2004

Speech quality prediction for voice over Internet protocol networks

Sun, Lingfen

SPEECH QUALITY PREDICTION FOR VOICE OVER INTERNET PROTOCOL NETWORKS

L. Sun

Ph.D. 2004

SPEECH QUALITY PREDICTION FOR VOICE OVER INTERNET PROTOCOL NETWORKS

by

LINGFEN SUN

A thesis submitted to the University of Plymouth
in partial fulfillment for the degree of

DOCTOR OF PHILOSOPHY

School of Computing, Communications and Electronics
Faculty of Technology

In collaboration with Acterna

January 2004

Speech Quality Prediction for Voice over Internet Protocol Networks

Lingfen Sun

Abstract

IP networks are on a steep slope of innovation that will make them the long-term carrier of all types of traffic, including voice. However, such networks are not designed to support real-time voice communication because their variable characteristics (e.g. due to delay, delay variation and packet loss) lead to a deterioration in voice quality. A major challenge in such networks is how to measure or predict voice quality accurately and efficiently for QoS monitoring and/or control purposes to ensure that technical and commercial requirements are met.

Voice quality can be measured using either subjective or objective methods. Subjective measurement (e.g. MOS) is the benchmark for objective methods, but it is slow, time consuming and expensive. Objective measurement can be intrusive or non-intrusive. Intrusive methods (e.g. ITU PESQ) are more accurate, but normally are unsuitable for monitoring live traffic because of the need for a reference data and to utilise the network. This makes non-intrusive methods(e.g. ITU E-model) more attractive for monitoring voice quality from IP network impairments. However, current non-intrusive methods rely on subjective tests to derive model parameters and as a result are limited and do not meet new and emerging applications.

The main goal of the project is to develop novel and efficient models for non-intrusive speech quality prediction to overcome the disadvantages of current subjective-based methods and to demonstrate their usefulness in new and emerging VoIP applications. The main contributions of the thesis are fourfold:

(1) a detailed understanding of the relationships between voice quality, IP network impairments (e.g. packet loss, jitter and delay) and relevant parameters associated with speech (e.g. codec type, gender and language) is provided. An understanding of the perceptual effects of

these key parameters on voice quality is important as it provides a basis for the development of non-intrusive voice quality prediction models. A fundamental investigation of the impact of the parameters on perceived voice quality was carried out using the latest ITU algorithm for perceptual evaluation of speech quality, PESQ, and by exploiting the ITU E-model to obtain an objective measure of voice quality.

(2) a new methodology to predict voice quality non-intrusively was developed. The method exploits the intrusive algorithm, PESQ, and a combined PESQ/E-model structure to provide a perceptually accurate prediction of both listening and conversational voice quality non-intrusively. This avoids time-consuming subjective tests and so removes one of the major obstacles in the development of models for voice quality prediction. The method is generic and as such has wide applicability in multimedia applications. Efficient regression-based models and robust artificial neural network-based learning models were developed for predicting voice quality non-intrusively for VoIP applications.

(3) three applications of the new models were investigated: voice quality monitoring/prediction for real Internet VoIP traces, perceived quality driven playout buffer optimization and perceived quality driven QoS control. The neural network and regression models were both used to predict voice quality for real Internet VoIP traces based on international links. A new adaptive playout buffer and a perceptual optimization playout buffer algorithms are presented. A QoS control scheme that combines the strengths of rate-adaptive and priority marking control schemes to provide a superior QoS control in terms of measured perceived voice quality is also provided.

(4) a new methodology for Internet-based subjective speech quality measurement which allows rapid assessment of voice quality for VoIP applications is proposed and assessed using both objective and traditional MOS test methods.

Author's Declaration

At no time during the registration for the degree of Doctor of Philosophy has the author been registered for any other University award.

This study was funded in part by Acterna.

Publications:

1. L. Sun and E. Ifeachor, New Models for Perceived Voice Quality Prediction and their Applications in Playout Buffer Optimization for VoIP Networks, in Proceedings of IEEE International Conference on Communications (IEEE ICC 2004), Paris, France, June 2004.

2. Z. Qiao, L. Sun, N. Heilemann, and E. Ifeachor, A New Method for VoIP Quality of Service Control Based on Combined Adaptive Sender Rate and Priority Marking, in Proceedings of IEEE International Conference on Communications (IEEE ICC 2004), Paris, France, June 2004.

3. E. Ifeachor and L. Sun, Learning Models for Non-intrusive Prediction of Voice Quality for IP Networks, Submitted to IEEE Transactions on Neural Networks.

4. L. Sun and E. Ifeachor, New Methods for Voice Quality Evaluation for IP Networks, in Proceedings of 18^{th} International Teletraffic Congress (ITC-18), Berlin, Germany, 31 August - 5 September 2003, pp. 1201 – 1210.

5. Z. Li, L. Sun, Z Qiao and E Ifeachor, Perceived Speech Quality Driven Retransmission Mechanism for Wireless VoIP, in Proceedings of IEE Fourth International Conference on 3G Mobile Communication Technologies, London, UK, June 2003, pp. 395 – 399.

6. L. Sun and E. Ifeachor, Prediction of Perceived Conversational Speech Quality and Effects of Playout Buffer Algorithms, in Proceedings of IEEE International Conference on Communications (IEEE ICC 2003), Anchorage, USA, May 2003, pp. 1 – 6.

7. L. Sun and E. Ifeachor, Perceived Speech Quality Prediction for Voice over IP-based Networks, in Proceedings of IEEE International Conference on Communications (IEEE ICC 2002) New York, USA, April 2002, pp.2573 – 2577.

8. L. Sun and E. Ifeachor, Subjective and Objective Speech Quality Evaluation under Bursty Losses, in Proceedings of On-line Workshop Measurement of Speech and Audio Quality in Networks (MESAQIN 2002), Prague, Czech Republic, Jan. 2002, pp.25 – 29.

9. L. Sun, G. Wade, B. Lines and E. Ifeachor, Impact of Packet Loss Location on Perceived Speech Quality, in Proceedings of 2nd IP-Telephony Workshop (IPTEL '01), Columbia University, New York, April 2001, pp.114 – 122.

10. L. Sun, G. Wade, B. Lines, D. Le Foll and E. Ifeachor, VoIP Speech Quality Simulation and Evaluation, in Proceedings of the Second International Network Conference (INC 2000), Plymouth, UK, July 2000, pp.157 – 164.

11. L. Sun, G. Wade, B. Lines, E. Ifeachor and D. Le Foll, End-to-end Speech Quality Analysis for VoIP, in Proceedings of IEE 16th UK Teletraffic Symposium on Management of Quality of Service, Harlow, UK, May 2000, pp. 23/1 – 23/6.

Signed Lingfen Sun

Date 22/3/2004

Acknowledgments

This thesis would not have been possible without the support and guidance of many people. First, I would like to thank my first supervisor and director of studies Professor Emmanuel Ifeachor for his professional guidance, encouragement and patience throughout this project, for the benefit of his wide knowledge and vision for this project and for the tremendous amount of time and efforts he has spent to ensure the high quality of my papers and this thesis. I would also like to thank my other supervisors Dr. Graham Wade and Dr. Benn Lines for their support and guidance during the project.

I would like to thank the members of Signal Processing and Multimedia Communication (SPMC) Research Group: Mr Brahim Hamadicharef, Mr Zizhi Qiao (William), Dr Nicolas Outram and Dr Julian Tilbury, the members of Network Research Group (NRG): Dr Steven Furnell, Mr Bogdan Ghita, Mr Paul Dowland, Mr Philip M Rodwell, Mr George Magklaras, Mr Harjit Singh and Miss Licha Mued for all the help, support and constructive discussions for this project. It is my great pleasure to have been working in the NRG and SPMC groups during my four years Ph.D studying in Plymouth. I would like to thank some MSc and BEng students, Mr. Chunshui Liu, Mr. Zhuoqun Li, Mr. Pin Hu and Mr. Nicolai Heilemann. I have benefited from the collaboration and discussion with them. I would also like to thank Mr. Dominique Le Foll from Acterna for his help and support during the project.

Lastly, this thesis is dedicated to my parents, my husband, my daughter, my brother and sisters for their endless love, support and encouragement.

Table of Contents

List of Tables

List of Figures

List of Abbreviations and Glossary

3G	Third Generation (wireless)
3GPP	3G Partnership Project (UMTS)
3GPP2	3G Partnership Project 2 (UMTS)
ACELP	Algebraic Codebook Excited Linear Prediction
ACR	Absolute Category Rating
AMR	Adaptive Multi-rate (speech codec)
ANN	Artificial Neural Network
CBR	Constant Bit Rate
CCI	Call Clarity Index
CDF	Cumulative Distribution Function
CELP	Code Excited Linear Prediction
CI	Confidence Interval
CLP	Conditional Loss Probability
DCR	Degradation Category Rating
DiffServ	Differentiated Services
DMOS	Degradation Mean Opinion Score
EMBSD	Enhanced Modified Bark Spectral Distortion
ESTI	European Telecommunications Standards Institute
FEC	Forward Error Correction
GPS	Global Positioning System
GSM	Global System for Mobile Communications
IETF	Internet Engineering Task Force

ICMP	Internet Control Message Protocol
iLBC	Internet Low Bit Rate Codec
INMD	In-service Non-intrusive Measurement Device
InterServ	Integrated Services
IP	Internet Protocol
ISDN	Integrated Services Digital Network
ITU	International Telecommunication Union
LPC	Linear Predictive Coding
MLE	Maximum Likelihood Estimator
MNB	Measuring Normalizing Blocks
MOS	Mean Opinion Score
MSE	Mean Square Error
NN	Neural Network
NS-2	Network Simulator version 2
PEAQ	Perceptual Evaluation of Audio Quality
PESQ	Perceptual Evaluation of Speech Quality
PESQ-LQ	Perceptual Evaluation of Speech Quality – Listening Quality
PSQM	Perceptual Speech Quality Measure
PSTN	Public Switched Telephone Network
QoE	Quality of Experience
QoS	Quality of Service
RED	Random Early Detection
RFC	Request for Comment
RMSE	Root Mean Square Error
RTCP	Real Time Transport Control Protocol (IETF)
RTP	Real Time Transport Protocol (IETF)

RTT	Round Trip Time
SCN	Switched Communication Network
SID	Silence Insertion Description
SIP	Session Initiation Protocol (IETF)
SNNS	Stuttgart Neural Network Simulator
SNR	Signal to Noise Ratio
TCP	Transmission Control Protocol
TIMIT	a speech data set from TI (Texas Instruments) and MIT
UDP	User Datagram Protocol
ULP	Unconditional Loss Probability
UMTS	Universal Mobile Telecommunications System
VAD	Voice Activity Detection
VoIP	Voice over Internet Protocol

Chapter 1

Introduction

This Chapter is organized as follows. The motivations behind the project are presented in Section 1.1. The research questions are given in Section 1.2. Project aims and objectives are outlined in Section 1.3. The major contributions are summarized in Section 1.4. In Section 1.5, a brief overview and the organisation of the thesis are given.

1.1 Motivations

IP networks are on a steep slope of innovation that will make them the long-term carrier of all types of traffic, including voice. However, such networks are not designed to support real-time voice communication because their variable characteristics (e.g. due to delay, delay variation and packet loss) lead to a deterioration in voice quality [1, 2]. A major challenge in such networks is how to measure or predict voice quality accurately and efficiently for QoS monitoring and/or control purposes to ensure that the technical and commercial requirements (e.g. service level agreements) are met.

In real-time voice communications, perceived speech quality[†], expressed as a Mean Opinion Score (MOS), is the key metric for quality of service as it provides a direct link to quality as perceived by the end user. MOS values may be obtained by subjective tests [3] or by objective perceptual methods, such as the latest ITU algorithm, the Perceptual Evaluation of Speech

[†]The terms of speech quality and voice quality are used interchangeably in this thesis.

Quality (PESQ) [4,5,6], the inherent problem in subjective MOS measurement is that it is time consuming, expensive, lack of repeatability, and cannot be used for long-term or large scale voice quality monitoring in an operational network infrastructure. This has made objective methods very attractive for meeting the demand for voice quality measurement in communications networks.

Unlike intrusive methods such as PESQ, non-intrusive, objective techniques are appropriate for monitoring live traffic because they do not need the injection of a reference signal and do not utilise the network. The ITU E-model [7], originally designed for conventional network planning [8], is the most widely used non-intrusive method. It may be used to predict end-to-end voice quality directly from IP network and/or terminal parameters [9, 10]. However, it is based on a complex set of fixed, empirical formulae and is applicable to a restricted number of codecs and network conditions (because subjective tests are required to derive model parameters) and this hinders its use in new and emerging applications.

Artificial neural networks-based models have recently been used to predict both speech and video quality from IP network parameters [11,12,13], but, as in the E-model, the previous neural network models rely on subjective tests to create the training sets. Unfortunately, subjective tests are costly and time-consuming and as a result the training sets are limited and cannot cover all the possible scenarios in dynamic and evolving networks, such as the Internet.

There is a need to develop a new methodology which provides efficient statistical or neural network models to measure/predict voice quality, non-intrusively, for both managed and best effort networks and for emerging applications. Such models can be used:

- for objective, non-intrusive, prediction/monitoring of end-to-end voice quality on live network, and to study error profile and IP network readiness for VoIP services;

- to optimise the quality of voice services in accordance with changing network conditions and to control the QoS and manage the utilisation of available resources.

1.2 Research Questions

This dissertation seeks to address the following questions/issues:

- What are the relationships between perceived speech quality, IP network impairments (e.g. packet loss, jitter and delay) and relevant parameters associated with speech (e.g. codec type, voiced/unvoiced, gender and languages)?

  This leads to a fundamental research to investigate the relationships between perceived voice quality, IP network impairments (e.g. packet loss, jitter and delay) and relevant parameters associated with speech (e.g. codec type, voiced/unvoiced, gender and language). A fundamental investigation of the impact of these parameters on perceived voice quality is undertaken using the latest ITU algorithm for perceptual evaluation of speech quality, PESQ, and a combined PESQ/E-model structure to obtain an objective measure of voice quality. Four modern codecs which are commonly used in VoIP and in emerging applications are used in the study (G.729, G.723.1, AMR and iLBC). This work will be discussed in Chapter 4.

- How should the perceived speech quality be measured/predicted non-intrusively and efficiently for VoIP networks?

  This leads to a new non-intrusive perceived speech quality prediction methodology. The novelty of the method is the exploitation of the latest intrusive ITU algorithm, PESQ, and the use of a combined PESQ/E-model, to provide a perceptually accurate prediction of voice quality (both listening and conversational), non-intrusively. This avoids time-consuming subjective tests and so removes one of the major obstacles in the development of models for voice quality prediction non-intrusively.

  Based on the new methodology, efficient non-linear regression models are developed to predict conversational voice quality for a variety of codecs, which is presented in Chapter 5. Further artificial neural network models are exploited for predicting both listening

and conversational voice quality based on PESQ and a combined PESQ/E-model structure. This is presented in Chapter 6.

- How should the perceived speech quality metric be exploited to monitor, optimize and control end-to-end speech quality?

  This leads to three main applications which use the newly developed voice quality prediction models (1) to monitor/predict voice quality for the current Internet, (2) to achieve perceived speech quality driven playout buffer optimization, and (3) to achieve perceived speech quality driven QoS control. Previous work on playout buffer optimization and QoS control is mainly based on individual network parameters (e.g. packet loss or delay). This approach is inappropriate as it does not provide a direct link to perceived speech quality. From QoS perspective, the optimization of playout buffer algorithms or the control of QoS should be determined by the likely perceived speech quality. The prediction of voice quality for current Internet using neural network models and regression models is described in Chapter 7. The perceived quality driven playout buffer optimization is presented in Chapter 8. The perceived quality driven QoS control with a combined send-bit-rate adaptive and priority marking control schemes is discussed in Chapter 9

- How should subjective MOS tests be conducted efficiently for VoIP applications?

  This leads to a new Internet-based methodology for subjective MOS tests. Unlike traditional MOS tests which have to be carried out in a sound-proof room following stringent test requirements, a methodology to conduct MOS tests under normal working environments through Internet is proposed. Controlled and uncontrolled Internet-based MOS tests are carried out and the results are compared with various objective test methods. The work is presented in Chapter 10.

1.3 Project Aim and Objectives

The main aims of the project are (1) to undertake a fundamental investigation to quantify the impact of network impairments and speech related parameters on perceived speech quality in IP networks, (2) to apply the results to develop novel and efficient models for non-intrusive speech quality measurement and prediction for VoIP applications, and (3) to apply the developed models to new and emerging applications in voice quality prediction/monitoring, voice quality optimization (e.g. jitter buffer optimization) and QoS control.

Specific objectives of the research are to:

- Undertake a fundamental investigation to quantify the impact of network impairments (e.g. packet loss, delay and jitter) and speech related parameters (e.g. codec, voiced/unvoiced, gender or language) on perceived speech quality in IP networks.
- Develop novel and efficient non-intrusive speech quality prediction models and methodology to predict perceived speech quality directly from IP network parameters and/or speech related parameters.
- Investigate the applications of the above models in areas such as speech quality prediction/monitoring, QoS performance optimization (e.g. jitter buffer optimization) and QoS control (e.g. rate adaptive QoS control or priority marking QoS control) in VoIP networks.

1.4 Contributions of Thesis

The contributions of the dissertation are the following:

1. A detailed understanding of the relationships between voice quality, IP network impairments (e.g. packet loss, jitter and delay) and relevant parameters associated with speech (e.g. codec type, gender and language). An understanding of the perceptual effects of the

key parameters on voice quality is important as it provides a basis for the development of efficient regression models and robust artificial neural network learning models.

A fundamental investigation to study the impact of key network parameters (i.e. loss rate, loss pattern and latency) and non-network parameters (e.g. codec type, gender and language) on perceived speech quality is undertaken using a combined PESQ/E-model scheme to obtain an objective measure of voice quality. Four modern voice codecs which are commonly used in VoIP and other emerging applications are selected (i.e. ITU G.729 [14] and G.723.1 [15], AMR [16] and iLBC [17]) for the study (this number can be readily expanded).

(The associated publications are [18, 19])

2. A new methodology to predict voice quality non-intrusively is presented. The novelty of the method is that it exploits the latest intrusive ITU algorithm, PESQ, and a combined PESQ/E-model structure to provide a perceptually accurate prediction of both listening and conversational voice quality non-intrusively, which avoids time-consuming subjective tests. The method is generic and as such has wide applicability in multimedia applications (e.g. objective, non-intrusive, prediction of end-to-end voice/audio/image/video quality; optimization of quality of multimedia services) and to other packet networks (e.g. ATM and managed IP networks). Efficient regression-based and neural network-based models are developed for predicting both listening and conversational voice quality for VoIP applications. The detailed contributions are:

   - New non-linear regression models for predicting conversational voice quality based on a novel combination structure of PESQ and E-model. Non-linear regression models for a variety of codecs (i.e. G.729, G.723.1, AMR and iLBC) for different network conditions are derived and their applications in voice quality prediction/-monitoring, playout buffer optimization and QoS control are given. (The associated publication is [20].)

- New learning models, based on supervised neural networks, with voice quality prediction accuracy close to the ITU PESQ/E-model (correlation coefficients of 0.94 for VoIP simulation system and real Internet VoIP trace data). These include models for non-intrusive, objective prediction of both listening and conversational voice quality. A key novelty is that the models have learning capability and this makes it possible for the models to adapt to changes in the network. Four modern codecs (i.e. G.729, G.723.1, AMR and iLBC) are chosen for the study and the impact of talkers (e.g. male or female) is considered. Their applications in voice quality prediction/monitoring are given. (The associated publication is [19] and the paper submitted for publication is [21].)

3. Three applications for the new perceived voice quality prediction models are investigated.

   - The newly developed efficient regression and neural network models are applied for voice quality prediction in the current Internet. Preliminary results show that in this application the neural network models have accuracy close to the ITU PESQ/E-model(correlation coefficient of 0.94) and the regression models have even higher accuracy with the correlation coefficient of 0.98. The results are based on Internet trace data measurements between UK and USA, UK and China, and UK and Germany. (The associated publication is [22] and the paper submitted for publication is [21].)

   - A new adaptive playout buffer algorithm and a new perceptual optimized playout buffer algorithm are presented. The use of minimum overall impairment as a criterion for buffer optimization is proposed. This criterion is more efficient than the traditional maximum Mean Opinion Score (MOS). It is shown that the delay behaviour of Voice over IP traffic is better characterized by a Weibull distribution than a Pareto or an Exponential distribution. The perceptual performance for existing jitter buffer algorithms and new proposed buffer algorithms are compared. The

results show that the proposed perceptual optimized buffer algorithm can achieve the optimum perceived speech quality compared with other algorithms for all the traces considered. The adaptive buffer algorithm can achieve sub-optimum perceived speech quality with low complexity. The results are based on Internet trace data collected internationally. (The associated publications are [22] and [20].)

- A new QoS control scheme that combines the strengths of rate-adaptive and priority marking control schemes is presented to provide a superior QoS control performance, in terms of perceived speech quality. An objective measure of perceived speech quality (i.e. objective MOS score) is used for adaptive control of sender behaviour (e.g. sender bit rate), as this provides a direct link to user perceived speech quality, unlike individual network impairment parameters (e.g. packet loss and/or delay). Our results show that the new combined QoS control method achieved the best performance under different network congestion conditions compared to separate adaptive sender rate and packet priority marking methods. (The associated publication is [23].)

4. A new methodology for Internet-based subjective speech quality measurement is presented. Unlike traditional MOS tests which have stringent test requirements (e.g. the use of a sound proof room), the Internet-based subjective tests aim to conduct MOS tests under normal working environments through Internet for Voice over IP applications. This method allows rapid assessment of voice quality. It is more efficient and close to reality than the traditional methods. Both uncontrolled and controlled Internet-based MOS tests are carried out and the results are compared with various objective test methods as well as by traditional MOS test methods. The results are quite promising. (The associated publications are [24, 25]).

1.5 Outline of Thesis

The outline of the thesis is shown in Figure 1.1 and described as follows:

Chapter 2 gives a brief background information about VoIP networks. The VoIP network connection, protocol and system structure are presented in Section 2.2. Perceived speech quality and factors affect speech quality are presented in Section 2.3. Voice coding technology and main codecs used in the thesis are introduced in Section 2.4. Network performance characteristics (e.g. packet loss and delay/delay variation) are presented in Section 2.5.

Chapter 3 summarizes state-of-the-art speech quality measurement/assessment methods including subjective and intrusive/non-intrusive objective methods. Section 3.2 introduces subjective tests. Section 3.3 presents intrusive speech quality measures such as PSQM, MNB, PESQ and PESQ-LQ. Section 3.4 discusses non-intrusive speech quality measures including parameter-based models(e.g. E-model and artificial neural network model) and signal-based models (e.g. vocal tract model), with emphasis on the former.

Chapter 4 investigates the impact of network impairments (e.g. packet loss rate, packet loss burstiness, packet loss location and end-to-end delay) and relevant parameters associated with speech (e.g. codec type, voiced/unvoiced, gender and language) on perceived voice quality. Section 4.2 introduces a VoIP simulation system set up for quality evaluation. In Section 4.3, the impact of packet loss location on perceived speech quality is evaluated. Section 4.4 shows how packet loss burstiness and packet size affect speech quality. Section 4.5 presents how different talkers (male or female) and seven different languages (e.g. English, Dutch, Chinese and Arabian) affect perceived voice quality. Section 4.6 discusses how end-to-end delay and codec types affect perceived voice quality.

Chapter 5 introduces a new methodology to predict voice quality, non-intrusively, based on an exploitation of the latest intrusive ITU algorithm, PESQ, and the use of a combined PESQ/E-model structure. Based on the new methodology, efficient regression-based and artificial neural network-based models are developed. The detailed non-linear regression models for

voice quality prediction are presented in this Chapter and the neural network-based models are presented in Chapter 6. The novel methodology for non-intrusive prediction of voice quality is presented in Section 5.2. The system structure and the procedures of using regression models for predicting voice quality are described in Sections 5.3 and 5.4, respectively. In Section 5.5, the non-linear regression models for different codecs (i.e. G.729, G.723.1, AMR and iLBC) under different network conditions are given.

Chapter 6 examines neural network-based models for predicting both listening and conversational voice quality. In Section 6.2, the neural network models for predicting listening-only voice quality is presented, which covers the simulation system structure, neural network database collection, neural network models and result analysis. Then the neural network models for predicting conversational voice quality are described in detail in Section 6.3.

Chapter 7 presents the perceived speech quality prediction using neural network and regression models for VoIP in the current Internet. In Section 7.2, the setup and approach for the Internet trace data measurement is introduced. In Section 7.3, the IP network performance analysis is carried out for the selected traces. The focus is on the characteristics of the delay/jitter and its distribution, and packet loss and its distribution. Speech quality prediction for selected traces using neural network models and regression models are presented in Sections 7.4 and 7.5, respectively. Performance analysis and comparison are given in Section 7.6.

Chapter 8 presents the method for applying newly developed speech quality prediction models for playout buffer optimization. The existing playout algorithms and their performance are analyzed in Section 8.2. In Section 8.3, a new adaptive playout buffer algorithm based on traditional jitter buffer algorithms is presented. In Section 8.4, a perceptual optimum playout buffer algorithm is given based on the derived speech quality prediction models, a minimum impairment criterion and Weibull delay distribution modeling. The performance analysis and comparison between existing buffer algorithms and newly proposed algorithms are presented in Section 8.5.

Chapter 9 introduces an application of perceived speech quality prediction for QoS control

by combining rate-adaptive control and priority marking control schemes. The adaptive AMR codec and its speech quality under packet loss is discussed in Section 9.2. The three QoS control schemes – the rate-adaptive, the priority marking and the new combined QoS control schemes are presented in Section 9.3. The simulation system and experiments are described in Section 9.4. The results and analysis are given in Section 9.5.

Chapter 10 describes a new Internet-based subjective speech quality measurement methodology. The uncontrolled and controlled Internet based MOS tests and quality evaluation/comparison are presented in Sections 10.2 and 10.3, respectively.

Chapter 11 reviews the work done, suggesting future work and presents the conclusions of the thesis.

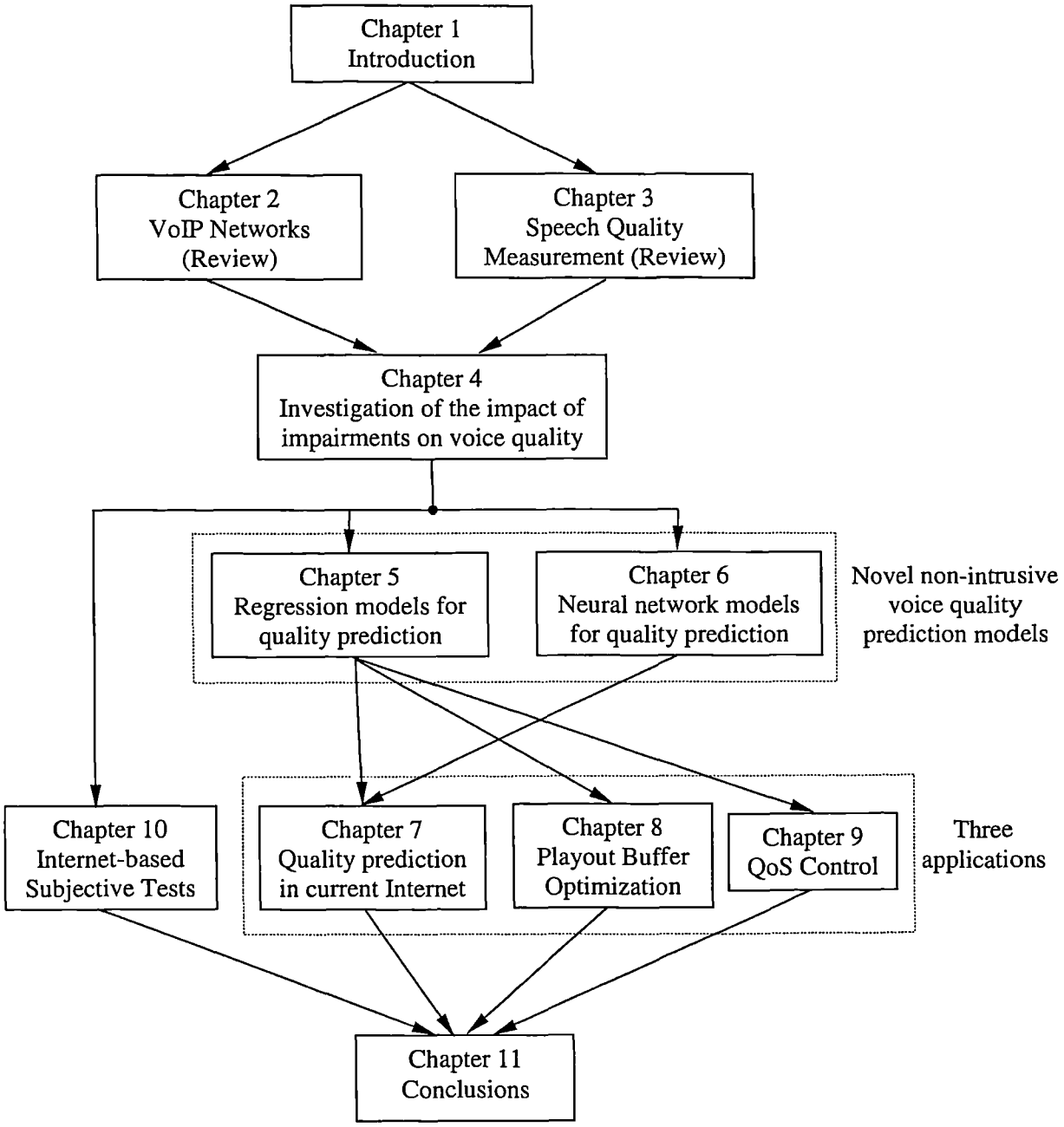


Figure 1.1: Thesis outline

Chapter 2

VoIP Networks

2.1 Introduction

The convergence of communications and computer networks has led to a rapid growth in real-time applications, such as Internet Telephony or Voice over IP (VoIP). However, IP networks are not designed to support real-time applications and factors such as network delay, jitter and packet loss lead to a deterioration in the perceived voice quality.

In this chapter, a brief background information about VoIP networks which is relevant to the thesis is summarized. The VoIP network, protocol and system structure are described in Section 2.2. Perceived speech quality or perceived quality of service (QoS) are presented in Section 2.3. Voice coding technology and main codecs used in the thesis (i.e. G.729, G.723.1, AMR and iLBC) are introduced in Section 2.4. Network performance characteristics (e.g. packet loss and delay/delay variation) are presented in Section 2.5. Section 2.6 summarises the chapter.

2.2 VoIP Networks

2.2.1 VoIP Network Connections

Common VoIP network connections normally include the connection from phone to phone, phone to PC (IP Terminal or H.323/SIP Terminal [26]) or PC to PC, as shown in Figure 2.1. The Switched Communication Network (SCN) can be a wired or wireless network, such as PSTN, ISDN or GSM.

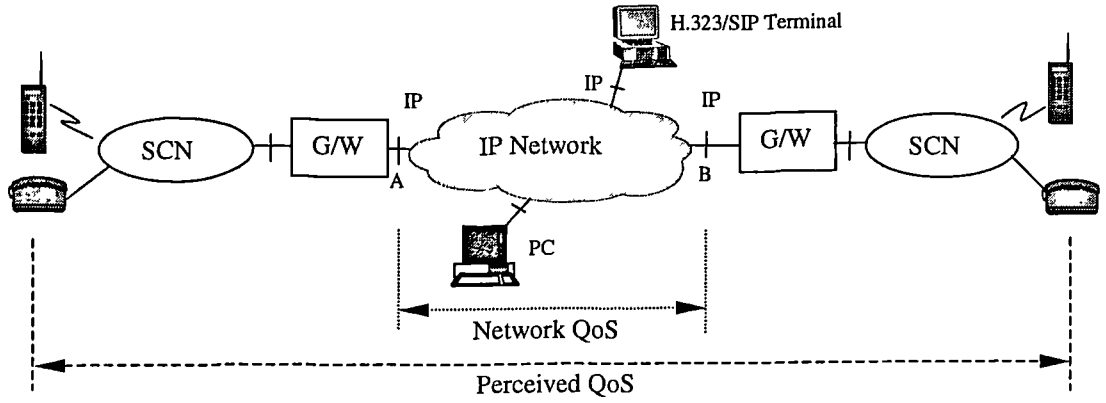


Figure 2.1: VoIP network connections

Perceived QoS or User-perceived QoS is defined as end-to-end or mouth to ear, as the quality perceived by the end user. It depends on the quality of the gateway (G/W) or H.323/SIP terminal and IP network performance. The latter is normally referred to as Network QoS, as illustrated in Figure 2.1.

As IP network is based on the "best effort" principle which means that the network makes no guarantees about packet loss rates, delays and jitter, the perceived voice quality will suffer from these impairments (e.g. loss, jitter and delay). There are currently two approaches to enhance QoS for VoIP applications. The first approach relies on application-level QoS mechanisms to improve perceived QoS without making changes to the network infrastructure. For example, different compensation strategies for packet loss (e.g. Forward Error Correction (FEC)) [27,28] and jitter [29, 30] have been proposed to improve speech quality even under poor network conditions.

The second approach relies on the network-level QoS mechanism and the emphasis is on how to *guarantee* IP Network performance in order to achieve the required Network QoS. For example, IETF is working on two QoS frameworks, namely Diffserv (the Differentiated Services) [31] and Intserv (the Integrated Services) [32] to support QoS in the Internet. IntServ uses the per-flow approach to provide guarantees to individual streams and is classified as a flow-based resource reservation mechanism where packets are classified and scheduled according to their flow affiliation. DiffServ provides aggregate assurances for a group of applications and is classified as a packet-oriented classification mechanism for different QoS classes. Each packet is classified individually based on its priority.

2.2.2 VoIP Protocol Architecture

Protocol Architecture

Voice over IP (VoIP) is the transmission of voice over network using the Internet Protocol. Here, we introduce briefly the VoIP protocol architecture, which is illustrated in Figure 2.2. The Protocols that provide basic transport (RTP [33]), call-setup signaling (H.323 [34], SIP [35]) and QoS feedback (RTCP [33]) are shown.

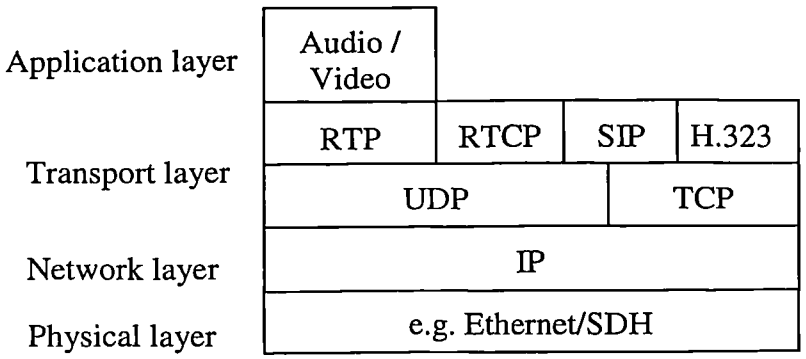


Figure 2.2: VoIP protocol architecture

In this thesis, we focus on voice transmission over Internet and the signaling part is not considered.

Real-Time Transport Protocol (RTP)

Currently, most interactive audio and video applications use the real-time transport protocol (RTP) for data transmission with real-time constraints. RTP runs on top of existing transport protocols, typically UDP, and provides real-time applications with end-to-end delivery services such as payload type identification and delivery monitoring. RTP provides transport of data with a notion of time to enable the receiver to reconstruct the timing information of the sender. Besides, RTP messages contain a message sequence number to allow applications to detect packet loss, packet duplication, or packet reordering.

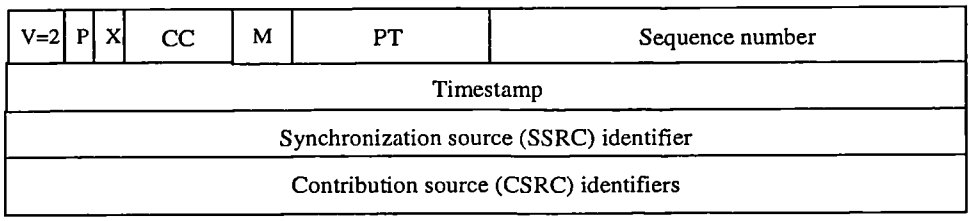


Figure 2.3: RTP header

An RTP message contains an RTP header followed by the RTP payload. An RTP message of version 2 is shown in 2.3. Some fields, which will be used in this report, are described briefly below.

- Payload type (PT): 7 bits. The payload type specifies the format of the RTP payload following the fixed header

- Sequence number: 16 bits. The sequence number counts the number of the RTP packets sent by the sender and is incremented by one for each transmitted packet. The sequence number allows the receivers to detect packet loss, packet duplication and to restore the packet sequence

- Timestamp: 32 bits. The timestamp reflects the sampling instant of the first data sample contained in the payload of RTP packets and is incremented by one for each data sample, regardless of whether the data samples are transmitted onto the network or are dropped

as silent. The timestamp helps the receivers to calculate the arrival jitter of RTP packets and synchronize themselves with the sender.

RTP is extended by the RTP control protocol (RTCP) that exchanges member information in an on-going session. RTCP monitors the data delivery and provides the users with some statistical functionality. The receivers can use RTCP as a feedback mechanism to notify the sender about the quality of an on-going session.

2.2.3 VoIP System Structure

Figure 2.4 shows a basic VoIP system (signalling part is not included), which consists of three parts – the sender, the IP networks and the receiver. At the sender, the voice stream from the voice source is first digitized and compressed by the encoder. Then, several coded speech frames are packetized to form the payload part of a packet (e.g. RTP packet). The headers (e.g. IP/UDP/RTP) are added to the payload and form a packet which is sent to IP networks. The packet may suffer different network impairments (e.g. packet loss, delay and jitter) in IP networks. At the receiver, the packet headers are stripped off and speech frames are extracted from the payload by depacketizer. Playout buffer is used to compensate for network jitter at the cost of further delay (buffer delay) and loss (late arrival loss). The de-jittered speech frames are decoded to recover speech with lost frames concealed (e.g. using interpolation) from previous received speech frames.

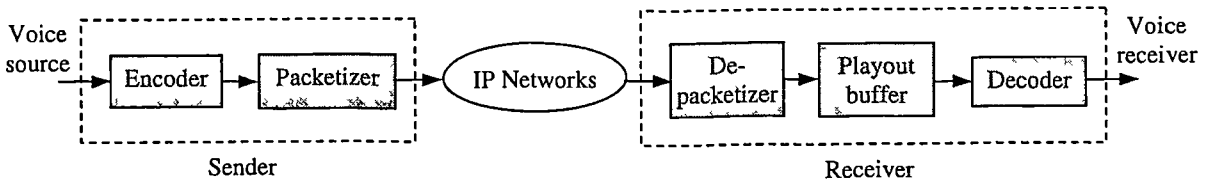


Figure 2.4: Conceptual diagram of a VoIP system

In the thesis, a simulated IP network (e.g. simulated packet loss and delay) and real trace data collected between UK and USA, UK and German, and UK and China will be used. Four

codecs (i.e. G.729, G.723.1, AMR and iLBC) are considered and different playout buffer algorithms are investigated.

2.3 Perceived Quality of Service (QoS)

In voice communications, perceived Speech Quality is the most important QoS metric, as it is related directly to the quality perceived by the end user. It is normally referred to as perceived Quality of Service (QoS) or Quality of Experience(QoE) for voice over IP applications.

Several factors influence perceived speech quality or QoE as shown in Figure 2.5 [36]. The network factors include network packet loss, network jitter and network delay which are the main parameters that determine Network QoS. The jitter buffer and codec located at the terminal side are application related. From an end-to-end point of view, the overall packet loss includes the network packet loss and late arrival loss dropped at the jitter buffer. The overall delay consists of the network delay and buffer delay which is the time spent in the jitter buffer. Except overall packet loss and overall delay, the end-to-end perceived speech quality or the QoE depends further on the codec and codec's packet loss concealment strategy (assuming no external packet loss concealment is used).

The characteristics of network packet loss, delay and delay variation (jitter) will be discussed in more detail in Section 2.5. The impact of overall packet loss, overall delay, and different codec type on perceived voice quality will be analyzed in Chapter 4. The impact from jitter buffer and different buffer algorithms on perceived speech quality will be discussed in Chapter 8.

Other factors affecting end-to-end perceived speech quality which are not shown in the figure are echo, noise, cross-talk, low(high) volume etc. [37]. These are not considered in the study.

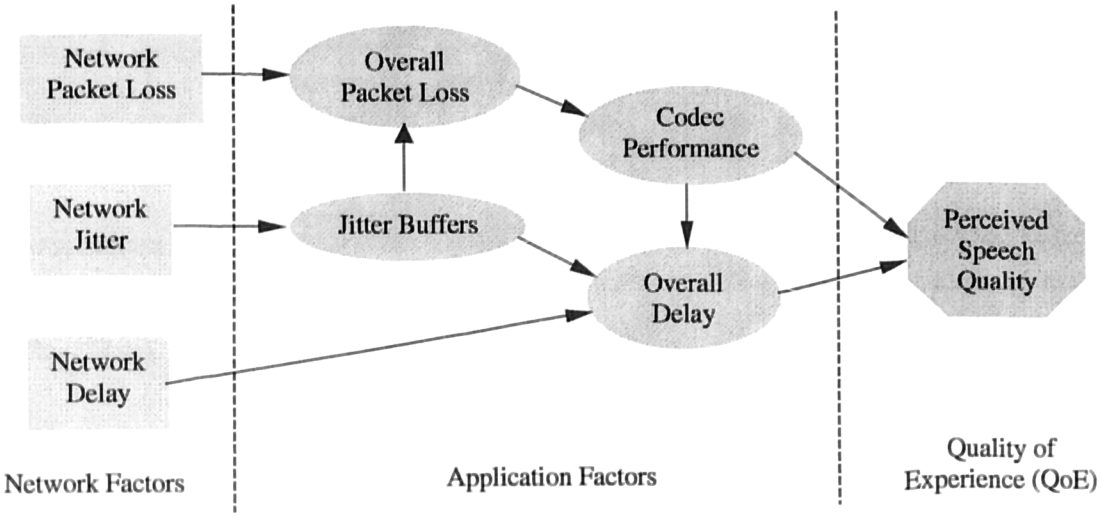


Figure 2.5: Inter-relationship of QoS factors

2.4 Voice Coding Techniques in VoIP Systems

2.4.1 Coding Basic Concept

In order to reduce bandwidth utilization in the transmission of speech signal, speech coding is employed to compress the speech signals. In general, speech coding techniques are divided into three categories: Waveform coders, Vocoders and Hybrid coders.

