.7267	15

Figure 4.14 $I_{e,mea}$ after interpolation – speech set 2

Figure 4.15 Calibrated MOS – speech set 2

Figure 4.16 MOS prediction error for G.723.1 B 5.3 kb/s – speech set 2

Figure 4.17 MOS prediction error for G.723.1 B 6.3 kb/s – speech set 2

Figure 4.18 MOS prediction error for G.729 – speech set 2

Figure 4.19 MOS prediction error – speech set 2

In conclusion, for speech set 2 (good speech samples), the calibrated MOS scores from measurement scatter around their predictions. The results show that the prediction formula demonstrates good accuracy. For vocoder G.729, in all loss rates from 0 to 50 percent, the error lies in $-0.10...+0.10$ MOS and the maximum error is 0.0962 MOS. For vocoder G.723.1 B 6.3 kb/s, the error lies in $-0.10...+0.26$ MOS and the maximum error is 0.2518 MOS; the prediction error is almost within 0.10 MOS when the loss rate is below 30 percent. For vocoder G.723.1 B 5.3 kb/s, the error lies in $-0.10...+0.28$ MOS and the maximum error is 0.2803 MOS; the prediction error is within 0.10 MOS when the loss rate is below 20 percent. After that, the prediction error becomes more pronounced.

For speech set 3, I_e pairs used in linear interpolation are given in Table 4.6, the coefficients are found to be: $a = 0.9680$ and $b = -1.7397$ for the interpolation line. The MOS prediction errors are illustrated in Figures 4.20, 4.21 and 4.22 respectively for three vocoders and summarized in Figure 4.23.

Table 4.6 I_e value pairs used in linear interpolation – speech set 3

Vocoders	MOS	$I_{e, mea}$	$I_{e, std}$
G.729	4.0700	11.9528	10
G.729 x2	3.6100	22.9151	20
G.729 x3	3.1280	32.6515	30
G.723.1 B 5.3 kb/s	3.6960	21.0461	19
G.723.1 B 6.3 kb/s	3.8500	17.5322	15

Figure 4.20 MOS prediction error for G.723.1 B 5.3 kb/s – speech set 3

Figure 4.21 MOS prediction error for G.723.1 B 6.3 kb/s – speech set 3

Figure 4.22 MOS prediction error for G.729 – speech set 3

Figure 4.23 MOS prediction error – speech set 3

For speech set 3 (bad speech samples), the prediction formula demonstrates good accuracy as well. For three vocoders, the prediction error is kept within 0.10 MOS when the packet loss rate is below 20 percent. The maximum prediction error is higher than that of speech set 2, and can achieve 0.38 MOS when packet loss rate is above 20 percent.

After examination with arbitrary test speech samples, we can conclude that our proposed formula holds, and can be used to give reliable MOS prediction independent of speech samples.

4.4 Summary

In this chapter, the effects of packet loss on speech quality are investigated. Two vocoders, G.723.1 B and G.729 are used in our simulation. We proposed a new formula to represent the relationship between packet loss rate and the equipment impairment factor I_e which is used in the E-model to predict the MOS score.

Error concealment techniques are used to recover the lost packet in packet voice communications. Several receiver-based error concealment methods, splicing, silence insertion and repetition, are compared. The results show the repetition method, which replaces the lost packet by the previous one, always gives the highest MOS score. It is adopted in our simulation due to its effectiveness and simple implementation.

In the present E-model, the available I_e values are limited in flexibility, accuracy and coverage. One frame per packet, random packet loss with repetition error concealment is simulated. We observe that the G.729 outperforms G.723.1 in speech quality especially when the packet loss rates are increasing.

A formula is proposed to predict the I_e . We model I_e value increases logarithmically with the packet loss rate. In the formula, two factors $C1$ and $C2$ are used, and different sets of $C1$ and $C2$ are found to represent different scenarios. In the derivation of I_e values, an interpolation stage is applied to keep the resulting values consistent with the ITU-T recommendation, and thus the bias from the reference signals is removed. These derived I_e values are stable and independent of speech sample used.

The proposed formula achieves good accuracy, which is validated by two different, arbitrary speech sample sets, one contains good speech samples, and the other contains bad speech samples. The prediction error is kept within 0.10 MOS when the packet loss rate is below 20 percent for all the vocoders. For the good speech samples, when the packet loss rate is above 20 percent, for vocoder G.729, the prediction error is still within 0.10 MOS up to 50 percent packet loss; for vocoder G.723.1, the prediction error becomes more pronounced with a maximum error of 0.28 MOS.

CHAPTER 5 EFFECT OF DELAY JITTER ON SPEECH QUALITY

The nature of packet voice transmission causes varying packet interarrival time. The jitter buffer is used at the receive side to smooth the traffic. Pareto-distributed network delay and fixed jitter buffer are simulated and their effects on speech quality are studied. We add a new parameter into the E-model and propose a new formula to present the impairment caused by delay jitter.

5.1 Delay Distribution

In VoIP, the voice packets may arrive at the receive side with a wide range of interarrival times even though they leave the send side with equal time intervals.

In our simulation, we focus on the effects of the network delay variation on speech quality as we explained in section 3.2.4. The other delays are fixed and have no effects on speech quality because the PAMS is a listening-only model.

5.1.1 Self-similarity in Network Delay

Modeling network delay distribution is a delicate problem. Empirical studies [47][48] on network traffic measurements from both Local Area Network (LAN) and Wide Area Network (WAN) demonstrated that actual network traffic is self-similar in nature. In [49][50][51], the researches showed the Internet packet delay also exhibits self-similarity.

Self-similar traffic displays the same or similar statistical properties over a wide range of time scales and therefore has long range dependence (LRD).

The traditional traffic model assumes the network traffic is memoryless. A Poisson process is memoryless and is widely used because of such statistical properties and mathematical simplicity. In the Poisson model [48], the interarrival times are both exponentially distributed and independent; the autocorrelation function also decreases exponentially. However, in self-similar traffic, the autocorrelation function decays more slowly than that of the exponential distribution, and the traffic is heavy-tailed. The Poisson model can't reflect the properties of network traffic.

A distribution is defined as heavy-tailed if: $P[T \geq t] \sim ct^{-\eta}$ with $\eta > 0$ as $t \rightarrow \infty$, which means that for some η and some constant c , the ratio $P[T \geq t] / ct^{-\eta}$ tends to 1 as $t \rightarrow \infty$ [48]. The degree of self-similarity can be defined by a single parameter H , known as the Hurst parameter. For self-similar traffic with long-range dependence, $0.5 < H < 1$, the degree of self-similarity increases as H increases. The parameter H can be measured from traffic.

5.1.2 Pareto Distribution

The Pareto distribution is a suitable model for network traffic that exhibits self-similarity [52][53]. The probability distribution function (PDF) and cumulative distribution function (CDF) of Pareto distribution are given:

$$f(t) = \frac{\beta a^\beta}{t^{\beta+1}} \quad (5.1)$$

$$F(t) = P[T \leq t] = 1 - (\alpha/t)^\beta \text{ with } \alpha > 0, \beta > 0 \text{ and } t \geq \alpha \quad (5.2)$$

where α , the location parameter, determines the location of distribution, and β , the shape parameter, determines the shape of distribution, just as their names imply. The complementary CDF, the mean and variance of the Pareto distribution are given in Equations 5.3, 5.4 and 5.5 respectively:

$$F^c(t) = P[T > t] = (\alpha/t)^\beta \text{ with } \alpha > 0, \beta > 0 \text{ and } t \geq \alpha \quad (5.3)$$

$$E[T] = \frac{\alpha\beta}{\beta-1} \text{ with } \beta > 1 \quad (5.4)$$

$$Var[T] = \frac{\alpha^2 \beta}{(\beta-1)^2(\beta-2)} \text{ with } \beta > 2 \quad (5.5)$$

From Equations 5.4 and 5.5, it is clear that when $\beta > 2$, the mean and variance of distribution are finite, the process is not self-similar; when $1 < \beta < 2$, the mean exists but the variance is infinite; and when $0 < \beta < 1$, both of them are infinite. The shape parameter β is associated with the Hurst parameter H by the following equation [52]:

$$\beta = 3 - 2H \quad (5.6)$$

The self-similar traffic can be generated by a Pareto distribution with $1 < \beta < 2$, when $\beta \rightarrow 1$, $H \rightarrow 1$, implies the degree of self-similarity is much intensified; when $\beta \rightarrow 2$, $H \rightarrow 0.5$, implies a less intensified self-similarity.

Figure 5.1 illustrates the complementary CDF of a Pareto distribution and a Poisson distribution that have the same mean. The Pareto distribution decays hyperbolically, while Poisson distribution decays exponentially, and the former decays more slowly than the latter.

Figure 5.1 Complementary CDF of Pareto distribution and Poisson distribution

In our simulation, we use a Pareto distribution to model the network delay. The parameter β is chosen from 1.2 to 1.9 with increments of 0.1, equivalent to a Hurst parameter H ranging from 0.90 to 0.55 with interval of 0.05. In addition, we assume that the average network delay is 50ms, i.e., the mean of the distribution is 50ms, and thus the parameter α can be determined from Equation 5.4. The general assumption of 50ms of average delay will not weaken our simulation results for the following two reasons:

1. The PAMS is a listen-only model; the MOS score is affected by the delay variation rather than the fixed delay.
2. For a given β , changing average delay will change α only, which means the location changes but the shape of distribution is not affected. The change of location is similar to the fixed delay change for all packets, which has no effect in PAMS measurement.

Figure 5.2 shows a network delay trace generated by a Pareto distribution with several long-delay spikes, and Figure 5.3 is the histogram which shows the heavy-tailed delay distribution.

The variability of delay is measured by delay jitter. The jitter depends on the delay distribution; it can be quite pronounced even when the average delay is low. Pareto distribution is used to simulate heavy-tailed network delay, and the resulting delay jitter is also heavy-tailed because it is defined as the difference between inter-departure and interarrival times. The PDF of delay jitter provides the information about individual jitter length and is shown in Figure 5.4. The jitter is a zero mean process and symmetric about the y -axis. Most jitters are bounded in a certain size, but the rest span over a long range.

Figure 5.2 Pareto-distributed delay trace

Figure 5.3 Histogram of Pareto delay

Figure 5.4 Jitter PDF

5.2 Jitter Buffer Algorithm

In the jitter buffer, the first packet is artificially held for a certain time before being played out so that sufficient packets are received for continuous playback. This holding time is the measure of the buffer size. Common buffer sizes range from 50 to 100 ms [34]. The buffer size can be fixed or adaptively adjusted during a call [39][54]. The buffer size adaptation is implemented by controlling its filling level, when the level of the buffer is increasing, some short pause sequences will be dropped deliberately to drain the buffer at a faster rate; when the level of the buffer is decreasing, some additionally short pause sequences will be inserted to drain the buffer at a slower rate [17]. Real-time applications require stringent end-to-end delay. When a packet arrives later than some threshold, it is considered as effectively lost and a gap may occur in the playout stream.

As explained in section 3.2.4, here we consider two fixed buffer policies in [55][56]: I-policy and E-policy. I-policy ignores the late arrival packets in order to preserve timing, while E-policy expands the playout time in order to preserve information.

Figure 5.5 illustrates how these two policies affect the playout of packets. We assume the voice packets are generated at 30 ms intervals corresponding to the vocoder G.723.1 frame size. They experience varying network delay during transmission, and arrive at the receive side with different interarrival times. Packets 2 and 3 arrive out of sequence, and packets 5, 6 and 7 arrive in clustering. The fixed buffer size is assumed to be 100 ms, which means the first packet is held for 100 ms before playout. In the jitter buffer, the arrival packets are reshuffled to the correct order and most of the delay jitter is absorbed. In I-policy, the packet 5 can't be played out as it arrives at 350 ms, which is later than its scheduled playout time of 330 ms, and thus a gap is introduced into the output stream.

However, in E-policy, when packet 5 is late, we wait for another 30 ms. All the packets starting with packet 5 can be played out but a gap between 330 ms and 360 ms is introduced.

Figure 5.5 I-policy and E-policy

In I-policy, the gap results from ignoring the late packet. If we use certain error concealment techniques to retrieve the late packet, the gap is filled and timing is not distorted. The problem of delay jitter becomes the problem of packet loss. The quality of speech is affected in the same way as with packet loss under some error concealment techniques, which we have already simulated in Chapter 4. Therefore, we will not focus on the I-policy.

In E-policy, the gap results from expanding the waiting time. This policy aims to minimize the packet loss at the cost of additional delay and some gaps are inserted into the output stream. We use E-policy in our simulation.

In E-policy, the effects will be analogous to insertion of pause into the silence interval in the adaptive buffer algorithm. The main difference is that, in our case, silence may be inserted between talkspurts as well because we disable VAD. The speech quality is expected to be a little worse than when using the adaptive buffer algorithm.

The vocoder can tolerate packet loss to some extent. Especially with certain error concealment methods, the reconstructed speech quality is still acceptable [38]. So, it is beneficial to accept a low packet loss rate to achieve a stringent delay budget instead of waiting for a quite long time. In our simulation, we expand the waiting time for one or two frame sizes. If the packet hasn't arrived yet, it is considered lost and concealed by the repetition method.

5.3 Simulation Results

The simulation diagram is shown in Figure 5.6. The network delay simulator generates a Pareto-distributed delay, the H ranges from 0.55 to 0.9, with increments of 0.05 to reflect the intensity of self-similarity in the packet delay, E-policy is used in the jitter buffer, and the buffer size changes from 30 ms to 100 ms with increments of 10 ms.

Figure 5.6 Simulation diagram for network delay

The simulation is repeated 10 times. In all cases, the standard deviation of the MOS score is within 0.25 MOS.

The simulation results for vocoder G.723.1 B and G.729 are illustrated in Figures 5.7, 5.8 and 5.9 respectively. In all cases, when the H increases, which implies the delay distribution is more heavy-tailed, the MOS score falls; when jitter buffer size increases, which means more delay jitter is absorbed, the MOS score rises. The measured MOS scores range from 3.0 to 3.8, much lower than the scenarios without packet loss. Even when the jitter buffer size increases to 100 ms, the speech quality is still not improved a lot. The delay jitter is a major impairment to speech quality.

Figure 5.7 MOS vs. H and buffer size for vocoder G.723.1 B 5.3 kb/s

Figure 5.8 MOS vs. H and buffer size for vocoder G.723.1 B 6.3 kb/s

Figure 5.9 MOS vs. H and buffer size for vocoder G.729

5.4 Formulation

We add a new parameter, jitter impairment factor I_j into the E-model to represent the effects of delay jitter on speech quality. We assume that I_j is additive in the psychological scale, thus the extended E-model function becomes:

$$R = R_o - I_s - I_d - I_e - I_j + A \quad (5.7)$$

where I_j is a function of degree of traffic self-similarity H and buffer size T . We assume H and T are independent, I_j can be represented by:

$$I_j = f_1(H) + f_2(T) \quad (5.8)$$

Further, we use a second order polynomial to represent f_1 , and make sure f_1 must increase when H increases within its range ($0.5 < H < 1$); we use an exponential function to represent f_2 , f_2 decays when T increases. The Equation 5.8 becomes:

$$I_j = C1 \cdot H^2 + C2 \cdot H + C3 + C4 \cdot e^{-T/K} \quad (5.9)$$

where $C1$, $C2$, $C3$, $C4$ are coefficients and K is time constant.