- Waveform coders: only explore the correlation in time-domain and frequency-domain and attempt to preserve the general shape of the signal waveform. e.g. G.711 PCM (64 Kb/s) and G.726 ADPCM (40/32/24/16 Kb/s)

- Voice coders (vocoders): based on simple (voiced/unvoiced) speech production model and no attempts are made to preserve the original speech waveform. The speech is synthetic. e.g. 2.4/1.2 Kb/s LPC

- Hybrid coders: incorporate the advantages of waveform coders and vocoders to achieve good speech quality at 4.8 to 16 Kb/s, includes all the modern codecs, e.g. G.729 CS-ACELP (8Kb/s), G.723.1 MP-MLQ/ACELP (6.3/5.3 Kb/s), AMR (Adaptive Multi-Rate, ACELP) and iLBC (Internet Low Bit Rate Codec).

In this thesis, we will focus on the modern hybrid coders such as G.729, G.723.1, AMR and iLBC which provide high speech quality at relatively low bit rates.

2.4.2 G.729, G.723.1 and AMR

The G.729 CS-ACELP (Conjugate Structure Algebraic Codebook Excited Linear Prediction, 8 Kbps) and G.723.1 (MP-MLQ/ACELP: Multipulse excitation with a maximum-likelihood-quantizer/Algebraic Codebook Excited Linear Prediction, Dual rate: 5.3/6.3 Kbps) are both standardized by ITU and have been used in VoIP applications. AMR (Adaptive Multi-Rate, ACELP) speech codec was developed by ETSI and has been standardized for GSM. It has been chosen by 3GPP as the mandatory codec. The AMR is a multi-mode codec with 8 narrow band modes with bit rates of 4.75, 5.15, 5.9, 6.7, 7.4, 7.95, 10.2 and 12.2 Kb/s. Mode switching can occur at any time (frame-based). AMR speech codec represents a new generation of coding algorithms which are developed to work with inaccurate transport channels. The flexibility on bandwidth requirements and the tolerance in bit errors of AMR codecs are not only beneficial for wireless links, but are also desirable for VoIP applications.

The three codec types belong to CELP (Codebook Excited Linear Prediction) analysis-by-synthesis hybrid codec. At each speech analysis frame, the speech signal is analysed to extract the parameters of the CELP model (Linear Prediction, or LP filter coefficients, adaptive and fixed codebooks' indices and gains). For stability and efficiency, LP filter coefficients are transformed into Line Spectral Frequencies, or LSF's for transmission. These parameters are then encoded and transmitted. At the decoder, the parameters are decoded and speech is synthesized by filtering the reconstructed excitation signal through the LP synthesis filter.

The major differences between the three codecs lie in the excitation signals, the partitioning of the excitation space (the algebraic codebook), delay and the way in which the coefficients of the filter are represented. For example, the G.729 uses two stage codebook structures for LSP parameters and gets the name "conjugate structure".

The frame information for G.729/G.723.1/AMR is shown at Table 2.1. The delay induced

at encoder is referred as algorithmic delay.

Table 2.1: Frame information for G.729/G.723.1/AMR

Codec	Algorithm	Bit Rates (Kb/s)	Frame length (ms)	Look-ahead (ms)	Algorithmic delay (ms)
G.729	CS-ACELP	8	10	5	15
G.723.1	MP-MLQ /ACELP	5.3/6.3	30	7.5	37.5
AMR	MR-ACELP	4.75 ~ 12.2	20	0	20

Three codecs all have voice activity detection and silence suppression processing. The frames are classified as normal speech frame, SID (Silence Insertion Description) frame and null frame (non-transmitted frame).

2.4.3 Internet Low Bit Rate Codec – iLBC

Recent work in speech coding has led to the development of a predictive speech coder with robustness to packet loss. The robustness to packet losses is obtained by a new design with a self-contained codec state within each speech frame, and frame-independent long-term prediction. The typical one is Internet Low Bit Rate Codec (iLBC) [38] which is a freeware speech compression algorithm developed by Global IP Sound (GIPS)[†]. Comparing with traditional code excited linear prediction (CELP), the iLBC can achieve better voice quality even under severe packet loss conditions. The frame length for iLBC is 20ms (15.2 Kb/s) and 30ms (13.33 Kb/s). The iLBC is currently considered for standardization in the audio/video transport (AVT) working group of the internet engineering task force (IETF) [38, 39].

2.4.4 Codec's Loss Concealment Algorithm

All four codecs (G.729, G.723.1, AMR and iLBC) have built-in loss concealment algorithms, which can interpolate the parameters for the loss frames from the parameters of the

[†]http://www.globalipsound.com

previous frames. For example, for the G.729 codec, the loss concealment algorithm works as below:

The line spectral pair coefficients of the last good frame are repeated. The adaptive and fixed codebook gain are taken from the previous frame but are damped to gradually reduce their impact.

If the last reconstructed frame was classified as voiced, the fixed codebook contribution is set to zero. The pitch delay is taken from the previous frame and is repeated for each following frame. If the last reconstructed frame was classified as unvoiced, the adaptive codebook contribution is set to zero and the fixed codebook vector is randomly chosen.

2.5 Network Performance Characteristics

2.5.1 Packet Loss and its Characteristics

Packet loss is a major source of speech impairment in Voice over IP networks. Such a loss may be caused by discarding packets in the IP networks (network loss) or by dropping packets at the gateway/terminal due to late arrival (late loss) as shown in Figure 2.5. Network loss is normally caused by congestion (router buffer overflow), routing instability such as route changes, link failure, and lossy links such as telephone modems and wireless links. Congestion is the most common cause of loss [40, 41].

The packet loss behaviour of IP networks can be represented as a Markov process because several of the mechanisms that contribute to loss are transient in nature (e.g. network congestion, late arrival of packets at a gateway/terminal, buffer overflow or transmission errors) [42,43,44], which is in fact why packet loss is bursty in nature. Several models [45,46,47] have been proposed for modelling network loss characteristics, which will be discussed briefly in the following sections.

Bernoulli Loss Model

In the Bernoulli loss model, each packet loss is independent (memoryless), regardless of whether the previous packet is lost or not. In this case, there is only one parameter, the average packet loss rate, which is the number of lost packets divided by the total number of transmitted packets in a trace.

2-state Gilbert Model

Most research in VoIP networks use a Gilbert model to represent packet loss characteristics [45, 48, 46]. In 2-state Gilbert model as shown in Figure 2.6, there are two states (state 0 and state 1). We define a random variable X as follows: $X = 0$ (state 0) is for a packet received (no loss) and $X = 1$ (state 1) is for a packet dropped (loss). p is the probability that a packet will be dropped given that the previous packet was received. q is the probability that a packet will be dropped given that the previous packet was dropped. Let π_0 and π_1 denote the state probability for state 0 and 1, as $\pi_0 = P(X = 0)$ and $\pi_1 = P(X = 1)$, respectively.

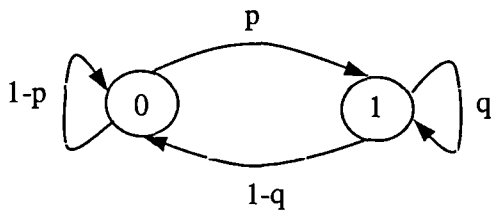


Figure 2.6: 2-state Gilbert model

The procedure to compute π_0 and π_1 is as follows. At steady state, we have:

$$\begin{cases} \pi_0 = (1-p) \cdot \pi_0 + (1-q) \cdot \pi_1 \\ \pi_0 + \pi_1 = 1 \end{cases} \tag{2.1}$$

Thus π_1, the unconditional loss probability (ulp), can be computed as follows:

$$\pi_1 = \frac{p}{p+1-q} \tag{2.2}$$

The *ulp* provides a measure of the average packet loss rate. q is also referred to as the conditional loss probability (*clp*).

The Gilbert model implies a geometric distribution of the probability for the number of consecutive packet losses k, that is the probability of a burst loss having length k, p_k, can be expressed as [46]:

$$p_k = P(Y = k) = (1 - q) \cdot q^{k-1} \tag{2.3}$$

Y is defined as a random variable which describes the distribution of burst loss lengths with respect to the burst loss events.

Based on Equation 2.3, the mean burst loss length $E[Y]$ can be calculated as:

$$E[Y] = \sum_{k=1}^{\infty} k \cdot p_k = \sum_{k=1}^{\infty} k \cdot (1 - q) \cdot q^{k-1} = \frac{1}{1 - q} \tag{2.4}$$

Note that $E[Y]$ is computed based on q, the conditional loss probability (*clp*) only, i.e. that the value of the mean burst loss length is dependent only on the loss behaviour of two consecutive packets.

The probability p and q can also be calculated from the loss length distribution statistics from a trace. Let o_i, $i = 1, 2, \ldots, n-1$ denote the number of loss bursts having length i, where $n - 1$ is the length of the longest loss bursts. Let o_0 denote the number of delivered packets. Then p, q can be calculated by the following equations [47]:

$$p = \frac{(\sum_{i=1}^{n-1} o_i)}{o_0} \qquad q = \frac{(\sum_{i=1}^{n-1} o_i \cdot (i-1))}{(\sum_{i=1}^{n-1} o_i \cdot i)} \tag{2.5}$$

When $p = q$, the 2-state Gilbert model reduces to a Bernoulli model.

In the thesis, unconditional loss probability (*ulp*) and conditional loss probability (*clp*) are used to describe packet loss performance for a 2-state Gilbert model, as in [47] and [49].

Alternatively, unconditional loss probability (ulp) and mean burst loss length are used as in [50] and [51], which may provide a more practical meaning. The transition probability from state 0 to state 1 (value p) can also be calculated from unconditional loss probability and mean burst loss length based on Equations 2.2 and 2.4.

Other Packet Loss Models

Other more complicated packet loss models include 3-state [50] or 4-state modified Markov models [52], 8-state Markov chain models [53] and general n^{th} order extended Gilbert model [46], loss run-length and no-loss run-length models [47]. The 3-state or 4-state modified markov models both introduce lossy periods (with high bursty) and lossless periods (in which no packets are lost). The 4-state one has been used in extended E-model [52, 51].

As the Internet changes so rapidly, there will be a continuous need to measure and further model the Internet parameters (e.g. packet loss) to obtain perceptually more accurate packet loss parameters for QoS monitoring or controlling purposes.

2.5.2 Delay and Delay Variation (Jitter)

Delay and delay variation (jitter) are the main network impairments that affect voice quality.

The end-to-end delay is the time elapsed between sending and receiving a packet. It mainly consists of the following components [40]:

- Propagation delay: depends only on the physical distance of the communications path and the communication medium. When transmitted over fiber, coax or twisted wire pairs, packets incur a one-way delay of 5 μs/km [41].

- Transmission delay: the sum of the time it takes the network interface to send out the packet. Typical wide-area Internet links have OC-12 (622 Mb/s) speed, so that a maximum-sized packet of 1,500 bytes suffers 20 μs of transmission delay at each hop.

- Queueing delay: the time a packet has to spend in the queues at the input and output ports before it can be processed. It is mainly caused by network congestion.
- Codec processing delay: including codec's algorithmic delay and lookahead delay.
- Packetization/de-packetization delay: the time needed to build data packets at the sender, as well as to strip off packet headers at the receiver.
- Playout buffer delay, the time waited at playout buffer at the receiver/terminal.

The ITU has recommended one-way delays no greater than 150ms for most applications [54], with a limit of 400ms for acceptable voice communications.

Jitter is the statistical variance of the packet interarrival time and is caused mainly by the queuing delay component. The IETF in RFC 1889 define the jitter to be the mean deviation (the smoothed absolute value) of the packet spacing change between the sender and the receiver [33]. According to RFC 1889, the interarrival jitter should be calculated continuously as each packet i is received. For one particular packet, the interarrival jitter J_i for the packet i is calculated thus:

$$J_i = J_{i-1} + (|D(i-1,i)| - J_{i-1})/16 \tag{2.6}$$

where D is the difference of the packet spacing.

We follow this definition for jitter calculation in the thesis.

2.6 Summary

The purpose of this chapter has been to present a background for VoIP networks, which underpins the work presented in subsequent chapters. The basic VoIP network connections, the protocol architecture, and the VoIP system structure have been described briefly. The concept of perceived QoS and factors affect speech quality (e.g. packet loss, delay and jitter), the voice coding technologies and the codecs used in the thesis (i.e. G.729, G.723.1, AMR and iLBC)

have been introduced. Finally the network performance characteristics (e.g. packet loss, delay and delay variation) are given.

Chapter 3

Speech Quality Measurement

3.1 Introduction

The need to measure speech quality is a fundamental requirement in modern communications systems for technical, legal and commercial reasons. Speech quality measurement can be carried out using either subjective or objective methods as shown in Figure 3.1. The Mean Opinion Score (MOS) is the most widely used subjective measure of voice quality and is recommended by the ITU [3]. A MOS value is normally obtained as an average opinion of quality based on asking people to grade the quality of speech signals on a five-point scale (Excellent, Good, Fair, Poor and Bad) under controlled conditions as set out in the ITU standard [3]. In voice communication systems, MOS is the internationally accepted metric as it provides a direct link to voice quality as perceived by the end user [37]. The inherent problem in subjective MOS measurement is that it is time consuming, expensive, lack of repeatability and cannot be used for long-term or large scale voice quality monitoring in an operational network infrastructure. This has made objective methods very attractive to estimate the subjective quality for meeting the demand for voice quality measurement in communication networks.

Objective measurement of voice quality in modern communication networks can be intrusive or non-intrusive. Intrusive methods are more accurate, but normally are unsuitable for monitoring live traffic because of the need for a reference data and to utilize the network. A typical intrusive method is based on the latest ITU standard, P.862, Perceptual Evaluation of

Speech Quality (PESQ) Measurement Algorithm [4]. This involves comparison of the reference and the degraded speech signals to obtain a predicted listening-only one-way MOS score, as shown in Figure 3.2(a).

Non-intrusive methods do not need the injection of a reference signal and are appropriate for monitoring live traffic. According to the difference of the inputs to the measurement unit, there are two categories of non-intrusive methods, which are signal-based (the inputs are single-end degraded speech signals) and parameter-based methods (the inputs are network or speech related parameters) as shown in Figure 3.2(b) and (c).

ITU-T E-model [7] is the most widely used parameter-based non-intrusive voice quality measurement method. It can predict the conversational MOS score directly from IP network and/or terminal parameters [52, 9]. Signal-based methods (e.g. vocal tract model [55]) aim to predict voice quality by analyzing directly the in-service speech signal (a degraded signal) without a reference signal.

In this Chapter, the subjective measurement methods are discussed in Section 3.2. Intrusive speech quality measurement methods are presented in Section 3.3. The non-intrusive measurement methods (mainly parameter-based methods) are introduced briefly in Section 3.4. This thesis focuses on parameter-based non-intrusive methods for predicting voice quality.

3.2 Subjective Speech Quality Measurement

Subjective methods are crucial for benchmarking objective methods. The ITU P.800 [3] describes several methods and procedures for conducting subjective evaluations of transmission quality. The most commonly used method is Absolute Category Rating (ACR) test which gives the Mean Opinion Score (MOS). Degradation Category Rating (DCR) is also used in some occasions, which gives Degradation Mean Opinion Score (DMOS).

MOS test is normally carried out under controlled conditions in a laboratory (e.g. in sound proof room). Great care is also required in defining the test conditions and interpreting the

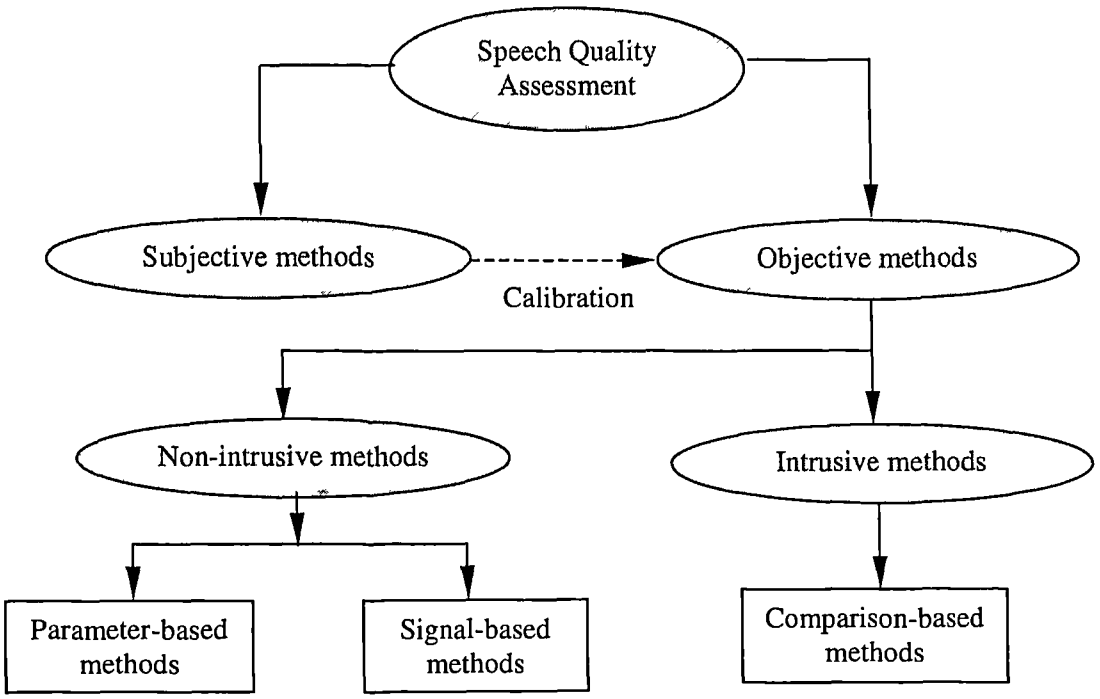


Figure 3.1: Classification of speech quality assessment methods

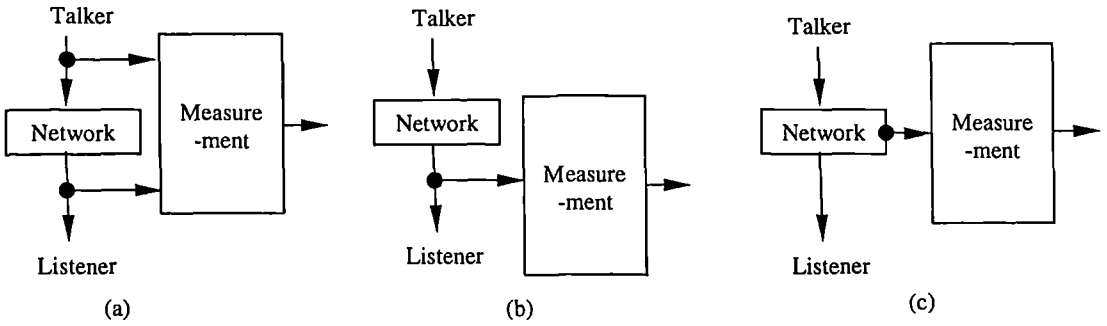


Figure 3.2: Three main categories of objective quality measurement: (a) Comparison-based methods, (b) Signal-based methods, (c) Parameter-based methods

results. This makes MOS test time consuming, expensive and stringent.

3.2.1 Absolute Category Rating (ACR)

For Absolute Category Rating (ACR) listening test, subjects (untrained listeners) are asked to rate the overall quality of a speech utterance being tested without being able to listen to the original reference. The rating of quality is based on an opinion scale as shown in Table 3.1. The average of opinion scores of the subjects gives the Mean Opinion Score (MOS).

Table 3.1: Opinion scale for ACR test

Category	Speech Quality
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

3.2.2 Degradation Category Rating (DCR)

When speech samples of good quality are evaluated, ACR tends to be insensitive, to the effect that small differences in quality are not detected. In such cases, Degradation Category Rating (DCR) is normally used. DCR procedure uses an annoyance scale and a quality reference. Subjects are asked to rate annoyance or degradation level by comparing the speech utterance being tested to the original (reference). The rating scales or the degradation levels are shown in Table 3.2.

Table 3.2: Opinion scale for DCR test

Score	Degradation level
5	Inaudible
4	Audible but not annoying
3	Slightly annoying
2	Annoying
1	Very annoying

The average of the opinion scores of subjects in DCR is called Degradation Mean Opinion Score (DMOS) [1].

In order to regulate the test and comparison between different subjective MOS tests, the ITU P.800 defines a detailed requirement for conducting subjective tests which range from the characteristics of the test material, test environment and test procedures. MOS tests are normally carried out in a restricted, double-walled, sound-proof room. It makes subjective test time-consuming, expensive and stringent. It encourages people working on different objective measurement methods.

3.2.3 Other Subjective Test Methods

Other subjective test methods were proposed recently for better assessing time-varying multimedia quality. Continuous scaled MOS tests [56] were carried out in which a slider was used for each signal to indicate subject's opinion of the voice quality. The work is mainly for subjective evaluation of multimedia services (e.g. audio-visual quality). Similarly, a Quality Assessment Slide (QUASS) [57] was used to continuously rate perceived quality along a specified dimension for audio-visual applications.

3.3 Intrusive Speech Quality Measurement

3.3.1 Introduction

Intrusive, objective speech quality measurement systems normally use two input signals, namely a reference (or original) signal and the degraded (or distorted) signal measured at the output of the network or system under test. They are referred as intrusive due to the injection of test signals and the need to utilize the network. They are more accurate to measure end-to-end perceived speech quality and are unsuitable for monitoring live traffic.

There are a variety of objective speech quality measurement methods, which are normally

classified into three major groups. The first group is time domain measures, such as Signal-to-Noise Ratio (SNR) and Segmental Signal-to-Noise Ratio (SNRseg). These methods are very simple to implement, but are not suitable for estimating the quality for low bit rate codec and modern networks. The second group is spectral domain measures, such as the Linear Predictive Coding (LPC) parameter distance measures and the cepstral distance (CD) [1] measure. These distortion measures are closely related to speech codec design and use the parameters of speech production models. Their performance is limited by the constraints of the speech production models used in codecs. In contrast to the spectral domain measures, the perceptual domain measures, are based on models of human auditory perception. They have been shown to be the most successful objective speech quality measures so far. These measures transform speech signal into a perceptually relevant domain such as bark spectrum or loudness domain, and incorporate human auditory models.

Typical perceptual measure methods are Perceptual Speech Quality Measure (PSQM) [58, 59], Perceptual Assessment of Speech Quality (PAMS) [60,61], Measuring Normalizing Blocks (MNB) [62, 63], Enhanced Modified Bark Spectral Distortion (EMBSD) [64, 65] and Perceptual Evaluation of Speech Quality (PESQ) [4,66] which is the latest ITU standard for assessing speech quality for communication systems and networks.

The basic structure of the perceptual measure methods is illustrated in Figure 3.3. It consists of two modules: perceptual transform module and cognition/judge module. The perceptual transform module transforms the signal into a psychophysical representation that approximates human perception. The cognition/judge model can map the difference between original (reference) and distorted (degraded) signals into estimated perceptual distortion or further to Mean Opinion Score (MOS) scale.

3.3.2 Perceptual Speech Quality Measure (PSQM)

PSQM was developed by PTT research (now KPN), the Netherlands, in 1994 [59]. PSQM was adopted as ITU-T Recommendation P.861 [58] and replaced by ITU-T Recommendation

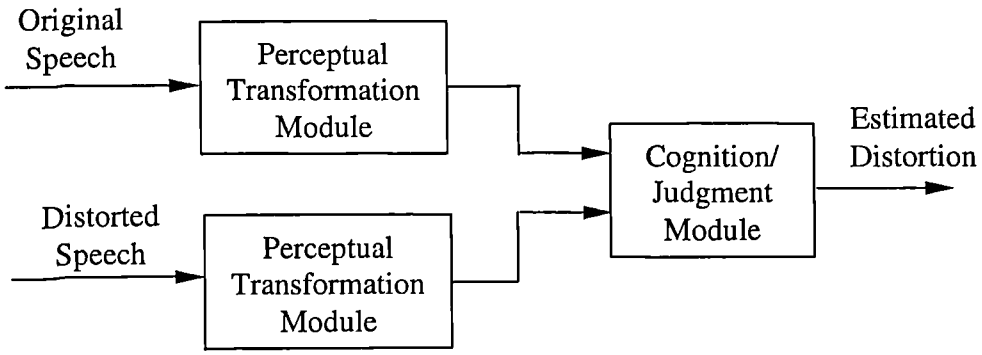


Figure 3.3: Basic structure of perceptual speech quality measurement

P.862 PESQ (see Section 3.3.5) in 2001.

PSQM transforms the speech signal into the loudness domain, modifying some parameters in the loudness calculation in order to optimize performance. PSQM applies a nonlinear scaling factor to the loudness vector of distorted speech. The scaling factor is obtained using the loudness ratio of the reference and the distorted speech. The difference between the scaled loudness of the distorted speech and loudness of the reference speech is called noise disturbance. The final estimated distortion is an average noise disturbance (ND) over all the frames processed. Silence portions have only a small weight in the calculation of distortion. The PSQM computes the distortion frame by frame, with the frame length of 256 samples (8 KHz sampling) with 50% overlap. The result is shown in noise disturbance as a function of time and frequency. The average noise disturbance is directly related to the quality of coded speech.

PSQM+ [67] was proposed by KPN to improve the performance of PSQM for loud distortions and temporal clipping. PSQM+ uses the same perceptual transformation module as PSQM. Comparing to PSQM, an additional scaling factor is introduced when the overall distortion is calculated. This scaling factor makes the overall distortion proportional to the amount of temporal clipping distortion. Otherwise, the cognition module is the same as PSQM.

3.3.3 Measuring Normalizing Blocks (MNB)

MNB was developed at the US Department of Commence in 1997 [62, 63]. It emphasizes the important role of the cognition module for estimating speech quality. MNB models human judgment on speech quality with two types of hierarchical structures. MNB employs two types of calculation in deriving a quality estimate: time measuring normalizing blocks (TMNB) and frequency measuring normalizing blocks (FMNB). Each TMNB integrates over frequency scales and measures differences over time intervals while the FMNB integrates over time intervals and measures difference over frequency scales. After calculating 12 MNBs, these MNBs are linearly combined to estimate overall speech distortion in Auditory Distance (AD) (as shown below). Unlike PSQM, MNB does not generate a distortion value for each frame since each MNB is integrated over frequency or time intervals. There are two MNB structures that offer relatively low complexity and high performance as estimators of perceived speech quality across a wide range of conditions and quality level. They are referred to as MNB structure 1 and MNB structure 2.

$$AD = \sum_{i=1}^{12} weight_i \cdot m(i)$$

These values are then passed through a logistic function to create L(AD). The logistic function is:

$$L(z) = \frac{1}{1 + e^{a \cdot z + b}}$$

L(AD) values range from zero to one and are positively correlated with perceived speech quality.

3.3.4 Enhanced Modified Bark Spectral Distortion (EMBSD)

The EMBSD was developed at Temple University, USA. [64]. The previous version is MBSD [65]. It can be classified as a perceptual domain measure that transforms the speech signal into a perceptually relevant domain which incorporates human auditory models. The MBSD is a modification of the BSD [68] in which the concept of a noise-masking threshold is incorporated, that differentiates audible and inaudible distortions. The MBSD assumes that loudness differences below the noise masking threshold are not audible and therefore are excluded in the calculation of the perceptual distortion. The MBD computes the distortion frame by frame, with the frame length of 320 samples using 50% overlap. The MBSD uses a simple cognition model to calculate the distortion value. The distortion value for an entire test speech utterance was obtained by averaging over all non-silence frames. The EMBSD is an enhancement of the MBSD measure in which some procedures of the MBSD have been modified and a new cognition model has been used. This new cognition model is based on post-masking effects.

3.3.5 Perceptual Evaluation of Speech Quality (PESQ)

PESQ [4,5,6] is the latest ITU standard for objective speech quality assessment for narrow-band telephony network and codecs. It was specifically developed to be applicable to end-to-end voice quality testing under real network conditions, such as VoIP, ISDN etc. It was developed by KPN Research, the Netherlands and British Telecommunications (BT), by combining the two advanced speech quality measures PSQM+ and PAMS.

Real systems may include filtering and variable delay, as well as distortions due to channel errors and low-bit-rate codes. Previous models, such as PSQM and MNB have not taken proper account of filtering, variable delay and short, localized distortions. PESQ addresses these effects with transfer function equalization, time alignment, and a new algorithm for averaging distortions over time. It makes the PESQ the only objective measurement algorithm suitable

for VoIP applications.

Unlike PSQM, MNB, and EMBSD, the objective MOS score at five grade scale (e.g., excellent = 5 and bad = 1) is directly calculated from PESQ algorithm. It makes the applications very convenient. Within the comparison test between PESQ and other objective test measurement algorithms, PESQ meets all requirements set on the 22 known ITU benchmark experiments, which cover 7 languages (e.g. English, French, German, Dutch, Swedish, Italian and Japanese) [69, 70]. The average correlation over all 22 known experiments is 0.935.

All these algorithms have been applied in our research. As PESQ algorithm (C source code) has just become available in 2001 from ITU-T, it was added to our system when it was available.

3.3.6 Perceptual Evaluation of Speech Quality - Listening Quality (PESQ-LQ)

PESQ-LQ [71] is the latest improvement on PESQ. As PESQ score may be between -0.5 and 4.5, while ACR listening quality MOS is on a 1-5 scale. PESQ-LQ was proposed to implement the mapping from P.862 PESQ score to an average P.800 ACR LQ MOS scale, in the range of 1 to 4.5. PESQ-LQ is defined as follows, where x is the P.862 PESQ score and y is the corresponding PESQ-LQ score:

$$y = \begin{cases} 1.0 & \text{for } x \leq 1.7 \\ -0.157268x^3 + 1.386609x^2 - 2.504699x + 2.023345 & \text{for } x > 1.7 \end{cases} \tag{3.1}$$

The function form of PESQ-LQ is shown in Figure 3.4.

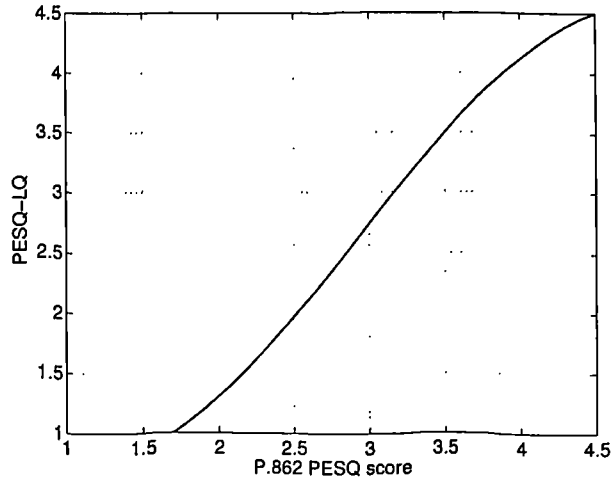


Figure 3.4: Mapping from PESQ score to PESQ-LQ

3.4 Non-intrusive Speech Quality Measurement

3.4.1 Introduction

Unlike intrusive methods described in Section 3.3, in which, a reference/test signal is injected into the tested system/network and live traffic has to be interrupted during the test, non-intrusive speech quality measurement methods do not need the injection of a reference signal and are appropriate for monitoring live traffic.

There are two categories of non-intrusive speech quality prediction methods. One is to predict speech quality directly from varying IP network impairment parameters (e.g. packet loss, jitter and delay) and non-IP network parameters (e.g. codec, echo, language and/or talker issues) as shown as Method 2 (parameter-based) in Figure 3.5. The purpose is to establish the relationship between perceived speech quality and network or non-network related parameters. Typical methods are E-model and artificial neural network (ANN) models, which are presented in Sections 3.4.2 and 3.4.3, respectively. Another approach is to predict speech quality directly from degraded speech signal (or in-service signal) using signal processing methods as in Method 1 (signal-based or output-based), Figure 3.5. The in-service speech signal can be derived directly from T1/E1 links as shown in the figure. Representative methods are INMD

(in-service, non-intrusive measurement devices)/CCI (call clarity index) [72, 73], vocal tract model [55] and machine speech recognition [74].

In our study, we focus on parameter-based non-intrusive speech quality prediction methods which predict speech quality directly from network and/or non-network parameters. The ITU E-model based and artificial neural network based methods will be described in detail in this section.

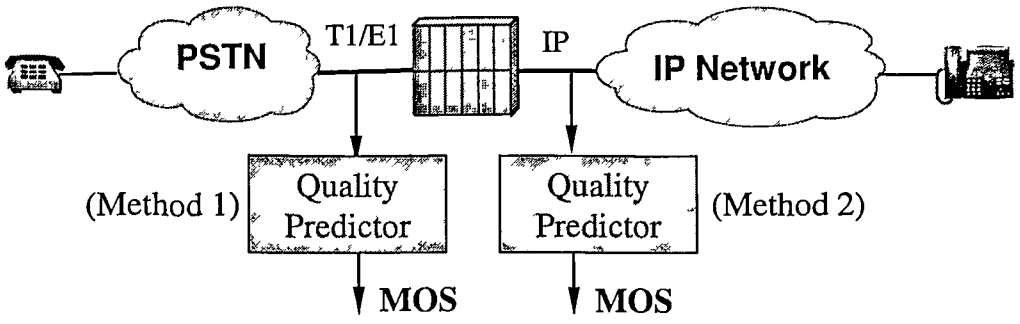


Figure 3.5: Non-intrusive speech quality measurement methods

3.4.2 E-model

The E-model abbreviated from the European Telecommunications Standards Institute (ETSI) Computation Model was developed by a working group within ETSI during the work on ETSI Technical Report ETR 250 [75]. It is a computational tool originally developed for network planning [8, 1, 7], but it is now being used to predict voice quality non-intrusively for VoIP applications [52, 9, 10].

The fundamental principle of the E-model is based on a concept established more than 20 years ago by J. Allnatt [76]: "Psychological factors on the psychological scale are additive". It is used for describing the perceptual effects of diverse impairments occurring simultaneously on a telephone connection. Because the perceived integral quality is a multidimensional attribute, the dimensionality is reduced to a one-dimensional so-called transmission rating scale, R [77]. On this scale, all the impairments are – by definition – additive and thus independent of one another.

The E-model combines the effect of the various transmission parameters into a rating factor, R (which lies between 0 and 100), and from this MOS scores can be derived. The rating factor R is given by:

$$R = R_0 - I_s - I_d - I_e + A \tag{3.2}$$

Where

R_0 : S/N at 0 dB_r point (groups the effects of noise)

I_s : impairments that occur simultaneously with speech (e.g. quantization noise, received speech level and sidetone level)

I_d: impairments that are delayed with respect to speech (e.g. talker/listener echo and absolute delay)

I_e: Effects of special equipment or equipment impairment (e.g. codecs, packet loss and jitter)

A : Advantage factor or expectation factor (e.g. 0 for wireline and 10 for GSM)

ITU G.109 [78] defines the speech quality classes with the Rating (R), as illustrated in Table 3.3. A rating below 50 indicates unacceptable quality.

Table 3.3: Speech quality classes according to E-model

R-value range	100 – 90	90 – 80	80 – 70	70 – 60	60 – 50
Speech transmission quality category	Best	High	Medium	Low	(very) Poor
User's satisfaction	Very satisfied	Satisfied	Some users dissatisfied	Many users dissatisfied	Nearly all users dissatisfied

MOS score can be converted from R value by using the equations in ITU G.107 [7].

For VoIP applications, the impact of IP network impairment (e.g. packet loss and jitter) is related with I_e value. As previous E-model considers only random loss (based on Bernoulli model), extended E-model [52] has been proposed to include more complicated packet loss

conditions (e.g. based on 4-state Modified Markov Model to cater for bursty packet losses). User memory recency effect is also taken into account [52,51]. User R factor and further MOS score can be obtained by complicated computational models.

The E-model (extended E-model) is attractive for non-intrusive voice quality prediction, but it has a number of limitations. For example, it is based on a complex set of fixed, empirical formulae and is applicable to a restricted number of codecs and network conditions (because subjective tests are required to derive model parameters) and this hinders its use in new and emerging applications. Further the E-model is a static model which cannot adapt to the dynamic environment of IP networks. The E-model is based on the assumption that the individual impairment factors defined on the transmission rating scale are independent of each other, this maybe not true [77]. This makes artificial neural network-based learning models very attractive. The neural network-based models will be discussed in the next section.

3.4.3 Artificial Neural Network Model

Unlike E-model which is a computational/mathematical model and is static, artificial neural networks (ANN) model can adapt to the dynamic environment of IP networks, such as the Internet, because of its ability to learn. An ANN model can be built up by learning the non-linear relationships between perceived speech quality (e.g. MOS score) and a variety of network or speech-related parameters.

Artificial neural networks (ANNs) have been very successful in tackling engineering problems such as speech and image recognition, adaptive control, detection, estimation, and telecommunication areas such as ATM call admission control and traffic flow control [79, 80]), traffic prediction for multimedia services [81, 82] and traffic control [83, 84].

The main applications of ANN can be classified into three categories:

i Pattern recognition or classification, e.g., image (e.g. handwriting) or speech recognition

ii Prediction, e.g. to learn the non-linear relationships between input and output of a neural

network.

iii Control and optimization

ANN models have been used to predict both speech and audio quality from perceptual speech or audio features (for comparison-based intrusive quality measurement). For example, neural network model was used to estimate the nonlinear mapping function between input perceptual parameters and output speech quality in [85]; in the new ITU Perceptual Evaluation of Audio Quality (PEAQ) [86], the cognition model is represented by an artificial neural network model. The output variables of the perceptual model are mapped to a prediction of the Subjective Difference Grade (SDG) via a multilayer neural network.

ANN models have recently been used to predict both speech and video quality from IP network parameters [11, 12, 13]. In packet switched networks (e.g. IP networks), the degradation of perceived speech quality is caused by a series of impairments in IP network, such as packet loss, jitter and delay as well as by impairments which are not IP network relevant, such as noise, echo, non-linear impairment via codec. ANN could be used to capture the non-linear mapping "built" by a group of human subjects, between the impairments and a 5-point-scale of quality levels (Mean Opinion Score, MOS, for speech). The basic structure of ANN model is illustrated in Figure 3.6. The input parameters to the ANN model can range from packet loss, packet size, delay, codec type, talker/language, and echo/noise etc. Network jitter has been resolved into packet loss and delay.

The ANN models in [11, 12, 13] are promising to predict voice or video quality from network or non-network parameters, but these rely on subjective tests to create the training sets. Unfortunately, subjective tests are costly and time-consuming and as a result the training sets are limited and cannot cover all the possible scenarios in dynamic and evolving networks, such as the Internet. The impact of a variety of network parameters (e.g. delay variation, loss rate, burstiness and loss pattern) and non-network related parameters (e.g. new codecs, gender or language) on perceived voice quality remains unclear. In addition, the development of previous

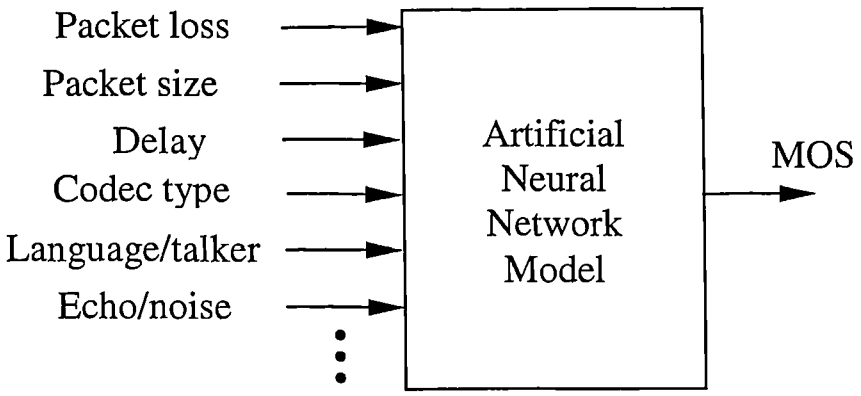


Figure 3.6: Conceptual diagram of ANN model for quality prediction

neural networks-based models was based on a limited number of codecs and can only predict one-way listening voice quality [11, 19] . There is a need for models to predict conversational quality to account for interactivity. Little attention has also been paid to talker dependency.

3.5 Summary

The purpose of this chapter has been to present the state-of-the-art subjective and objective speech quality measurement methods. The subjective voice quality measurement (e.g. MOS), and most importantly, the objective voice quality measurement including both intrusive (e.g. PESQ and MNB) and non-intrusive voice quality measurement (e.g. E-model and neural network models) have been described. The features of intrusive and non-intrusive methods are also given.

Chapter 4

Study of the Impact of Network and Other Impairments on Speech Quality

4.1 Introduction

In Section 2.3, the perceived speech quality and main impairments which may affect voice quality in VoIP networks have been introduced. These impairments include network impairments (e.g packet loss, jitter and delay) and speech-related impairments (e.g. codec type, echo, and gender/language). In order to understand the relationship between voice quality, IP network impairments and relevant parameters associated with speech, a fundamental investigation is undertaken to quantify the impact of network impairments and speech related parameters on perceived speech quality in IP networks, by using the latest ITU algorithm for perceptual evaluation of speech quality, PESQ, and a combined PESQ/E-model structure to obtain an objective measure of voice quality. An understanding of the perceptual effects of these key parameters on voice quality is important as it provides a basis for the development of new and efficient non-intrusive speech quality prediction models for voice quality prediction/monitoring and for QoS optimization and control.

In this chapter, a VoIP simulation system for speech quality analysis is introduced in Section 4.2. The impact of packet loss location and packet loss bursty (e.g. loss patterns and packet size) on perceived speech quality is analyzed thoroughly in Sections 4.3 and 4.4 respec-

tively. Impact of different talkers and languages on perceived speech quality is presented in Section 4.5. The impact of end-to-end delay and codec type on perceived speech quality is described in Section 4.6. Section 4.7 summarises the chapter.

4.2 VoIP Platform

In order to understand the impact of network impairment on the perceived speech quality for VoIP applications, a VoIP simulation platform, as shown in Figure 4.1, has been set up. It offers a window for analysis and evaluation of perceived speech quality for VoIP system. It can be easily expanded due to the modular structure. The basic functions include:

- **File**: Select the reference speech file, select the frame-length for display
- **Edit**: Edit the waveform, keep time-alignment
- **Play**: play the reference and degraded speech files for subjective listening test
- **Codec**: choose codec type from G.729, G.723.1 and AMR, choose bit rate for G.723.1 and AMR, enable/disable VAD etc.
- **Network**: choose network simulation conditions, including loss simulation (Bernouli model, 2-state Gilbert model, any-position burst loss simulation), packet size (the number of speech frames within a packet), initial seeds for random number generation, end-to-end jitter simulation (insert and drop speech frames at any position)
- **Evaluate**: choose evaluate methods including PSQM/PSQM+, MNB, EMBSD and PESQ, show the evaluate results in text format and/or visually as shown in Figure 4.1 together with the loss location.
- **Others**: include showing envelop of the time-domain waveform, amplify/reduce the waveform etc.

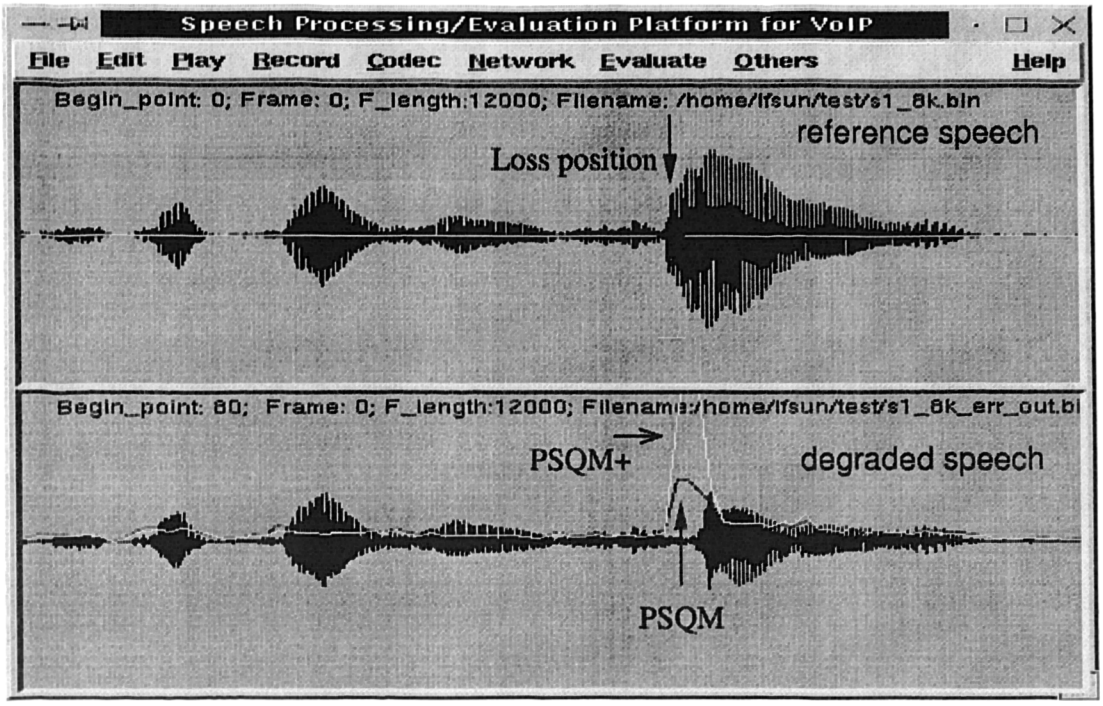


Figure 4.1: Speech processing and evaluation platform for VoIP

This platform has also been used for teaching purposes in the University to demo the concept of VoIP networks, impairments and voice quality measurement.

4.3 Impact of Packet Loss Location on Perceived Speech Quality

4.3.1 Introduction

Impact of packet loss location on perceived speech quality is related with codec's loss concealment performance. When a loss occurs the decoder derives the parameters for the lost frame from the parameters of previous frames to conceal the loss. The loss also affects subsequent frames because the decoder takes a finite time (the convergence time) to re-synchronize its state to that of the encoder. Recent research has shown that for some codecs (e.g. G.729) loss concealment works well for a single frame loss, but not for consecutive or burst losses, and

that the convergence times are dependent on speech content. Further, the effectiveness of a loss concealment algorithm is affected by which part of speech is lost (e.g. voiced or unvoiced). For example, it has been shown that loss concealment for G.729 works well for unvoiced frames, but for voiced frames it only works well after the decoder has obtained sufficient information [48]. Further, the decoder fails to conceal the loss of voiced frames at an unvoiced/voiced transition. Thus, the location of packet loss in relation to different parts of speech is important.

In most studies [28, 48], the analysis of loss concealment performance and convergence times is based on the mean square error (MSE) and signal-to-noise ratio (SNR) criteria. Subjective or perceptual-based objective methods are only used for overall quality assessment under stochastic loss simulations. The perceptual impact of loss concealment algorithms or convergence times for different loss locations is still unknown. It is important to understand the effects of loss location and loss pattern on perceived speech quality, for different types of codec, to allow a more accurate measurement of voice quality. This could be helpful in setting up more efficient speech recovery system and for the development of perceptually relevant packet loss metrics which could be valuable in non-intrusive VoIP measurement.

The IETF has recently proposed a set of new metrics for packet loss [87]. This includes loss constraint distance (i.e. distance threshold between two losses) and "noticeable" loss rate (i.e. percentage of lost packets with loss distances smaller than loss constraint distance). For the same loss rate, different loss patterns may have different effects on perceived speech. In VoIP applications, the loss constraint is related to the convergence times of the decoder. However, it is still unclear how to determine the loss constraint threshold and whether (or how) the threshold is related to codec type, burst size or speech.

The aims of the study in this Section are two fold: (1) to investigate the impact of loss location on perceived speech quality and hence the loss concealment performance of codecs, and (2) to investigate the relationships between convergence times and loss locations/speech content, codec type or loss size.

4.3.2 Simulation System

In order to investigate the impact of packet loss location on perceived speech quality, and the relationships between convergence time and loss location, we set up a simulation system. This includes speech encoder/decoder, loss simulation, perceptual quality measure and convergence time analysis, as shown in Figure 4.2. For codecs, we have a choice of G.729, G.723.1 and AMR. The standard 16 bit, 8 kHz sampled speech signal is processed by the encoder first. Then the parameter-based bit stream is sent to the decoder without frame losses (speech quality degradation in this case is only due to codec). The bitstream is also sent to the loss simulation module where the loss position and frame loss size can be selected. After loss simulation the bit stream is processed by the decoder to obtain the degraded speech signal with loss. The overall perceptual speech quality is measured between the reference speech signal and the degraded speech signal with loss by calculating the perceptual distance values using the PSQM+, MNB and EMBSD algorithms. The perceptual distance for each frame is also measured between the degraded speech without loss and the degraded speech with loss using PSQM+. This eliminates coding impairment from the computation. The convergence time is calculated using the normal Mean Square Error (MSE) method [28].

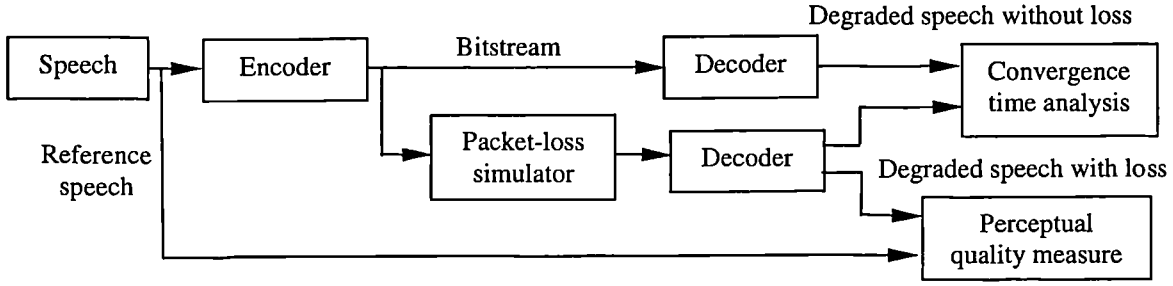


Figure 4.2: Simulation system for analysis of loss location and convergence time

Loss simulation for each codec differs from the loss specification in the codecs. For G.729, if a parameter byte in the bit stream is set to zero, the frame is treated as a loss by decoder and loss concealment is initiated automatically. For AMR, there is an extra byte for the transmit/receive frame type. For a lost frame, there is only a need to set the type as a BAD/ERASED

frame. For G.723.1, a loss location mark file is created and serves as the input to the decoder.

4.3.3 Impact of Loss Location on Perceived Speech Quality

In this experiment, the impact of loss position on the overall perceptual speech quality or the performance of loss concealment under different loss locations is investigated. The PSQM+, MNB and EMBSD perceptual distance values are calculated for the whole test speech sentence (about 6 seconds), while only one loss is produced each time and the loss position moves smoothly from left to right. The move is one frame each time and the frame size is decided by the chosen codec. At each loss location, the frame loss size can change by one, two, three or four frames to simulate different packet size or burst loss size.

The waveform for the first talkspurt for the test sentence "Each decision show (s)" is shown in Figure 4.3. It consists of four voiced segments - V (1) to V (4) corresponding to the vowels 'i', 'i', 'ə' and 'au'. The voiced segments are separated by unvoiced segments. Four voiced segments, V(1) to V(4), which can be decided by pitch delay as shown in Figure 4.4 (for G.729 codec).

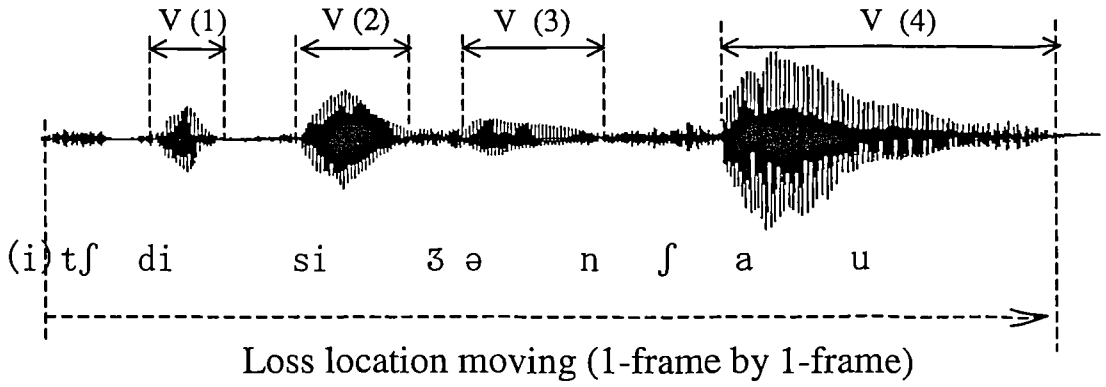


Figure 4.3: Speech waveform for the 1^{st} talkspurt of test sentence

The overall perceptual distance values for PSQM+, MNB and EMBSD for G.729 are shown in Figure 4.5, Figure 4.6 and Figure 4.7 respectively. The values (using PSQM+) for G.723.1 (6.3 Kb/s) and AMR (12.2 Kb/s) are shown in Figure 4.8 and Figure 4.9. In all the figures, the horizontal scales are in the unit of frames. As the frame sizes are 10, 20 and 30ms for G.729,

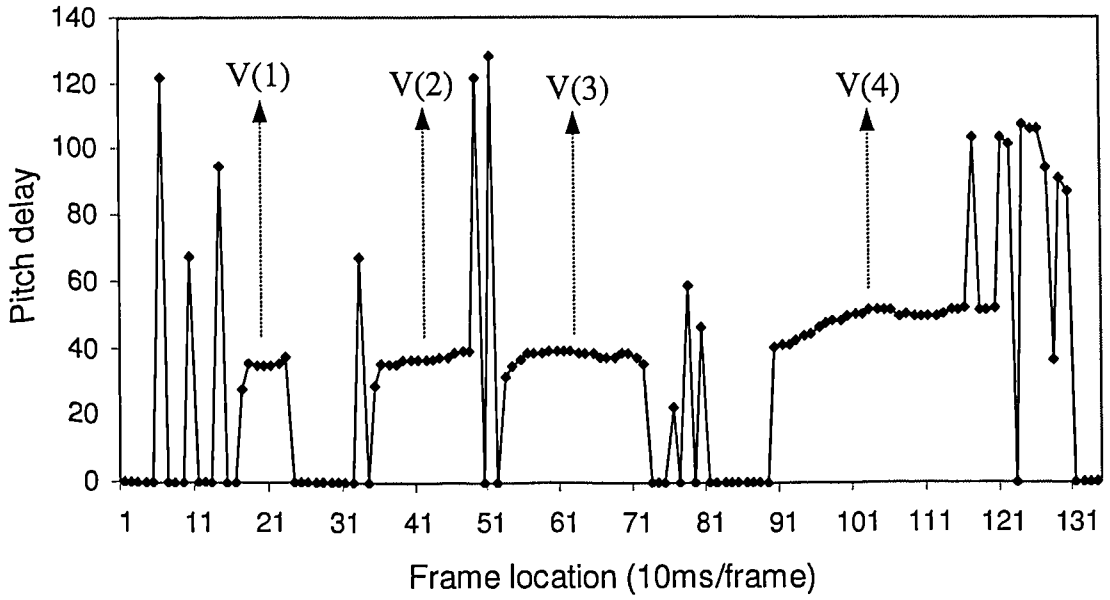


Figure 4.4: Pitch delay for the speech waveform

AMR and G.723.1, respectively, the total number of frames for the test segments shown are 134, 67 and 45.

Examination of Figure 4.5 shows that the perceptual distance value varies between 1.4 and 2.4 as the loss location moves from left to right. In the PSQM+, a change in perceptual distance indicates a change in perceptual speech quality (the smaller the distance, the better the perceived quality). Similar changes in perceived speech quality can also be seen for the MNB (Figure 4.6) and EMBSD (Figure 4.7), as well as for the different codecs (Figure 4.8 and Figure 4.9. It is evident that the same loss condition (one packet loss for the whole test speech segment) causes an obvious variation in overall perceived speech quality, but the variation is dependent on speech content. A loss at unvoiced speech segments shows little impact on perceived speech quality (almost the same perceptual distance values as for no-loss cases). However, a loss at voiced segments has different effects on perceived speech quality depending on its location within the voiced segment. At the beginning of a voiced segment, it has the most severe impact (the peaks in the figures). At the end of voiced segments, the impact is small. In the middle voiced segments, perceptual distances change depending on the codec and frame loss size.

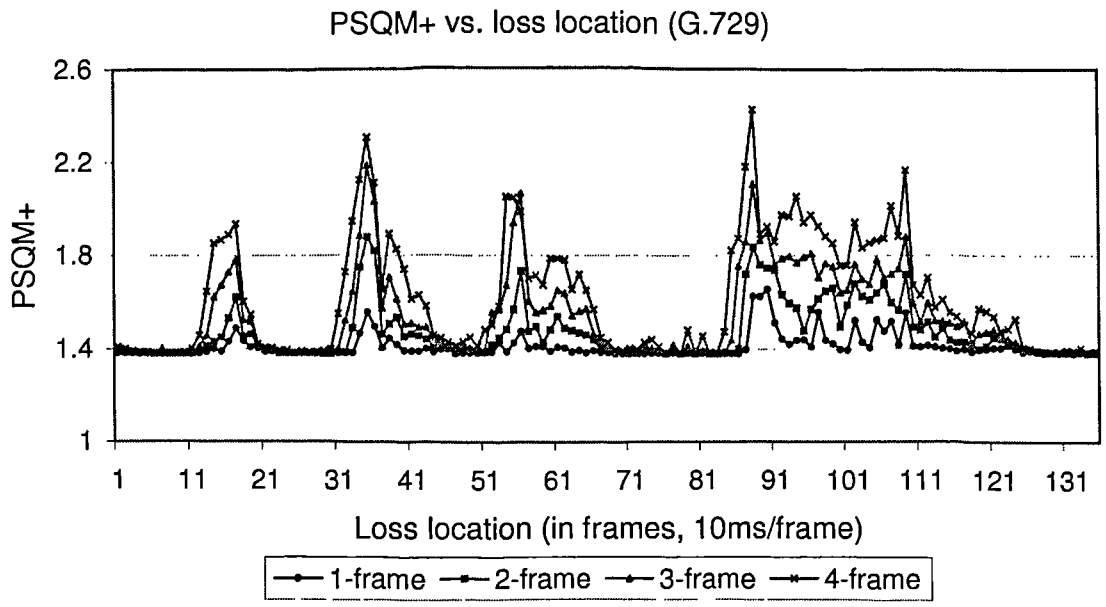


Figure 4.5: Overall PSQM+ values vs. loss location for G.729

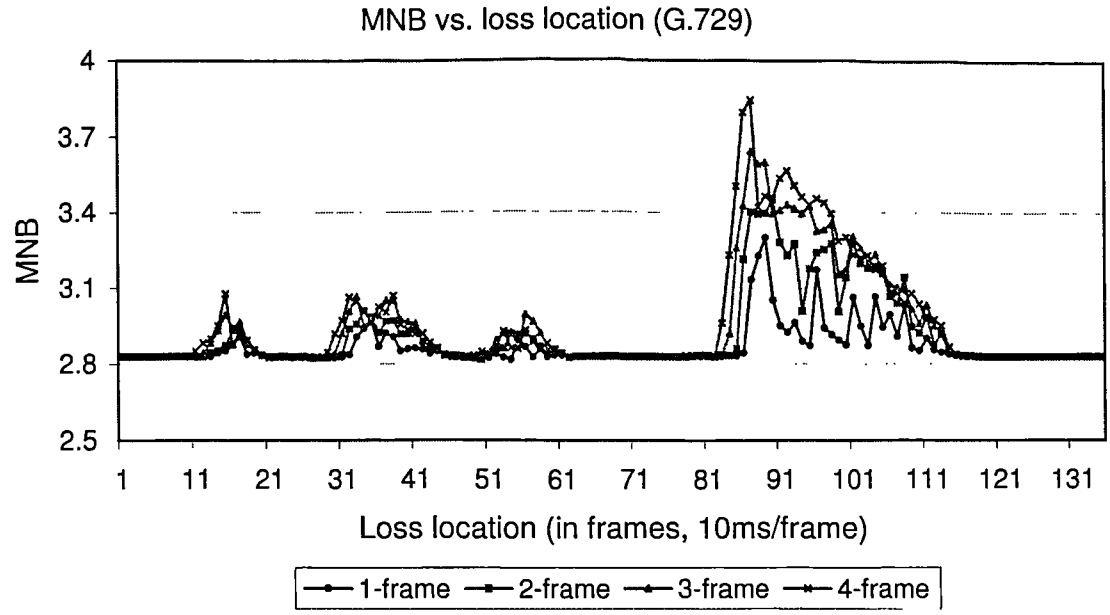


Figure 4.6: Overall MNB value vs. loss location for G.729

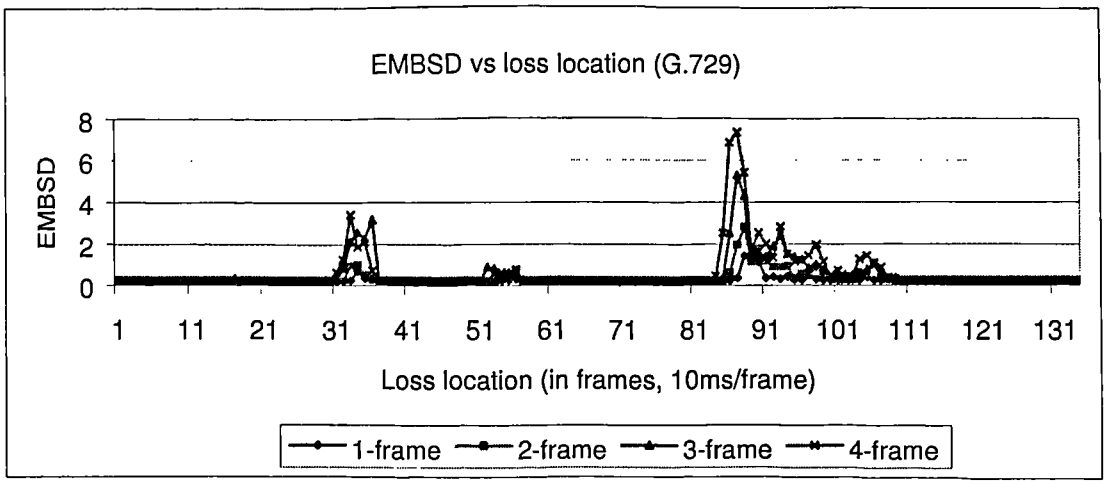


Figure 4.7: Overall EMBSD value vs. loss location for G.729

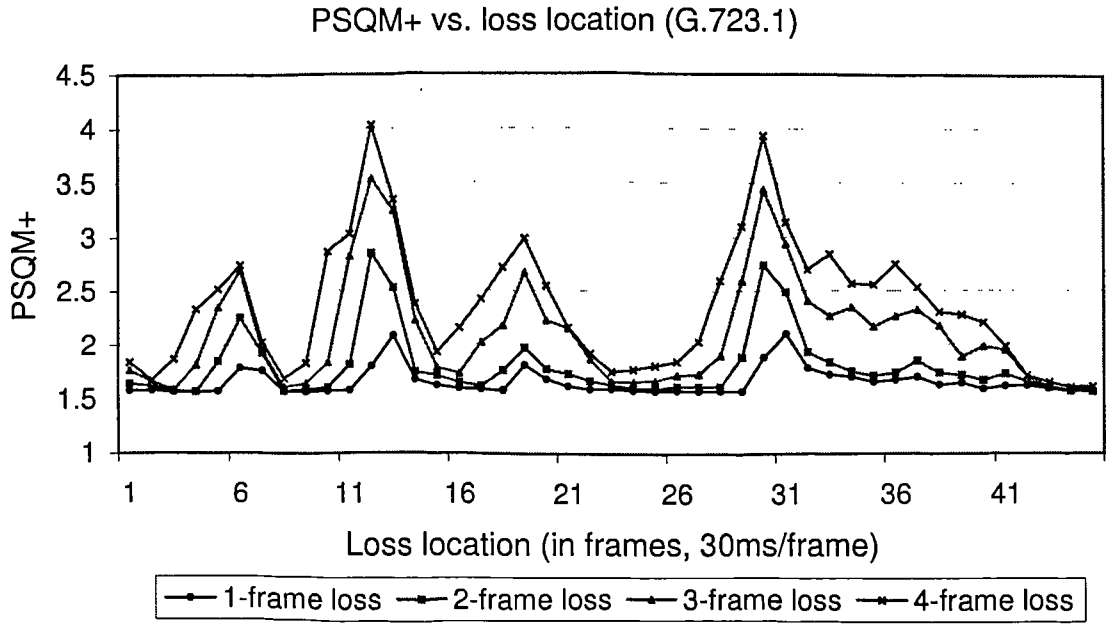


Figure 4.8: Overall PSQM+ value vs. loss location for G.723.1 (6.3 Kb/s)

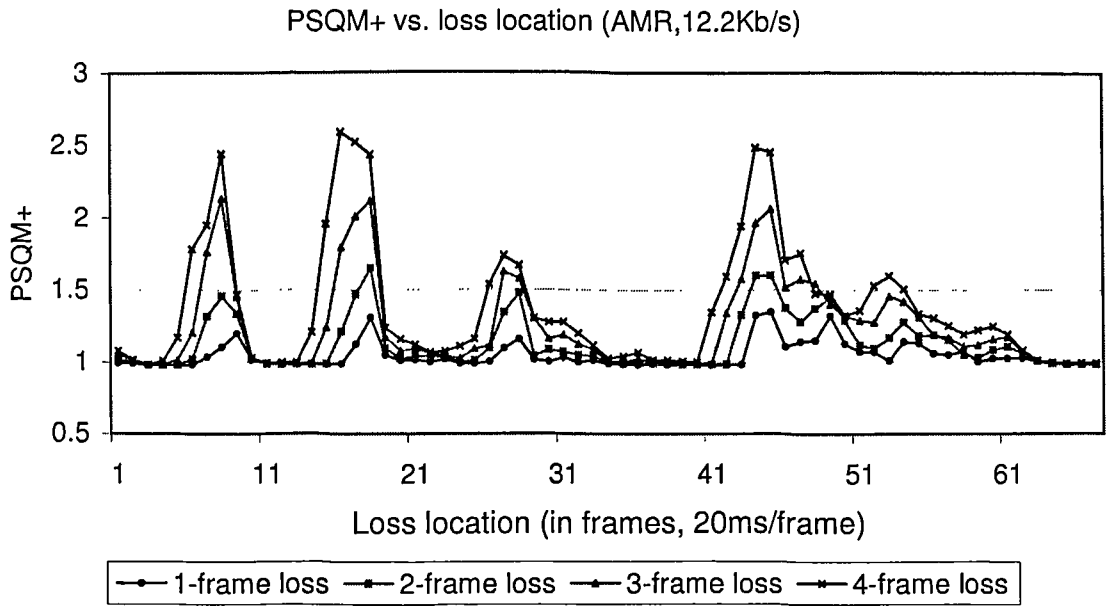


Figure 4.9: Overall PSQM+ value vs. loss location for AMR (12.2 Kb/s)

We explain this phenomenon from two perspectives:

(i). From the perspective of the codec or loss concealment algorithms

In the case of a loss at the beginning of voiced segment, as the previous frame is clearly an unvoiced frame or an unvoiced/voiced transition frame. The loss concealment algorithm will conceal the loss using the filter coefficients and the excitation for an unvoiced sound. It causes the lost frame to be concealed using the unvoiced features. In other words, during the unvoiced to voiced transition period, the shape of the vocal tract is in transition (not stable), and the LP filter coefficients will change rapidly for each frame. The excitation signal is also changing from unvoiced to voiced. The loss concealment algorithm cannot conceal properly for the loss at this transition stage.

For a loss during the stationary part of a voiced segment, the loss concealment algorithm will conceal the current frame with the gain further reduced from the previous frame (adaptive codebook gain). The line spectral pair coefficients (or LP filter coefficients) of the last good frame are repeated. In other words, the vocal tract is at a stable stage (after the transition) and keeps the same shape. The LP filter coefficients are very stable during this stage. If the pitch delay does not change much within a short time period, a small loss can be concealed perfectly

using the parameters of the previous frames. However, when there is an increase in burst loss size or frame size, it is difficult to conceal the losses adequately. The loss concealment performance degrades depending on the features in the voiced segments.

(ii). From the perspective of the perceptual quality measurement algorithms

The signal energy is very important for the overall perceived speech quality for all the perceptual algorithms. If a reference signal frame has a large signal energy (e.g. the beginning of a voiced segment), and the degraded signal has a very small energy (due to improper loss concealment), this will cause a significant increase in the perceptual distance. For a loss during the voiced segment, the degraded signal will normally have a rather large energy. Perceptual distance will vary for different loss size and loss location.

Of the three perceptual measurement methods (PSQM+, MNB and EMBSD), the PSQM+ provides perceptual distance values for most parts of the speech segment. The EMBSD and MNB only show the variations in perceived speech quality for frames with high energy. A loss at the unvoiced or voiced segments with small energy (see Figure 4.5) has no impact on perceived speech quality (flat line area in Figure 4.6 and Figure 4.7. This is due to the different processing methods for silence and non-silence frames in the perceptual quality measurement algorithms. For EMBSD, the perceptual distance for an entire test speech segment is obtained by averaging over all non-silence frames (which are defined as the frames with the energy of the reference speech and the degraded speech both above their preset thresholds). For a loss at short and small energy voiced segments (e.g. voiced segment 1), the degraded speech with a loss has a limited energy. This is not taken into account by the EMBSD in the overall perceptual distance calculation. A similar phenomenon exists for the MNB. The PSQM+ also classifies the frames as silence or non-silence. But it calculates all perceptual distances for silence or non-silence frames and uses different weighting factors for the overall perceptual distance calculation. Thus PSQM+ also gives the perceptual distance value for a loss during small energy.

4.3.4 Impact of Loss Location on Convergence Time of the codec

This experiment was carried out to analyze the convergence time and its relationship to speech content or loss position. The convergence time is calculated by comparing the difference between the degraded signal without loss and the degraded signal with loss (as shown in Figure 4.2). First the MSE method [28] is used to calculate the convergence time for each loss position for a speech waveform such as that shown in Figure 4.3. Here the convergence time is defined as the time between the first good frame received after a burst of lost frames and the frame with its MSE value below a threshold (1% of the maximum MSE value seen so far). The convergence time for G.729 is shown in Figure 4.10 in units of frames (10ms/frame). From the figure, we can see that the convergence times are almost the same for different loss sizes. It shows a good linear relationship for loss at the voiced segments. It is at a maximum at the beginning of the voiced segments and decreases gradually to a minimum at the end of the voiced segments. The convergence time for a loss at the unvoiced segments appears stable. Similar results were also obtained for the AMR and G.723.1 codecs. It seems that the convergence time is only related to the speech content and not to codec and frame loss size.

We analyze further the convergence time based on perceptual distance. We measured the frame-based PSQM+ values between degraded speech without loss and degraded speech with loss. We choose two voiced segments in Figure 4.3. One with only voiced part (V(2)) and another one with the adjacent unvoiced part (V(4)). We change loss positions from the beginning to the end of the waveforms. The perceptual distance variation curves for selected loss positions are shown in Figure 4.11 and Figure 4.12 in the unit of frames (here it is the frame of PSQM+ calculation, which is 32ms frame size with 50% overlapping resulting in 16 ms real frame size). Curves 1 to 5 (Figure 4.11 and 1 to 12 (Figure 4.12) correspond to the loss position from left to right. The loss position for each curve corresponds to the first non-zero point in the curve. The duration of the frames with non-zero (or over a threshold) perceptual distance is related to the convergence time.

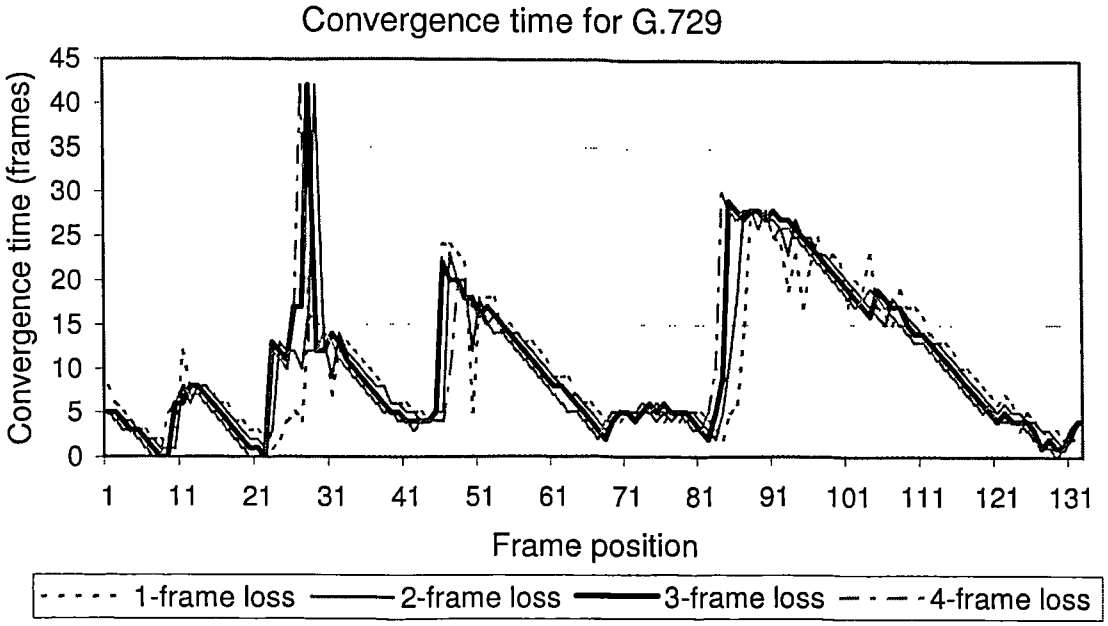


Figure 4.10: Convergence time vs. loss location for G.729

From Figure 4.11 and Figure 4.12, we can see that if a loss occurs during a voiced segment, then the convergence time is almost the remainder of the length of that voiced segment from the loss point (curve 1 to 5 in Figure 4.11 and curve 6 to 12 in Figure 4.12). The perceptual distance itself changes significantly with changes in the location of loss while the influence of the loss seems only limited to the voiced segment. The convergence times are almost the same as for a loss at unvoiced parts (curves 1 to 5 in Figure 4.12). The PSQM+ curves vary in a similar way. We also tested other voiced segments and obtained similar results. The convergence time is more closely related to speech content and less affected by frame loss size and codec type. The convergence time is constrained by the duration of the voiced segments.

We have investigated the impact of loss positions on perceived speech quality and the relationships between the convergence time and loss locations. Preliminary results show that a loss at unvoiced speech segment has almost no obvious impact on perceived speech quality. However, a loss at the beginning of voiced segments has the most severe impact on perceived speech quality. The impact of loss position on perceived speech or the loss concealment performance of three modern codecs (G.729, G.723.1 and AMR) have also been compared and

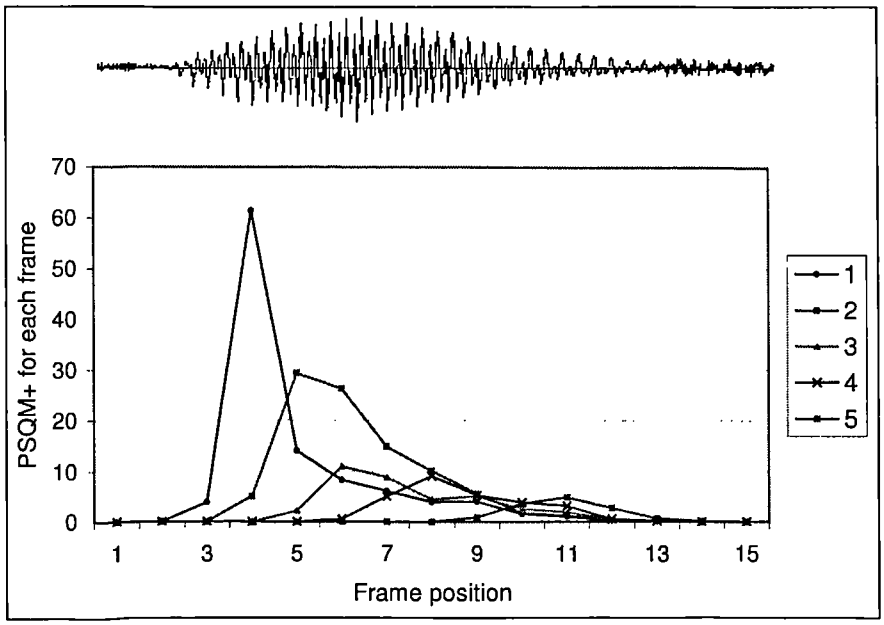


Figure 4.11: PSQM+ for voiced segment 2 (G.729, 2-frame loss)(curves 1 – 5 correspond to 5 loss locations from left to right)

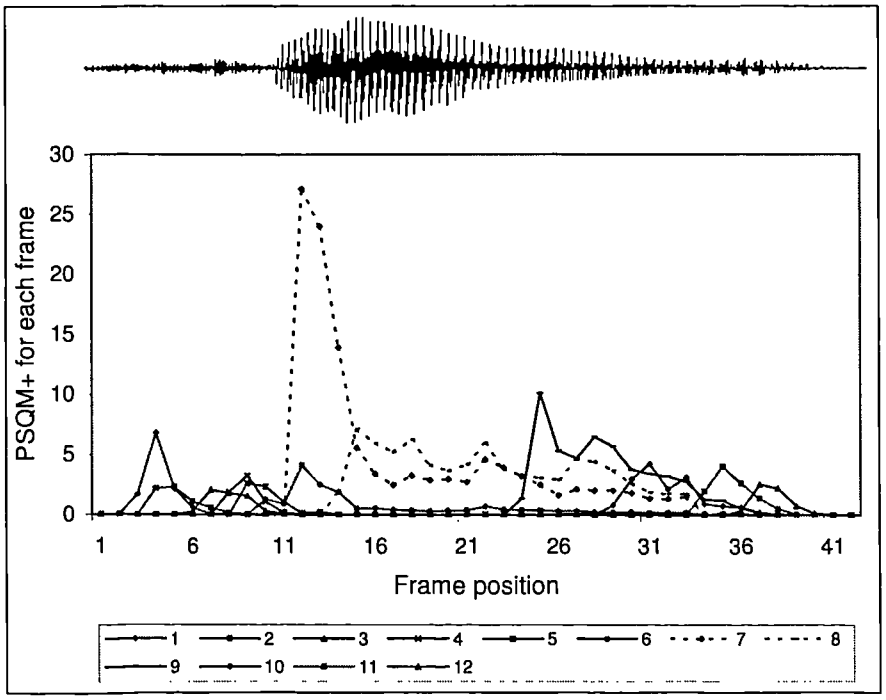


Figure 4.12: PSQM+ for voiced segment 4 (G.729, 2-frame loss)(Curves 1 to 12 correspond to 12 loss locations from left to right)

analyzed. Three different perceptual speech quality measurement algorithms (PSQM+, MNB and EMBSD) are compared for the purpose of loss location analysis. We have analyzed the convergence times for different loss locations and different codecs by taking into account the normal MSE and perceptual PSQM+ measure. The results show that the convergence time is affected mainly by speech content (e.g. it is very stable within unvoiced segment whereas it varies but constrained by the duration of the voiced segments).

4.4 Impact of Packet Loss Bursty on Perceived Speech Quality

4.4.1 VoIP Simulation System

This section is to investigate how packet loss bursty, packet loss pattern and packet size affect the perceived speech quality. A block diagram of the set up that was used to generate the data for the study is depicted in Figure 4.13. The system similar with Figure 4.2, includes a speech database, an encoder/decoder, a packet loss simulator, a speech quality measurement module. The speech database is taken from the ITU-T dataset [88]. Four modern codecs were chosen for the study. These are G.729 CS-ACELP (8 Kbps), G.723.1 MP-MLQ/ACELP (5.3/6.3 Kbps), Adaptive Multi-Rate (AMR) codecs with eight modes (4.75 to 12.2 Kbps) and iLBC (internet Low Bit-rate Codec) with two models (13.33 and 15.20 Kbps).

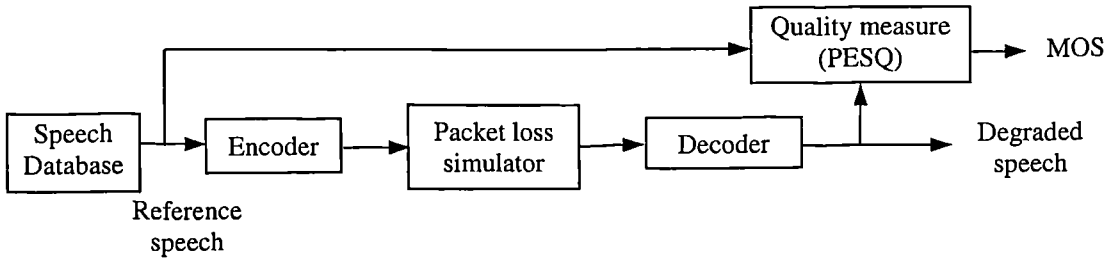


Figure 4.13: Conceptual diagram of VoIP system for speech quality analysis

Instead of one loss generated each time, a 2-state Gilbert model was used to simulate packet

loss (see Figure 2.6). The Gilbert model is well known to represent the packet loss behaviour of a real network, even after the late arrival loss due to jitter is taken into account (if a packet arrives too late, it will be discarded by jitter buffer) [46]. In the figure, State 0 is for a packet received (no loss) and State 1 is for a packet dropped (loss). p is the probability that a packet will be dropped given that the previous packet was received. q is the probability that a packet will be dropped given that the previous packet was dropped. q is also referred to as the conditional loss probability (clp). The probability of being in State 1 is referred to as unconditional loss probability (ulp). The ulp provides a measure of the average packet loss rate.

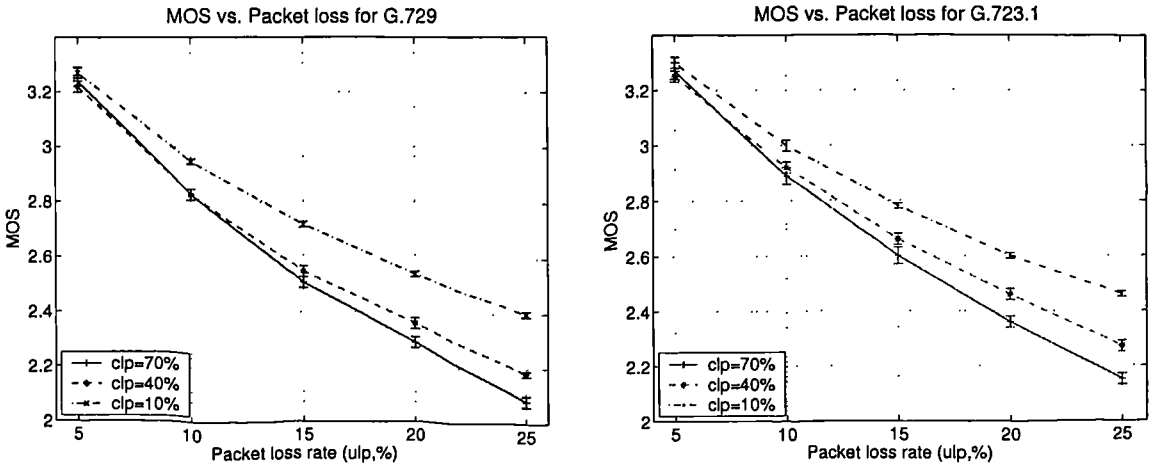
The conditional loss probability (clp) and unconditional loss probability (ulp) are used in the section to characterize the end-to-end packet loss behavior of the network (including network packet loss and jitter buffer loss).