5.4.1 Derivation of Jitter Impairment Factor

The jitter impairment factor I_j can be derived from the same steps as described in section 4.2.2.

After scale transformation, the measured MOS is converted to jitter impairment factor $I_{j,mea}$. The $I_{j,mea}$ values for three vocoders are shown in Figures 5.10, 5.11 and 5.12 respectively.

Figure 5.10 $I_{j,mea}$ values from scale transformation for vocoder G.723.1 B 5.3 kb/s

Figure 5.11 $I_{j,mea}$ values from scale transformation for vocoder G.723.1 B 6.3 kb/s

Figure 5.12 $I_{j,mea}$ values from scale transformation for vocoder G.729

Then, the linear interpolation is applied to the $I_{j,mea}$ to remove the bias from the speech samples. The linear interpolation line is given in Equation 4.3 where $a = 0.8532$ and $b = 8.2640$.

It is important to note that the interpolated $I_{j,mea}$ consists of two parts: one part is from the delay jitter impairment I_j ; the other part is from the vocoder impairment without packet loss. The latter is 19, 15 and 10 for G.723.1 B 5.3 kb/s, 6.3 kb/s and G.729 respectively. They should be subtracted. Figures 5.13, 5.14 and 5.15 show the I_j after linear interpolation and subtraction of vocoder impairment.

Figure 5.13 I_j values for vocoder G.723.1 B 5.3 kb/s

Figure 5.14 I_j values for vocoder G.723.1 B 6.3 kb/s

Figure 5.15 I_j values for vocoder G.729

It is interesting to see that G.729 is the most sensitive to delay jitter as the I_j is the largest among the three. One explanation for this can be that each G.729 packet contains a smaller portion of speech (10 ms compared with 30 ms of G.723.1). When they are separated by gaps, the rhythm of the speech is more disrupted and hence the perceived quality is worse.

5.4.2 Formulation of Jitter Impairment Factor

We use formula 5.9 to represent I_j values as a function of H and T . In the formula, H is the Hurst parameter from 0.55 to 0.9; T is buffer size from 30 ms to 100 ms. $C1$, $C2$ and $C3$ are used to adjust the shape of the polynomial; $C4$ and K are used to adjust the exponential decay. The fitness of the formula is examined by root mean square error σ

and correlation coefficient ρ as well. All the coefficients and time constant are calculated by linear regression. The results are summarized in Table 5.1.

Table 5.1 Coefficients and time constant for prediction model

Vocoder	$C1$	$C2$	$C3$	$C4$	K	ρ	σ
G.723.1 B 5.3 kb/s	-8.3	22.3	-1.1	9.0	40	0.9350	0.6937
G.723.1 B 6.3 kb/s	-23.7	45.4	-6.8	9.7	36	0.9478	0.7799
G.729 8.0 kb/s	-15.5	33.5	4.4	13.6	30	0.9499	0.7836

The I_j decays exponentially at a rate of 30-40 ms with increasing jitter buffer size. On the other hand, I_j decreases parabolically with decreasing H .

5.4.3 MOS Prediction under Delay Jitter

Under the assumption of Pareto-distributed network delay and the fixed buffer algorithm (E-policy), I_j can be predicted from our proposed Equation 5.9 with the constants given in Table 5.1. Then we substitute I_j and I_e into Equation 5.7 to calculate overall rating R in the E-model, and finally transform R into a MOS score. Figure 5.16 shows a predicted MOS surface for vocoder G.729, when other parameters are set to their default values. The predicted MOS surface for G.723.1 is similar.

Figure 5.16 Predicted MOS as a function of H and T for G.729

5.5 Validation

As our prediction formula in Chapter 4 achieves good accuracy no matter speech set 2 or set 3 used in validation. In this section, we only use speech set 2 to validate the proposed formula 5.9. The simulation is carried out under the same testing conditions. The measured MOS scores for G.723.1 B 5.3 kb/s, 6.3 kb/s and G.729 are illustrated in Figures 5.17, 5.18 and 5.19 respectively. These figures resemble Figures 5.7, 5.8 and 5.9 for speech set 1.

Figure 5.17 MOS vs. H and buffer size for vocoder G.723.1 B 5.3 kb/s – speech set 2

Figure 5.18 MOS vs. H and buffer size for vocoder G.723.1 B 6.3 kb/s – speech set 2

Figure 5.19 MOS vs. H and buffer size for vocoder G.729 – speech set 2

All the MOS scores of speech set 2 are calibrated before comparison. First, the scale transformation is used to get $I_{j,mea}$ values from the measured MOS; then $I_{j,mea}$ is linearly interpolated by $I_j = 0.8726 \cdot I_{j,mea} + 5.8340$, and finally the interpolated I_j is substituted into the E-model to get MOS scores. These calibrated MOS scores are compared with prediction from our proposed formula, the prediction error is defined by Equation 4.7. The error surfaces are illustrated in Figures 5.20, 5.21 and 5.22.

The error surfaces show wide irregularities of MOS prediction error. But all the errors are kept within $-0.10 \dots +0.20$ MOS for the three vocoders. We can conclude that our model has achieved good accuracy in predicting MOS degradation caused by delay jitter.

Figure 5.20 MOS prediction error for vocoder G.723.1 5.3 kb/s

Figure 5.21 MOS prediction error for vocoder G.723.1 6.3 kb/s

Figure 5.22 MOS prediction error for vocoder G.729

5.6 Summary

In this chapter, the effects of delay jitter on speech quality are examined. We introduce a new parameter, jitter impairment factor, I_j into the E-model. We give a formula to predict I_j based on simulation and measurement, then the MOS score can be calculated from the extended E-model function.

Today, it is widely accepted that Internet traffic is self-similar. The degree of self-similarity is represented by the Hurst parameter H , and Pareto distribution can be used to model such traffic. In our simulation, the network delay is assumed to follow Pareto distribution.

At the receive side, a jitter buffer is used to hold the first packet for a certain time to ensure continuous play back. Such holding time is the size of the jitter buffer. Our simulation focuses on one of the fixed buffer algorithms, E-policy. Gaps are introduced into the output audio stream, thus the speech quality is degraded.

In the simulation, the H is chosen from 0.55 to 0.9 and jitter buffer size is chosen from 30 ms to 100 ms. Vocoders G.723.1 B and G.729 are used in the simulation with one frame per packet. The results show that speech quality falls if the degree of traffic intensity increases or jitter buffer size decreases. Also, vocoder G.729 is more sensitive to delay jitter than vocoder G.723.1.

As the E-model doesn't include the impairment from delay jitter, we introduce a new parameter, I_j to reflect such impairment. A formula is then proposed to predict the I_j as a function of H and T . These I_j values are then substituted into the extended E-model function to predict MOS score. In the derivation of I_j values, linear interpolation is also applied. The proposed formula is validated by another set of speech samples. In all conditions, the prediction error is within $-0.10...+0.20$ MOS.

As an important observation from our simulation results, the impairment from delay jitter is more annoying than that of packet loss. For example, for vocoder G.723.1 B 6.3 kb/s, the measured MOS scores range from 3.3 to 3.7 MOS, like the results of 5 to 10 percent random packet loss. Even if the jitter buffer is set to 100 ms, the speech quality still can't be improved a lot. The results imply that the jitter buffer is not large enough to compensate the heavy-tail distributed delay.