When this experiment was carried out, the latest ITU perceptual measurement algorithm, the Perceptual Evaluation of Speech Quality (PESQ), was just available from ITU-T. So only PESQ is chosen for quality evaluation in the experiment. Unlike PSQM/MNB/EMBSD, PESQ gives the objective MOS score in a 5-grade scale directly.

4.4.2 Impact of Packet Loss Burstiness on Perceived Speech Quality

We first investigated how packet loss affects perceived speech quality. Packet loss is the dominant impairment in IP networks, but its impact on perceived voice quality is still unclear (because of the use of jitter buffer, loss here includes both network loss and late arrival loss due to jitter). A fixed packet size was set for different codec. Different network ulp and clp were chosen and the corresponding MOS score was calculated. To account for a wide range of possible type of packet loss patterns and locations, 300 different initial seeds for random number generation were chosen for each pair of ulp and clp. The average MOS score and 90% Confidence Internal (definition see Appendix A) were calculated. The results for G.729 and G.723.1 (6.3 Kb/s mode) are shown in Figures 4.14(a) and 4.14(b). The length of the test speech sentence was about 12 seconds. The packet size for G.729 and G.723.1 was 2 and 1

frames/packet, respectively. No VAD was activated.



(a) for G.729 (b) for G.723.1

Figure 4.14: MOS vs. Packet Loss

From Figure 4.14(a) and Figure 4.14(b), it can be seen that the *clp* has an obvious impact on the perceived speech quality even for the same average loss rate (*ulp*). When burst loss increases (*clp* increasing), the MOS score decreases and the variation of the MOS score (shown in CI) also increases. This is because losses may occur more concentrated with high burst losses and this results in large variation in the MOS scores due to the locations of the losses, whereas it may occur evenly in low burst loss cases which results in small deviations. There is only a small difference between the results for G.729 and G.723.1, when *ulp* is 10%, and *clp* is from 40% to 70%. Similar results were obtained for AMR and iLBC.

If we consider the deviation in the MOS score due to different loss locations and loss patterns, it is possible to get two cases, which show the different results. For example, one may show the perceived quality decreasing with the increasing of bursty losses, and another one may show the opposite.

We have seen non-consistent results from different sources by subjective tests (e.g. MOS score may increase or decrease when bursty loss increases [89,90]). From [12], it was reported that "the MOS tends to improve as the number of consecutively lost packets increases" in Spanish, whereas, "it shows almost no variation and goes down a little bit" as the number

of consecutively lost packets increases (burst loss increases) in French. It was explained as language dependency. That is, given the same network parameter variations, the MOS scores differ according to language.

According to our analysis based on objective tests, it seems that these non-consistent results are mainly due to packet loss location, not other reasons. For subjective tests, the influence from different loss locations/patterns could be easily missing, as it is almost impossible to cover all the cases by a subjective test. This is one advantage of using objective methods for a thorough analysis of the impact of bursty loss on perceived speech quality.

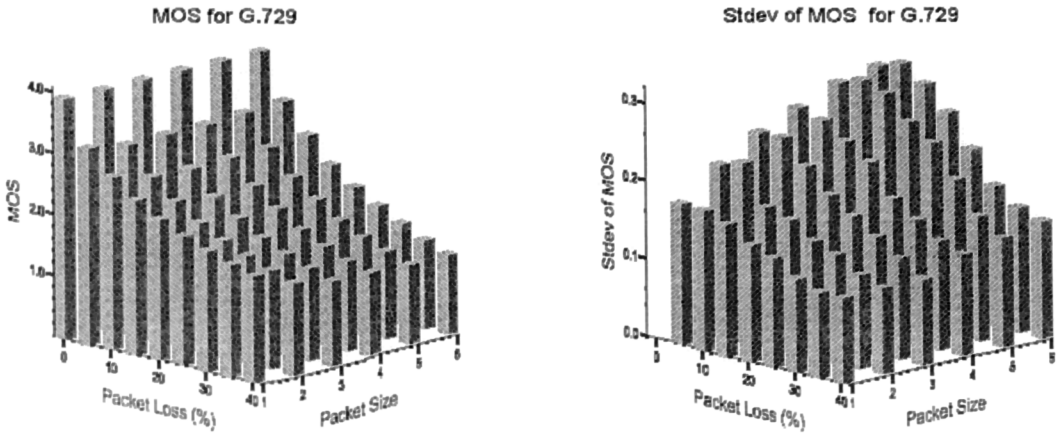
4.4.3 Impact of Packet Size on Perceived Speech Quality

This experiment investigated how packet size affects perceived speech quality. A fixed *clp* (40%) was set and *ulp* was changed from 0% to 40% in 5% increment. Packet size was changed from 1 to 6 frames/packet. As before, 300 different initial seeds were generated randomly. The average MOS and the standard deviation of MOS scores for G.729, G.723.1 (6.3 Kb/s) and AMR (12.2 Kb/s) are shown in Figures 4.15(a),4.15(b), 4.16(a), 4.16(b), 4.17(a) and 4.17(b), respectively. From these figures, it can be seen that the packet size has in general no obvious influence on average perceived speech quality values for a given packet loss rate, but the deviation in speech quality for the same network loss rate depends on packet size and codec. When packet loss rate is lower and packet size is larger, the higher values of the standard deviation of MOS scores means larger deviation in speech quality for the same network conditions. The deviation in speech quality is due to different packet loss locations. It will affect the accuracy for non-intrusive voice quality prediction.

The impact of packet size on perceived speech quality has been reported by literatures [2,91], which are both based on subjective listening test. In [2], it was first pointed out that the results for the male voice showed no significant effect of packet size on the perceived quality for a given packet loss ratio, while for the female voice, a small effect of packet size was found. The difference between male and female was most probably caused by coincidental place of packet

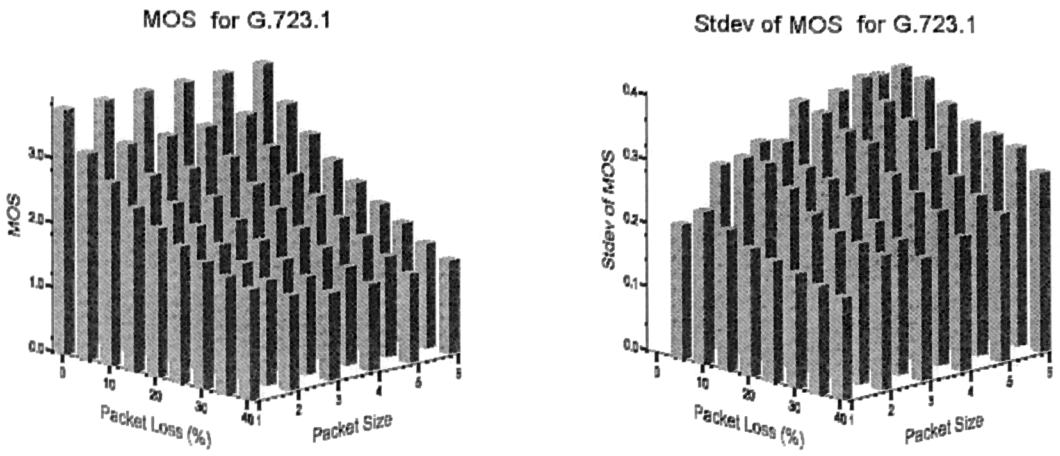
losses. In [91], it was further confirmed that the speech quality does not in general depend on the number of speech frames/packet (packet length) for a given overall packet loss ratio for G.729. However the test was only based on 4 subjects, it is difficult to achieve statistical significance of the test results but to indicate possible trends.

Our experiment based on PESQ gave a thorough analysis of the impact of packet size on the average of MOS score and the deviation in MOS score. It confirmed and extended the previous results and gave a full picture of the influence of packet size on perceived speech quality.



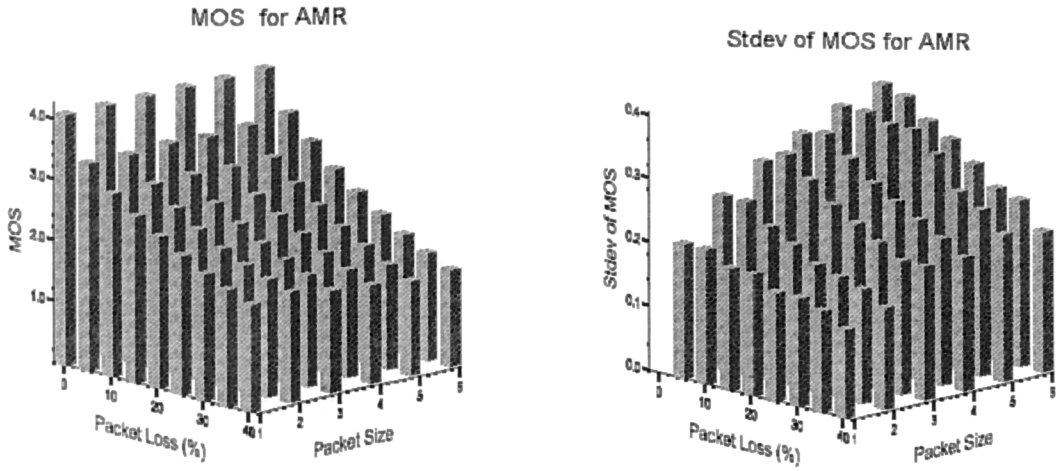
(a) (b)

Figure 4.15: Average MOS and Stdev of MOS for G.729



(a) (b)

Figure 4.16: Average MOS and Stdev of MOS for G.723.1 (6.3Kb/s)



(a) (b)

Figure 4.17: Average MOS and Stdev of MOS for AMR (12.2Kb/s)

4.5 Impact of Talkers/Languages on Perceived Speech Quality

4.5.1 Introduction

In this section, we investigated how different speakers or languages affect the perceived speech quality.

Language dependency in VoIP systems has recently been reported [12]. French, Spanish and Arabic have been tested. Speech quality language dependency has been observed under certain situations. That is, given the same network parameter variations, the MOS scores differ according to the languages. As the work is based on subjective tests and only limited languages are considered. Little attention has been paid to talker dependency.

Two experiments were carried out to investigate whether different talkers (male or female) or languages have an effect on perceived voice quality for the same network conditions. The first experiment is based on TIMIT dataset [92] and the second one is based on ITU-T dataset [88]. The detailed procedure of the experiment and the results are described below.

4.5.2 Experiments and Result Analysis for TIMIT dataset

This experiment was based on DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus Training Data Set [92]. TIMIT is a speech recognition database with train and test data set. It contains total 2342 sentences spoken by 630 speakers from 8 major dialect regions of the United States. It has three sentence types as follows:

- Dialect (SA), total 2 sentences, spoken by 630 speakers
- Phonetically Compact (SX), total 450 sentences, spoken by 7 speakers
- Phonetically Diverse (SI), total 1890 sentences, spoken by 1 speaker

For the purpose of talker dependency analysis, we chose speech data with the same contents spoken by different speakers. "sa1" and "sa2" are chosen as they are the only two sentences in the data set which are spoken by all the speakers. As "sa1" and "sa2" are too short (about 2.5 ~ 3 s), we grouped them to form a longer speech file (about 10s) with the format of "sa1 + silence + sa2 + silence + sa1". "silence" period is added between two short sentences. 6 talkers (3 male and 3 female) from the train set (dialect 1 and 2) were chosen. We kept the file name as same as that in TIMIT dataset for the purpose of easier check/comparison in the future. The name started with letter "f" (for female) or "m" (for male), and was followed by 3 letters, which were initials of the speaker. All speech files were pre-converted to –26 dBov (active speech level) according to ITU P.56 [93]. The activity factor for all speech files was about 0.82.

The experiment system was similar with Figure 6.1. We chose one speech reference file from the 6 composite speech files mentioned above. We changed *ulp* (*unconditional loss probability*) in packet loss simulator from 0 to 30 % in 5% increment, and *clp (conditional loss probability)* was set to 10%, packet size was set to 2 for G.729 (no VAD). For each combination of *ulp* and *clp*, one initial seed was created randomly, and then perceived speech quality (MOS score) is calculated between the distorted and reference speech using PESQ algorithm. As before, 300 different random initial seeds were generated and the average MOS score was calculated. The average MOS scores for the six talkers for G.729 are illustrated in Figure 4.18.

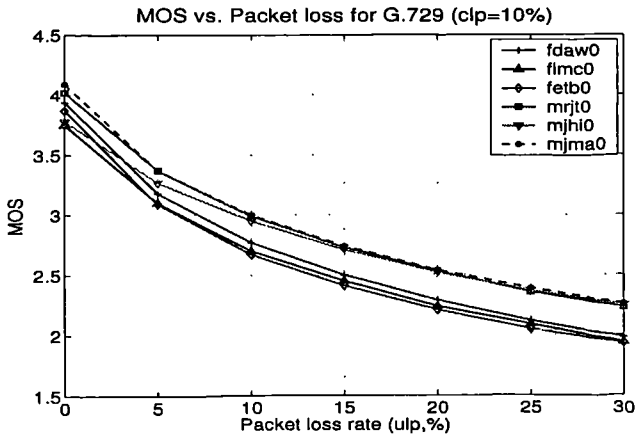


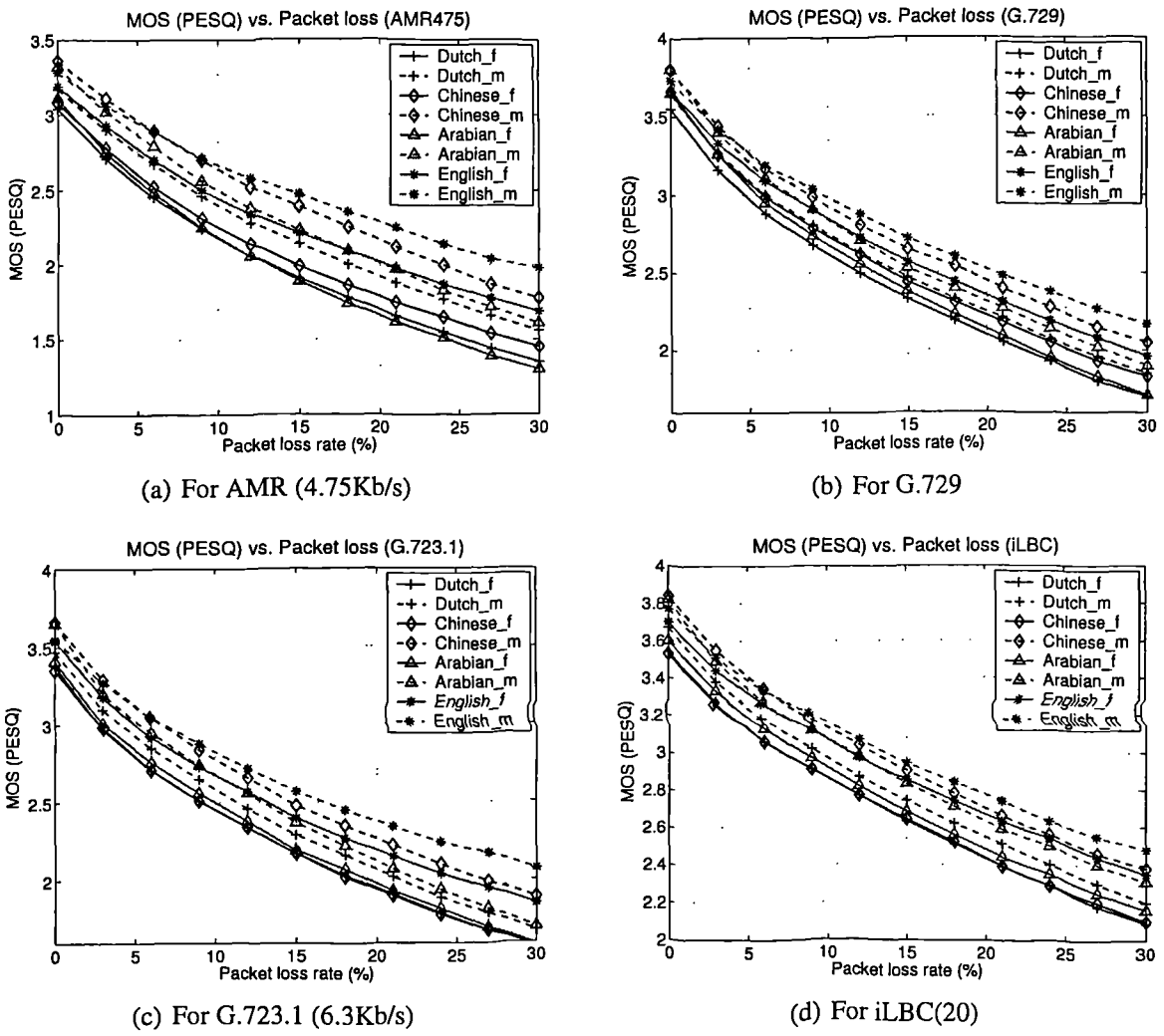
Figure 4.18: MOS vs. loss rate for different talkers

From inspection of Figure 4.18, it can be seen that the impact of different talkers on perceived speech quality appears to depend mainly on the gender of the talker (male or female), irrespective of the dialect/accent. The quality for the female talker tends to be worse than that of male talker for the same network impairments. This effect is more obvious when loss increases.

4.5.3 Experiments and Result Analysis for ITU-T Dataset

In last section, we showed that the gender of a speaker (male or female) has an effect on perceived voice quality. The results were based on English language speech files taken from TIMIT database and suggest that voice quality for female speakers is more susceptible to packet loss than for male speakers. In this study, we have extended the work further to consider the impact of gender on voice quality for more speakers and seven different languages using the ITU-T database [88]. Reference speech files for sixteen speakers (8 females and 8 males) in each of the seven languages (English, Dutch, Chinese, Arabic, American English, French and German) were chosen for the study. For each language, the average MOS score was calculated for all male speakers and similarly for all eight female speakers. To remove the influence of packet loss location [18], the MOS score was obtained by averaging over 20 random seeds for each sentence. Bernoulli packet loss model was used for simplicity. The results for G.729, G.723.1, AMR (4.75Kb/s) and iLBC (15.2Kb/s) are shown in Figures 4.19 and 4.20.

From inspection, the results confirm that gender (male or female) has an obvious impact on perceived voice quality for all the languages considered. Voice quality for the female talker tends to be worse than that of the male talker for the same network impairments. This is more obvious as packet loss increases (the difference in MOS values between the different gender is about 0.3 to 0.4). The impact of language is also obvious. For example, English has the highest MOS score and Dutch the lowest for the same network conditions for all four codecs considered. The difference in MOS score between them is about 0.5 (for the same gender). The difference between English and American English is minor for all four codecs for the same gender.



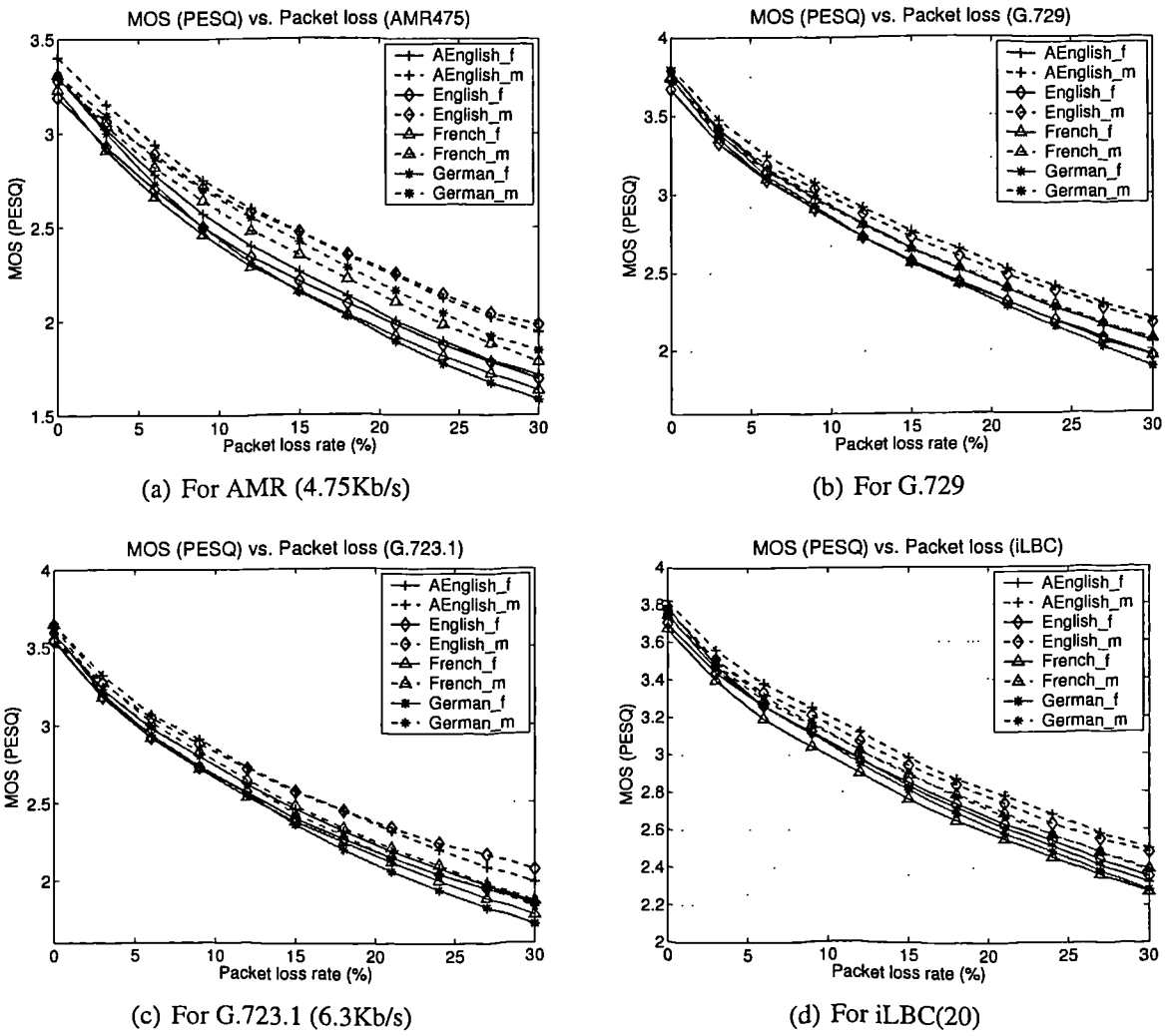
(a) For AMR (4.75Kb/s)

(b) For G.729

(c) For G.723.1 (6.3Kb/s)

(d) For iLBC(20)

Figure 4.19: MOS vs. loss rate for ITU-T dataset (1)



(a) For AMR (4.75Kb/s)

(b) For G.729

(c) For G.723.1 (6.3Kb/s)

(d) For iLBC(20)

Figure 4.20: MOS vs. loss rate for ITU-T dataset (2)

The reason for talker (male or female) and language dependency is likely to be due to the codec algorithms. As the G.729, G.723.1, AMR and iLBC are all CELP-based codecs, the use of linear predictive model of speech production can lead to variations in codec performance with different talkers and languages [94].

4.6 Impact of Delay and Codec on Perceived Speech Quality

End-to-end delay affects only interactivity, or it affects conversational voice quality. The impact of end-to-end delay on speech quality for different codecs can be derived from E-model as follows:

Assuming a perfect echo cancellation, impairment effect due to delay, I_d, can be calculated by several complex equations in ITU-T G.107 [7] and shown in Figure 4.21.

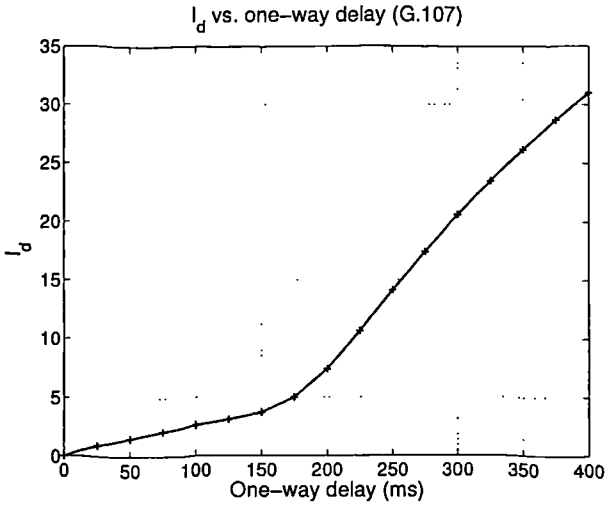


Figure 4.21: I_d vs. one-way delay from G.107

The R factor can be simplified as Equation 4.1.

$$R = R_0 - I_d - I_e \tag{4.1}$$

The default value for R_0 is 93.2 [7]. I_e accounts for equipment impairment. If assuming no packet loss, I_e only represents codec impairment itself. For G.711, $I_e = 0$. For other codecs,

$I_e > 0$.

The MOS score can be further derived from R value (detailed procedures will be explained later in Chapter 5). The MOS vs. one-way delay (d) is shown in Figure 4.22 for G.711, AMR, iLBC, G.729 and G.723.1 (all without packet loss). The method to obtain the equipment impairment value (i.e. I_e) for all the codecs (with or without loss) will be detailed in Chapter 5.

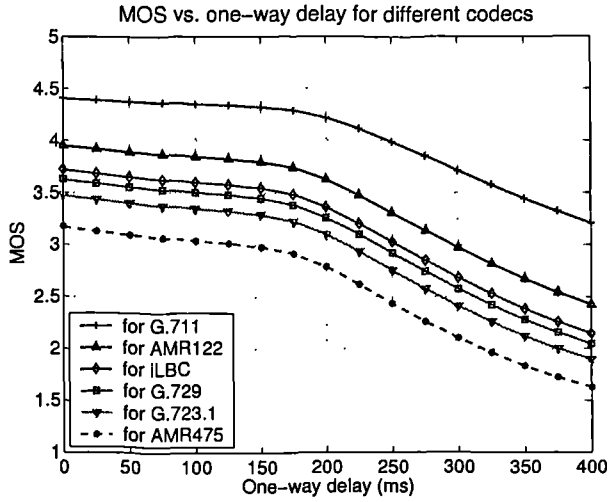


Figure 4.22: MOS vs. one-way delay (d)

From Figure 4.22, it can be seen that the impact of delay on voice quality is small when delay is less than bout 170ms, but increases noticeably when the delay is greater than 170ms. The extent of the impact of delay is also related to the codec used.

In order to compare the impact of different codecs on voice quality under different packet loss conditions, we carried out a test for different packet loss rate (Bernoulli packet loss model was used here for simplicity). Each MOS score was obtained by averaging over all eight male and eight female speech samples (English) from the ITU-T data set (20 random seeds were chosen to avoid the influence from loss location). The results of MOS vs. packet loss for different codecs are shown in Figure 4.23.

From the figure, it can be seen that the iLBC codec gives the best voice quality when packet loss rate is high (over 4 %). The AMR (12.2 Kb/s) codec has the highest MOS score when packet loss rate is zero, whereas the AMR (4.75 Kb/s) codec has the lowest quality regardless

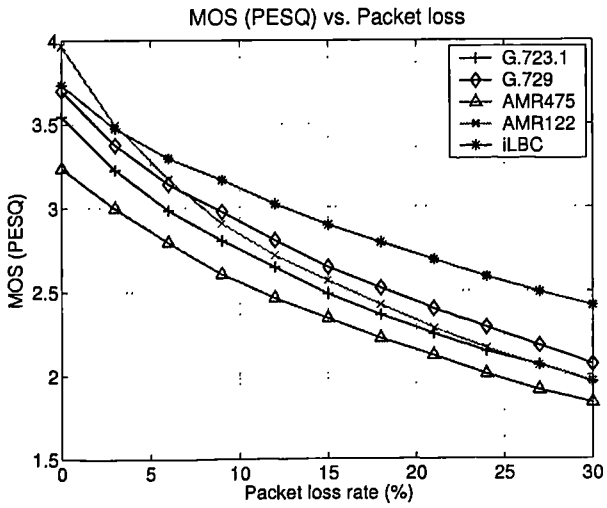


Figure 4.23: MOS vs. packet loss for different codecs

of packet loss rate. Quite clearly, different codecs (or different modes of codecs) have different effects on voice quality under packet loss conditions.

4.7 Summary

In this chapter, a fundamental investigation between voice quality and network impairments (e.g. packet loss location, packet loss burstiness, packet size, delay) and relevant parameters with speech (e.g. codec type, voice/unvoiced, gender and language) has been undertaken using objective measurement (e.g. PESQ) based on a VoIP simulation system. Four modern codecs (i.e. G.729, G.723.1, AMR and iLBC) which are commonly used in VoIP and in emerging applications are used in the study. The results show that the packet loss burstiness, loss locations/patterns, packet size, delay, codec, gender of talkers and language all have an impact on perceived speech quality. This helps to determine parameters for developing statistical or neural network models for voice quality prediction, non-intrusively, which are investigated in Chapter 5 and 6, respectively.

Chapter 5

Regression-based Models for Non-intrusive Speech Quality Prediction

5.1 Introduction

In Chapter 3, different intrusive and non-intrusive speech quality measurement methods have been analyzed and compared. Intrusive methods (e.g. PESQ) can provide an accurate measurement of end-to-end speech quality, but cannot be used for monitoring live traffic because of the injection of a reference signal and the utilizing of the network. Non-intrusive methods (e.g. E-model or neural network models) are appropriate for monitoring/prediction of voice quality for live traffic. However, the methods are based on subjective tests to derive relevant parameters for E-model [95] or for neural network models' training. As subjective tests are time-consuming, slowly, and expensive, as a result, the current E-model is only applicable to a restricted number of codecs and network conditions. The training sets for neural network models are limited and cannot cover all the possible scenarios in dynamic and evolving networks, such as the Internet. Research has been carried out on instrumental derivation of equipment impairment factors [77]. But the work is still limited to consider purely codec (in single or in tandem operation).

Considering the efficiency/accuracy of the intrusive methods and the applicability of the non-intrusive methods for monitoring/prediction of voice quality for live traffic, a new non-

intrusive perceived speech quality prediction methodology is proposed and presented in this chapter. The novelty of the method is the exploitation of the latest intrusive ITU algorithm, PESQ, and the use of a combined PESQ/E-model structure, to provide a perceptually accurate prediction of voice quality, non-intrusively. This avoids time-consuming subjective tests and so removes one of the major obstacles in the development of models for voice quality prediction non-intrusively.

Based on the new methodology, efficient non-linear regression models are developed to predict conversational voice quality non-intrusively for a variety of codecs in this chapter. This easily extends the current E-model to new codecs and new network conditions as the method is based on intrusive objective method (e.g. PESQ) instead of time-consuming subjective tests. The method is also more efficient than E-model, as it provides a direct non-linear regression function between voice quality and network impairments and avoids a set of complex statistical models/functions as in current E-model. The artificial neural network models for predicting both listening and conversational voice quality based on the new methodology will be presented in Chapter 6.

The structure of the Chapter is as follows. A novel non-intrusive voice quality prediction methodology is presented in Section 5.2. The system structure of non-linear regression-based models for predicting voice quality is depicted in Section 5.3. The procedures to derive the regression-based models for voice quality prediction using AMR codec as an example is introduced in Section 5.4. The non-linear regression models for other codecs (e.g. G.729, G.723.1 and iLBC) are presented in Section 5.5. Section 5.6 summarises the Chapter.

5.2 Novel Non-intrusive Prediction of Voice Quality

5.2.1 A Novel Scheme for Non-intrusive Voice Quality Prediction

Figure 5.1 depicts a simplified, conceptual diagram of the proposed novel scheme for non-intrusive prediction of voice quality in IP networks. The Non-linear regression or ANN-based learning models are used to predict end-to-end, conversational voice quality (Predicted MOSc), non-intrusively, from network parameters (e.g. packet loss and delay) and non-network parameters (e.g. codec type and gender of speaker). In practice, IP packets transporting voice data through the network are captured at a monitoring point which may be at any suitable location (e.g. at the gateway) [96, 52]. Network parameters (e.g. delay, packet loss and jitter) and other relevant parameters (e.g. codec type and gender) are then extracted from analysis of the headers (e.g. RTP headers) or voice payload if necessary. The parameters are then applied to the learning model or regression model to provide a prediction of voice quality.

The emphasis in this thesis is on the new techniques that underlie the regression or learning model which is at the heart of the scheme. A novelty of the scheme is the use of a combined PESQ/E-model to provide an objective measure of conversational voice quality (Measured MOSc) which is then used to generate appropriate data for curve fitting (for regression models) or for neural network training (see next chapter for details). Another important novelty is that the scheme exploits the latest intrusive ITU-T algorithm, PESQ, to provide a perceptually accurate prediction of voice quality, non-intrusively. This avoids time-consuming subjective tests and removes one of the major obstacles in the development of models for non-intrusive monitoring and prediction of voice quality.

The benefits of the new method for non-intrusive applications include that

- It is generic and based on end-to-end, intrusive measurement of speech quality (in this case, using PESQ). Thus, it can be easily applied to other applications, such as audio (e.g. using ITU-T Perceptual Evaluation of Audio Quality (PEAQ) [86]), image (e.g. using a

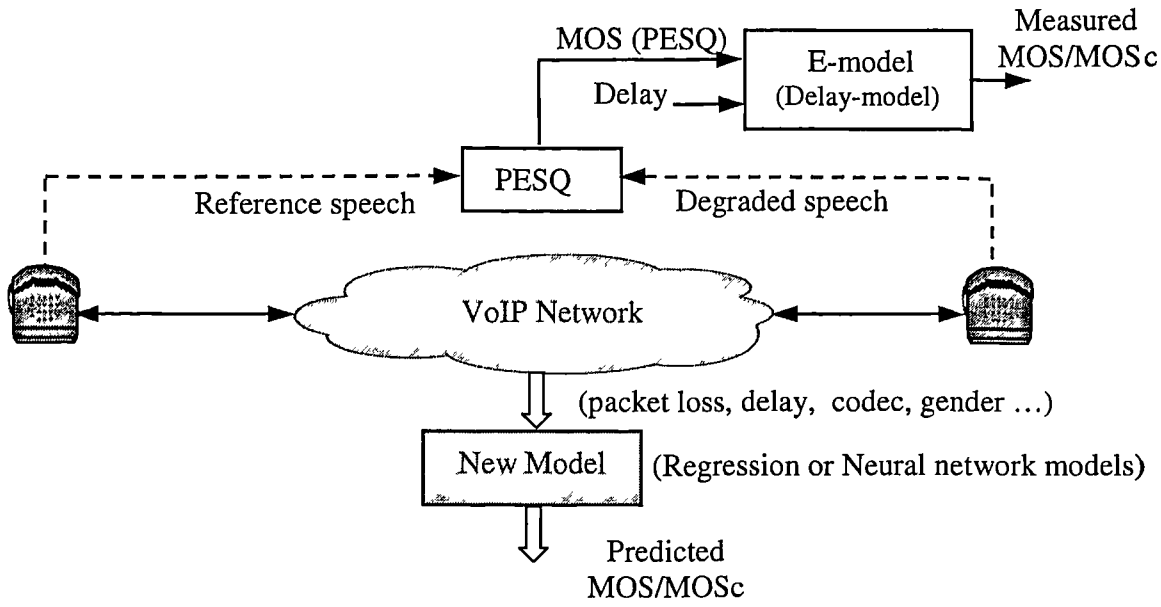


Figure 5.1: Conceptual diagram of the new scheme for non-intrusive prediction of voice quality

universal image quality index [97]) and video (e.g. using Video Quality Metric, VQM, the latest ANSI objective video quality standard T1.801.03-2003 [98]), provided the neural networks models are retrained (new features may need to be added) or regression models are re-derived.

- It avoids expensive and time-consuming subjective tests.

- It can be easily applied to new voice codecs (e.g. over 4.8Kb/s [4]), new packet loss conditions (e.g. new packet loss burstiness patterns) or different speakers/languages.

- For neural network-based models, it has learning ability and so can adapt to changing network conditions.

The non-linear regression and neural network models based on the new methodology are generic and as such have wide applicability. For example, the models can be used:

- for objective, non-intrusive, prediction or monitoring of end-to-end voice quality on live network, and to study error profile and IP network readiness for VoIP services.

- to optimize the quality of voice services in accordance with changing network conditions and to control the QoS and manage the utilization of available resources.

5.2.2 Prediction of Conversational Voice Quality

A key feature of the new scheme depicted in Figure 5.1 is that it avoids time-consuming subjective tests. This is made possible by the exploitation of PESQ and E-model to provide an objective measure of conversational quality which is used to derive regression models or to generate training data for neural networks training.

Figure 5.2 illustrates how a measure of conversational voice quality is obtained using a combined PESQ/E-model structure. PESQ is an accurate and reliable method for speech quality measurement, but it is an intrusive method and can only predict one-way listening quality. It does not consider the impact of end-to-end delay which is important for interactivity in voice communications. The approach in Figure 5.2 exploits the accuracy of PESQ and the delay model of the E-model.

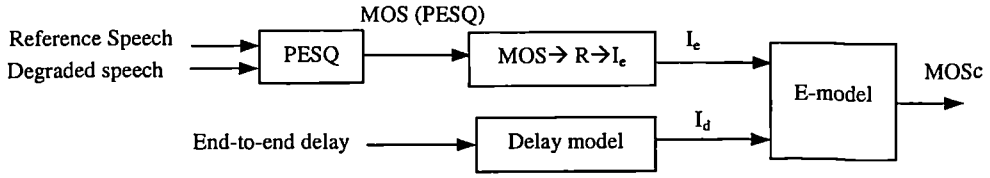


Figure 5.2: Measurement of conversational voice quality using a combined PESQ/delay model

As shown in the figure, the listening MOS score is obtained directly from the PESQ algorithm by comparing the reference and the degraded speech samples. The MOS is converted to a rating factor (the R factor) [7] and then to an equipment impairment value (I_e). The conversational MOS scores, MOSc, is obtained by combining the I_e value and the effects of end-to-end delay (the I_d value). The procedures to derive MOSc are as follows:

(1). Convert voice quality from MOS(PESQ) to I_e

The ITU-T G.107 [7] defines the relationships between the R to MOS as in Equation 5.1

and shown in Figure 5.3.

$$MOS = \begin{cases} 1 & \text{for } R \leq 0 \\ 1 + 0.035R + R(R-60)(100-R)7 \times 10^{-6} & \text{for } 0 < R < 100 \\ 4.5 & \text{for } R \geq 100 \end{cases} \tag{5.1}$$

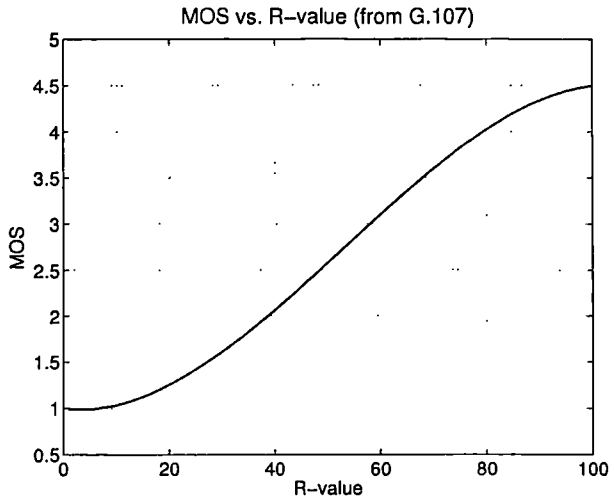


Figure 5.3: MOS vs. $R-$value for G.107

However, Equation 5.1 cannot be inverted directly to obtain the R values because it covers the R values between 0 and 6.5, which maps to MOS scores below 1. Thus, the R values are normally restricted to the range [6.5, 100] , with R values below 6.5 assigned as MOS $= 1$ before inversion. Candono's Formula [99] can be used to obtain the R-values from the MOS score as shown in Equation 5.2.

$$R = \frac{20}{3}\left(8 - \sqrt{226}\cos\left(h + \frac{\pi}{3}\right)\right) \tag{5.2a}$$

with

$$h = \frac{1}{3}\arctan 2\left(18566 - 6750MOS, 15\sqrt{-903522 + 1113960MOS - 202500MOS^2}\right) \tag{5.2b}$$

As Equation 5.2 are very complicated, a simplified 3^{rd} order polynomial fitting is used here to obtain the equation for mapping from MOS to R values (see Equation 5.3). The fitting curve and original curve from G.107 are both shown in Figure 5.4.

$$R = 3.026MOS^3 - 25.314MOS^2 + 87.060MOS - 57.336 \tag{5.3}$$

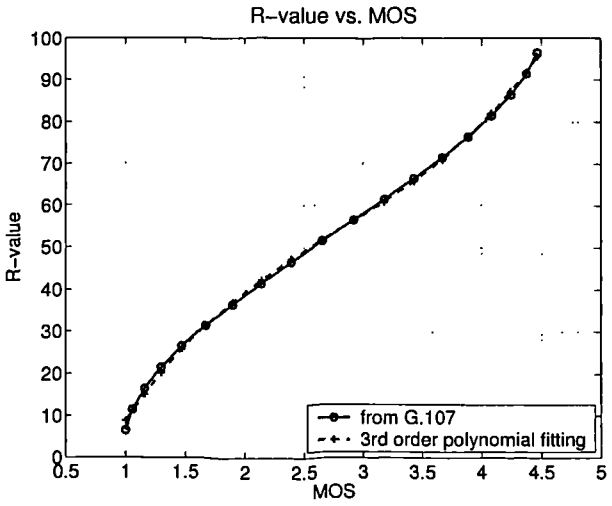


Figure 5.4: R value vs. MOS

If we consider only the equipment impairment, R can be converted to I_e as in Equation 5.4.

$$I_e = R_0 - R \tag{5.4}$$

The default value for R_0 is 93.2 [7].

(2). Obtain I_d from one-way delay d

The delay impairment factor, I_d, represents all impairments due to delay of voice signals, and includes impairments due to Listener Echo, Talker Echo and Absolute delay as shown in

Equation 5.5 [7]:

$$I_d = I_{dte} + I_{dle} + I_{dd} \tag{5.5}$$

The factor I_{dte} gives an estimate for the impairments due to Talker Echo. The factor I_{dle} represents impairments due to Listener Echo and the factor I_{dd} represents the impairment caused by too-long absolute one-way mouth-to-ear delay, T_a. The delay related with Listener Echo is T_r, the average, round trip delay in the four-wire loop and the delay related with Talker Echo is T, the average, one-way delay from the receive side to the point in the end-to-end path where a signal coupling occurs as a source of echo.

Following the same assumption for IP-based transport and VoIP application [9], the one-way delay d can be expressed as:

$$d = T_a = T = T_r/2 \tag{5.6}$$

Now I_d can be expressed as a function of one-way delay d. Assuming only the default values listed in G.107 [7] are used, the relationship of I_d versus one-way delay (d) from G.107 is shown in Figure 5.5 (curve from G.107). The details for the complicated computation equations to obtain I_d can be found in G.107 [7] (totally about 10 equations are used there). I_d can also be calculated using a more simplified equation (Equation 5.7) by curve fitting as provided in [9]. The curve from the simplified model is also shown in Figure 5.5.

$$\begin{aligned} I_d &= 0.024d + 0.11(d - 177.3)H(d - 177.3) \\ where &\begin{cases} H(x) = 0 \text{ if } x < 0 \\ H(x) = 1 \text{ if } x \geq 0 \end{cases} \end{aligned} \tag{5.7}$$

In order for a more accurate fit to the curve from G.107, a 6^{th} degree polynomial fit function

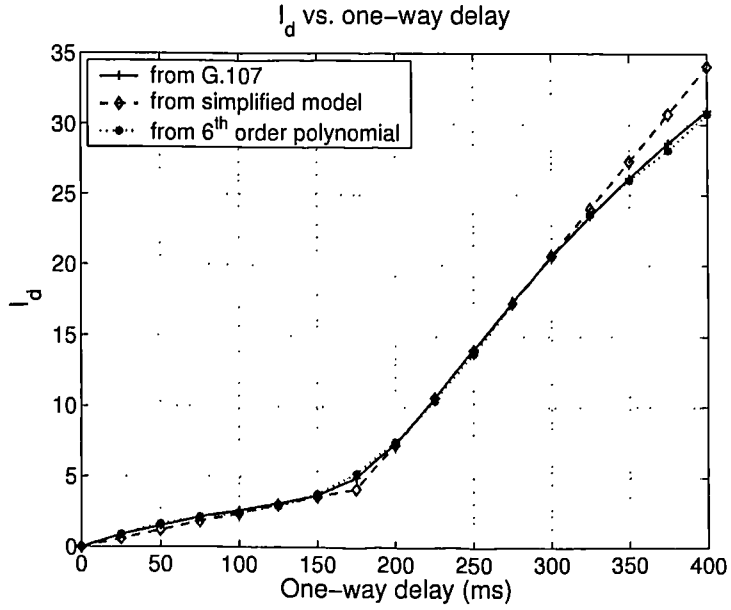


Figure 5.5: I_d vs. Delay

is provided as Equation 5.8 and the curve is also shown in the Figure 5.5.

$$I_d = 1.618{\cdot}10^{-13}d^6 - 1.765{\cdot}10^{-10}d^5 + 6.447{\cdot}10^{-8}d^4 - 8.221{\cdot}10^{-6}d^3 + 0.0002315d^2 + 0.0352d - 0.02434 \tag{5.8}$$

Depending on the application, Equation 5.8 (more accurate) or Equation 5.7 (more simple) can be used to represent the delay impairment I_d under one-way delay d.

(3). Obtain MOSc from I_d and I_e

Considering I_d and I_e, E-model R factor can be simplified as Equation 5.9.

$$R = R_0 - I_d - I_e \tag{5.9}$$

From R, the conversational MOS score (MOSc) can be calculated using Equation 5.1. Overall, the conversational MOS score (Measured MOSc score in Figure 5.1) can be obtained from PESQ MOS score and end-to-end delay d.

5.3 System Structure of Regression-based Models

Figure 5.6 illustrates how to use PESQ and E-model to derive regression-based models for voice quality prediction in VoIP applications. Information about the codec, packet loss rate and delay is suitably transformed by the I_e and I_d models and then processed by the E-model to produce a MOS value (Figure 5.6(a)). The MOS value is a prediction of what the perceived voice quality would be under these conditions. The I_e model is codec dependent and can be derived from PESQ [4] (and also the new PESQ-LQ [71]) as shown in Figure 5.6(b). This avoids time-consuming subjective tests.

In Figure 5.6(b), the reference speech files are first encoded and then processed in accordance with the network impairments parameter values and then decoded to generate the degraded speech. The degraded speech and the reference speech are then processed by PESQ (or PESQ-LQ) to provide a MOS value, which is a measure of voice quality. The MOS values can then be suitably transformed to give measured I_e values. As shown later, given a set of measured I_e values for a codec we can then derive an I_e model for the codec using regression techniques without the need for subjective tests.

We will illustrate this for four modern codecs which are relevant for VoIP - G.729 (8 Kb/s), G.723.1 (6.3 Kb/s), AMR (the highest mode, 12.2 Kb/s and the lowest, 4.75 Kb/s) and iLBC (15.2 Kb/s). In the study, the reference speech database was taken from the ITU-T data set [88]. Packet loss was generated from 0% to 30%, in an incremental step of 3% and Bernoulli loss model was used for simplicity. PESQ-LQ (Listening Quality), the latest improvements on PESQ algorithm, is also included for comparison.

5.4 Procedures for Regression-based Models

As an example, we first derived the I_e value for a new codec for VoIP applications using PESQ (the AMR at the highest mode of 12.2 Kb/s. I_e model does not exists for AMR codecs at present in public domain). The procedures are as follows:

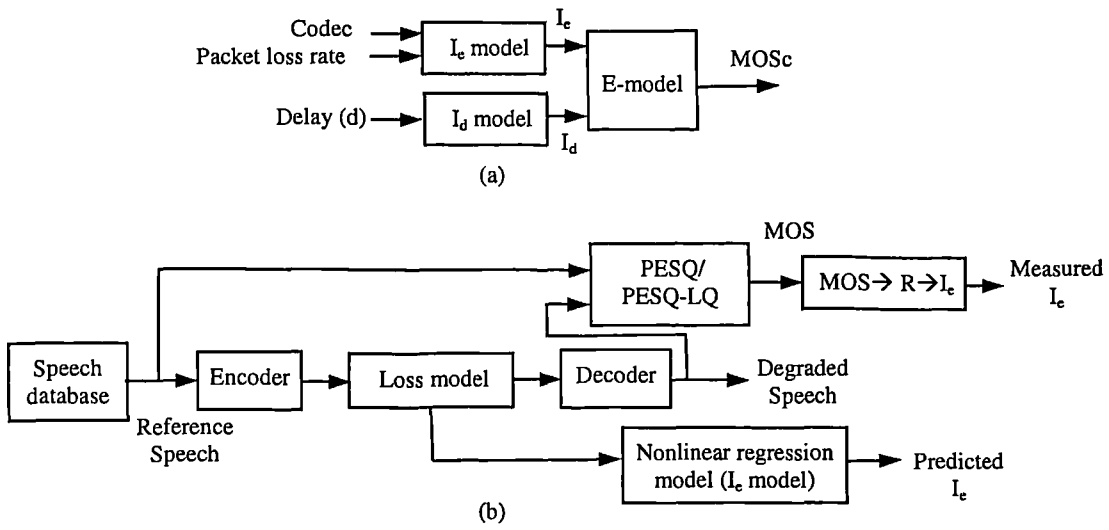


Figure 5.6: (a) An illustration of how to predict voice quality using the E-model, (b) Prediction of I_e model using the PESQ.

Step 1: Obtain MOS (PESQ) vs. packet loss rate for the AMR codec.

For each speech sample in the ITU-T data set for British English, a MOS (PESQ) score is obtained by averaging over 30 different packet loss locations (via different random seed setting) in order to remove the influence of packet loss location. Further, the MOS score for one packet loss rate is obtained by averaging over all speech samples (a total of 16 samples, consisting of 8 males and 8 females), so that the influence of gender is removed (We did not consider the gender issue for regression-based models only for simplicity). The relationships between the average MOS and packet loss rate for AMR codec are shown in Figure 5.7.

Step 2: Convert the MOS vs. packet loss rate to I_e vs. packet loss rate

The relationship between the MOS vs. packet loss rate in Figure 5.7 can now be converted to the Equipment impairment I_e (measured I_e in Figure 5.6(b)) vs. packet loss rate via Equations 5.3 and 5.4. The derived curves for I_e versus packet loss rate ρ are shown in Figure 5.8 (the curve from PESQ). A logarithm fitting function, similar as in [9]), can be derived as Equation 5.10 by curve fitting (e.g. by using Matlab's Curve Fitting Tool). The fitting curve is also

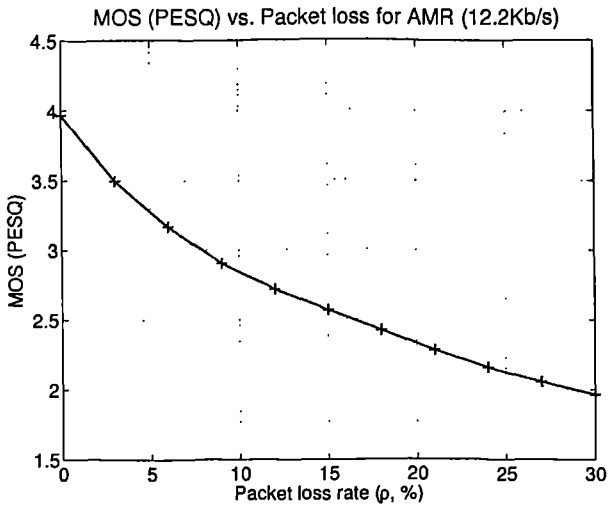


Figure 5.7: MOS vs. packet loss rate ρ for AMR codec

shown in Figure 5.8 (the one from fitting).

$$I_e = 16.68 \ln(1 + 0.3011\rho) + 14.96 \tag{5.10}$$

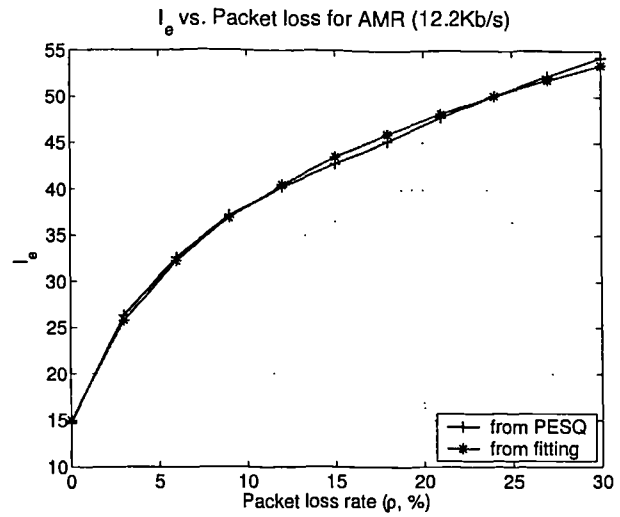


Figure 5.8: I_e vs. packet loss rate ρ for AMR codec

The goodness of the fit is: SSE=2.83, R^2=0.998, and $RMSE$=0.5947.

Step 3: Calculate the MOS for AMR codec (12.2 Kb/s mode)

Considering I_d (Equation 5.8 or Equation 5.7) and I_e (Equation 5.10), E-model's R factor

can be simplified as Equation 5.11.

$$R = R_0 - I_d - I_e \tag{5.11}$$

The conversational MOS score (MOSc) can be calculated from R using Equation 5.1 for a given random packet loss rate and end-to-end delay. The MOS vs. packet loss rate and delay (using 6^{th} order polynomial or simplified delay model) are shown in Figure 5.9 and 5.10, respectively. It can be seen that the relationship of MOS vs. loss rate and delay are non-linear.

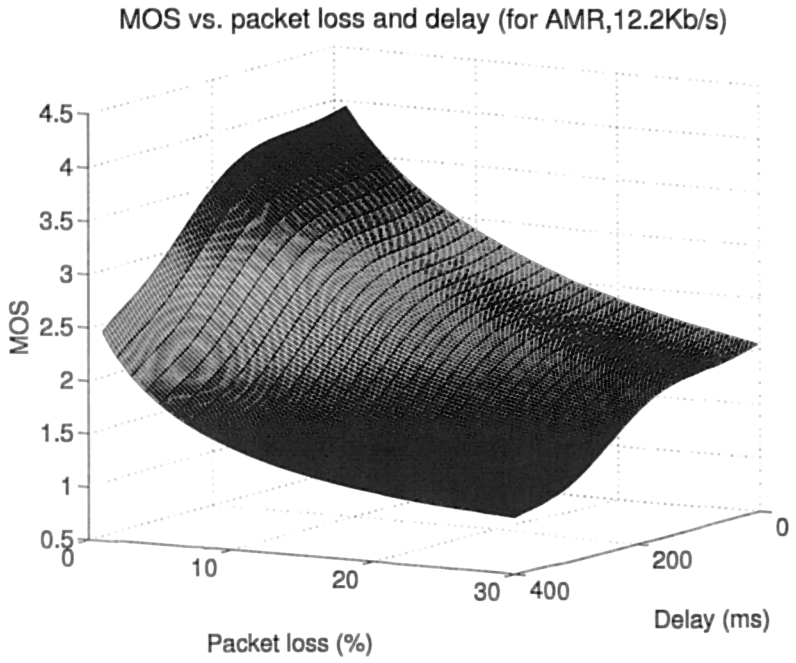


Figure 5.9: MOS vs. packet loss and delay (using 6^{th} order polynomial model)

Overall, by using the model for I_d (Equation 5.8 or Equation 5.7) and the model for I_e (Equation 5.10), voice quality can be predicted using the E-model as shown in Figure 5.6(a).

The MOS vs. packet loss and delay (from the simplified delay model, Equation 5.7) as shown in Figure 5.10 is used for calculating MOSc value from packet loss and delay for other codecs and other applications in the thesis.

The MOS vs. packet loss and delay (from 6^{th} order polynomial model, Equation 5.8) as shown in Figure 5.9 is used to derive a direct nonlinear regression model from packet loss and

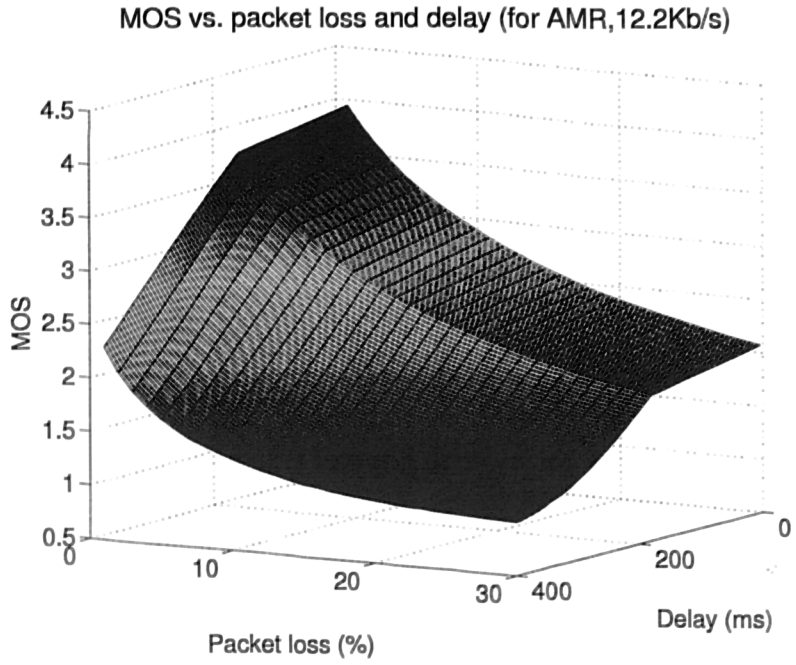


Figure 5.10: MOS vs. packet loss and delay (using simplified model)

delay to MOSc which is described in Step 4. This is more accurate than using simplified delay model (see Figure 5.5) and the surface is more smooth and easier for nonlinear fitting.

Step 4 (Optional): Surface fitting for nonlinear mapping from packet loss and delay to MOSc

From Figure 5.9, a nonlinear regression surface fitting can be conducted to obtain the non-linear function from packet loss, delay to MOSc.

For example, a simplified cubic polynomial function can be used for a surface fitting. The equation derived is as below:

$$MOS = 3.797 - 0.164\rho + 2.689 \cdot 10^{-3} d + 6.322 \cdot 10^{-3} \rho^2 - 2.851 \cdot 10^{-5} d^2 - 9.579 \cdot 10^{-5} \rho^3 + 3.446 \cdot 10^{-8} d^3 \quad (5.12)$$

The error surface of MOS value fitting is depicted in Figure 5.11. The absolute error is within ± 0.2 of MOS scale. The Standard Error of Mean is $1.24 \cdot 10^{-3}$.

The accuracy of the fitting can be improved when more complex non-linear regression models are used.

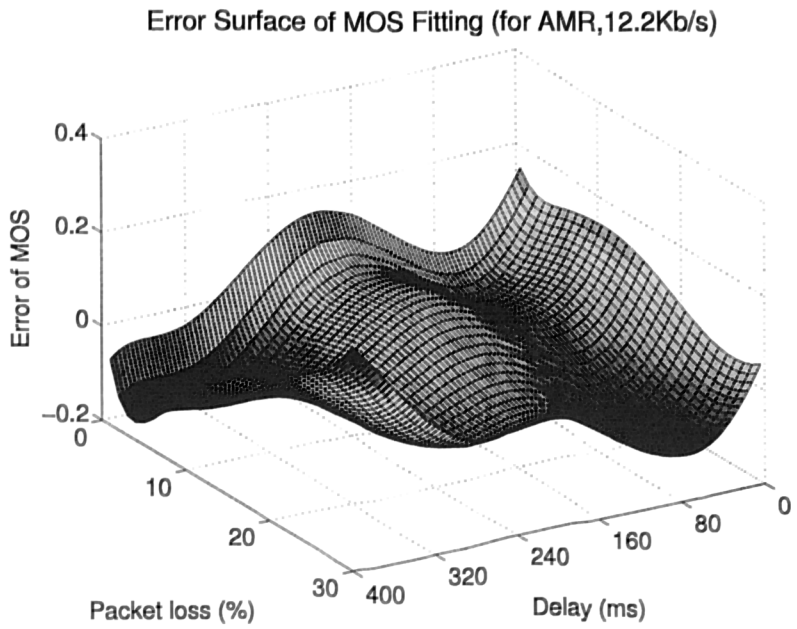


Figure 5.11: Error surface for MOS fitting for AMR (12.2Kb/s)

In the thesis, the work is focused on obtaining MOSc from packet loss rate (ρ) and end-to-end delay (d) using Step 1 to 3 as described above. This method can be extended to other speech codecs, emerging or new ones (above 4.8Kb/s [4]) and packet loss patterns (e.g. burst packet loss) which will be described in the next section.

5.5 Non-linear Regression Models for Different Codecs

Following the procedures in Section 5.4, we extended the non-linear regression models for other codecs, i.e. AMR(L, 4.75 Kb/s), G.729 (8 Kb/s), G.723.1 (6.3 Kb/s) and iLBC (15.2 Kb/s). The results for AMR(H, 12.2 Kb/s) is also included for comparison. Parameters for PESQ-LQ is also derived together with those for PESQ algorithm.

For each speech sample in the ITU-T data set for British English, a MOS (PESQ or PESQ-LQ) score is obtained by averaging over 30 different packet loss locations (via different random seed setting) in order to remove the influence of packet loss location. Further, the MOS score for one packet loss rate is obtained by averaging over all speech samples (a total of 16 sam-

ples, consisting of 8 males and 8 females), so that the influence of gender is removed (We did not consider the gender issue for regression-based models only for simplicity). The relationships between the average MOS and packet loss rate for each of the four codecs are shown in Figure 5.12.

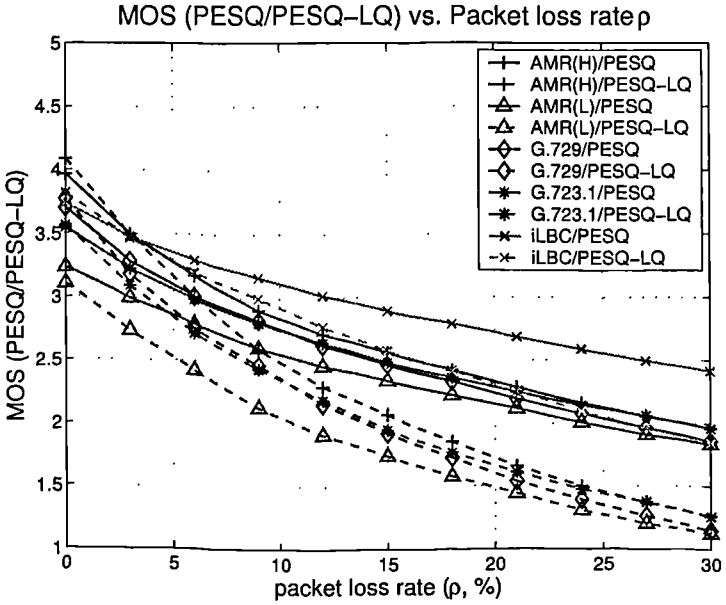


Figure 5.12: MOS vs. packet loss rate ρ

From Figure 5.12, it can be seen that PESQ-LQ has a much lower MOS score when the packet loss rate is high. iLBC shows the best voice quality when packet loss rate is high (over 4%). AMR (H, 12.2 Kb/s) has the highest MOS score when packet loss rate is zero. AMR (L, 4.75 Kb/s) has the lowest quality no matter with or without loss.

The relationship between the MOS vs. packet loss rate in Figure 5.12 can now be converted to the Equipment impairment I_e (measured I_e in Figure 5.6(b)) vs. packet loss rate via Equations 5.3 and 5.4. The derived curves for I_e versus packet loss rate ρ are shown in Figure 5.13.

5.5.1 Obtain Non-linear Regression Models for Different Codecs

From Figure 5.13, a non-linear regression model (similar to the logarithm fitting function in [9]) can be derived for each codec based on the PESQ or PESQ-LQ data by the least squares

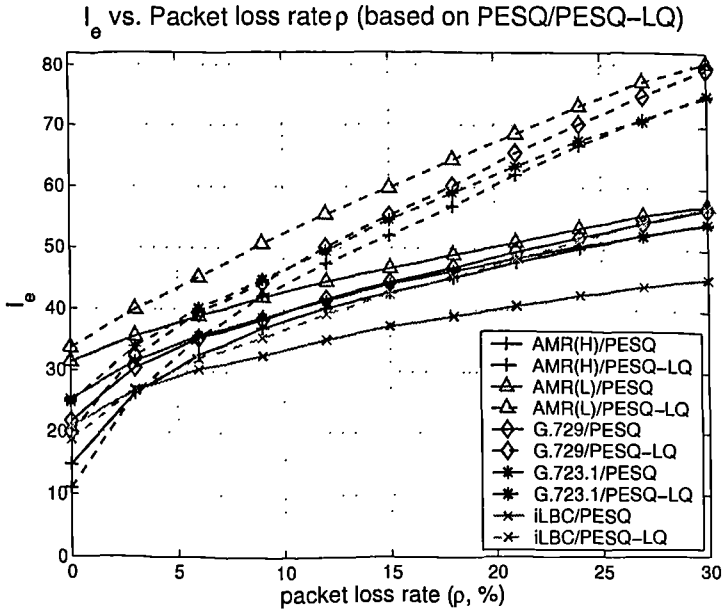


Figure 5.13: I_e vs. packet loss rate ρ

method and curve fitting. The derived I_e model has the following form:

$$I_e = a \ln(1 + b\rho) + c \tag{5.13}$$

where ρ is the packet loss rate in percentage. The parameters (a, b and c) for different codecs under PESQ and PESQ-LQ are shown in Tables 5.1 and 5.2, respectively.

Table 5.1: Parameters of regression models for different codecs (PESQ)

Parameters	AMR (H)	AMR (L)	G.729	G.723.1	iLBC
a	16.68	30.86	21.14	20.06	12.59
b*100	30.11	4.26	12.73	10.24	9.45
c	14.96	31.66	22.45	25.63	20.42

Table 5.2: Parameters of regression models for different codecs (PESQ-LQ)

Parameters	AMR (H)	AMR (L)	G.729	G.723.1	iLBC
a	40.0	93.66	63.20	60.09	31.72
b*100	12.11	2.16	4.84	4.17	7.22
c	12.2	33.82	21.71	25.79	19.65

These models can be combined with the delay model (e.g. Equation 5.7) to obtain the

predicted MOSc score as described in Section 5.4. The method and the derived models shown in Tables 5.1 and 5.2 can be readily extended to other codecs or network conditions (e.g. burst packet loss).

5.5.2 Equipment Impairments with Packet Loss

In some applications (e.g. perceived buffer optimization which will be described in Chapter 8), it is important to know the relationship of equipment impairment (e.g. I_e) with packet loss, or to know the robustiness of codec with packet loss. In this section, we derive the relationship of I_e with packet loss after removing the impairment from codec itself.

In Figure 5.13, the I_e value for zero packet loss rate represents the impairment for the codec itself. The AMR (4.75 Kb/s) has the largest codec impairment (the largest I_e value or the lowest MOS score at zero packet loss), the AMR (12.2 Kb/s) has the lowest I_e value (or highest MOS score). G.729 and iLBC codecs have similar MOS values at zero packet loss rate, but iLBC has the best overall MOS values of all four codecs at high packet loss rates (over 3%).

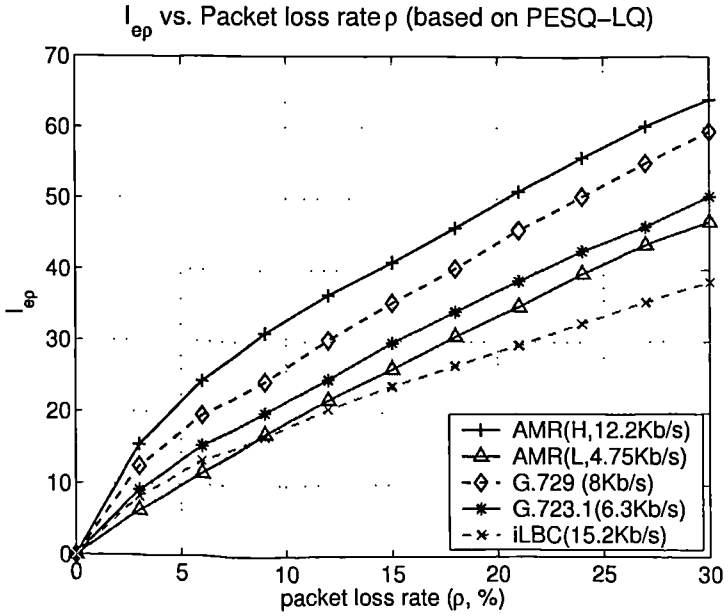


Figure 5.14: I_{ep} vs. packet loss rate p

Considering that the effect of codec impairment (without loss) is fixed for any codec, I_e can

be viewed as consisting of two main components: $I_e = I_{ec} + I_{ep}$, where I_{ec} is the impairment without loss and I_{ep} the impairment with loss. The I_{ep} vs. packet loss rate for PESQ-LQ is shown in Figure 5.14.

Figure 5.14 illustrates the ability of a codec to cope with network packet loss. From the curves, the iLBC has the lowest slope, whereas, the AMR (H, 12.2 Kb/s) has the highest. This further shows that the iLBC has an obvious high robustness to packet loss as claimed in iLBC project's website (a speech codec suitable for robust voice communication over IP) [100] . AMR (12.2 Kb/s) has the highest MOS score under zero packet loss condition (as shown in Figure 5.12), but it has the least ability to cope with packet loss (quality decreases sharply as packet loss increases). The G.723.1, G.729 and AMR (4.75 Kb/s) have similar ability coping with packet loss, with the curves in between.

From Figure 5.14, it is clear that to use only one curve (or model) as suggested in [101] to represent all codecs is inappropriate. Obviously with emerging new network codecs (with even higher robustness to loss), the diversity in the ability of codecs to cope with packet loss will be even larger. Thus, we recommend to use different models for each codec in VoIP applications for accurate parameter optimization or quality monitoring/control purposes. In Section 8.4, we will show using I_{ep} vs. packet loss rate to derive a minimum impairment criterion for buffer algorithm optimization.

5.6 Summary

In the chapter, a novel non-intrusive voice quality prediction methodology has been presented. The novelty of the method is that it explores the latest intrusive measurement (e.g. PESQ) and a combination structure of PESQ/E-model for non-intrusive voice quality prediction. It can be efficiently used to develop statistical and neural network based models for voice quality prediction. The statistical models (i.e. the non-linear regression models) have been presented in this chapter and neural network models will be discussed in Chapter 6.

In the chapter, the regression models for a variety of codecs (i.e. G.729, G.723.1, AMR and iLBC) for speech quality prediction for VoIP networks have been developed. The system structure and detailed procedures to derive the model are given. These models can be easily and efficiently used for voice quality prediction which will be discussed in Chapter 7 and voice quality optimization (e.g. buffer algorithm optimization) which will be described in Chapter 8.

Chapter 6

Neural Network-based Models for Non-intrusive Speech Quality Prediction

6.1 Introduction

In Chapter 5, the proposed new non-intrusive voice quality prediction methodology has been presented and detailed efficient non-linear regression models for predicting voice quality based on the new methodology are given. The non-linear regression models have greatly extended the current E-model for new applications (e.g. new codecs and new network conditions) and can be easily used for voice quality monitoring/prediction or for voice quality optimization. However, the regression models, like the E-model, are inconvenient (each model exists for each codec, packet size, random or burst packet loss) and static (it cannot adapt to the changing network conditions such as the Internet). This makes artificial neural network models more attractive for predicting voice quality non-intrusively, as neural network models have the learning ability which can adapt to the changing network conditions.

Neural networks-based models have recently been used to predict speech quality from IP network parameters [11, 12], but these rely on subjective tests to create the training sets. Unfortunately, subjective tests are costly and time-consuming and as a result the training sets are limited and cannot cover all the possible scenarios in dynamic and evolving networks, such as the Internet. In addition, the development of previous neural networks-based models was based

on a limited number of codecs and can only predict one-way listening voice quality [11]. There is a need for models to predict conversational quality to account for interactivity. Little attention has also been paid to talker dependency.

The work on neural network modelling for predicting both listening and conversational voice quality based on the new non-intrusive voice quality prediction methodology are presented in the chapter. A key novelty for neural network models is that the models have learning capability, a new concept in QoS monitoring, and this makes it possible for the models to learn the non-linear relationships between voice quality and impairment parameters and to adapt to changes in the network. Another novelty is the exploitation of the latest intrusive ITU algorithm for perceptual evaluation of speech quality to provide a perceptually accurate prediction of voice quality, non-intrusively.

The structure of the chapter is as follows. In Section 6.2, the neural network models for predicting listening voice quality (PESQ MOS score) is presented. In Section 6.3, the neural network models for predicting conversational voice quality (MOSc score) is discussed. Section 6.4 summarises the chapter.

6.2 Neural Network Models to Predict Listening Voice Quality

6.2.1 Simulation System Structure

A block diagram of the speech quality prediction system that was used in the study is depicted in Figure 6.1. It is a PC-based software system that allows the simulation of key processes in voice over IP. It enables the simulation of a variety of network conditions and objective measurement of the effects on perceived speech quality. The system includes a speech database, an encoder/decoder, a packet loss simulator, a speech quality measurement module, a parameter extraction and an ANN model. The speech database is taken from the TIMIT data

set [92]. Speech files from different male and female talkers are chosen to generate a database for ANN model development

Three modern codecs were chosen for the study. These are G.729 CS-ACELP (8 Kbps), G.723.1 MP-MLQ/ACELP (5.3/6.3 Kbps) and Adaptive Multi-Rate (AMR) codecs with eight modes (4.75 to 12.2 Kbps).

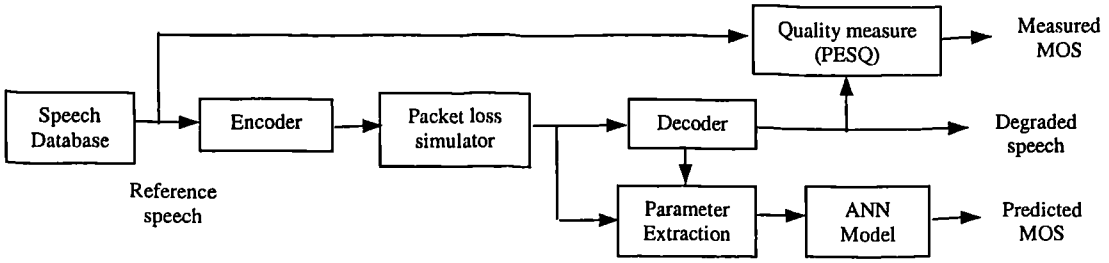


Figure 6.1: System structure for speech quality analysis and prediction

A 2-state Gilbert model was used to simulate packet loss (see Figure 2.6).

In our system, the latest ITU perceptual measurement algorithm, PESQ, is used to measure the perceived speech quality under different network conditions and for different talkers/languages. The PESQ compares the degraded speech with the reference speech and computes an objective MOS value in a 5-point scale. In the study, the MOS score obtained from the PESQ is referred to as the 'measured MOS' to differentiate it from the 'predicted MOS' obtained from the ANN model. The Parameter Extraction module is used to extract salient information from the IP network and the decoder (including the codec type and network packet loss). In real VoIP applications, codec type and packet loss would be parsed from the RTP header. After processing, the information is fed to the ANN model to predict speech quality.

As a network packet payload may include a normal speech frame (speech talkspurt) or a silence frame. The number of silence frames depends on whether VAD (Voice Activity Detection) is activated or not at encoder side. If VAD is activated, silence frame only represents SID (Silence Insertion Description) frame. Packet loss during silence period or small signal energy segment has no or very small impact on perceived speech quality.

Here we combined the information from decoder's VAD indicator and network packet loss,

and calculated the *ulp* and *clp* according to Gilbert model only during speech talkspurt. In this case, State 1 in Figure 2.6 represents loss during talkspurt, and State 0 represents no loss or loss during silence. We used *ulp(VAD)* and *clp(VAD)* to differentiate them from the simulated network *ulp* and *clp*. The benefit is that it can always count the packet loss which are perceptually relevant no matter whether or not VAD is used, or what kinds of VAD is used in the system. Another benefit is that the calculation of *ulp(VAD)* and *clp(VAD)* can be frame-based, which can include the impact of different packet size. It may save one input parameter for the neural network analysis. The frame size depends on codec used. It is 10 ms for G.729, 20 ms for AMR and 30 ms for G.723.1.

The pitch delay can be extracted from decoder and the gender can be decided according to a preset threshold for pitch delay between the male and female. In this stage of the research, we just set the gender value according to the speech file we chose.

6.2.2 Artificial Neural Network Model

An important objective of our study is to develop neural-networks based models to learn the non-linear relationships between the key impairment parameters and perceived voice quality. The use of learning models is necessary because the relationships are not explicit. Unlike conventional models which are static, e.g. the E-model, a neural networks based model can also be re-trained to learn new relationships for IP networks which are continually changing.

For simplicity, a three-layer, feed-forward neural net architecture and the standard back-propagation learning algorithm were used (see Figure 6.2). Four variables were identified as inputs to the neural network model, namely: codec type, gender, *ulp(VAD)* and *clp(VAD)*. The predicted MOS score was the only output(see Figure 6.3).

For a three-layer feed-forward neural net, the network is made up of the input layer, the hidden layer and the output layer. Input data is fed to the input layer and processing is done layer by layer up to the output layer. Activation function of a node controls the output signal from the node. To start with, a given set of randomized values of the weights and biases are

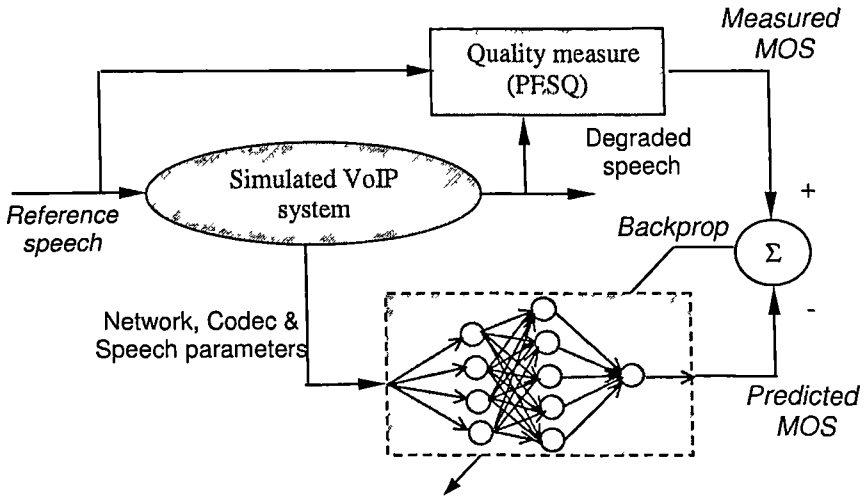


Figure 6.2: Conceptual diagram of the training process for neural network model (for listening quality prediction)

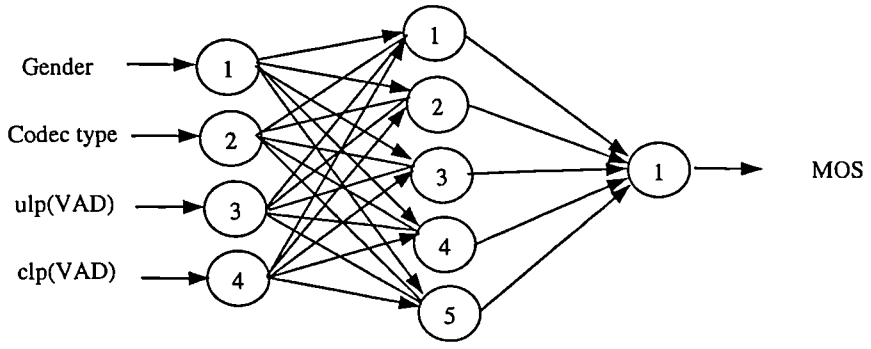


Figure 6.3: Schematic diagram of an artificial neural network

assigned to the network. The connection weights are then updated to decrease the difference (error) between the network output and the desired output using certain minimization algorithm. The process is repeated until the error falls below a specified limit. The neural net is then said to have been trained. Outputs of the hidden and output layers are generated using the asymmetric sigmoid activation function. Input and output values are scaled from 0 to 1 using the minimum and maximum values in the training data.

ulp(Real) and *clp(Real)* are generated from the Gilbert model and represent the contribution from packet loss. In this context, the Gilbert model serves as a means of pre-processing the received packet streams to capture and represent the underlying features of packet loss before it is applied to the neural networks to facilitate learning. It allows the packet loss behaviour of IP networks to be represented as a Markov process because several of the mechanisms that contribute to loss are transient in nature (e.g. network congestion, late arrival of packets at a gateway/terminal, buffer overflow or transmission errors), which is in fact why packet loss is bursty in nature [46]. An attraction is that it provides a compact representation of the loss behaviour of IP networks which can be used directly as inputs to the learning models.

The Stuttgart Neural Network Simulator (SNNS) package [102] was used for neural network training and testing. The neural network was trained to learn the non-linear relationship between four input variables and one output variable.

6.2.3 Neural Network Database Collection

In order to train and test the neural networks, a database was generated from two talkers (one male and one female) from the TIMIT database for the three codecs, G.729, G.723.1 (6.3Kb/s) and AMR (4.75 Kb/s). For dual-mode G.723.1 and eight-mode AMR, only one mode was chosen for simplicity. To enhance generalisation, the training data set was carefully designed to contain representative examples of key impairment conditions that may be encountered in real IP networks. The network unconditional loss probability (*ulp*) was set to 0, 10, 20, 30 and then to 40 % and conditional loss probability (*clp*) was set to 10, 50 and 90 % respectively

to simulate different bursty loss conditions. The packet size was set to 1 to 5 frames/packet for all three codecs. For each case, an initial seed was generated randomly to cater for a range of possible loss patterns. In order to compare the results from real network loss and talkspurt-based network loss, the real loss rate at the end of test sentence (*ulp(Real)/clp(Real)*) and loss rate during talkspurt (*ulp(VAD)/clp(VAD)*) were calculated at the same time. The difference between *ulp(Real)/clp(Real)* and *ulp/clp* is due to pseudo-random number generation and limited length of test sentences (10s). For each case, the speech quality (MOS score) between the reference and degraded speech file was calculated using PESQ algorithm. An example of the dataset was shown in Table 6.1 (only selected samples are shown).

Table 6.1: Variables used in ANN database generation

Codec type	Gender	*ulp (%)*	*clp (%)*	Packet size	*ulp(Real)*	*clp(Real)*	*ulp(VAD)*	*clp(VAD)*	MOS
1	0	10	90	1	0.14	0.90	0.10	0.86	2.9
1	0	20	90	1	0.16	0.90	0.14	0.89	2.6
1	0	40	90	2	0.41	0.95	0.38	0.94	1.5
2	1	20	50	1	0.18	0.47	0.07	0.52	2.1
3	0	40	90	3	0.31	0.96	0.27	0.96	2.2

Note:

- Codec type: G.729 = 1, G.723.1 = 2, AMR = 3.
- Gender: male = 0, female = 1
- Packet size: 1, 2, 3, 4 and 5

The calculated *ulp(VAD)*, *clp(VAD)*, the codec type, the gender, and the calculated MOS score formed one sample for neural network database as shown in Table 6.2. A total of 362 samples (patterns) were generated. 70% of the samples were chosen randomly as the training set and the remaining 30% as the testing set. All input and output variables are normalized to [0, 1] before neural network processing.