CHAPTER 6 COMBINED EFFECTS OF PACKET LOSS AND DELAY JITTER ON SPEECH QUALITY

In this chapter, we examine the combined effects of packet loss and delay jitter on speech quality. In the previous two chapters, the effects are investigated separately, and the degradation in MOS scale is transformed into a corresponding equipment impairment factor and jitter impairment factor in the extended E-model formula. We assume these two factors are additive. In this chapter we will do the additivity check of the combined effects.

6.1 Simulation Results

First, we overview the results obtained in Chapter 4 and Chapter 5. In Chapter 4, we propose a formula 4.4 to predict the equipment impairment factor I_e as a function of random packet loss. In Chapter 5, we introduce a new parameter I_j into the E-model to represent the delay jitter impairment, and propose a formula 5.9 to predict I_j as a function of buffer size and delay distribution. The formula is derived under the Pareto distributed network delay and fixed buffer size using E-policy.

We assume these two factors I_e and I_j are additive when the effects of packet loss and delay jitter on speech quality are combined. The additivity is checked here.

The simulation block diagram is shown in Figure 6.1.

Figure 6.1 Simulation block diagram for combined effects

The simulation of combinations of all testing parameters in Chapter 4 and 5 is an enormous task which involves nearly one million pairs of speech samples. We simplify the simulation as follows:

- Packet loss percentage: 0, 0.5, 1, 2, 3, 5, 7.5, 10, 15 and 20, as the prediction error is more than 0.10 MOS if packet loss rate reaches 20 percent.
- Buffer size T : 50, 70 and 90 ms.
- Hurst parameter H : 0.6, 0.75 and 0.9.

The simulation is repeated twice under the same conditions. The standard deviation of MOS is mostly within 0.20 MOS with the maximum of 0.40 MOS.

The simulation results are given in Tables 6.1, 6.2 and 6.3 for vocoders G.723.1 B 5.3 kb/s, G.723.1 B 6.3 kb/s and G.729 respectively

Table 6.1 Combined effects for vocoder G.723.1 B 5.3 kb/s — measured MOS

Loss %		0	0.5	1	2	3	5	7.5	10	15	20
	T (ms)										
H = 0.90	50	3.3075	3.2540	3.1913	3.2440	3.1569	3.0465	2.7162	2.5163	2.2773	2.0130
	70	3.4582	3.3110	3.2962	3.2185	3.2405	3.0027	2.8835	2.7068	2.5262	2.2605
	90	3.4512	3.5223	3.4599	3.2967	3.2878	3.1305	2.9250	2.6022	2.3857	2.1207
H = 0.75	50	3.4100	3.3232	3.3293	3.2643	3.0565	3.1425	2.7645	2.6360	2.4013	2.0425
	70	3.4030	3.4482	3.2502	3.0680	3.1285	2.9780	2.8163	2.7197	2.5995	2.1447
	90	3.4998	3.2717	3.2177	3.2073	3.2128	3.0953	2.9638	2.6125	2.5257	2.0685
H = 0.60	50	3.4797	3.3988	3.3560	3.2142	3.2357	3.1275	2.8498	2.7637	2.3855	2.1375
	70	3.4950	3.4370	3.3320	3.2285	3.1827	3.0637	2.8665	2.6830	2.3748	2.1615
	90	3.5665	3.3782	3.4400	3.2355	3.4154	3.2964	3.0975	2.7553	2.3750	2.3447

Table 6.2 Combined effects for vocoder G.723.1 B 6.3 kb/s — measured MOS

Loss %		0	0.5	1	2	3	5	7.5	10	15	20
	T (ms)										
H = 0.90	50	3.4213	3.3447	3.2835	3.2153	3.0170	3.0225	2.7790	2.7135	2.4915	2.2215
	70	3.5320	3.3928	3.3778	3.1945	3.2317	3.1367	2.8758	2.8128	2.4840	2.1592
	90	3.6107	3.5685	3.3585	3.2745	3.2975	3.0957	3.0599	2.7540	2.5425	2.1025
H = 0.75	50	3.4450	3.2563	3.4295	3.2135	3.0540	3.0617	2.8292	2.6475	2.3068	1.9410
	70	3.3698	3.3790	3.4555	3.2808	3.2698	2.9362	2.8497	2.6908	2.4835	2.1318
	90	3.6260	3.5168	3.4510	3.3713	3.2145	3.1395	2.8520	2.7722	2.6380	2.2190
H = 0.60	50	3.5595	3.5763	3.5235	3.3832	3.3682	3.0520	2.9405	2.7958	2.5570	2.2185
	70	3.7035	3.4845	3.4283	3.2938	3.3853	3.2530	3.0565	2.7975	2.4803	2.2960
	90	3.6699	3.6455	3.5892	3.5692	3.2565	3.3466	3.1178	2.9388	2.6560	2.2563

Table 6.3 Combined effects for vocoder G.729 — Measured MOS

Loss %		0	0.5	1	2	3	5	7.5	10	15	20
	T (ms)										
H = 0.90	50	3.1505	3.1232	3.1080	3.0667	3.0907	2.7340	2.6578	2.6522	2.2283	2.1223
	70	3.2885	3.2832	3.1103	3.0857	3.0250	3.0368	2.7620	2.6790	2.4058	2.1765
	90	3.2862	3.2367	3.2138	3.1875	3.2083	2.9440	2.7390	2.6060	2.4073	2.1290
H = 0.75	50	3.2210	3.2123	2.9960	2.9453	2.9680	2.8998	2.6614	2.5980	2.3365	2.1607
	70	3.3250	3.3135	3.2345	3.2075	3.1158	2.8945	2.7865	2.6927	2.4013	2.1730
	90	3.4594	3.4342	3.2755	3.0938	3.1558	2.9290	2.7625	2.5842	2.5080	2.2185
H = 0.60	50	3.4005	3.2840	3.3693	3.1688	3.1008	2.9050	2.8870	2.7050	2.3963	2.2542
	70	3.3468	3.3603	3.3312	3.3180	3.1917	3.0505	2.7763	2.6703	2.5320	2.1710
	90	3.5097	3.4390	3.4455	3.3365	3.1910	3.1393	3.0435	2.8780	2.4190	2.2770

To visualize the combined effects, one group of data, $H = 0.75$ for vocoder G.729 is plotted in Figure 6.2.

Figure 6.2 Combined effects for vocoder G.729 with $H = 0.75$

It clearly demonstrates that the MOS score falls with increasing packet loss rate or increasing jitter buffer size. On the other hand, the MOS falls with increasing intensity of delay self-similarity.

6.2 Additivity Check

In this section, we examine the difference between predicted MOS by our formula and measured MOS after calibration.

The measured MOS scores in Tables 6.1, 6.2 and 6.3 are calibrated before comparison. In brief, they are first transformed into the rating factor R scale to get the summed impairments of packet loss and delay jitter, then linearly interpolated, and finally converted back to the MOS score. The MOS score after calibration is considered as stable, independent of the speech samples. It is slightly different from the measured MOS.

Table 6.4 shows the calibrated MOS for vocoder G.729 with $H = 0.75$ as an example. The MOS scores are also illustrated in Figure 6.3.

Table 6.4 Combined effects for vocoder G.729 with $H = 0.75$ — calibrated MOS

Loss %		0	0.5	1	2	3	5	7.5	10	15	20
Buffer (ms) $H = 0.75$	50	3.0281	3.0205	2.8333	2.7898	2.8093	2.7508	2.5476	2.4938	2.2725	2.1237
	70	3.1191	3.1090	3.0398	3.0163	2.9367	2.7463	2.6540	2.5742	2.3273	2.1341
	90	3.2380	3.2156	3.0757	2.9176	2.9713	2.7758	2.6336	2.4821	2.4176	2.1727

Figure 6.3 Combined effects for vocoder G.729 with $H = 0.75$ — calibrated MOS

Our proposed MOS prediction formula is given by:

$$MOS = \begin{cases} 1 & R < 0 \\ 1 + 0.035R + R(R - 60)(100 - R)7 \cdot 10^{-6} & 0 < R < 100 \\ 4.5 & R > 100 \end{cases} \quad (6.1)$$

where:

$$R = Ro - Is - Id - Ie - Ij + A \quad (6.2)$$

$$Ie = Ie_{opt} + C1 \cdot \ln(1 + C2 \cdot loss_rate) \quad (6.3)$$

$$Ij = C1 \cdot H^2 + C2 \cdot H + C3 + C4 \cdot e^{-T/K} \quad (6.4)$$

The factors in Equations 6.3 and 6.4 are given by Tables 4.2 and 5.1 respectively.