Table 6.2: An example of ANN database for MOS prediction

Codec type	Gender	*ulp(VAD)*	*clp(VAD)*	MOS
1	0	0.10	0.86	2.9
1	0	0.14	0.89	2.6
1	0	0.38	0.94	1.5
2	1	0.07	0.52	2.1
3	0	0.27	0.96	2.2

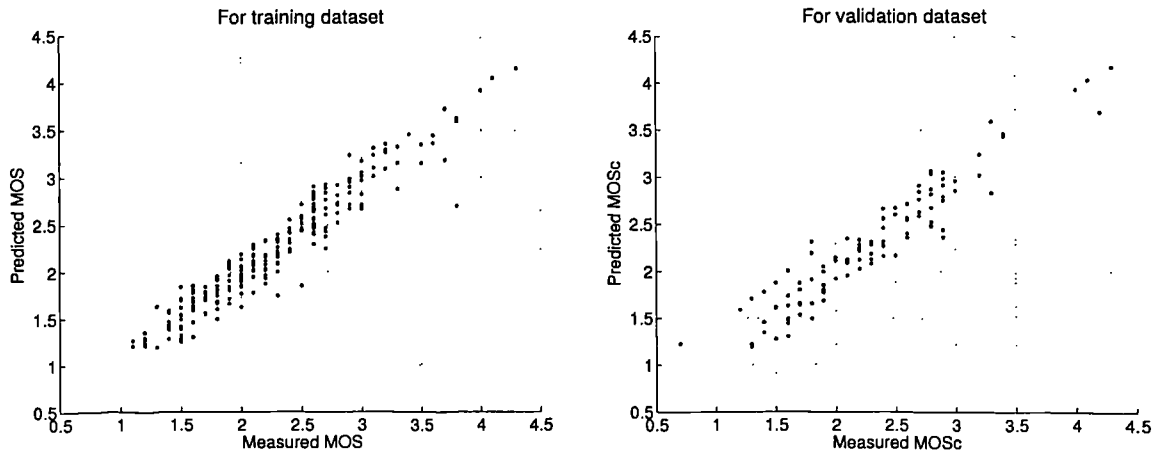
The neural network was used to learn the non-linear relationship between four input variables (e.g. *ulp(VAD)*, *clp(VAD)*, codec type, gender) and one output variable (MOS score) from the training set. Then, it was tested using the new (unseen) samples from the test set.

6.2.4 Speech Quality Prediction Results Analysis

Different network structures (e.g. the number of neurons in the hidden layer and the parameters of Standard Backpropagation learning algorithm [103]) were investigated to determine a suitable architecture for ANN model. Comparing the predicted MOS score from the ANN model and the measured MOS using PESQ algorithm, we obtained a maximum Correlation Coefficient (ρ) of 0.967 and an average error of 0.12 for the training set. For the testing set, ρ was 0.952 and the average error was 0.15. The learning rate (η) was 0.4 and the maximum difference (d_{max}) was 0.01 for a 4-5-1 net (see Figure 6.3. The scatter diagrams of the predicted versus the measured MOS scores for the training and validation data sets are illustrated in Figures 6.4(a) and 6.4(b). Increasing the number of neurons in the hidden layer did not improve the prediction accuracy. However, when *ulp(Real)/clp(Real)* was used instead of *ulp(VAD)/clp(VAD)*, the Correlation Coefficients for the training and testing datasets both dropped by 2-3 percent. This suggested that *ulp(VAD)/clp(VAD)* are better for speech quality prediction than *ulp(Real)/clp(Real)*. We also investigated the effect of including packet size as an input to the neural net (i.e. 5 inputs) and obtained similar results. This suggested that packet size might not be necessary as an input to the neural network.

As the training and testing data sets were from the same talkers, we further generated a vali-

date data set from another male and female talkers and set the different network loss conditions (*ulp:* 5, 15, 25, 35%, and *clp:* 30, 70%). Packet size was still set to 1 to 5. A total of 210 new patterns were generated and used to validate the trained net. We obtained ρ of 0.946 and an average error of 0.19. It suggested that the designed neural network model works well for speech quality prediction in general.



(a) For training data set (b) For validation data set

Figure 6.4: Predicted MOS vs. measured MOS for training and validation sets

The correlation coefficients obtained from the training, testing and validating datasets are between 0.946 to 0.967. It seems difficult to improve the performance further from neural network side. We think this is mainly due to the following two reasons. (1). *ulp(VAD)/clp(VAD)* is still not accurate enough to express perceptual relevant loss information for some loss patterns/locations; (2). Objective MOS scores from PESQ may not be as accurate as subjective MOS scores for some loss conditions. Our subjective test results in Section 10.2 have also confirmed that PESQ shows higher sensitivity than subjects in higher burstiness conditions, especially in the case of missing words, whereas, it shows lower sensitivity than subjects in lower burstiness cases for G.729.

6.3 Neural Network Models to Predict Conversational Voice Quality

The neural network models developed in the previous section can only predict one-way listening voice quality (PESQ MOS score). There is a need for models to predict conversational quality to account for interactivity. In this section, we extend the neural network models to predict conversational voice quality (expressed as MOSc score).

6.3.1 Artificial Neural Network Model

The procedure of neural network modelling for conversational voice quality prediction is similar to the one described in the previous section for listening quality.

For simplicity, a three-layer, feed-forward neural net architecture and the standard back-propagation learning algorithm were used (see Figure 6.5). The difference between the figure and Figure 6.2 is that conversational quality (MOSc) is used instead of listening quality (MOS). Further the input variables used for neural network models are different (with more variables for conversational quality prediction).

Six variables were identified as inputs to the neural network model, namely: delay, codec type, gender, packet size, *ulp(Real)* and *clp(Real)*. The predicted MOSc score was the only output. *ulp(Real)* and *clp(Real)* are generated from the Gilbert model and represent the contribution from packet loss.

The Stuttgart Neural Network Simulator (SNNS) package [102] was used for neural network training and testing. The neural network was trained to learn the non-linear relationship between six input variables and one output variable.

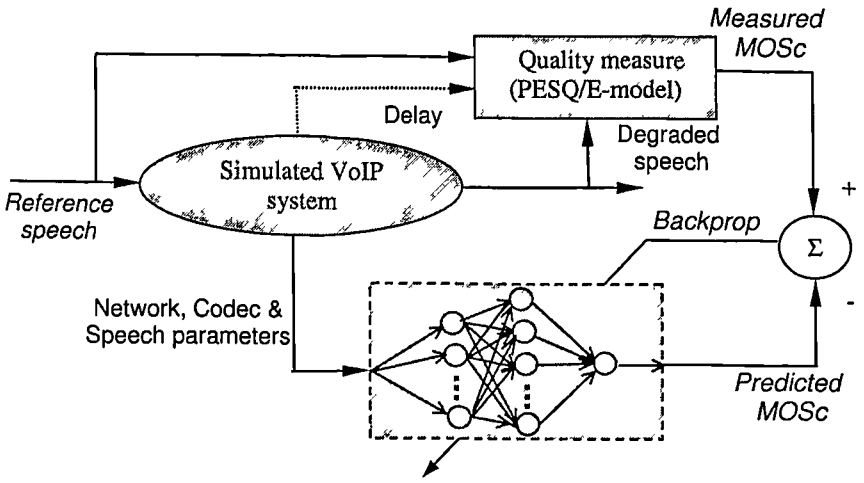


Figure 6.5: Conceptual diagram of the training process for neural network model (for conversational quality prediction)

6.3.2 Training, Validation and Test Database

In order to train and test the neural networks, a database was generated from two talkers (one male and one female) from the ITU-T database for the four codecs, G.729, G.723.1 (6.3Kb/s), AMR (4.75 Kb/s) and iLBC (15.2Kb/s). For dual-mode G.723.1, iLBC and eight-mode AMR, only one mode was chosen for simplicity. To enhance generalisation, the training data set was carefully designed to contain representative examples of key impairment conditions that may be encountered in real IP networks. The network unconditional loss probability (*ulp*) was set to 0, 10, 20, 30 and then to 40 % and conditional loss probability (*clp*) was set to 0, 20, 50 and 80 % respectively to simulate different bursty loss conditions. The packet size was set to 1 to 3 frames/packet for all four codecs. The end-to-end delay was varied between 100 and 400 ms as follows: 100, 150, 200, 300 and 400 ms. For each case, an initial seed was generated randomly to cater for a range of possible loss patterns. The state transitions were counted according to Gilbert model in Figure 2.6. Considering the pseudo-random number generation and limited length of test sentences (about 10s), the packet loss rate at the end of test sentence (*ulp(Real)/clp(Real)*) were calculated and used as inputs to neural network models. For each case, the speech quality (MOS score) between the reference and degraded

speech file was first calculated using PESQ algorithm, and then the MOS (PESQ) score was converted to conversational MOS score (MOSc) after taking into account the impact of delay (Using the method described in Section 5.2).

Each dataset, consists of a MOS score and associated impairment parameters. The *ulp(Real)*, *clp(Real)*, codec type, gender, delay, packet size and the calculated MOSc score formed one sample for neural network database. An example of a dataset is shown in Table 6.3 (only selected samples are shown).

Table 6.3: An example of ANN database for MOSc prediction

Delay (ms)	Codec type	Gender	Packet size	$ulp(Real)$	$clp(Real)$	MOSc
100	1	0	1	0.29	0.68	1.94
100	1	1	2	0.05	0.62	3.16
100	3	0	2	0.23	0.85	2.45
200	1	1	1	0.30	0.23	2.01
200	2	1	2	0.16	0.93	2.56
300	3	0	3	0.16	0.88	1.65
300	4	1	1	0.19	0.63	1.81
400	1	0	1	0.06	0.32	1.69

Note: the inputs were coded as follows:

- Delay: 100, 150, 200, 300, 400 (ms)
- Codec type: G.729 = 1, G.723.1 = 2, AMR = 3, iLBC = 4.
- Gender: male = 0, female = 1
- Packet size: 1, 2 and 3
- $ulp(Real)$ and $clp(Real)$ are actual measured unconditional loss probability (ulp) and conditional loss probability (clp)
- MOSc is measured conversational voice quality by using PESQ/E-model

A total of 2400 samples (or data sets) were generated. 80% of the samples were chosen randomly as the training set and the remaining 20% as the validation set. All input and output variables were normalized to [0, 1] before neural network processing. The neural network was used to learn the non-linear relationship between six input variables and one output variable from the training set. A training process is completed when the neural network stopping criteria is reached (e.g certain amount of training epoches or certain error target for validation data set are reached).

In order to verify the generalization of the trained neural net, an unseen test data set was created from a randomly chosen male and female talkers and a set of different network loss conditions (*ulp*: 0, 7, 14 %; *clp*: 10, 60%; delay: 120, 350 ms; packet size:1 to 3 for all codecs). A total of 264 new patterns were generated and used to test the trained net.

The training, validation and test sets were used to tune/optimise the neural networks to minimise the prediction errors.

6.3.3 Neural Network Size and Speech Quality Prediction Results Analysis

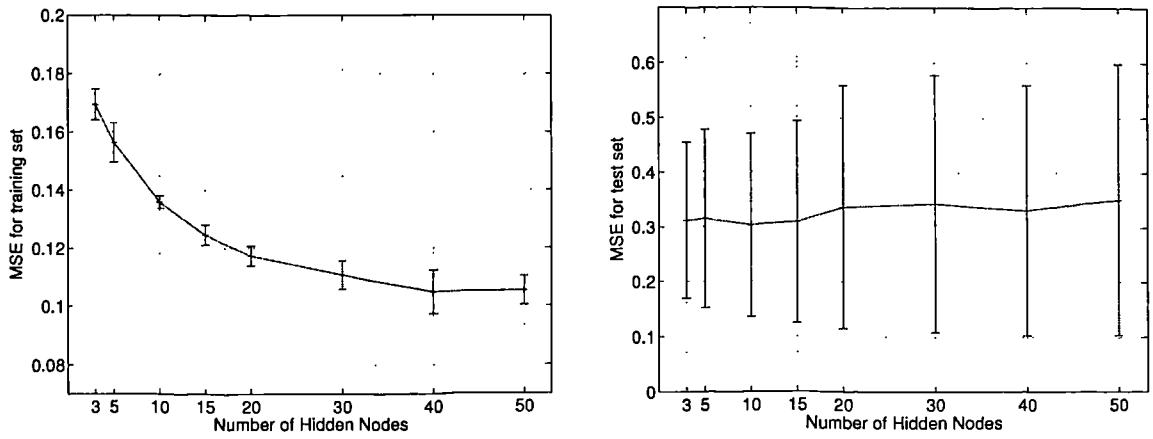
Different network structures (e.g. the number of hidden neurons and parameters of Standard Backpropagation learning algorithm [103]) were investigated to determine a suitable architecture for ANN model.

The network size was first investigated for generalization [104]. The trained networks had the following architecture: 6:m_h:1, where m_h is for hidden neuron number was varied from 3 to 50. Each configuration of the network was tested with ten simulations, each with a different starting condition (random weights). All networks were trained for an identical number of stochastic updates (e.g. 20000). The learning rate (η) was set to 0.4 and the maximum difference (d_{max}) was 0.01 for the training. The results for training and test dataset (in terms of error MSE (mean square error) and correlation coefficient ρ between measured and predicted

MOSc) was shown in Table 6.4 and Figure 6.6 (shown in mean MSE and standard deviation of MSE).

Table 6.4: The impact of network size on training and test data set

m_h	3	5	10	15	20	30	40	50
Training MSE	0.1695	0.1564	0.1359	0.1246	0.1173	0.1107	0.1048	0.1052
Training ρ	0.9555	0.9649	0.9721	0.9765	0.9794	0.9821	0.9836	0.9835
Test MSE	0.3117	0.3149	0.3024	0.3083	0.337	0.3431	0.331	0.3509
Test ρ	0.9273	0.9337	0.9365	0.9329	0.9251	0.9133	0.9102	0.9171



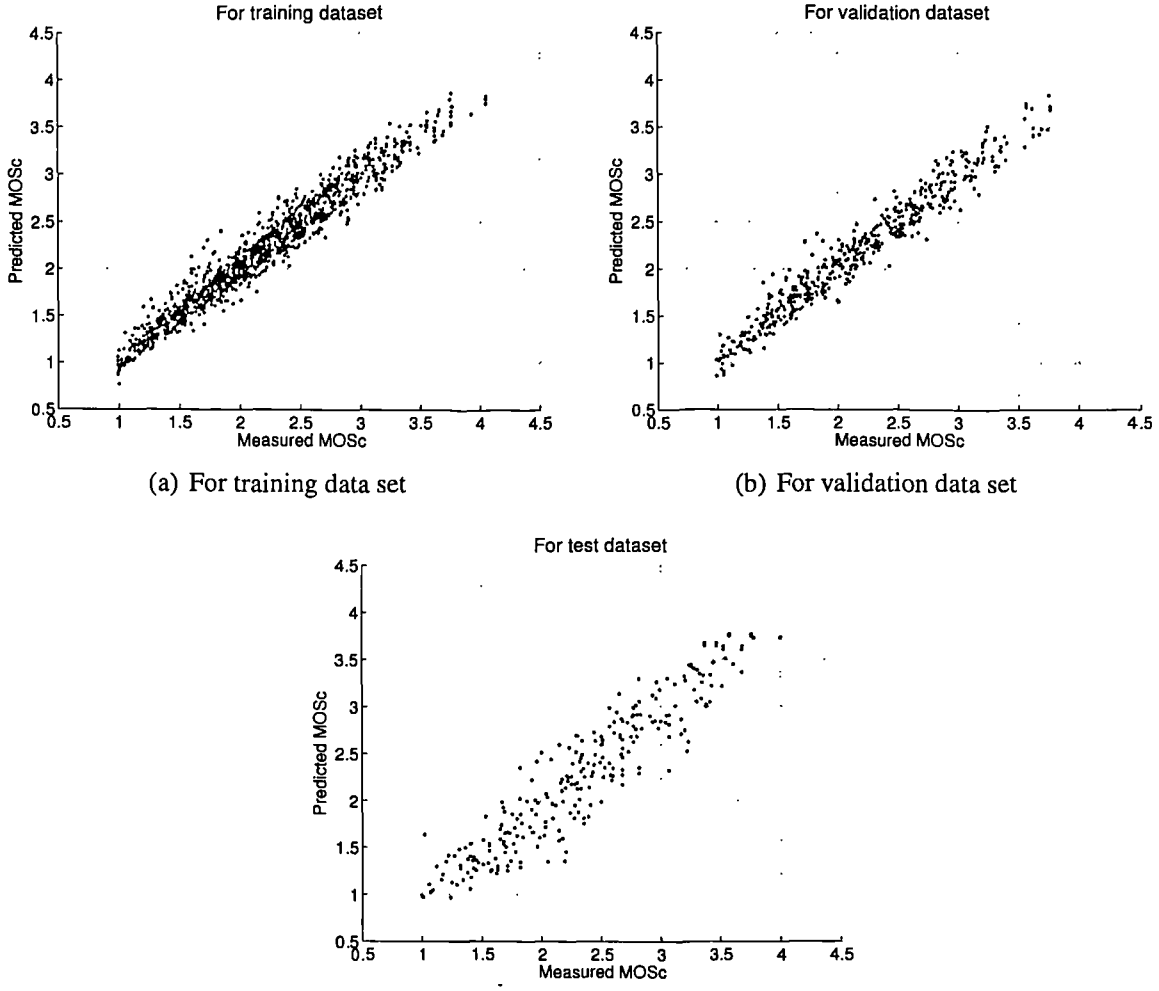
(a) for training set (b) for test set

Figure 6.6: The error (MSE) for training and test data set for different network size

From Figure 6.6 and Table 6.4, it can be seen that the error for the training set decreases with increase in the number of hidden neurons, however, the error for the test set decreases slightly when m_h varies from 5 to 10, but increases with further increase in m_h. The results suggest that an ANN model with 10 hidden nodes has the best generalization performance (with the minimum test error) for this application. Larger networks (hidden neuron number over 10) resulted in worse generalization due to overfitting [105].

For a 6-10-1 net, we obtained a correlation coefficient (ρ) of 0.97 and MSE of 0.154 for the training set, ρ of 0.97 and MSE of 0.152 for the validation set. For the test set, ρ was 0.94 and MSE was 0.28. The scatter diagrams of the predicted versus the measured MOSc scores

for the training, validation and test data sets are illustrated in Figure 6.7(a), 6.7(b) and 6.7(c). The results suggested that the designed neural network model works well for speech quality prediction in general.



(a) For training data set

(b) For validation data set

(c) For test data set

Figure 6.7: Predicted MOSc vs. measured MOSc for training, validation and test sets

6.4 Summary

In this Chapter, neural network models for predicting both listening and conversational voice quality, non-intrusively, have been developed. The models are based on intrusive methods (e.g. PESQ for listening quality and a combined PESQ/E-model structure for conversa-

tional quality) for neural network training, which avoids time-consuming subjective tests. The simulation system structure, the NN dadabase generation, and the chosen of neural network structure/parameters are described. Preliminary results show that both the listening and conversational neural network models have accuracy close to the ITU PESQ or PESQ/E-model (correlation coefficient of 0.946 and 0.940 for the test set, respectively) based on a VoIP simulation system. The models will be verified using real Internet trace data, which will be presented in Chapter 7.

Chapter 7

Perceived Speech Quality Prediction for VoIP in Internet

7.1 Introduction/Motivation

In Chapter 5 and Chapter 6, nonlinear regression and neural network models have been developed for predicting voice quality non-intrusively. As the work so far is only based on a VoIP simulation system (e.g. simulated packet loss), it is unclear whether the characteristics of parameters used for voice quality assessment in the simulated system (e.g. 2 state Gilbert model) are realistic because IP networks continue to grow and evolve. It is also unclear how accurate the developed models are for predicting voice quality in the current Internet.

In order to characterise the behaviour of current VoIP networks in terms of the key parameters (i.e. packet loss, jitter and delay) and to further verify the neural network and regression models for voice quality prediction for real Internet trace data, the VoIP trace data were collected from different international Internet links (e.g. between UK and Germany, between UK and USA and between UK and China). Based on the trace data, the network performance (e.g. packet loss, delay and jitter) and network parameter modelling (e.g. packet loss model) are analysed. Then, the voice quality is predicted from the network parameters and speech related parameters by using either neural network models and regression based models.

The structure of the Chapter is as follows. The method of Internet trace data collection

is presented in Section 7.2. The IP network performance and modelling are analyzed in Section 7.3. The speech quality prediction using neural network models and regression models are presented in Section 7.4 and 7.5, respectively. The performance analysis and comparison between neural network and regression models are given in Section 7.6. Section 7.7 concludes the chapter.

7.2 Internet Trace Data Measurement

7.2.1 Related Work

Many studies and measurements have been conducted on the Internet. The most large scale measurements in the Internet were carried out by Vern Paxson in 1996 [106] and 1997 [44] where he carried out the study on routing behavior and dynamics of the Internet by tracing TCP bulk transfers. He observed loss probabilities that ranged between 0% and 65%.

Mukherjee [107] in 1992 found that end-to-end one way packet delays were well modelled using a shifted gamma distribution for all three network paths studied (all within US), but the parameters of the distribution depended on the path and time of day. The work was based on a Round trip delay measurement using ICMP.

Work dedicated solely to measurement of audio data include Jean Bolot in 1993 [42] where he performed end-to-end packet delay and loss measurements in the Internet for a single link from INRIA in France to the University of Maryland in the U.S. The tests were based on measuring round trip delays of small UDP probe packets sent at regular time intervals. They found that the interarrival time distribution for Internet packets is consistent with an exponential distribution.

More VoIP related measurements were carried out by Olof hagsand et al [108] focusing on interarrival variance or jitter, and Jiang [46] and Rosenberg [109] where tests are mainly based on USA or between USA and Germany.

Borella [43] analyzed a month of Internet UDP packet loss statistics for speech transmission using three different sets of client/server host pairs. Their results exhibit packet loss that is highly bursty.

Yajnik and Moon et al [45] examine the temporal dependence of packet loss for unicast and multicast traffic collected over 128 hours. They evaluate the accuracy of three models of increasing complexity: the Bernoulli model, the 2-state Markov chain model and the k-th order Markov chain model. Out of the 38 trace segments considered, the Bernoulli model was found to be accurate for 7 segments considered, the Bernoulli model was found to be accurate for 10 segments.

Fujimoto et al [110] analyzed the characteristics of the tail part of packet delay distributions by measuring Round Trip Time and one-way delay (GPS was used for time synchronization). They found that the Pareto distribution is most appropriate as the model of one-way delay distribution (tail part), as well as RTT distributions.

Recently, Athina [10] carried out delay and loss measurements over the backbone networks of Internet Service Providers in the US. GPS was used to synchronize the clocks of senders and receivers. Although backbone networks are known to be sufficiently provisioned to cause negligible degradation on data traffic, their study shows that a large number of the Internet paths exhibited poor VoIP performance, mainly due to high delay and high delay variability.

Clearly, Internet measurement is a research topic that is continuously evolving along with the evolution of the Internet as well as the evolution of the applications. It is also impossible to study the delay and loss characteristics for all possible connections.

In order to understand the characteristics of current Internet in terms of the key parameters (i.e. packet loss, jitter and delay) and to predict voice quality for the current Internet using developed regression and neural network models, we collected network information internationally between UK and USA, between UK and China, and between UK and German during the study.

The motivations for performing our tests are:

- to present an up-to-date series of results on today's Internet and to present the characteristics and modelling of packet loss and delay,

- to evaluate the readiness of today's Internet for VoIP applications on both good links and rough links internationally,

- to verify the developed neural network models and non-linear regression models for real Internet VoIP voice quality prediction,

- to modify and improve current playout buffer algorithms according to real Internet trace data.

7.2.2 Measurement Approach and Trace Data Collection

A UDP/IP probe tool [111] is used to collect and measure the main network parameters that affect voice quality (i.e. packet loss, delay and delay variation). In VoIP applications, UDP/IP is used to transport speech over the network. The tool provides a convenient and yet effective way to measure relevant impairment parameters of paths of IP networks as would be experienced by actual UDP traffic carrying speech. Similar tools have been used for experimental assessment of end-to-end behaviour of Internet in the past [112, 42, 43] and more recently for speech quality prediction [10].

The structure of UDP/IP trace data collection is shown in Figure 7.1. It consists of 4 application processes: Source, Echo, Sink and Logger. Typically, the Source and the Sink processes run on the same local host (e.g. Sender Host), while the Echo process runs on a remote host (e.g. Receiver Host). The Logger process may run on any host. Here, it runs on the Sender Host, for convenience. The Source process generates packets and sends them to the Echo process, via the Internet. The echo process receives the packets and sends them back. By changing the packet size, packet interval, total sent packets, different VoIP applications can be emulated (e.g. packet interval of 10 ms for G.729 codec and 30 ms for G.723.1 codec).

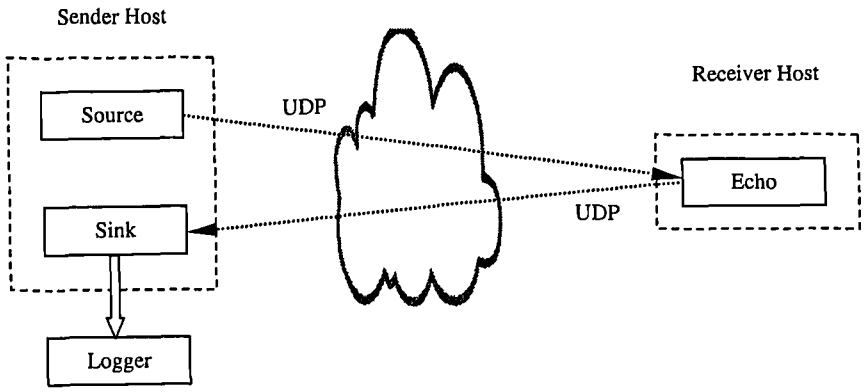


Figure 7.1: Structure of UDP trace data collection

As for the RTP protocol, a time stamp and a sequence number are written to each packet sent from Source process. Echo and Sink processes write a timestamp when they receive the packet. An example of the trace data is shown in Figure 7.2. From the sequence number, the packets that have been lost in the network can be deduced. From the timestamps, the network delay and delay variations can be calculated. In our experiments, the size of the probe packets is set to 32 bytes. The interval between successive packets is 30 ms, which is similar to G.723.1.

Internet weather station data file
Generated on Sat Apr 13 13:22:12 2002
recv host is 128.59.15.46
send host is 141.163.75.216
pktsize 36 (bytes) pktgap 30000 (microsecond) runtime 7200 (seconds)

Packet interval = 30 ms

Seq. Num.	Source_timestamp	Echo_timestamp	Sink_timestamp
1	0.869845	119.684693	0.995452
2	0.891684	119.716319	0.995663
3	0.921623	119.735033	1.013999
4	0.951612	119.767939	1.046485
5	0.981975	119.795492	1.074011
6	1.011660	119.825114	1.103706
.....			

Figure 7.2: An example of UDP trace data

Based on the University of Plymouth, another four sites around the world were chosen for Internet measurement. The other four sites are Columbia University (CU), USA; Beijing University of Posts & Telecommunications (BUPT), China (Northern China); Darmstadt University of Technology (DUT), Germany and Nanchang Telecommunication Bureau (NCT) China (Southern China). These sites were selected because they are international connections with different delay characteristics. The measurement setup is shown in Figure 7.3 and five sites are shown in Table 7.1.

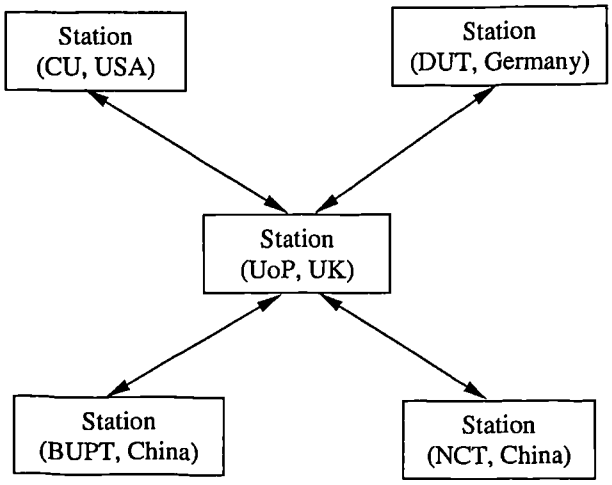


Figure 7.3: Internet measurement setup

Table 7.1: Locations of stations for trace data collection

Station Number	Station Location	IP Address
1	Uni. of Plymouth (UoP), UK	141.163.75.216
2	Columbia Uni. (CU), USA	128.59.15.46
3	Darmstadt Uni. of Technology (DUT), Germany	130.83.245.100
4	Beijing Uni. of Posts & Telecomm. (BUPT), China	202.112.101.135
5	Nanchang Telecomm. Bureau (NCT), China	218.65.107.2

Trace data were collected among four pairs (from Plymouth to other four sites). All of the data were collected during April and June of 2002.

In order to have a better understanding of the collected traces, the routes from UoP to other four destinations are summarised briefly here. The route from UoP to BUPT (China) was via JANET (UK's Education and Research Network) directly to CERNET(China Education and

Research Network). The route from UoP to NCT (China) was via JANET, to SPRINTLINK NET, then to CHINANET (China Telecommunication Network). The route from UoP to CU (USA) was via JANET, GÉANT (the pan-European Research and Education Network), then to NYSERNET (New York State's academic and research network). The route from UoP to DUT (Germany) was via JANET, LINX (London Internet Exchange) to TELIA net.

7.2.3 Trace Data Preprocess - Clock Synchronization and Drift

As the send and receive clocks are not synchronized in trace data collection, the first step in determining one-way delay is to remove the time difference between the clocks (expressed as δ) at the two hosts. We estimate δ by measuring the minimum Round Trip Time (RTT_{min}) over a short interval.

If the RTT_{min} is obtained when a packet is sent at T_1, received at T_2, and echoed back at T_3, where T_1, T_2, and T_3 are all local times measured by the sender and receiver, then δ can be calculated as:

$$\delta = \frac{RTT_{min}}{2} - (T_2 - T_1) \tag{7.1}$$

Here we assume that the links are symmetric.

Later, for any packet sent at T_s and received at T_r, both local times, the one-way delay D can be calculated as:

$$D = T_r - T_s + \delta \tag{7.2}$$

Further we remove clock drift (or clock skew), which is caused by the two clocks running at different frequencies. We use a linear regression method [46, 113] to calculate a drift rate and then remove the drift from the one-way trace data.

7.2.4 Trace Data Preprocess – Speech Talkspurt/Silence On/Off Model

It is known that speech can be modelled as a process that alternates between talkspurts and silences and follows an exponential distribution [114]. The lengths of talkspurts and silence periods are related to the Voice Activity Detection (VAD) threshold and hangover time configured in codec or terminal. For the purpose of our simulations, an exponential distribution with a mean of 1.5sec for both talkspurts and silences is selected as in [10, 115].

With a random seed chosen for the exponential distribution, the data (continuous in time scale) is processed to contain talkspurt and silence part (similar to a real VoIP trace).

7.3 IP Network Performance Analysis

7.3.1 Delay/Jitter and its Distribution

Distribution of end-to-end delay is an important component for network performance monitoring and modelling [107] or for jitter buffer algorithms optimization [116].

The basic information of delay/jitter/loss for the 5 selected traces from 5 directions is shown in Table 7.2. Delay is for the average network delay and jitter is calculated according to Equation 2.6. The loss parts (right three columns) will be explained later.

Table 7.2: Basic information for trace data #1 to #5

Trace	Trace Description	Start-time (Sender)	Delay (ms)	Jitter (ms)	Loss (ulp %)	Loss (clp %)	$E[Y]$
1	BUPT → UoP	17:30pm, 07/06/02,Fri	153	16.2	1.1	38.4	1.62
2	UoP → CU	13:22pm, 13/04/02,Sat	46	0.8	0.3	82.8	5.81
3	UoP → BUPT	9:11am, 11/06/02,Tue	186	19.5	14.3	42.4	1.74
4	UoP → DUT	18:44pm, 10/06/02,Mon	16	0.7	4.4	17.3	1.21
5	UoP → NCT	11:02am, 30/05/02,Thu	150	0.2	0.2	64.9	2.85

From Table 7.2, it can be seen that the traces between UoP and BUPT (China) suffer large delay and delay variation (jitter) (regarded as rough traces). The trace from UoP to NCT (China) has large delay but small jitter. The traces from UoP to CU (USA) and from UoP to DUT (Germany) experience low delay and delay variation (regarded as good traces).

The delay characteristics can be described by the Cumulative Distribution Function (CDF) of delay, which is defined as $F(x) = P(X \leq x)$. The CDFs for five selected traces in Table 7.2 are shown in Figure 7.4. Delay is normalized for comparison (shift to the minimum delay).

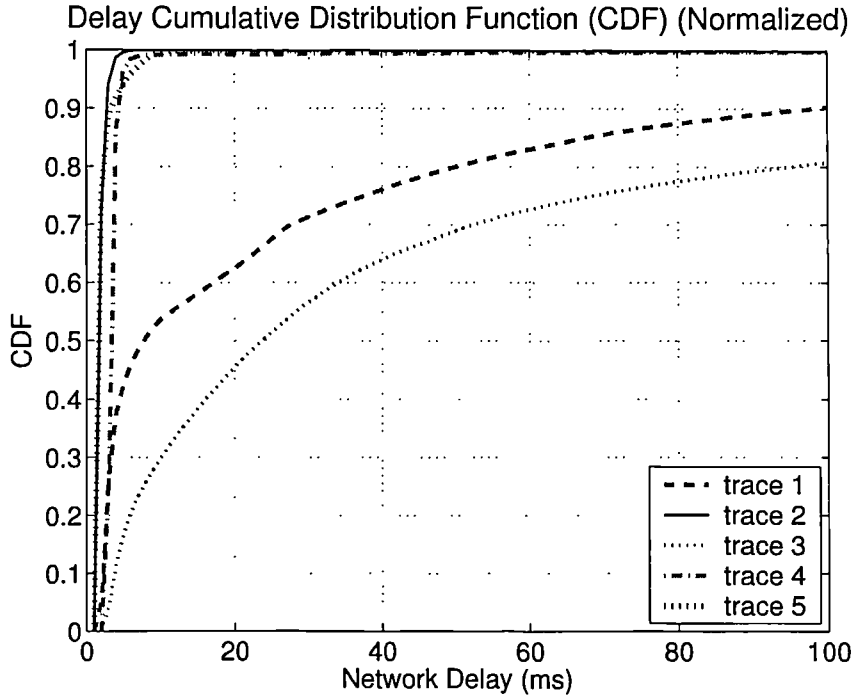
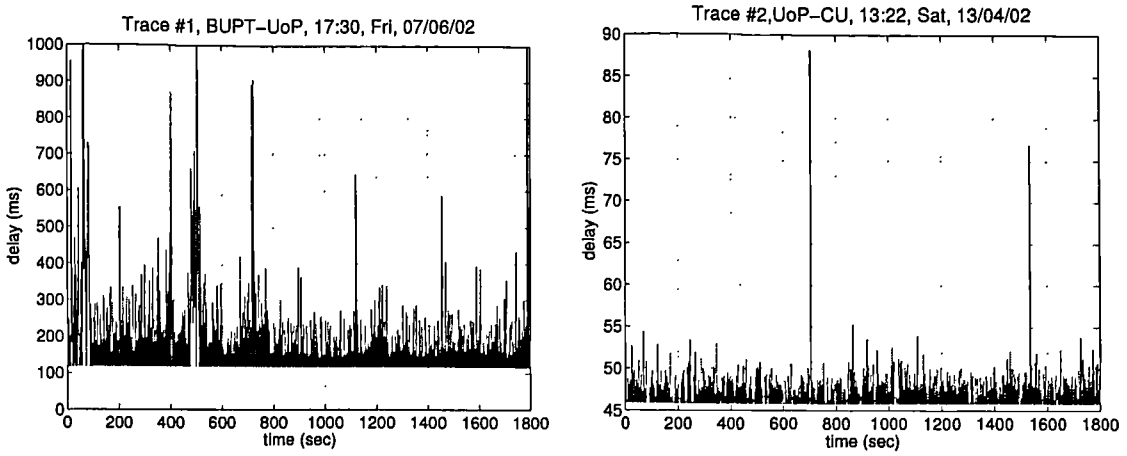


Figure 7.4: Delay cumulative distribution function (CDF) for 5 traces

From the figure, it is very clear that traces #1 and #3 have high delay variation, whereas traces #2, #4, and #5 have low delay variation. As one example for high or low delay variation traces, Figure 7.5 also illustrates trace #1 (high delay variation) and trace #2 (low delay variation).



(a) Trace #1 (b) Trace #2

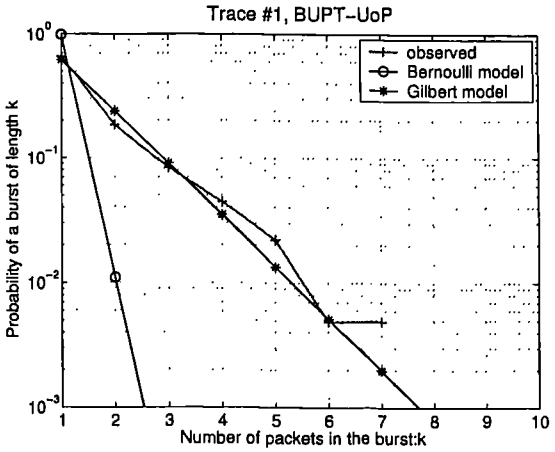
Figure 7.5: Trace data #1 and #2

7.3.2 Packet Loss and its Distribution

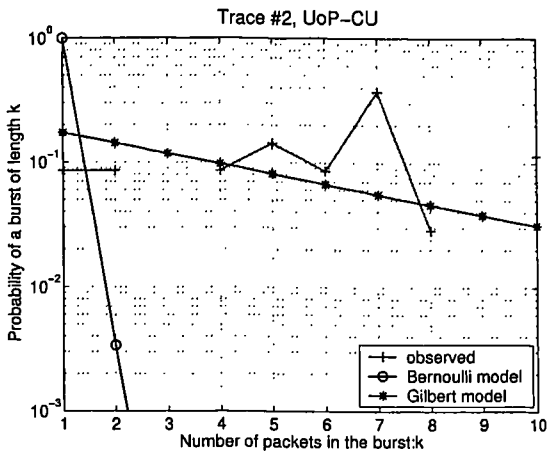
As described in Section 2.5, packet loss for Internet trace data can be characterized by a statistical model (e.g. Bernoulli model or Gilbert model). In this section, we analyze the packet loss characteristics for the Internet trace data collected using Bernoulli and 2-state Gilbert model.

For selected traces #1 to #5, the probability for burst length k was calculated using both Bernoulli model and 2-state Gilbert model (detail see Section 2.5). The observed value from the trace data was also calculated. The results are shown in Figure 7.6. The unconditional loss probability (ulp), conditional loss probability (clp), and mean burst loss length ($E[Y]$) are calculated and tabulated in Table 7.2.

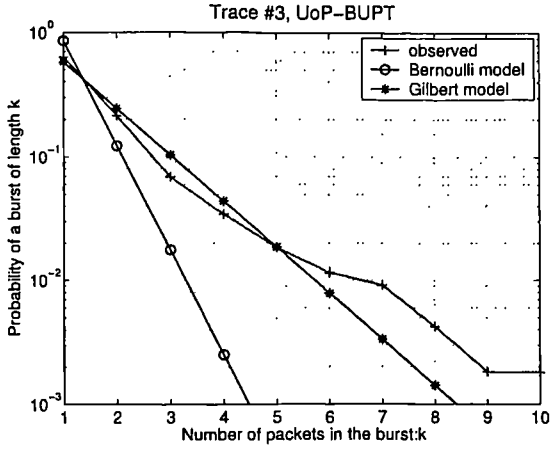
From the figure, it can be seen that 2-state Gilbert model has a better fit for packet loss characterization than Bernoulli model. This confirms the result in [46]. We also observe that clp is greater than ulp ($clp > ulp$) for all five traces, as mentioned in [42]. This shows that all five traces are quite bursty. The worst one is the trace from UoP to CU where the mean burst loss length reaches 5.81, whereas, the best one is the trace from UoP to DUT where the mean burst loss length is only 1.21.



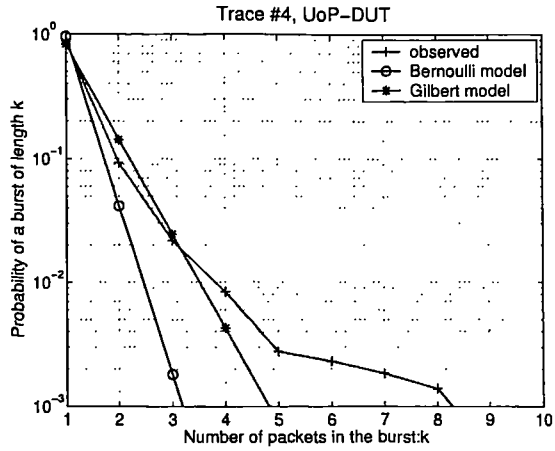
(a) Trace #1



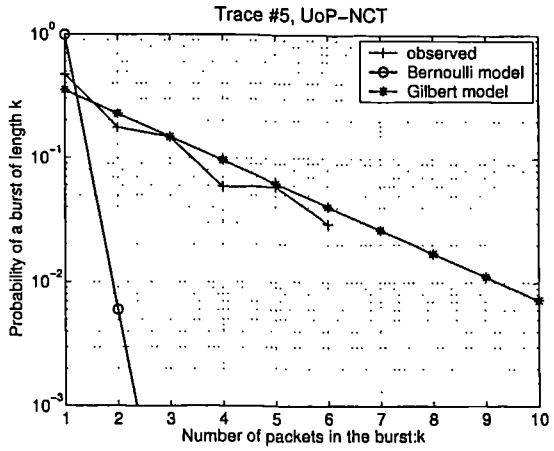
(b) Trace #2



(c) Trace #3



(d) Trace #4



(e) Trace #5

Figure 7.6: The burst loss distribution for traces #1 to #5

7.4 Perceived Speech Quality Prediction Using NN Models

As shown in Figures 7.5(a), 7.5(b) and 7.4, Internet trace data shows different amount of delay variation or jitter (high or low). Jitter buffer (or playout buffer) at the receiving end is normally used to compensate jitter. This causes further delay (buffer delay) and additional packet loss (buffer loss). The tradeoff between packet loss and delay is mainly decided by the playout buffer algorithm used in the receiving end. Different playout buffer algorithms will result in different end-to-end perceived speech quality even under the same network conditions.