The prediction error is the difference between MOS prediction and real measured MOS after calibration. The errors are given in Tables 6.5, 6.6 and 6.7. The worst case happens with vocoder G.729, the error surface for G.729 with $H = 0.75$ is plotted in Figure 6.4.

Table 6.5 Prediction errors for vocoder G.723.1 B 5.3 kb/s

Loss %		0	0.5	1	2	3	5	7.5	10	15	20
	T (ms)										
H = 0.90	50	-0.037	-0.038	-0.030	-0.168	-0.179	-0.248	-0.152	-0.150	-0.234	-0.242
	70	-0.117	-0.035	-0.069	-0.092	-0.199	-0.157	-0.242	-0.260	-0.396	-0.406
	90	-0.079	-0.191	-0.182	-0.128	-0.208	-0.235	-0.245	-0.140	-0.247	-0.260
H = 0.75	50	-0.060	-0.031	-0.084	-0.118	-0.024	-0.263	-0.126	-0.186	-0.277	-0.209
	70	-0.002	-0.089	0.0386	0.1062	-0.033	-0.068	-0.117	-0.204	-0.395	-0.249
	90	-0.056	0.098	0.0988	0.0176	-0.074	-0.136	-0.211	-0.081	-0.302	-0.156
H = 0.60	50	-0.037	-0.012	-0.021	0.0133	-0.092	-0.162	-0.111	-0.208	-0.181	-0.213
	70	0.0012	0.006	0.0527	0.0536	0.0069	-0.054	-0.072	-0.087	-0.121	-0.185
	90	-0.032	0.0894	-0.011	0.0794	-0.165	-0.224	-0.238	-0.116	-0.091	-0.311

Table 6.6 Prediction errors for vocoder G.723.1 B 6.3 kb/s

Loss %		0	0.5	1	2	3	5	7.5	10	15	20
	T (ms)										
H = 0.90	50	-0.057	-0.045	-0.046	-0.092	-0.021	-0.212	-0.213	-0.342	-0.464	-0.479
	70	-0.103	-0.034	-0.076	-0.020	-0.152	-0.257	-0.242	-0.375	-0.408	-0.381
	90	-0.144	-0.161	-0.028	-0.059	-0.179	-0.190	-0.369	-0.294	-0.429	-0.307
H = 0.75	50	-0.030	0.0811	-0.126	-0.041	-0.003	-0.196	-0.207	-0.238	-0.263	-0.200
	70	0.0896	0.0263	-0.096	-0.046	-0.136	-0.034	-0.170	-0.222	-0.362	-0.316
	90	-0.110	-0.066	-0.061	-0.095	-0.057	-0.178	-0.141	-0.261	-0.464	-0.363
H = 0.60	50	-0.030	-0.100	-0.106	-0.085	-0.171	-0.082	-0.197	-0.261	-0.376	-0.345
	70	-0.109	0.0348	0.0316	0.0471	-0.133	-0.202	-0.242	-0.209	-0.260	-0.363
	90	-0.049	-0.081	-0.083	-0.167	0.0115	-0.253	-0.264	-0.299	-0.380	-0.302

Table 6.7 Prediction errors for vocoder G.729

Loss %		0	0.5	1	2	3	5	7.5	10	15	20
	T (ms)										
H = 0.90	50	0.062	0.0031	-0.062	-0.170	-0.320	-0.238	-0.403	-0.584	-0.508	-0.619
	70	0.0066	-0.071	0.0021	-0.120	-0.198	-0.433	-0.428	-0.547	-0.604	-0.617
	90	0.0418	0.0033	-0.054	-0.175	-0.323	-0.320	-0.376	-0.453	-0.577	-0.551
H = 0.75	50	0.0625	-0.012	0.0974	-0.003	-0.152	-0.318	-0.346	-0.481	-0.548	-0.606
	70	0.036	-0.035	-0.044	-0.163	-0.213	-0.249	-0.389	-0.500	-0.547	-0.567
	90	-0.050	-0.109	-0.046	-0.031	-0.214	-0.245	-0.336	-0.377	-0.609	-0.579
H = 0.60	50	0.0021	0.0235	-0.129	-0.096	-0.167	-0.224	-0.442	-0.479	-0.514	-0.610
	70	0.1133	0.0208	-0.030	-0.161	-0.180	-0.284	-0.284	-0.388	-0.572	-0.487
	90	0.0011	-0.016	-0.098	-0.143	-0.145	-0.327	-0.479	-0.532	-0.446	-0.550

Figure 6.4 Prediction error for vocoder G.729 with $H = 0.75$

It can be seen that the prediction error is negative in general. The error is increasing when the packet loss rate is increasing. When the packet loss rate is no more than 5 percent, the error doesn't exceed -0.26 MOS for vocoder G.723.1 and -0.35 MOS for vocoder G.729. However, if the packet loss rate increases to 20 percent, the error can reach -0.40 to -0.60 MOS.

6.3 Discussion

The predicted MOS is lower than the calibrated real measurement. Our model gives good prediction for packet loss up to 5 percent. In other cases, the model provides a lower bound in MOS prediction. Yet we can apply a certain correction factor γ to I_e or I_j or to both of them to match the calibrated real measurement. However, this will invalidate the prediction for separated effects.

The prediction model is pessimistic in predicting the combined effects of packet loss and delay jitter on speech quality, although the prediction for the separated effect is proven accurate. The assumed additive properties of I_e and I_j doesn't hold well. The combined impairments are less than the addition of individual impairment, so the real quality is not degraded so much.

The reasons why the additivity of I_e and I_j doesn't hold well are two-fold. Firstly, the effects of delay jitter become packet loss with timing distortion when a jitter buffer is used. Secondly, the effects of packet loss on speech quality don't satisfy additivity.

1. In simulations of effects of delay jitter, the packets are buffered at the receive side. In E-policy, gaps may be inserted into the talkspurt or silence, and too late packets are treated as lost. In the adaptive buffer algorithm, the silence period

between two talkspurts is artificially shortened or elongated, and too late packets are also considered as lost. In either case, the delay jitter would behave in a similar way to packet loss with timing distortion.

2. If the MOS score drops at a constant rate with the packet loss, the additivity of packet loss on the MOS scale will hold. However, the MOS score tends to drop at a decreasing rate when the packet loss rate increases.

We denote packet loss rate with p , then MOS score drop rate is given by $d MOS / d p$. In [43][44], the MOS is modeled as a logarithmic function of packet loss.

This tendency is clearly shown by $d MOS / d p \sim -\frac{1}{1+p}$. When p increases, the MOS drop rate decreases. In [7], ITU-T provides some provisional I_e values for the E-model. The tendency is not so direct, but can be observed after transforming into the MOS scale as shown in Figure 6.5. It is also supported by our simulation results in Figures 4.5, 4.12 and 4.13.

The tendency suggests that the MOS score drop for a certain packet loss rate is smaller than the sum of two MOS score drops whose corresponding loss rates, if added, equal that certain packet loss rate. This is illustrated in Figure 6.6.