In this section, we investigate whether we can use the neural network models developed in Section 6.3 to predict conversational speech quality from real Internet trace data and to verify how good the models are for predicting voice quality. We assume an adaptive playout buffer algorithm [22] is used for postprocessing of the trace data for voice quality prediction. The algorithm adjusts the buffer at the beginning of each talkspurt(Playout buffer and buffer algorithms will be described in detail in Chapter 8).

Figure 7.7 shows the systematic structure to obtain measured MOSc from PESQ/E-model (Figure 7.7 (a)) and predicted MOSc using neural network models (Figure 7.7 (b)) .

For every 9 sec trace data (9 sec was chosen because it is within the recommended length for PESQ algorithm [4]), the actual packet loss (including network packet loss and late arrival loss) and actual end-to-end delay (including network delay and buffer delay) were calculated based on the adaptive playout buffer algorithm [22]. An average of actual delay for the 9 sec trace data was also calculated and sent to delay model to get the delay impairment I_d. According to the actual packet loss patterns, the degraded speech was generated by G.723.1 codec and compared with reference speech to obtain the PESQ MOS score as shown in Figure 7.7 (a). The conversational speech quality (MOSc) was then derived from the PESQ MOS and actual delay as described in Section 5.2. This gives the Measured MOSc which is used to verify the performance of the developed neural networks models. The actual unconditional loss probability (ulp) and conditional loss probability (clp) for each 9 sec trace segment were calculated using

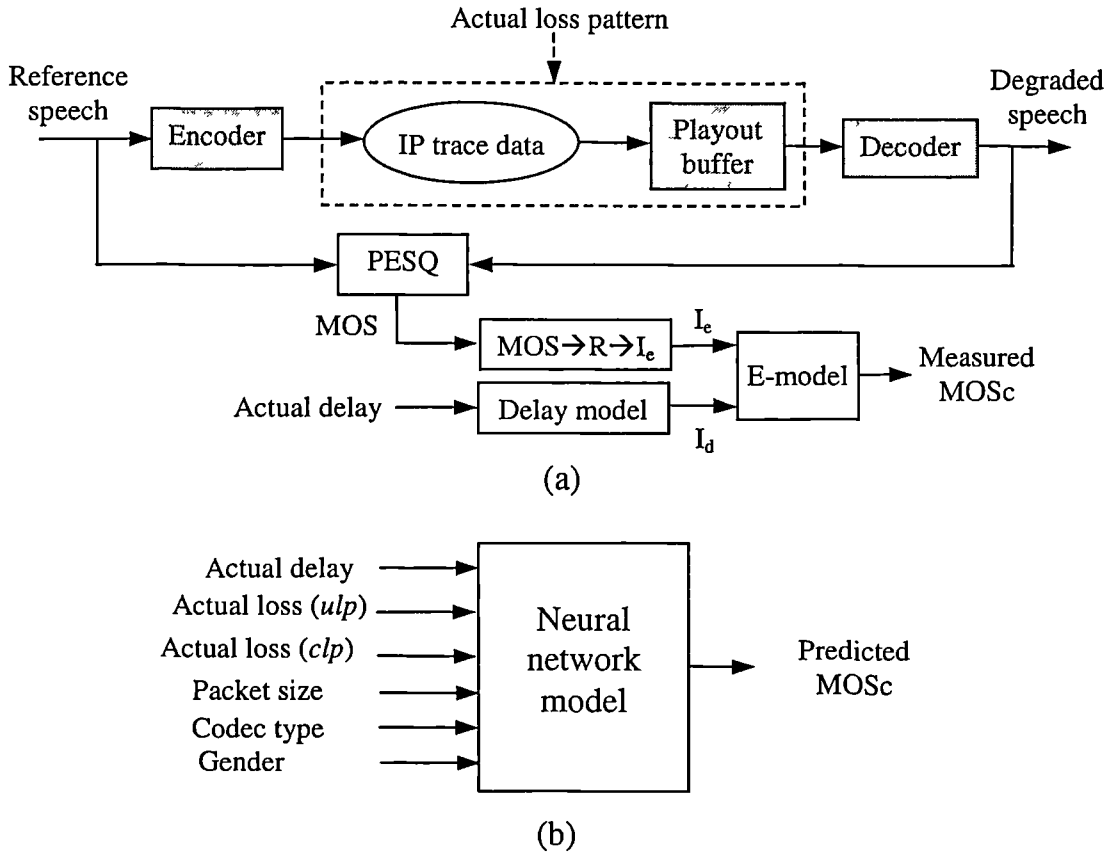


Figure 7.7: Systematic structure to obtain measured and predicted MOSc from trace data

the Gilbert model as in Figure 2.6 (details see Section 2.5). The codec type here was G.723.1 (codec type '2' in neural network inputs) and packet size was '1' for this trace data. The gender was determined directly from speech sample for simplicity (in practice, gender would be determined from the pitch of the decoded speech). The predicted MOSc was obtained by applying the parameters (e.g.*ulp*, *clp*, an average end-to-end delay, codec type, packet size, gender) to a trained neural network (i.e. net-6-10-1) as shown in Figure 7.7 (b).

For each trace data (length of 30 minutes), 99 test segments were generated for voice quality prediction. Four traces (trace #1 to trace #4 with two representing high and two representing low delay variation) were selected for neural network model verification. There was a total of 396 samples tested using real Internet trace data which represent a high diversity in IP network performance (see Table 7.2). The scatter diagram of the predicted versus the measured MOSc scores for Internet trace data (#1 to #4) using a trained neural network is illustrated in Figure 7.8. Preliminary results show that the correlation coefficient of 0.94 and MSE of 0.23 were obtained. This demonstrates that the neural network model works well for speech quality prediction for real Internet VoIP trace data in general.

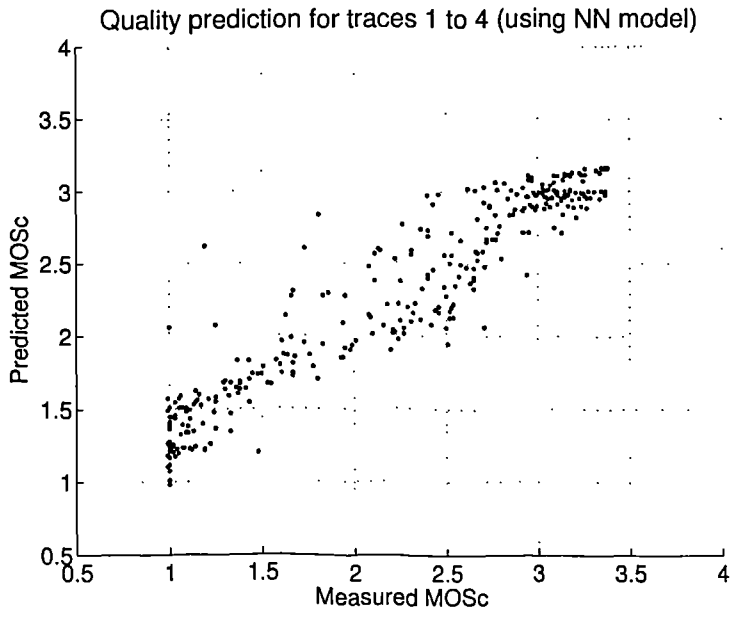


Figure 7.8: Predicted MOSc vs. measured MOSc for trace data (#1 to #4) using NN model

7.5 Perceived Speech Quality Prediction Using Regression Models

In last section, we verified the method of speech quality prediction using neural network models for real VoIP trace data. The speech quality can also be predicted directly using non-linear regression models developed in Chapter 5.

As the collected Internet VoIP trace data is for G.723.1 codec (30ms packet interval) with packet size of one, thus, the relationship of I_e vs. packet loss rate ρ (in percentage) can be decided from Table 5.1 in Chapter 5 and shown as below:

$$I_e = 20.06 \ln(1 + 0.1024\rho) + 25.63 \tag{7.3}$$

Following the same procedures in Section 7.4, the actual packet loss rate (ρ) (including network loss and late arrival loss) was calculated from every 9 second trace segment (only average packet loss rate was calculated from Bernoulli model for simplicity). The average actual delay (d) for the 9sec trace segment was also calculated and then I_d can be obtained from Equation 5.7. From I_e and I_d, the predicted MOSc can be obtained from Equations 5.11 and 5.1. Overall the predicted conversational voice quality (MOSc) can be obtained from packet loss rate, codec type, packet size, and delay using regression model as shown in Figure 7.9.

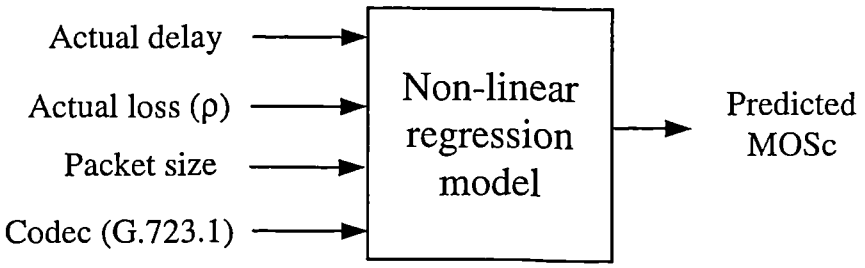


Figure 7.9: Systematic structure to obtain predicted MOSc using regression model

Similarly there was a total of 396 samples generated from Internet trace data #1 to #4. The predicted MOSc was calculated using non-linear regression model and the measured MOSc

was obtained by applying PESQ/E-model directly as shown in Section 7.4. The scatter diagram of the predicted versus the measured MOSc scores for Internet trace data (#1 to #4) using regression model is illustrated in Figure 7.10. Preliminary results show that the correlation coefficient of 0.98 and MSE of 0.12 were obtained. In this application, it seems that the non-linear regression model achieves higher accuracy of voice quality prediction than the neural network model.

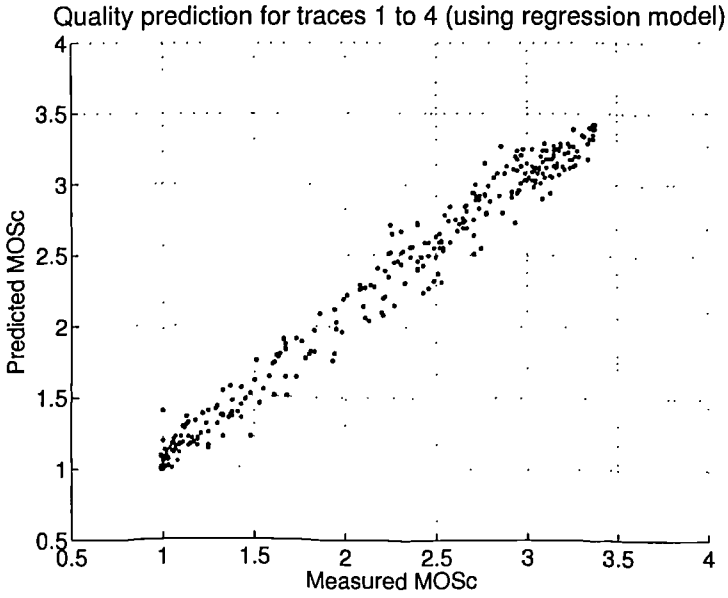


Figure 7.10: Predicted MOSc vs measured MOSc for trace data (#1 to #4) using regression model

7.6 Performance Analysis/Comparison between NN and Regression Models

In order to compare with measured MOSc from PESQ/E-model and predicted MOSc from the neural network model and from the non-linear regression model, the average measured MOSc, predicted MOSc using the regression model, and predicted MOSc using the neural network model are summarized in Table 7.3 for trace data #1 to #4. The network delay, network loss and actual end-to-end delay and packet loss are also shown in the table.

Table 7.3: Comparison of measured against predicted MOSc for trace data #1 to #4

Trace	Network Delay (ms)	Network Loss (%)	Actual Delay (ms)	Actual Loss (%)	MOSc (Measured)	MOSc (NN-predicted)	MOSc (Regression-predicted)
1	153	1.2	205	10.1	2.27	2.33	2.35
2	46	0.3	58	0.4	3.33	3.13	3.37
3	186	14.2	309	22.3	1.22	1.47	1.27
4	16	4.2	71	4.3	2.90	2.84	2.96

In general, trace #2 has the best network performance and the highest perceived speech quality (MOSc = 3.33) and trace #3 has the worst network performance and the lowest perceived speech quality (MOSc = 1.22). The predicted MOSc using the neural network model and the regression model both are quite close to the measured MOSc obtained directly from PESQ/E-model with higher accuracy from the regression model (correlation coefficient of 0.98) than that from the neural network model (correlation coefficient of 0.94).

It is noticed that the regression model has higher accuracy than the neural network model. The possible reasons are as following:

- The regression model used here is the one developed only for G.723.1 codec with packet size of one, whereas, the neural network model applied here is a general model suitable for four codecs (G.723.1, G.729, AMR and iLBC) and different packet size (packet size of 1 to 5). The generality of the neural network model causes the lower accuracy for speech quality prediction while compared with the regression model.

- There is no difference between regression model and PESQ/E-model method in considering the impact from end-to-end delay (both using Equation 5.7). The only difference between them is at the calculation of I_e, where PESQ/E-model obtains I_e value from PESQ algorithm, whereas the regression model calculates I_e using a simplified regression function (see Equation 7.3). However, only five delay values (100, 150, 200, 300, 400ms) are considered in the training set for the neural network model, this also contributes to the lower accuracy of the neural network model while compared with the

regression one.

In general, there are both advantages and disadvantages for neural network and non-linear regression models for voice quality prediction. These are summarized in Table 7.4.

Table 7.4: Comparison between neural network and regression models

Models	Advantages	Disadvantages
Neural network	learning ability with adaptability and generality, robust	low accuracy, more complex
Non-linear regression	simple and straight-forward, high accuracy for specific scenario	static, lack of generality, application inconvenient (one equation for one scenario)

7.7 Summary

In this chapter, the measurement, collection and preprocess of Internet trace data has been presented. The trace data from international links between UK and USA, UK and China, and UK and Germany have been chosen for analysing the IP network performance (e.g. delay, jitter, packet loss and their distributions). Results show that different traces have different delay and delay variation (e.g. the trace between UK and USA has lower delay and delay variation, whereas, the trace between UK and China (BUPT) has higher delay and delay variation). All the traces show that 2-state Gilbert model has a better fit for packet loss characterisation than Bernoulli model. The neural network models and regression models developed in previous chapters have been used to predict voice quality from real Internet traces. Preliminary results show that both regression and neural network models can predict voice quality well. Non-linear regression models can achieve higher accuracy of voice quality prediction (with correlation coefficient of 0.98) while compared with neural network models (with correlation coefficient of 0.94). The reason behind and the performance analysis and comparison between neural network and regression models were also presented.

Chapter 8

Perceived Speech Quality Prediction for Buffer Optimization

8.1 Introduction/Motivation

In this chapter, an application of the voice quality prediction models on perceived quality driven jitter buffer optimization is investigated. Nonlinear regression models are used for simplicity.

In Voice over IP (VoIP) applications, delay, jitter (i.e. delay variation) and packet loss are the main network impairments that affect perceived speech quality. Jitter can be partially compensated for by using a playout buffer at the receiving end, but this introduces further delay (buffer delay) and additional packet loss (packets arriving after their playout times will be dropped by the receiver). A tradeoff is necessary between increased packet loss and buffer delay to achieve a satisfactory result for any playout buffer algorithm. For example, the longer the buffer delay, the lower the late arrival loss and vice versa.

In the past, the choice/design of buffer algorithms was largely based on buffer delay and loss performance (e.g. a design objective could be to achieve a minimum average end-to-end delay for a specified packet loss rate [117, 118, 29, 119] or minimum late arrival loss [117]. This approach is inappropriate as it does not provide a direct link to perceived speech quality. From QoS perspective, the choice of the best buffer algorithm for a given situation should be

determined by the likely perceived speech quality. The importance of this is now starting to be recognised [10, 120, 22]. For example, in [120], perceived voice quality is used to control the playout buffer in order to maximise the MOS values in terms of delay and loss. The concept of perceptual optimization has also been extended to other QoS control problems, such as joint playout buffer/FEC control [101] to maximise MOS values in terms of delay, loss and rate.

However, current methods of perceptual optimization are based on assumptions about perceived voice quality which are inappropriate. In [120], the method is based on the assumption that the effects of packet loss and delay on voice quality are linearly additive on the MOS scale which is doubtful. A further assumption is that the relationship between MOS and packet loss for codecs is linear which is not correct for most codecs. It has also been suggested in [101] that one equation may be used to represent the impairments due to packet loss for all codecs. This may not be appropriate, especially for newer codecs.

In all perceptual-based buffer design/optimisation and QoS control for VoIP, voice quality is used as the key metric because it provides a direct link to user perceived QoS. However, this requires an efficient and accurate objective way to measure perceived voice quality. Most current methods [101] [121] use the ITU-T E-model [7] to predict voice quality, but the E-model requires subjective tests to derive model parameters which is time-consuming and often impractical. As a result, the E-model is only applicable to a limited number of codecs and network conditions. It is also inevitable that discontinuities exist in subjective results [9] because only a limited range of scenarios can be tested for. PESQ [4] gives a good measure of voice quality, but it is not appropriate for optimisation because of the overhead involved in its use in real-time.

In Chapter 5, novel methods to predict voice quality non-intrusively based on a combination structure of PESQ and E-model have been proposed and two models (e.g. statistical nonlinear regression model and neural network-based model) have been developed. The developed nonlinear regression models are used in this chapter for perceived quality driven playout buffer optimization.

For perceived buffer design, it is important to understand the delay distribution modeling as it is directly related to buffer loss (or late arrival loss). The characteristics of packet transmission delay over Internet can be represented by statistical models which follow Normal, Exponential, Pareto and Weibull distributions depending on applications. For example, the delay distribution for Internet packets (for a UDP traffic) has been shown to be consistent with an Exponential distribution [42], whereas, Pareto distribution may be the most appropriate one to represent the tail delay characteristics for streaming media [116]. As delay characteristics may change with networks and applications, it is unclear what the appropriate delay distribution modelling is the best fit for current VoIP traffic (with on/off pattern). This motivated us to investigate the delay distribution modelling for VoIP trace data described in the previous chapter.

In this Chapter, the existing four jitter buffer algorithms are examined using the method of perceptual speech quality analysis. An adaptive playout buffer algorithm which can adapt to the most suitable traditional buffer algorithm according to network delay and delay variation is proposed. Further perceived quality driven playout buffer algorithm is investigated. It is proposed to use minimum overall impairment as a criterion for buffer optimization or QoS control. This criterion is more efficient than using traditional maximum MOS score. It is also shown that the delay characteristics of Voice over IP traffic is better characterized by a Weibull distribution than a Pareto or an Exponential distribution. Based on the developed new regression models for voice quality prediction, the Weibull delay distribution model and the minimum impairment criterion, a perceptual optimization playout buffer algorithm is proposed and performance is compared with other jitter buffer algorithms.

The structure of the Chapter is as follows. In Section 8.2, the existing four playout buffer algorithms and their performance are analyzed. In Section 8.3, a new adaptive playout buffer algorithm based on traditional jitter buffer algorithms is presented. In Section 8.4, a perceptual optimum playout buffer algorithm is described based on the speech quality prediction models, a minimum impairment criterion and Weibull delay distribution modeling. The performance

analysis and comparison between existing buffer algorithms and newly proposed algorithms are conducted in Section 8.5. Section 8.6 summarises the chapter.

8.2 Existing Playout Algorithms and Performance Analysis

8.2.1 Existing Playout Buffer Algorithms

A playout buffer can be fixed or adaptive. A fixed buffer cannot adapt to changing network delay conditions and this may result in poor speech quality. Thus, we have focused on adaptive buffer algorithms and adjust the buffer at the beginning of each talkspurt [117, 118, 29].

The notations used to describe buffer algorithms are defined in Figure 8.1. For packet i, we define t_i as the send time; a_i and p_i as the arriving and playout times, respectively. n_i represents network delay and d_i is the actual end-to-end delay or "playout delay". b_i is the buffer delay.

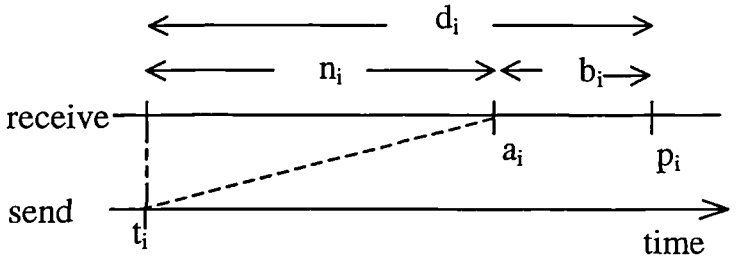


Figure 8.1: Timing associated with packet i

We first implemented four algorithms proposed by Ramachandran et al [117]. These four algorithms maintain a running estimate of the mean and variation of network delay, i.e. $\hat{d}_i$ and $\hat{v}_i$, seen up to the arrival of the i^{th} packet. If packet i is the first packet of a talkspurt, its playout time p_i is computed as:

$$p_i = t_i + \hat{d}_i + \mu \times \hat{v}_i \tag{8.1}$$

where μ is a constant and $\hat{v}_i$ is given by:

$$\hat{v}_i = \alpha \hat{v}_{i-1} + (1 - \alpha) \left| \hat{d}_i - n_i \right| \tag{8.2}$$

n_i is the network delay of the i^{th} packet.

The playout delay for subsequent packets (e.g. packet j) in a talkspurt is kept the same as $d_j = d_i$.

The four algorithms differ only in the computation of $\hat{d}_i$.

- Algorithm 1 ("exp-avg"):

This algorithm estimates the mean delay through an exponentially weighted average.

Algorithm 1 "exp-avg" [117]

$\hat{d}_i = \alpha\hat{d}_{i-1} + (1-\alpha)n_i$
with $\alpha = 0.998002$

- Algorithm 2 ("fast-exp" [117]):

This algorithm is similar to the first, except it adapts more quickly to increases in delays by using a smaller weighting factor as delays increase:

Algorithm 2 "fast-exp"

if $(n_i > \hat{d}_{i-1})$ **then**
  $\hat{d}_i = \beta\hat{d}_{i-1} + (1-\beta)n_i$
else
  $\hat{d}_i = \alpha\hat{d}_{i-1} + (1-\alpha)n_i$
end if
with $\beta = 0.75$ and $\alpha = 0.998002$ as before.

- Algorithm 3 ("min-delay" [117]):

This algorithm is more aggressive in minimizing delays. It uses the minimum delay of all packets received in the current talkspurt.

Algorithm 3 "min-delay"

Let S_i be the set of all packets received during the previous talkspurt.
$\hat{d}_i = \min_{j \in s_i}\{n_j\}$

- Algorithm 4 ("spk-delay" [117]):

This algorithm contains a spike detection algorithm. During a spike, the delay estimate tracks the delays closely, but after a spike, it is similar to Algorithm 1 (with $\alpha = 0.875$ under Normal mode).

Algorithm 4 "spk-delay"

For every packet i received, calculate the network delay n_i
if ($mode$ == NORMAL) **then**
  if ($abs(n_i - n_{i-1}) > abs(\hat{v} \times 2 + 800)$) **then**
    /* detected beginning of a spike */
    $var = 0$
    $mode$ = SPIKE
  end if
else
  $var = var/2 + abs((2n_i - n_{i-1} - n_{i-2})/8)$
  if ($var \leq 63$) **then**
    /* end of a spike */
    $mode$ = NORMAL
  end if
end if

At the beginning of a talkspurt
if ($mode$ == NORMAL) **then**
  $\hat{d}_i = 0.125 \times n_i + 0.875 \times \hat{d}_{i-1}$
else
  $\hat{d}_i = \hat{d}_{i-1} + n_i - n_{i-1}$
end if

There are other more complicated algorithms, which can achieve better spike detection than Algorithm 4, such as those mentioned in [118]. As our purpose here is not to find a better algorithm for spike detection, those algorithms are not covered.

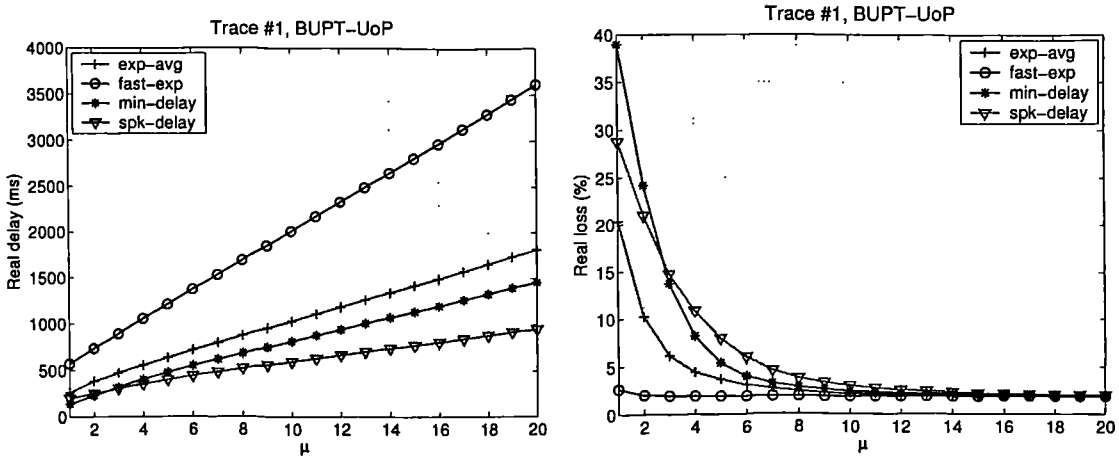
8.2.2 Performance Analysis of Buffer Algorithms

We selected four traces from our collected trace data set for jitter buffer analysis (trace #1 to #4 in Table 7.2, in which traces #1 and #3 are traces with large delay/jitter and traces #2 and #4 are with small delay/jitter).

In the first experiment, we investigated how the buffer algorithm parameters affect perceived speech quality using MOSc metric. We assume no limitations in buffer size and adapt μ in Equation 8.1 from 1 to 20, as in [118]. In comparing with the existing performance metrics, we also include the performance of average playout delay (or real delay) and average loss rate

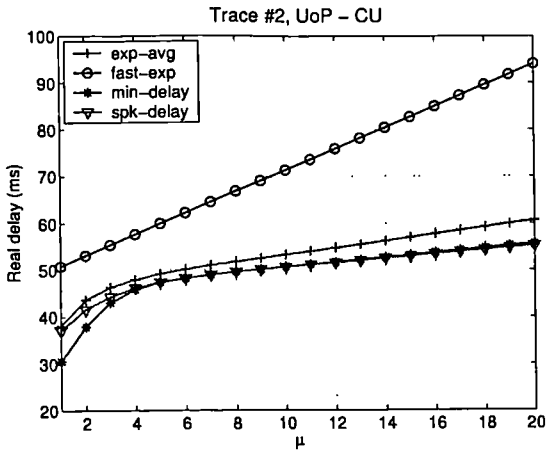
(or real loss).

The real delay and loss vs. μ for traces #1 and #2 are shown in Figure 8.2(a) to 8.2(d), respectively. It is clear that the "fast-exp" algorithm has the lowest real loss rate but the highest real delay for both traces, as it adapts more quickly to increase in delay. The "min-delay" algorithm has the lower real delay and higher real loss for both traces, as it targets at minimum delay. The results for the other two algorithms are between that of the "fast-exp" and the "min-delay".

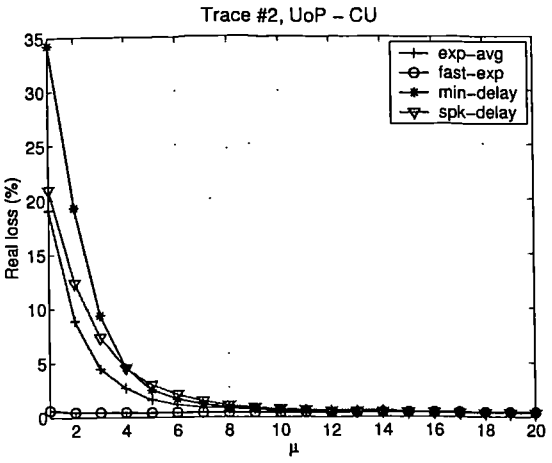


(a) Delay vs. μ for trace #1

(b) Loss vs. μ for trace #1



(c) Delay vs. μ for trace #2

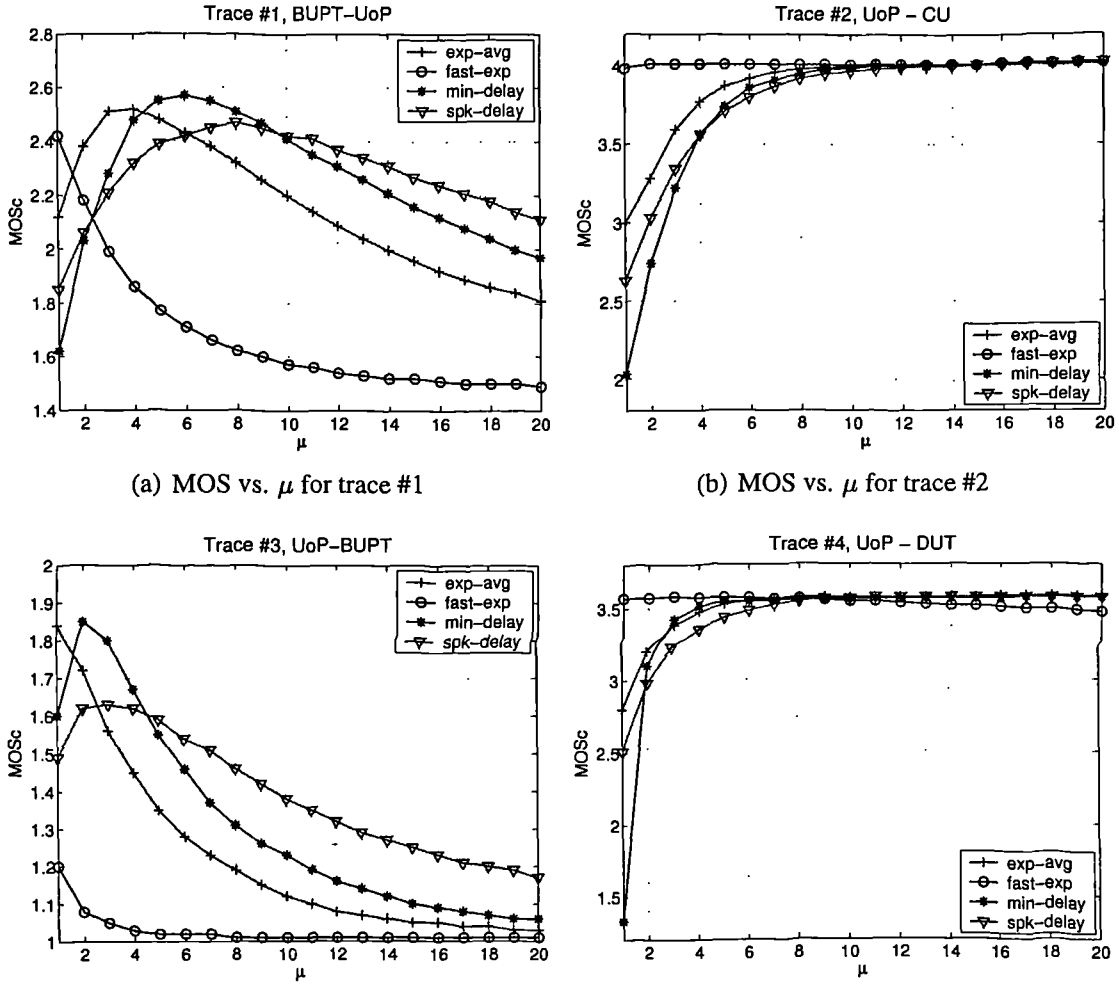


(d) Loss vs. μ for trace #2

Figure 8.2: Performance comparison of playout buffer algorithms for traces #1 and #2

Four buffer algorithms show similar trends at real delay and loss metrics for traces #1 and

#2 (similar results obtained for traces #3 and #4). However, the combined effect on perceived quality shows a big difference for two categories of traces (see Figure 8.3(a) to 8.3(d). There is an obvious similarity within the same category of traces (e.g. trace #1 and #3, trace #2 and #4). This suggests that the perceived performance of the four buffer algorithms for different parameters is mainly affected by the end-to-end delay/jitter of the trace data.



(a) MOS vs. μ for trace #1

(b) MOS vs. μ for trace #2

(c) MOS vs. μ for trace #3

(d) MOS vs. μ for trace #4

Figure 8.3: Performance comparison of playout buffer algorithms for traces #1 to #4

For small delay/jitter traces, the MOSc score can achieve its "optimum" value when μ is set within a proper range for a certain algorithm (e.g. $2 < \mu < 10$ for the "fast-exp" algorithm, and $\mu > 10$ for other three algorithms). The reason behind this is that the end-to-end delay for these two traces does not affect MOSc score, as the overall end-to-end delay is near or less

than 100 ms, with the I_d in Equation 5.7 on Page 79 near to zero. In this case, MOSc is only affected by packet loss and codec.

The performance of four algorithms differs slightly for traces #1 and #3 (see Figure 8.3(a) and 8.3(c). It seems that the "min-delay" algorithm can reach the maximum MOSc value for both traces #1 and #3 at different μ values (e.g. μ=6 for trace #1 and μ=2 for trace #3). This maximum MOSc score represents the best overall tradeoff between delay and loss for the selected traces. As the two traces have both large end-to-end delay (over 100ms), delay has a major effect on the perceived speech quality. The "min-delay" algorithm can achieve its good performance as it induces lower buffer delay among the four algorithms. For the "exp-avg" and the "spk-delay" algorithms, there also exists a μ value to achieve a maximum MOSc score, although this maximum value is lower than that of the "min-delay" algorithm. For the "fast-exp" algorithm, MOSc scores just decrease monotonously with μ increasing. This suggests that the impact on speech quality due to buffer delay induced by this algorithm is much higher than the benefits induced by lower late arrival loss.

8.3 A Novel Adaptive Playout Buffer Algorithm

From the performance analysis on these two categories of traces, we find that there is no 'best' algorithm/parameter, which can always achieve the 'best' MOSc value for all the traces. However, there is a best algorithm among the four, which is more suitable for each category of traces. For example, the "fast-exp" algorithm is preferred for low delay trace/path within a wide range of μ value (μ within 1 to 10), whereas, the "min-delay" algorithm seems better for a longer delay trace/path under a certain μ value (μ=6 for trace #1 and μ=2 for trace #3). It suggests that a different algorithm or μ value should be chosen for different traces to achieve an "optimum" perceived quality. Based on this, we propose a modified buffer algorithm, which can adapt to the preferred algorithm (e.g. "fast-exp" or "min-delay") automatically according to the running estimate of mean network delay $\hat{d}_i$. The algorithm (abbreviated as "adaptive") is

as follows:

Algorithm 5 Adaptive Buffer Algorithm ("adaptive")

1: **if** $(\hat{d}_i \geq delay\_threshold)$ **then**
2: $\quad \hat{d}_i = \min_{j \in s_i} \{n_j\}$
3: **else if** $(n_i > \hat{d}_{i-1})$ **then**
4: $\quad \hat{d}_i = \beta \hat{d}_{i-1} + (1-\beta) n_i$
5: **else**
6: $\quad \hat{d}_i = \alpha \hat{d}_{i-1} + (1-\alpha) n_i$
7: **end if**

Considering the impact of delay on MOSc (imperceptible when delay is under 150ms [122]), we first set the *delay_threshold* (mean delay) to 150ms and calculate the MOSc score under different μ values (as before). The "adaptive" algorithm can adapt to the "fast-exp" for traces #2 and #4 and to the "min-delay" for traces #1 and #3 (in most cases). The result is the same as that of their adapted algorithms in Figure 8.3).

In order to see how delay threshold affects MOSc, we also set *delay_threshold* to 170, 190, 210 and 250ms and calculate the MOSc score for trace #3 (its average network delay is 186ms as in Table 7.2). The "adaptive" algorithm swaps between the "fast-exp" and the "min-delay" algorithms according to the change of end-to-end delay. The results for the "adaptive", the "min-delay" and the "fast-exp" algorithms are shown in Figure 8.4. When the threshold is 170ms, the result of "adaptive" algorithm is similar to that of the "min-delay". With the increase of threshold, the results move towards the direction of the "fast-exp" algorithm and the maximum MOSc score becomes lower.

The results suggest that the "adaptive" algorithm can adapt to the best algorithm for four traces to achieve the best perceived speech quality under the selected delay threshold.

We further investigated how to choose and adapt the parameters (e.g. μ value) to keep the buffer algorithm to achieve the "optimum" perceived quality all the time. The stages in the adaptation strategy are as follows:

(1). The best μ value (corresponding to the maximum MOSc score) is searched for each test segment (e.g. 9 sec), and this best μ value will be used in next segment for the calculation

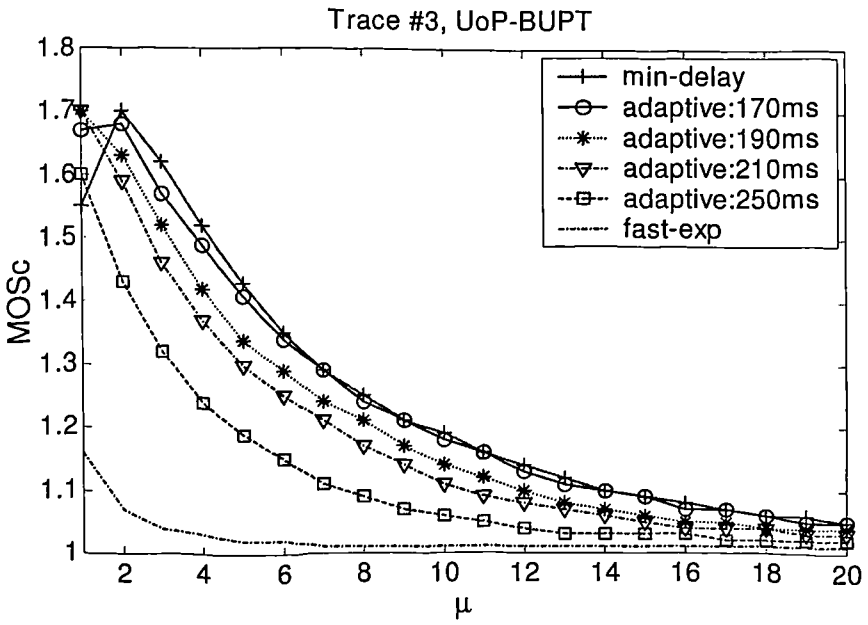


Figure 8.4: Performance comparison for trace #3

of playout time (p_i) in Equation 8.1.

(2). For each segment, also calculate $MOS_{\max}$ which is the maximum $PESQ$ MOS score with only network packet loss.

(3). Search μ ($\mu_{i+1} = \mu_i + 1$, $\mu_{i-1} = \mu_i - 1$), until

$(((MOSc_{u_i} \geq MOSc_{u_{i-1}}) \wedge (MOSc_{u_i} \geq MOSc_{u_{i+1}})) \vee (MOSc_{u_i} = MOS_{\max}))$, then, μ_i is the best one for the segment. For the first segment, the search starts from $\mu = 1$, for other segment, the search starts from the best μ of the previous segment. If $(MOSc_{u_i} = MOS_{\max})$, the lowest μ met this criterion is selected, as this suggests an "optimum" MOSc score with the lowest end-to-end delay.

We implemented this parameter adjustment scheme on the four traces. The preliminary results show that an overall MOSc score increased obviously for traces #1 and #3. For traces #2 and #4, MOSc scores can always stay on $MOS_{\max}$.

8.4 A Perceptual Optimization Playout Buffer Algorithm

8.4.1 Optimum Voice Quality and Minimum Impairment Criterion

In Chapter 5, we have derived non-linear regression models for predicting voice quality for different codecs. These models can be used for perceptual jitter buffer optimization which is presented in this section.

For perceptual-based buffer optimization, the aim is to achieve an optimum end-to-end voice quality (e.g. in the term of MOS score). Considering the relationship of voice quality (in MOS score) and impairments (e.g. packet loss and delay), the problem of an optimum voice quality can be converted to an issue of minimum impairment.

We define an overall impairment function I_m which is a function of delay d and packet loss ρ, with $I_m = f(d,\rho) = I_d + I_{e\rho}$. As described in Sections 5.4 and 5.5.2, the E-model's R factor can be simplified as Equation (8.3) (if ignoring other impairments such as echo).

$$R = 93.2 - I_d - I_e = 93.2 - I_d - (I_{ec} + I_{e\rho}) = (93.2 - I_{ec}) - I_m \tag{8.3}$$

As MOS increases monotonously with R (see Figure 5.3 on page 77), a maximum R value corresponds to a maximum MOS score. Further when maximum R is obtained, it corresponds to a minimum impairment function, I_m.