Figure 6.5 Predicted MOS vs. packet loss from I_e values in [7]

Figure 6.6 Additivity of packet loss

Figure 6.6 is the curve of MOS vs. packet loss for vocoder G.729 by our prediction model. The MOS drops are represented by lines a , b and c for 6, 10 and 16 percent packet loss respectively. If the drop rate is constant, then $a + b$ should be equal to c as 16 percent is the sum of 6 percent and 10 percent. However, since the MOS drop rate is decreasing, it is obviously $a + b > c$.

When we evaluate the combined effects of packet loss and delay jitter, the delay jitter behaves in a similar way to packet loss with timing distortion. However, the effects of packet loss don't satisfy additivity, and addition always overestimates the MOS degradation. As a result, our prediction model doesn't hold well and prediction error is mainly negative.

CHAPTER 7 CONCLUSIONS

This thesis investigates the effects of packet loss and delay jitter on speech quality in the VoIP scenario, and extends the coverage of ITU-T Rec. G.107, the E-Model. The vocoder G.723.1 and vocoder G.729 are used in the simulation. The 5-point MOS score is used as an index of speech quality and is measured by the DSLA which implements the PAMS algorithm. The reference speech samples are carefully selected to meet the testing requirements. The simulation is repeated several times under each testing condition and the average is used.

For the effects of packet loss on speech quality, one frame per packet and random packet loss are simulated. In the E-model, the impairment from packet loss is represented by equipment impairment factor I_e . Currently only limited I_e values are available from ITU-T standard. We first compare the performance of several receiver-based error concealment methods: splicing, silence insertion and repetition. The repetition method, which replaces the lost packet with the previous one, is then selected for its effectiveness and simple implementation. It shows that the vocoder G.729 always outperforms vocoder G.723.1, especially when the packet loss rates are high. In addition, the MOS score drops at a decreasing rate as the packet loss rate increases.

We propose a formula to predict I_e as a function of packet loss. In the derivation of these I_e values, linear interpolation is applied to remove the bias of speech samples, making the measurement consistent with the ITU-T recommendation. A logarithmic function is used to model the relationship between I_e and packet loss rate. Two factors

$C1$ and $C2$ are used to adjust the shape of the curve. And different sets of $C1$ and $C2$ are found to fit the I_e values from different testing scenarios.

The proposed formula achieves good accuracy. For repetition method, when the packet loss rate is below 20 percent, the prediction error is within 0.10 MOS; when the packet loss rate increases up to 50 percent, the prediction error is within 0.10 MOS, 0.28 MOS for vocoders G.729 and G.723.1 respectively.

The formula is very useful for MOS prediction. For example, if we want the MOS higher than 3.0 in listening-only model, the tolerable packet loss rate is 10.8, 12.2 and 13.3 percent for vocoders G.723.1 B 5.3 kb/s, G.723.1 B 6.3 kb/s and G.729 respectively when a lost packet is concealed by the repetition method.

The impairment from delay jitter is not included in the E-model. To represent its effects on speech quality, we introduce a new parameter, jitter impairment factor I_j . The nature of packet transmission results in varying packet interarrival times. We use Pareto distribution to simulate network delay, as it is widely accepted today that the Internet traffic is self-similar. In the jitter buffer, we use the E-policy, a fixed buffer size algorithm. Gaps are introduced into the output audio stream because of an expansion of the waiting time, and some packets are treated as lost because of too late arrivals, thus the speech quality is degraded.

The speech quality is affected by two factors: H , degree of self-similarity of network delay, and T , the jitter buffer size. The simulation results show: 1) the MOS score falls as H increases or T decreases, 2) the vocoder G.729 is more sensitive to delay jitter when using 10 ms packets compared with 30 ms packets for vocoder G.723.1, and 3) the impairment from delay jitter is more annoying than that of packet loss. In the simulation,

even if the jitter buffer is set to 100 ms, the speech quality still can't be improved a lot. The results imply that the jitter buffer is not large enough to compensate for delay with heavy-tailed distribution. In real-time applications, we must accept a certain percent of packet loss to meet end-to-end delay budgets.

We propose a formula to predict the I_j as a function of H and T . These I_j values are then substituted into the extended E-model formula to calculate the MOS score. In the derivation of I_j values, linear interpolation is also applied. The proposed formula achieves good accuracy as well; the prediction error is within $-0.10...+0.20$ MOS.

The I_e and I_j are derived from separate simulations; we assume they are additive in the E-model when combining the effects of packet loss and delay jitter. The results show our formula works well for low packet loss rates up to 5 percent. When the packet loss increases to 20 percent, the prediction error can reach $-0.40...-0.60$ MOS.

The prediction error is negative in general, so the predicted MOS is lower than the real MOS. The prediction model gives a lower bound for the MOS score. To understand why the additivity works well only under low packet loss rate conditions, we show that the additivity for packet loss doesn't hold, as the MOS score tends to drop at a decreasing rate when the packet loss rate increases; additivity always overestimates the total degradation. On the other hand, the problem of delay jitter becomes the problem of packet loss with timing distortion when the jitter buffer is used. Thus, the addition of I_e and I_j lowers the MOS prediction. Although our simulation is based on a fixed buffer size algorithm, it can be extended to the case of an adaptive buffer algorithm, where the timing is distorted by adjusting the length of silence intervals and too late packets are effectively lost.

During the research, several ideas have arisen but not yet been implemented, due to time limitations. They are mainly concerned with including more testing conditions.

These ideas are summarized as follows:

1. Developing a joint model to effectively represent the combined effects of packet loss and delay jitter by examining the cross correlations between these two impairments as they are not separated in reality.
2. Measuring the speech quality by PESQ.
3. Investigating the effects of packet loss with other concealment methods, such as the built-in and interpolative error concealments. Also, investigating the speech quality by using different number of frames per packet or other languages than English.
4. Investigating the performance of other coders, such as G.711 (PCM, 64 kb/s), G.722 (broadband, Sub-band ADPCM, 48/56/64 kb/s) and G.728 (LD-CELP, 16 kb/s). A trend of VoIP is that the bandwidth consumption and algorithm complexity are no longer big issues as the increasing bandwidth capacities and processing powers.
5. Investigating the effects of delay jitter by using adaptive buffer algorithms.

REFERENCES

- [1] Uyless Black, *Voice Over IP*, Prentice Hall, Upper Saddle River, New Jersey, 2000.
- [2] Daniel Minoli and Emma Minoli, *Delivering Voice over IP Networks*, John Wiley & Sons, New York, 1998.
- [3] Zarak Systems, "Using PSQM to test packetized speech," http://www.zarak.com/ftp/white_PSQM.pdf, March 2000.
- [4] ITU-T Rec. P.800, *Methods for subjective determination of transmission quality*, 1996.
- [5] ETSI EG 201 377-1 V1.1.1, *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*, 1999.
- [6] ITU-T Rec. G.107, *The E-model, a computational model for use in transmission planning*, 2000.
- [7] ITU-T Rec. G.113, *Transmission impairments due to speech processing*, 2001.
- [8] J.G. Beerends and J.A. Stemerdink, "A perceptual audio quality measure based on psychoacoustic sound presentation," *J. Audio Eng. Soc.*, vol. 40, no. 12, pp. 963-978, December 1992.
- [9] 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.
- [10] S. Voran, "Objective estimation of perceived speech quality. I. Development of the measuring normalizing block technique II. Evaluation of the measuring normalizing block technique," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 4, pp. 371-382, 383-390, July 1999.

- [11] M.P. Hollier, M.O. Hawksford and D.R. Guard, "Characterisation of communications systems using a speech-like test stimulus," *J. Audio Eng. Soc.*, vol. 43, no. 12, pp. 1008-1021, December 1993.
- [12] M.P. Hollier, M.O. Hawksford and D.R. Guard, "Error activity and error entropy as a measure of psychoacoustic significance in the perceptual domain," *IEE Proceeding-Vision, Image and Signal Processing*, vol. 141, no. 3, pp. 203-208, June 1994.
- [13] A.W. Rix, J.G. Beerends, A.P. Hekstra and M.P. Hollier, "PESQ – the new ITU standard for end-to-end speech quality assessment," *109th Audio Engineering Society Convention*, pre-print no. 5260, September 2000.
- [14] ITU-T Rec. P.861, *Objective quality measurement of telephone-band (300-3400 Hz) speech codecs*, 1998.
- [15] PsyTechnics Group, *British Telecommunications*, "PAMS usage guidelines," February 2000.
- [16] ITU-T Rec. P.862, *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*, 2001.
- [17] ITU-T Rec. G.108, *Application of the E-model: A planning guide*, 1999.
- [18] N.O. Johannesson, "The ETSI computation model: a tool for transmission planning of telephone networks," *IEEE Communications Magazine*, vol. 35, no. 1, pp. 70-79, January 1997.
- [19] ITU-T Rec. G.109, *Definition of categories of speech transmission quality*, 1999.
- [20] *Digital Speech Level Analyzer, User Guide*, Revision 3.3, Malden Electronics Ltd., 2000.
- [21] John Anderson, "Methods for measuring perceptual speech quality," Agilent Technologies, <http://cp.literature.agilent.com/litweb/pdf/5988-2352EN.pdf>, 2001.
- [22] ITU-T Rec. P.830, *Subjective performance assessment of telephone-band and wideband digital codecs*, 1996.

- [23] A.W. Rix, A. Bourret and M.P. Hollier, "Models of human perception," *BT Tech. J.*, vol. 17, no. 1, pp. 24-34, January 1999.
- [24] A.W. Rix, R. Reynolds and M.P. Hollier, "Robust perceptual assessment of end-to-end audio quality," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 39-42, October 1999.
- [25] A.W. Rix and M.P. Hollier, "The perceptual analysis measurement system for robust end-to-end speech quality assessment," *Proceedings of ICASSP 2000*, vol. 3, pp. 1515-1518, June 2000.
- [26] Walter J. Goralski and Matthew C. Kolon, *IP Telephony*, McGraw-Hill, New York, 1999.
- [27] Andreas S. Spanias, "Speech coding: a tutorial review," *Proceedings of the IEEE*, vol. 82, no. 10, pp. 1541-1582, October 1994.
- [28] ITU-T Rec. G.723.1, *Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s*, 1996.
- [29] ITU-T Rec. G.729, *Coding of speech at 8 kbit/s using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP)*, 1996.
- [30] R. Salami *et al.*, "Design and description of CS-ACELP: a toll quality 8 kb/s speech coder," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 116-130, March 1998.
- [31] ITU-T Rec. G.729 – Annex A, *Reduced complexity 8 kbit/s CS-ACELP speech codec*, 1996.
- [32] H. Schulzrinne, "RTP: A transport protocol for real-time applications," <http://www.ietf.org/rfc/rfc1889.txt>, January 1996.
- [33] ITU-T Rec. G.114, *One-way transmission time*, 1996.
- [34] M. Hassan, A. Nayandoro and M. Atiquzzaman, "Internet telephony: services, technical challenges, and products," *IEEE Communications Magazine*, vol. 38, no. 4, pp. 96-103, April 2000.
- [35] Thomas J. Kostas *et al.*, "Real-time voice over packet-switched networks," *IEEE Network*, vol.12, no. 1, pp. 18-27, January/February 1998.

- [36] C.A. Fulton and San-Qi Li, "Delay jitter first-order and second-order statistical functions of general traffic on high-speed multimedia networks," *IEEE/ACM Transactions on Networking*, vol. 6, no. 2, pp. 150-163, April 1998.
- [37] Ramachandran Ramjee *et al.*, "Adaptive playout mechanisms for packetized audio applications in wide-area networks," *Proceedings of INFOCOM 1994*, vol. 2, pp. 680-688, 1994.
- [38] J. Rosenberg, Lili Qiu and H. Schulzrinne, "Integrating packet FEC into adaptive voice playout buffer algorithms on the Internet," *Proceedings of INFOCOM 2000*, vol. 3, pp. 1705 –1714, 2000.
- [39] W.A. Montgomery, "Techniques for packet voice synchronization," *IEEE Journal on Selected Areas in Communications*, vol. SAC-1, no. 6, pp. 1022-1028, December 1983.
- [40] Ralf Steinmetz, "Human perception of jitter and media synchronization," *IEEE Journal on Selected Areas in Communications*, vol. 14, no. 1, pp. 61-72, January 1996.
- [41] 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.
- [42] B.W. Wah, Xiao Su and Dong Lin, "A survey of error-concealment schemes for real-time audio and video transmissions over the Internet," *Proceedings of International Symposium on Multimedia Software Engineering*, pp. 17-24, December 2000.
- [43] L. Yamamoto and J.G. Beerends, "Impact of network performance parameters on the end-to-end perceived speech quality," *Expert ATM Traffic Symposium*, Greece, 1997.
- [44] Bart Duysburgh *et al.*, "On the influence of best-effort network conditions on the perceived speech quality of VoIP connections," *Proceedings of ICCCN 2001*, pp. 334-339, 2001.

- [45] ITU-T Rec. P.833, *Methodology for derivation of equipment impairment factors from subjective listening-only tests*, 2001.
- [46] U. Jekosch, S. Moller and A. Raake, "Extensions of the E-model," *ITU-T Study Group 12 Workshop on QoS and user-perceived transmission quality in evolving networks*, <http://www.itu.int/itudoc/itu-t/workshop/qos/s6p4.pdf>, Dakar, Senegal, October 2001.
- [47] W.E. Leland *et al.*, "On the self-similar nature of Ethernet traffic (extended version)," *IEEE/ACM Transactions on Networking*, vol. 2, no. 1, pp. 1-15, February 1994.
- [48] V. Paxson and S. Floyd, "Wide area traffic: the failure of Poisson modeling," *IEEE/ACM Transactions on Networking*, vol. 3, no. 3, pp. 226-244, June 1995.
- [49] M.S. Borella *et al.*, "Self-similarity of Internet packet delay," *ICC 1997*, vol. 1, pp. 513-517, 1997.
- [50] M.S. Borella and G.B. Brewster, "Measurement and analysis of long-range dependent behavior of Internet packet delay," *Proceedings of INFOCOM 1998*, vol. 2, pp. 497-504, 1998.
- [51] Qiong Li and D.L. Mills, "On the long-range dependence of packet round-trip delays in Internet," *ICC 1998*, vol. 2, pp. 1185 –1191, 1998.
- [52] J. Gordon, "Pareto process as a model of self-similar packet traffic," *GLOBECOM 1995*, vol. 3, pp. 2232 –2236, 1995.
- [53] F. Yegenoglu, F. Faris and O. Qadan, "A model for representing wide area Internet packet behavior," *Proceedings of IPCCC 2000*, pp. 167-173, 2000.
- [54] C. Weinstein and J.W. Forgie, "Experience with speech communication in packet networks," *IEEE Journal on Selected Areas in Communications*, vol. SAC-1, no. 6, pp. 963-980, December 1983.
- [55] W.E. Naylor and L. Kleinrock, "Stream traffic communication in packet switched networks: destination buffering considerations," *IEEE Transactions on Communications*, vol. COM-30, no.12, pp. 2527-2534, December 1982.

- [56] Sanjay Jha and Michael Fry, “Continuous media playback and jitter control,” *Proceedings of the Third IEEE International Conference on Multimedia Computing and Systems*, pp. 245–252, 1996.

APPENDIX A: DETAILED ALGORITHM OF THE E-MODEL

For all the input parameters used in E-model, their default values and permitted ranges are listed in Table A.1 from ITU-T Recommendation G.107 and G.108.

A.1 Default values and permitted ranges for the 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	TELRL	dB	65	5...65
Weighted Echo Path Loss	WEPL	dB	110	5...110
Mean one-way Delay of the Echo Path	T	msec	0	0...500
Round Trip Delay in a 4-wire Loop	Tr	msec	0	0...1000
Absolute Delay in echo-free Connections	Ta	msec	0	0...500
Number of Quantization Distortion Unit	qdu	-	1	1...14
Equipment Impairment Factor	Ie	-	0	0...40
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				

Only those parameters indicated by bold letters are usually subject to planning. The other parameters can be set to default values in calculations.

The transmission rating factor R is given by:

$$R = R_o - I_s - I_d - I_e + A \quad (\text{A.1})$$

R_o represents in principle the basic signal-to-noise ratio, including noise sources such as circuit noise and room noise. The factor I_s is a combination of all impairments which occur more or less simultaneously with the voice signal. The factor I_d represents the impairments caused by delay and the equipment impairment factor I_e represents impairments caused by low bit rate codecs. The advantage factor A allows for compensation of impairment factors when there are other advantages of access to the user. The term R_o and the I_s and I_d values are subdivided into further specific impairment values.

The basic signal-to-noise ratio R_o is defined by:

$$R_o = 15 - 1.5(SLR + N_o) \quad (A.2)$$

The term N_o [in dBm0p] is the power addition of different noise sources:

$$N_o = 10 \lg \left[10^{\frac{N_c}{10}} + 10^{\frac{N_{os}}{10}} + 10^{\frac{N_{or}}{10}} + 10^{\frac{N_{fo}}{10}} \right] \quad (A.3)$$

N_c [in dBm0p] is the sum of all circuit noise powers, all referred to the 0 dBr point.

N_{os} [in dBm0p] is the equivalent circuit noise at the 0 dBr point, caused by the room noise P_s at the send side:

$$N_{os} = P_s - SLR - D_s - 100 + 0.004(P_s - OLR - D_s - 14)^2 \quad (A.4)$$

where $OLR = SLR + RLR$. In the same way the room noise at the receive side is transferred into an equivalent circuit noise N_{or} [in dBm0p] at the 0 dBr point.

$$N_{or} = RLR - 121 + P_{re} + 0.008(P_{re} - 35)^2 \quad (A.5)$$

The term P_{re} [in dBm0p] is the "effective room noise" caused by the enhancement of P_r by the listener's sidetone path:

$$Pre = Pr + 10 \lg \left[1 + 10^{\frac{(10-LSTR)}{10}} \right] \quad (A.6)$$

Nfo [in dBm0p] represents the "noise floor" at the receive side,

$$Nfo = Nfor + RLR \quad (A.7)$$

The factor Is is the sum of all impairments which may occur more or less simultaneously with the voice transmission. It can be divided into three further specific impairment factors:

$$Is = Iolr + Ist + Iq \quad (A.8)$$

$Iolr$ represents the decrease in quality caused by too-low values of OLR and is given by:

$$Iolr = 20 \left[\left\{ 1 + \left(\frac{X}{8} \right)^8 \right\}^{\frac{1}{8}} - \frac{X}{8} \right] \quad (A.9)$$

where:

$$X = OLR + 0.2(64 + No - RLR) \quad (A.10)$$

Ist represents the decrease caused by non-optimum sidetone,

$$Ist = 10 \left[1 + \left(\frac{STMRO - 12}{5} \right)^6 \right]^{\frac{1}{6}} - 46 \left[1 + \left(\frac{STMRO}{23} \right)^{10} \right]^{\frac{1}{10}} + 36 \quad (A.11)$$

where:

$$STMRO = -10 \lg \left[10^{\frac{-STM}{10}} + e^{\frac{-T}{4}} 10^{\frac{-TEL}{10}} \right] \quad (A.12)$$

I_q represents the impairments caused by quantization distortion from a pure PCM process:

$$I_q = 15 \lg [1 + 10^Y + 10^Z] \quad (\text{A.13})$$

where:

$$Y = \frac{Ro-100}{15} + \frac{46}{8.4} - \frac{G}{9} \quad (\text{A.14})$$

$$Z = \frac{46}{30} - \frac{G}{40} \quad (\text{A.15})$$

and:

$$G = 1.07 + 0.258Q + 0.0602Q^2 \quad (\text{A.16})$$

$$Q = 37 - 15 \lg(qdu) \quad (\text{A.17})$$

Id , the impairment factor representing all impairments due to delay of voice signals, is further subdivided into the three factors $Idte$, $Idle$ and Idd :

$$Id = Idte + Idle + Idd \quad (\text{A.18})$$

the factor $Idte$ estimates the impairments due to talker echo and is given by:

$$Idte = \left[\frac{Roe - Re}{2} + \sqrt{\frac{(Roe - Re)^2}{4} + 100} - 1 \right] (1 - e^{-T}) \quad (\text{A.19})$$

where:

$$Roe = -1.5(No - RLR) \quad (\text{A.20})$$

$$Re = 80 + 2.5(TErv - 14) \quad (\text{A.21})$$

$$TERV = TELR - 40 \lg \frac{1 + \frac{T}{10}}{1 + \frac{T}{150}} + 6 e^{-0.3T^2} \quad (\text{A.22})$$

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.20) is replaced by $TERVs$, where:

$$TERVs = TERV + \frac{Ist}{2} \quad (\text{A.23})$$

For $9 \text{ dB} \leq STMR \leq 15 \text{ dB}$:

the above Equations (A.18) to (A.21) apply.

For $STMR > 15 \text{ dB}$:

$Idte$ is replaced by $Idtes$, where:

$$Idtes = \sqrt{Idte^2 + Ist^2} \quad (\text{A.24})$$

The factor $Idle$ represents impairments due to listener echo:

$$Idle = \frac{Ro - Rle}{2} + \sqrt{\frac{(Ro - Rle)^2}{4} + 169} \quad (\text{A.25})$$

where:

$$Rle = 10.5 (WEPL + 7) (Tr + 1)^{-0.25} \quad (\text{A.26})$$

The factor Idd represents the impairment caused by too-long absolute delay Ta , which occurs even with perfect echo canceling.

For $Ta < 100\text{ms}$:

$$Idd = 0$$

For $Ta > 100\text{ms}$:

$$Idd = 25 \left\{ \left(1 + X^6 \right)^{1/6} - 3 \left(1 + \left[\frac{X}{3} \right]^6 \right)^{1/6} + 2 \right\} \quad (\text{A.27})$$

with:

$$X = \frac{\lg \left[\frac{Ta}{100} \right]}{\lg 2} \quad (\text{A.28})$$

APPENDIX B: *I_e* VALUES FROM ITU-T

The ITU-T Recommendation G.113 appendix I provides up-to-date information on available values of the Equipment Impairment Factor, *I_e*. It is intended to be updated regularly.

B.1 Provisional planning values for the equipment impairment factor *I_e* under conditions of random packet loss, vocoders G.729 A + VAD and G.723.1-A + VAD

% Packet Loss	G.729A + VAD	G.723.1.A + VAD 6.3 kbit/s
0	11	15
0.5	13	17
1	15	19
1.5	17	22
2	19	24
3	23	27
4	26	32
5	Note 2	Note 2
8	36	41
16	49	55
NOTE 1 – Number of frames per packet: • G.729-A + VAD: 2; • G.723.1-A + VAD: 1. NOTE 2 – No values have been available for these conditions		