Using Equation 5.13 on page 88 and Equation 5.7 on page 79, I_m can be further expressed as:

$$I_m = f(d,\rho) = I_d + I_{e\rho} = 0.024d + 0.11(d - 177.3)H(d - 177.3) + a\ln(1 + b\rho) \tag{8.4}$$

where a and b are codec related constants, as shown in Table 5.1 on page 88. d is the play-out delay, including network delay (d_n) and buffer delay (d_b). ρ, the end-to-end packet loss, consists of network packet loss (ρ_n) and buffer loss or late arrival loss (ρ_b).

It is a trade-off between delay and packet loss for any buffer algorithm. When playout delay $d \uparrow$ (delay impairment $I_d \uparrow$), then buffer loss or late arrival loss $\rho_b \downarrow$ (buffer loss impairment $I_{e\rho} \downarrow$). When $d \downarrow$ ($I_d \downarrow$), then $\rho_b \uparrow$ ($I_{e\rho} \uparrow$). An optimum playout delay d can be obtained when minimum impairment I_m is reached.

A minimum impairment criterion for buffer optimization is set and defined in Table 8.1.

Table 8.1: Definition of a minimum impairment criterion

Given:	network delay d_n, network loss ρ_n and codec type
Required to estimate:	an optimized playout delay d_{opt}
Such that:	minimum I_m can be reached

Obviously seeking for a minimum I_m is more efficient than for traditionally seeking for a maximum MOS, as it is not necessary to convert I_m to R and then to MOS (a 3^{rd} order polynomial) for each buffer adaptation/calculation.

8.4.2 Playout Delay and Delay Distribution Modeling

The relationship between d and ρ_b can be described by delay Cumulative Distribution Function (CDF) which is defined as $F(x) = P(X \leq x)$. For a playout delay d, the buffer loss ρ_b can be calculated as $\rho_b = P(X \geq d) = 1 - F(d)$.

To understand the delay distribution for current VoIP traffic, we investigated the delay distribution modeling for the VoIP trace data #1 to #4 in Table 7.2 on page 115. We experimented with Exponential, Pareto and Weibull distributions using Matlab curve fitting tool. The definition of CDF for three distributions are listed in Table 8.2. The RMSE (Root Mean Square Error) for the four selected traces for different approximation models are tabulated in Table 8.3. The empirical and fitted CDF for trace 1 and 3 (two high delay variation traces) are illustrated in Figures 8.5 and 8.6.

Table 8.2: Definition of several cumulative probability distributions

Distribution	Exponential	Pareto	Weibull
CDF: F(x)	$1 - e^{-(x-\mu/\beta)}$	$1 - (k/x)^{\alpha}$	$1 - e^{-((x-\mu)/\alpha)^{\gamma}}$

Table 8.3: RMSE of different distribution functions for different traces

Traces	Exponential	Pareto	Weibull
Trace 1	0.04467	0.03916	0.005607
Trace 2	0.0007858	0.0007389	0.0007233
Trace 3	0.05228	0.03398	0.01064
Trace 4	0.01926	0.02029	0.004269

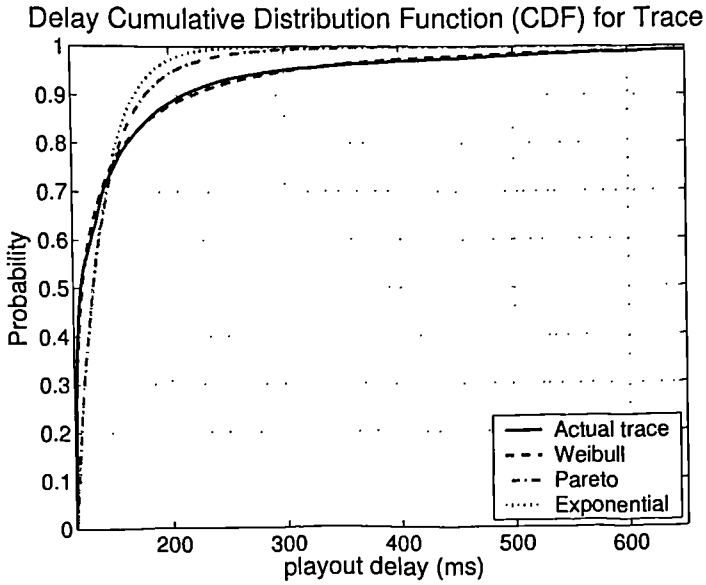


Figure 8.5: Empirical and fitted CDF for trace #1 (Weibull: μ = 116, α = 15.9, γ = 0.4451; Pareto: k = 116, α = 5.277; Exp: μ = 116, β = 23.47)

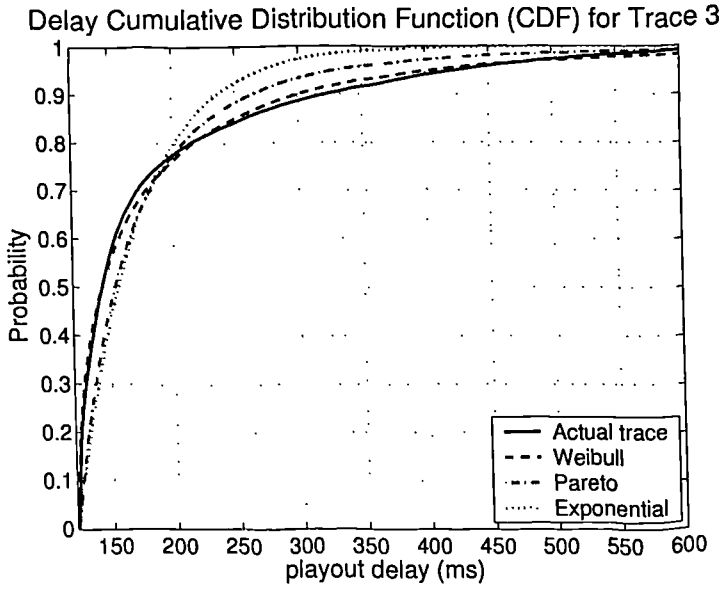


Figure 8.6: Empirical and fitted CDF for trace #3 (Weibull: μ = 122, α = 40.96, γ = 0.5674; Pareto: k = 122, α = 3.004; Exp: μ = 122, β = 49.41)

From Table 8.3, Figures 8.5 and 8.6, it can be seen that Weibull distribution achieved the best fit for all four traces (with the lowest RMSE) when compared with Pareto and Exponential distribution. As a result, we use Weibull distribution to represent delay distribution in the perceptual-based buffer design.

8.4.3 Perceptual Optimization of Playout Delay

Given network packet loss ρ_n (in percentage) and playout delay d, the buffer loss (ρ_b) for a Weibull Distribution can be calculated in the following Equation.

$$\rho_b = (1 - \rho_n/100)P(X \geq d) = (1 - \rho_n/100)e^{-((d-\mu)/\alpha)^\gamma} \tag{8.5}$$

Replacing ρ_b of Equation 8.5 into Equation 8.4, overall impairment factor, I_m, can be depicted as follows:

$$I_m = 0.024d + 0.11(d-177.3)H(d-177.3) + a\ln\left[1 + b[\rho_n + (100 - \rho_n)e^{-((d-\mu)/\alpha)^\gamma}]\right] \tag{8.6}$$

For a given trace segment, the Weibull Distribution location parameter μ equals to the minimum network delay d_n, the scale parameter α and shape parameter γ can be estimated using maximum-likelihood-estimator (MLE) method [123] (we use Matlab's weibfit function directly in the simulation for simplicity). The optimum playout delay (d_{opt}) can be obtained by searching for a playout delay d which meets the minimum impairment criterion. Figure 8.7 shows an example of impairment I_m vs. playout delay d for a trace segment (with 1000 packets) selected from trace #1. In order to see how different codecs and objective measurement methods (e.g. PESQ/PESQ-LQ) affect playout delay optimization, Figure 8.7 also shows the I_m vs. d for AMR122 and iLBC (two codecs with the most different packet loss features as in Figure 5.14 on page 89) using PESQ and PESQ-LQ, assuming with the same network traces. It is obvious that the optimum playout delay differs according to which codec and which objective

perceived quality method are used. The iLBC/PESQ has the smallest optimum playout delay (d_1) and AMR122/PESQ-LQ has the largest one (d_4). The minimum impairment values obtained also differ for different codecs, with iLBC (PESQ) the lowest I_m (the highest MOS) and AMR122 (PESQ-LQ) the highest I_m(the lowest MOS). It is clear that the minimum impairment criterion can be applied to seek an optimized playout delay for a chosen codec (e.g. G.723.1) or objective measurement (e.g. PESQ).

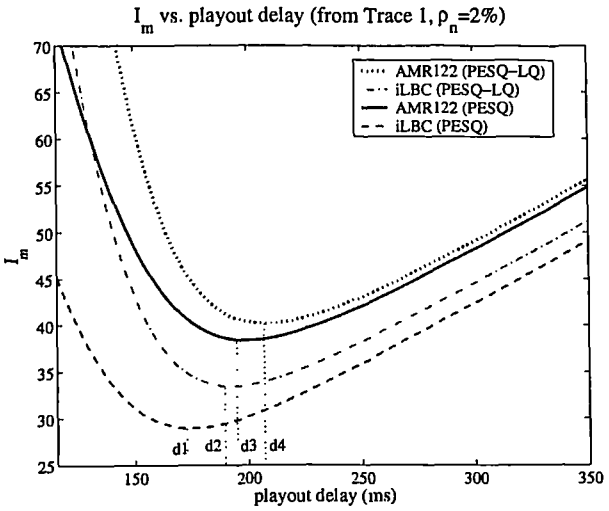


Figure 8.7: Optimization of playout delay

8.4.4 A Perceptual Optimization Playout Buffer Algorithm

In Section 8.4.1, we have derived Equation 8.4 which relates impairment (I_m) with playout delay (d) and network packet loss (ρ_n) for a given trace. This can be used directly for perceived jitter buffer algorithm optimization. For simplicity, we only use the equation for G.723.1 codec to show the concept of perceptual optimization buffer design.

As network traces show high possibility of "spike" which is defined as a number of packets that have significantly higher delays than the rest. The "spike" state can be regarded as an exceptional state in the trace data (seen as a short-term delay characteristics) and the remaining

"non-spike" state can be analysed in long-term delay distribution. Several algorithms exist for spike detection. For example, Ramachandran et al [117] proposed to use $(n_i - n_{i-1}) > threshold$ (see Algorithm 4 on page 131) as the detection of a start of a spike (n_i is the network delay for i^{th} packet). This accounts for the spike with a sudden increase of delay. However through the analysis of our collected Internet trace data, we notice that large amounts of spike is with gradual increase which cannot be detected by the above algorithm. Considering spikes with sudden or gradual increase, we follow the spike detection based on $(n_i > threshold)$ as in [118]. The proposed perceptual optimum buffer algorithm (P-optimum) is shown in Algorithm 6.

Algorithm 6 Perceptual Optimum Buffer Algorithm ("P-optimum")

For every packet i received, calculate the network delay n_i
if $mode$ == SPIKE **then**
  if $n_i \leq tail \times old\_d$ **then**
    /* the end of a spike */
    $mode$ = NORMAL
  end if
else if $n_i > head \times d_i$ **then**
  /* the beginning of a spike */
  $mode$ = SPIKE
  /* save d_i to detect the end of a spike later */
  $old\_d = d_i$
else
  /* normal model*/
  - update delay records for the past W packets
end if

At the beginning of a talkspurt
if $mode$ == SPIKE **then**
  /* estimated playout delay d_i */
  $d_i = n_i$
else
  /* normal mode */
  - obtain (μ, α, γ) in Weibull distribution for the past W packets
  - search playout delay d for $d_i = d_{opt}$ which meets $\rightarrow min(I_m)$
end if

Depending on the current mode, the playout delay for the next talkspurt is estimated dif-

ferently in each mode as shown in Algorithm 6. In spike-detection mode, the delay of the first packet of a talkspurt becomes the estimated playout delay for the talkspurt. Otherwise, the perceptually optimized playout delay based on the delay distribution of the last W packets (in *NORMAL* mode) is used. The large the W value, the less responsive the scheme to adapt. The *head* and *tail* parameters are used to set the threshold for spike detection.

8.5 Performance Analysis and Comparison

In order to compare with other jitter buffer algorithms, we also implemented "exp-avg", "fast-exp", "min-delay", "spk-delay" and "adaptive" algorithms (with different threshold). The results are shown in Table 8.4 for the above mentioned four traces. The window size W is set to 1000. The *head* is 4 and the *tail* is 2, as suggested in [118]. During the experiment, we changed the window size W from 100 packets (3 sec) to 10,000 packets (300 sec, as suggested by [118] and [120]), we noticed that the performance (the overall MOS score) does not show a big difference within the range. We chose W of 1000 (30 sec), as it is an appropriate duration for the I_m or MOS calculation and has higher computation efficiency than the longer window length.

From Table 8.4, it can be seen that "P-optimum" algorithm obtained almost the optimum MOS scores among all the five traces. Our previous proposed "adaptive" algorithm achieved sub-optimum results. The remaining buffer algorithms can achieve good results only in some traces, but not for all. It has to be mentioned that P-optimum has the highest complexity, whereas the others including "adaptive" have the similar low complexity.

Table 8.4: Performance comparison for different buffer algorithms

Trace	Buffer algorithms	Loss ρ (%)	Delay d (ms)	MOS
Trace 1	Exp-avg	4.9	298.5	2.01
	Fast-exp	1.5	750.8	1.00
	Min-delay	9.4	208.8	2.34
	Spk-delay	10.4	225.0	2.18
	Adaptive	9.0	208.1	2.37
	P-optimum	10.5	188.2	2.43
Trace 2	Exp-avg	1.8	27.3	3.28
	Fast-exp	0	35.9	3.44
	Min-delay	1.7	27.3	3.29
	Spk-delay	3.4	24.9	3.15
	Adaptive	0	35.9	3.44
	P-optimum	0.1	44.5	3.42
Trace 3	Exp-avg	18.2	432.4	1.01
	Fast-exp	14.3	1408.6	1.00
	Min-delay	22.1	312.7	1.30
	Spk-delay	23.8	325.4	1.22
	Adaptive	22.1	299.8	1.35
	P-optimum	32.0	171.1	1.80
Trace 4	Exp-avg	5.9	24.0	2.97
	Fast-exp	4.3	94.4	2.99
	Min-delay	5.3	23.0	3.01
	Spk-delay	7.6	21.9	2.86
	Adaptive	4.3	72.8	3.02
	P-optimum	5.1	34.4	3.02

8.6 Summary

In this Chapter, the performance of different existing buffer algorithms is analyzed using the proposed voice quality prediction methods (in terms of MOSc score) for the newly collected Internet trace data. Results show that end-to-end delay/delay variation, in general, has a major effect on the selection of buffer algorithms/parameters. For large to medium end-to-end delay/delay variation, a buffer algorithm that aims for a minimum delay is preferred, whereas, for small end-to-end delay/delay variation, an algorithm that targets a minimum loss is better. Based on this, a new adaptive buffer algorithm has been proposed. Results show that it can achieve a better perceived quality for all the traces considered. Further, the perceptual optimized jitter buffer algorithm has been investigated. The minimum overall impairment is used as a criterion for buffer optimization. This criterion is more efficient than using traditional maximum Mean Opinion Score (MOS). It is also shown that the delay characteristics of Voice over IP traffic is better characterized by a Weibull distribution than a Pareto or an Exponential distribution. Based on the nonlinear regression models for voice quality prediction, the Weibull delay distribution model and the minimum impairment criterion, a perceptual optimization playout buffer algorithm has been proposed and performance is compared with other jitter buffer algorithms. Preliminary results show that the proposed perceptual optimum buffer algorithm can achieve the optimum perceived voice quality compared with other algorithms under all network conditions considered. The adaptive algorithm can achieve sub-optimum perceived voice quality with low complexity.

As the work is based on the buffer adaptation at the beginning of each talkspurt, it cannot adapt to any delay changes during a talkspurt. Future work can extend the idea to consider buffer adaptation during a talkspurt [124] in order to achieve a best trade-off among delay, loss and end-to-end jitter.

Chapter 9

Perceived Speech Quality Prediction for QoS Control

9.1 Introduction

In Chapter 8, the application of perceived voice quality prediction for playout buffer optimization has been presented. In this Chapter, the application of perceived voice quality prediction in Quality of Service Control is investigated. Here, the perceived speech quality is used as a control metric to control the send behavior (i.e. the sender bit rate of codecs) instead of using traditional individual network parameters (e.g. packet loss, jitter or delay) for QoS control. Further a combined control scheme which combines the strength of adaptive bit rate control and priority marking control is investigated.

QoS control mechanisms for VoIP should aim to make optimum use of available network/terminal resources and to minimise the effects of network impairments on voice quality. Several approaches exist to realise QoS control, but most seek to control the information flow from the audio/video sources, adaptively, in accordance with significant changes in the network. An important class of QoS control technique involves rate control (i.e. QoS control is achieved by automatically adjusting the send bit rate depending on network congestion conditions). However, current rate control mechanisms [125, 126, 127] are based largely only on network impairments such as packet loss rate or delay during congestion. The strategy is

to control the sender behaviour, using the network impairments, from the receiver or the network node but this may not be sufficient to provide optimum QoS, in terms of the voice quality delivered, because the control information is directly linked to user perceived quality.

A second important class of QoS control techniques exploits knowledge of the fact that different parts of speech have different perceptual importance and so do not contribute equally to the overall voice quality [128, 129]. In this approach, voice packets that are perceptually more important are marked, i.e. given priority, and so are less likely to be dropped than packets that are of less perceptual importance, if there is congestion. The priority marking based QoS schemes are open loop and do not make use of changes in the network impairments.

The main objective of this Chapter is to investigate the possibility of combining rate adaptation control technique with priority marking, to exploit the advantages of the two approaches to provide a robust control scheme which delivers optimum QoS in terms of voice quality. In rate control schemes, the cost of adapting the data flow to changes in the network is that some packets may be dropped randomly when congestion occurs and this will increase the packet loss rate. However, in priority marking schemes important packets are dropped less and delayed less. Thus, the combined scheme should provide improved overall user perceived quality. Differentiated Service (DiffServ) architecture [31] is used to implement the scheme and employs different queuing methods, the most important of which is a variation of random early drop queue (RED queue). RED not only gives different packets different drop probabilities, it also gives the receiver hints about whether congestion has occurred or is about to occur. With a proper feedback mechanism, this information can be used to control the send bit rate.

The main contributions of this Chapter are twofold. First, we propose a new QoS control scheme that combines the strengths of the adaptive rate control technique and speech priority marking QoS technique to provide a superior QoS control performance than hitherto possible. Second, we propose the use of an objective measure of perceived speech quality (i.e. objective MOS score [22]) instead of individual network impairments (e.g. packet loss and/or delay) to control sender behaviour as this provides a direct link to user-perceived speech quality.

Preliminary results show that by exploiting the strengths of both methods, the new scheme achieved the best perceived quality compared to rate-adaptive, marking and no control schemes under different network congestion conditions. The results are based on extensive simulation in an environment that integrates the NS-2 network simulator [130], adaptive speech codec (the AMR codec [16]) and an objective perceived speech quality measurement system, which is based on the ITU-T speech quality evaluation standard [4].

In this Chapter, Section 9.2 briefly introduces the AMR codec and its features under network packet loss. Section 9.3 presents the three QoS control schemes – the rate-adaptive, the priority marking and the new combined QoS control schemes. The simulation system and experiments are described in Section 9.4. The results and analysis are given in Section 9.5, and Section 9.6 summarises the chapter.

9.2 Adaptive AMR Codec and Its Speech Quality Under Packet Loss

AMR (Adaptive Multi-Rate) speech codec was developed by ETSI and has been standardized for GSM. It has been chosen by 3GPP as the mandatory codec. The AMR is a multi-mode codec with eight modes (MR475 to MR122) with bit rates between 4.75 to 12.2 Kb/s. Mode switching can occur at any time (frame-based). We first analysed the AMR codec's speech quality under different packet loss conditions using the ITU PESQ algorithm. The relationships between MOS (obtained with PESQ) and packet loss rates (between 0 to 5%) for the eight AMR modes are shown in Figure 9.1. In order to avoid the influence of packet loss locations [18], the MOS scores were obtained by averaging over 50 different loss patterns (50 different seeds in random number generation) for each loss rate (Bernoulli loss model used for simplicity). A mixture of male and female speech sample was chosen to minimise the influence of gender.

From Figure 9.1, it can be seen that the difference in perceptual quality between the highest

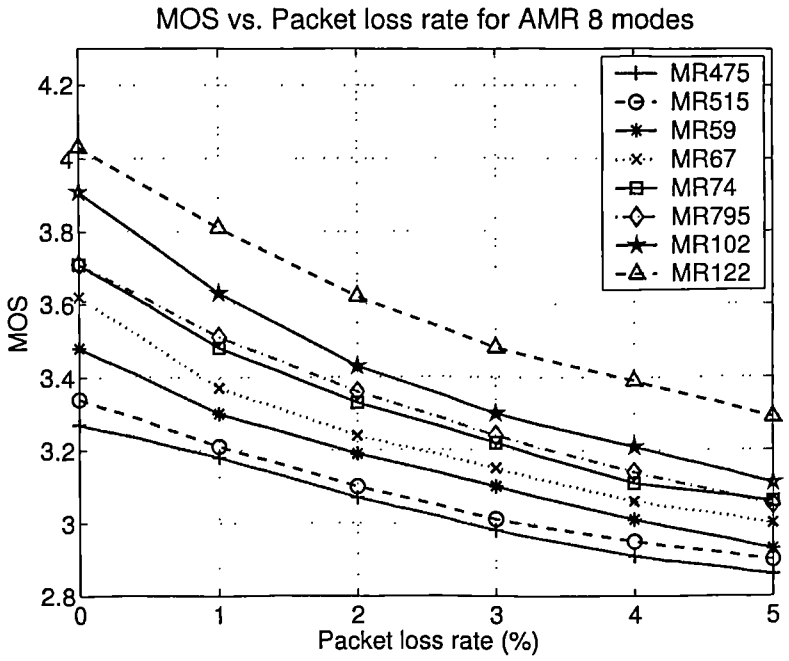


Figure 9.1: MOS vs packet loss rate for AMR eight modes

mode (MR122, 12.2 kBit/s) and the lowest mode (MR475, 4.75 kBit/s) is approximately 0.5 points on the 5-point MOS scale. A similar level of degradation occurs when increasing the packet loss rate from 0 to 3% for the same AMR mode. Roughly speaking, a reduction in loss rate of about 3% (by reducing the bit rate) is accompanied by an improvement in the overall perceived speech quality. This forms the basis for adjusting AMR codec sender bit rate to support perceived QoS.

Besides the feature of variable sender bit rate which enables it to adapt to different network conditions, the AMR codec also shows the perceptual difference for packet loss under different locations due to its built-in concealment algorithm. In Section 4.3, we showed that loss location has a severe effect on the perceived speech quality for the AMR codec (as well as G.729 and G.723.1 codecs). Loss at unvoiced speech segments has little impact on perceived quality. However, loss at the beginning of voiced segments has the most severe impact on speech quality. This suggests that the overall speech quality can be improved under the same network congestion conditions if perceptually important speech segments (e.g. those at the beginning of voiced segments) can be protected (e.g. by being given a higher priority marking). This

provides a basis for priority marking control scheme.

9.3 QoS Control Schemes

9.3.1 Rate-adaptive QoS Control Scheme

The adaptive rate QoS control scheme based on AMR codec is shown in Figure 9.2. In the scheme, the send rate of the AMR codec is adjusted in accordance with the network conditions to achieve the best possible QoS. The bit rate control mechanism is based on individual network parameters (e.g. packet loss rate and delay) or on the predicted, perceived speech quality (e.g. MOS score). In VoIP applications, the feedback information can be sent via RTCP reports. The specification of RTP [33] stipulates that the RTCP traffic does not exceed 5% of the whole traffic and that the time between the reports is at least 5sec.

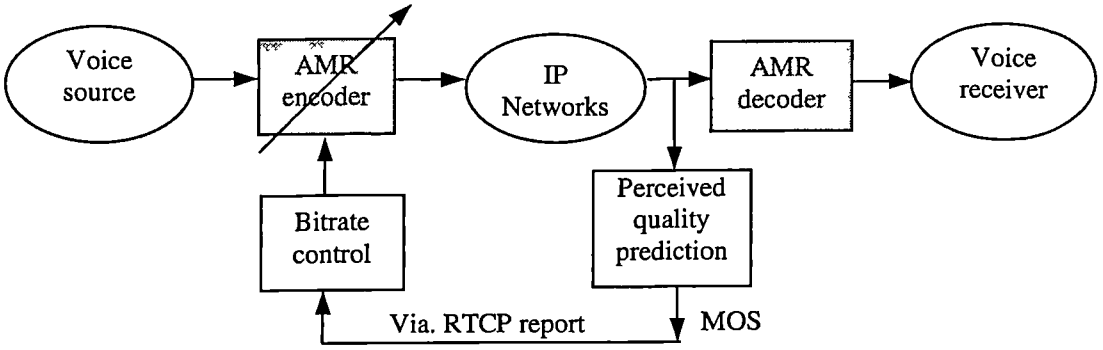


Figure 9.2: Rate-adaptive QoS control scheme

The two important modules in Figure 9.2 are the bit-rate control module at the sender side and the perceived speech quality prediction module at the receiver side. Approximately, every 5sec (the time interval between RTCP reports), the perceived speech quality (i.e. MOS score) is predicted from network parameters (e.g. packet loss and delay) by a combined PESQ/E-model based method which is described in Section 5.2.

The bit rate control module is used to adapt the sender bit rate in accordance to the feedback information. The adaptive algorithm used in the module follows the 'additive increase/

multiplicative decrease' concept that has been successfully employed in other, such as TCP or ABR [127]. The basic idea is that the AMR codec can reduce its bit-rate (if possible) when there is a network congestion and increase its bit-rate when no congestion is detected. The predicted MOS score is compared with the existing MOS and the controller will choose the best step to change or keep existing AMR rate. The control mechanism used is presented in Section 9.4.

9.3.2 Priority Marking QoS Control Scheme

In rate-adaptive QoS control scheme, it is assumed that all the packets within a flow are equally important. Previous research [128] has shown that some speech segments are more important than others. This phenomenon has been proven to exist for codecs such as G.729, G.723.1 and AMR (detailed analysis see Section 4.3). This forms the basis for the priority marking control scheme as shown in Figure 9.3. Each speech frame is marked differently depending on whether it is perceptually important or not. For example, the priority-marking module marks the beginning of a voiced segment (e.g. the first 5 or 10 frames of a voiced segment for the AMR codec) as high priority (e.g. marked as a 'premium' class), while others are marked as perceptually unimportant (e.g. marked as a 'best-effort' class). When there is network congestion, the perceptually unimportant frames have a higher drop probabilities. This protection scheme results in a lower loss probability for packets with high priority and can lead to a better perceived QoS compared to no-control scheme.

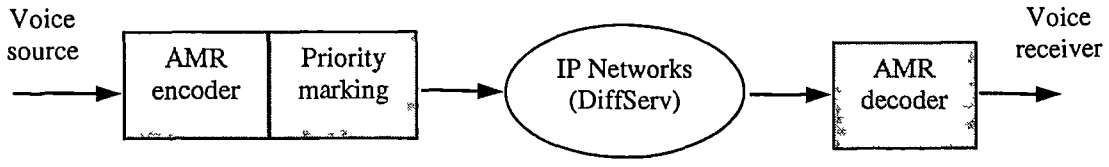


Figure 9.3: Priority marking QoS control scheme

Priority marking QoS control scheme can be implemented in networks that support Differentiated Services (DiffServ) architecture [31], e.g. a simplified 2-bit marking DiffServ imple-

mentation [131].

9.3.3 Combined Rate-Adaptive and Priority Marking QoS Control Scheme

As discussed above, the rate-adaptive QoS control scheme is based mainly on an objective MOS score at the receiver and packet loss contributed significantly to the measured or predicted MOS score. To reduce the long delay caused by simple over buffered drop tail queue, the network operators commonly use the RED queue or a similar queue management method to provide a better congestion notification and control. The use of RED queue method in the system makes it logical to try to link send rate adaptive control with packet priority marking because it is relatively easy to set different queues or virtual queues in a RED queue management system to provide different treatment for different priority packets.

Priority marking should reduce the loss or delay of important packets. An important goal is to investigate whether the overall perceived speech quality can be improved further by combining rate-adaptive and priority marking control schemes. This is the motivation of the proposed combined QoS control scheme, which is shown in Figure 9.4.

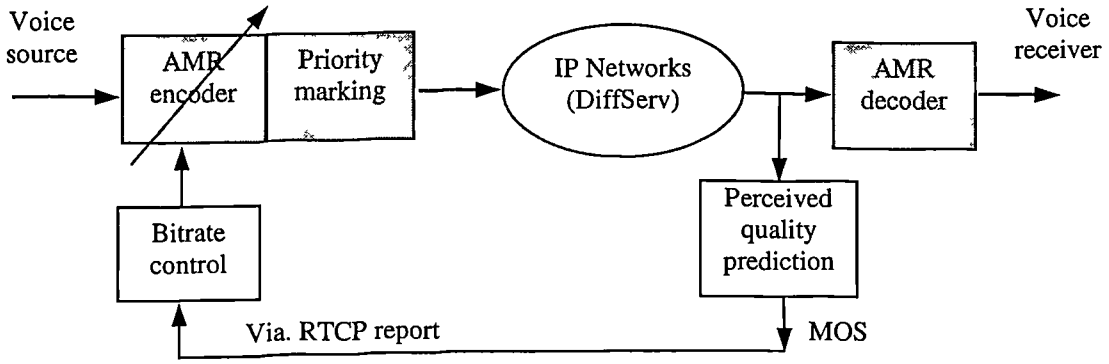


Figure 9.4: Combined QoS control scheme

As shown in Figure 9.4, the bit rate of the AMR codec is adjusted in accordance with the objective, predicted MOS, and, at the same time, the perceptually important segments of speech are protected by priority-marking. Potentially, this should make it possible to optimize

the perceived speech quality for VoIP applications using AMR codec.

9.4 Simulation Systems and Experiments

9.4.1 Simulation System

The combined QoS control scheme was set up as shown in Figure 9.5. This consists of three main parts: (I). a NS-2 network simulator to simulate multiple VoIP flows and IP networks with congestion; (II). a VoIP simulation system to simulate a VoIP flow, which includes an AMR encoder/marker, loss simulator, decoder, and control modules, and (III) a perceived quality evaluation system to provide a measure of the overall speech quality and quantify the performance of each control method.

We simulated a simple bottleneck network topology using NS-2, as shown in Figure 9.5(I). A total of N adaptive AMR sources were simulated for VoIP traffic (this assumed that the available bandwidth was shared among these UDP sources). All the sources were set as constant bit rate (CBR) UDP source (in order to match with the simulation of VoIP flow in part (II), as VAD for AMR codec was not activated there). The sender bit rate (plus header) was set according to the required bit rate for adaptive AMR codec (CBR source can change the send rate by request but still using the name CBR). All flows sent by traffic source was traced and the loss information collected and sent back to the loss simulator in part (II). A VoIP flow was simulated via encoder/marking, loss simulator and decoder. The loss information was also sent to the quality prediction module to obtain a MOS score. The MOS score was then fed back to the send side for bit rate control. A single hop of 2Mbit/s bandwidth, representing a bottleneck link, was set in a DiffServ enabled IP network. With the increase in the number of simultaneous users sharing the bottleneck link, we were able to investigate the performance of different QoS control methods under different network congestion situations. The overall performance of each of the different QoS control methods is evaluated by the evaluation system in party (III).

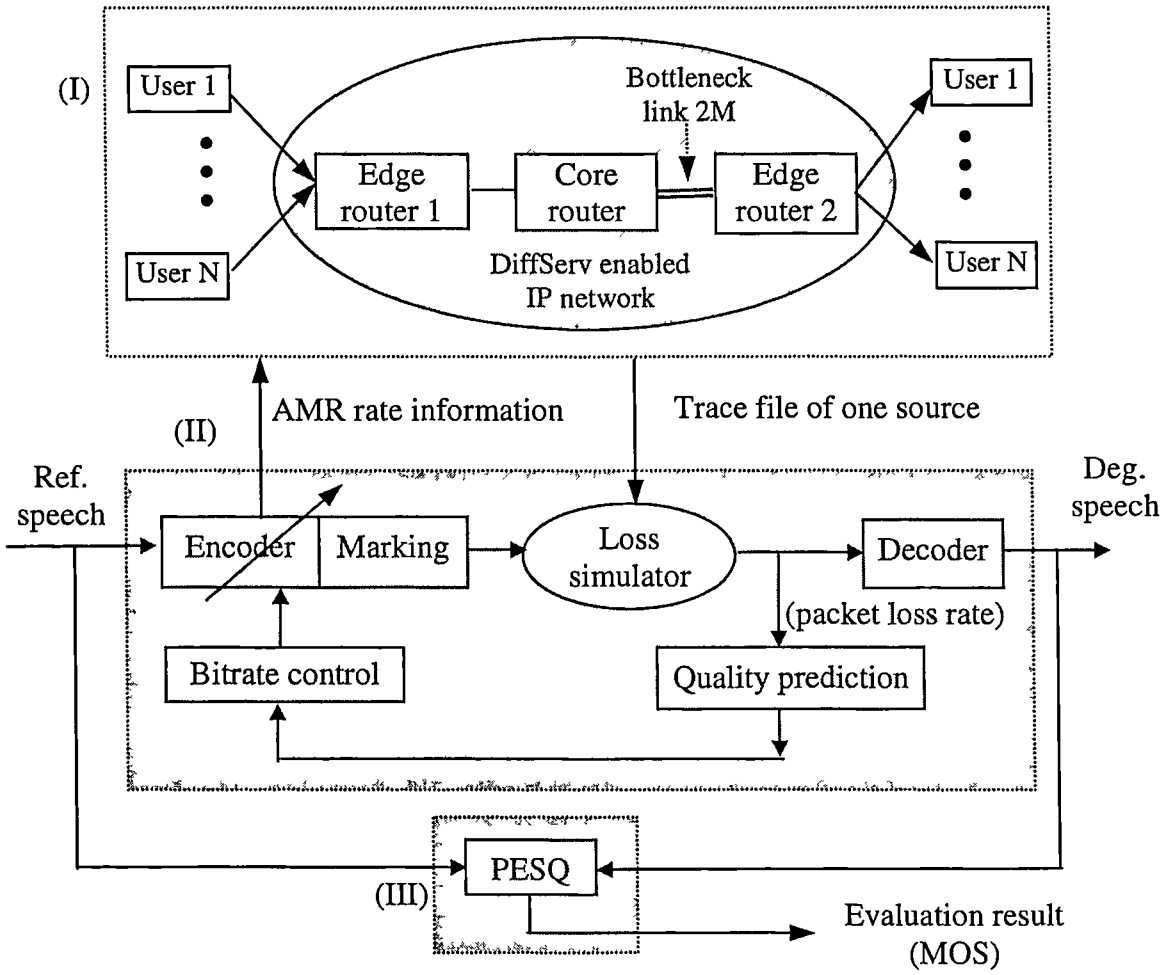


Figure 9.5: Simulation system for combined QoS control scheme

Perceived speech quality prediction is based on the ITU PESQ (for simplicity, only PESQ MOS score is used directly in order to prove the concept). For every loop (e.g. every 5 seconds), speech quality (MOS) is calculated in the 'quality prediction' module, depending on the network packet loss and AMR mode. The measured speech quality (e.g. using PESQ) is also used to evaluate the overall quality of the control schemes as shown in Figure 9.5 (III).

9.4.2 Priority Marking and Loss Simulation

Every frame generated from the AMR encoder was marked as perceptually important or unimportant, depending on the information from the AMR coder. In the simulation, the parameter that indicates whether it is voiced/unvoiced for each frame was extracted directly from the decoder's voiced_hangover flag for simplicity. References [128] and [129] give a more detailed explanation of packet marking.

The priority-marking scheme can be readily implemented in DiffServ supported networks. DiffServ is implemented in NS-2 version higher than NS2.1b8a and our simulation used NS2.1b9a to support the DiffServ simulation part [132]. For simplicity, the DiffServ policer used for the simulation is a Time Sliding Window with 2 colour marking policer (built-in function supported by NS-2). It uses CIR (Committed Information Rate) and a drop precedence of two levels. The basic idea is that a lower precedence is used when the CIR is exceeded. The default scheduling mode is Round Robin.

9.4.3 Perceived Speech Quality Driven Rate-adaptive Control Simulation

The bit rate control module aims to detect the optimal bit rate settings that would yield the best perceived speech quality under a given network condition. We use perceived speech quality MOS score as a control metric to drive the control mechanism. MOS is predicted at the receiver side for each RTCP interval (e.g. 5 seconds) and then sent back via the RTCP report.

A predicted MOS is then calculated and compared with reported MOS. If the MOS after rate adaptive control are predicted better, the rate will be changed otherwise it will stay unchanged. This mechanism can get the balance between rate reduction and congestion decrease. A MOS driven rate adaptive control loop pseudo code is shown in Figure 9.6.

```
(1)  For each RTCP report
(2)  { bitrate_old = bitrate_new;
(3)  MOS_old = MOS_new; //get the new status
(4)  obtain MOS_new from RTCP report;
(5)  // compare MOS scores
(6)  if (MOS_new > MOS_max) goto NOCHANGE
(7)  else if ((MOS_new − MOS_old > th1) && (bitrate_old ! = bitrate_max))
(8)       bitrate_new = next_higher_bitrate ; //increase the bit rate
(9)  else if ((MOS_old − MOS_new) > th2)
(10)      bitrate_new = next_half_lower_bitrate; //halve the bit rate
(11) else if (( th1 < (MOS_old − MOS_new) < th2 ) && (bitrate_old ! = bitrate_min))
(12)      bitrate_new = next_lower_bitrate; //decrease the bit rate
(13) else
(14) NOCHANGE: bitrate_new = bitrate_old; // no change of bit rate
(15)
(16) //Predict MOS after rate change
(17) obtain MOS_predicted from PESQ{AMR_rate,lossrate_predicted}
(18) if (MOS_new > MOS_predicted)
(19) bitrate_new = bitrate_old; // no change of bitrate
(19) else send bitrate_new to sender; //control the encoder
(20) }
```

Figure 9.6: Control loop pseudo code

In the simulation, MOS prediction is based on the ITU PESQ measurements for a given AMR rate and packet loss rate as shown in Equation 9.1 for simplicity. This can be improved by directly using regression models/method developed in Chapter 5.

$$MOS = PESQ\{AMR\,rate, loss\,rate\} \tag{9.1}$$

The predicted packet loss in Equation 9.1 is based on the following:

$$lossrate = (MR \times N - BW)/(MR \times N) \times 100\% \tag{9.2}$$

Where loss rate is the predicted packet loss rate for next control step, MR is the next step AMR transmission rate (RTP/UDP/IP header inclusive [†]), N is the number of users and BW is the bandwidth they are shared. This is a simplified equation and does not consider the effect of a limited buffer size and the distribution of arriving packets, but this will not affect the main control idea using predicted MOS score.

In the simulation, the threshold1 ($th1$) in Figure 9.6 was set to 0.2 in order to avoid unnecessary fluctuation, and the threshold2 ($th2$) was set to 0.5 to indicate an obvious decrease in perceived speech quality. The maximum MOS (MOS_max) score achievable for AMR codec was set to 4. For the AMR codec, the maximum bit rate, $bitrate\_max$ was set to 12.2Kbit/s and the minimum bit rate, $bitrate\_min$, was set to 4.75Kbit/s. For every control loop, the modified sender bit rate was sent back to NS-2 simulator to adjust the source bit rate. For simplicity, we assume that all N sources in NS-2 use the same AMR sender bit rate at the beginning and are adjusted to the same bit rate when adaptation occurs.

9.4.4 Simulation of Combined Rate-adaptive and Priority Marking Method

The modules described above can be integrated to support the simulation of the new combined quality of service control method. Each user's packets are traced and recorded for evaluation. The packet size and packet loss information is used to process a reference speech to get a degraded speech. The degraded speech is then compared with reference speech using PESQ to get the evaluation result. The results of the simulation are discussed in the next section.

9.5 Results and Analysis

In order to investigate how the QoS control schemes affect perceived speech quality under different network conditions, we simulated different network congestion scenarios using the

[†]E.g. for AMR 12.2 Kb/s, the transmission rate is $(40 \times 8 + 244)/20 = 28.2Kb/s$.

NS-2 network simulator. The bandwidth of the bottleneck link was set to a fixed value (2Mbit/s) with a delay of 1ms. The number of the streams sharing the link was increased from a small number to a large number to simulate different congestion scenarios. The starting point was 70 streams sharing the bottleneck when there is no congestion at all. The number of users was increased from 70 to 140 in steps of 5. By reaching 140 users, almost every stream suffered from a very high loss rate and all the control methods were unable to cope with the impairments well. (Packet loss rate for non-control scheme was measured more than 40%). The latest investigation about PESQ method's performance in high packet loss situation [71] suggested that MOS score is much lower from PESQ result so we stopped increasing the user number after 140 as the result is meaningless.

In order to compare the performance between the different QoS control schemes and a 'no control scheme', we also implemented the priority marking, the rate-adaptive and no control schemes. For the priority marking and no control schemes, the send bit rate of the AMR codec was set to a fixed mode (12.2 Kb/s). For rate-adaptive-only control method, the bottleneck link was set to a non-DiffServ link with the same delay parameters and the rest of the system remained the same. The simulation was carried out using the same scenarios as described previously. The number of simultaneous users was increased from 70 to 140 as before. Figure 9.7 compares the results for all four schemes.

The results show that for 70 simultaneous users, i.e. when there is no congestion, all four methods have the same performance. The MOS scores represent the highest score obtainable from an AMR codec.

In general, as shown in Figure 9.7, the drop of the speech quality follows the similar pattern for all four schemes because they were all suffering from the packet loss occured in the bottleneck link (see Figure 9.5). For the adaptive rate control scheme, the drop of speech quality is less steep compared with the "non-control" scheme. This is because the MOS driven rate adaptation can choose the best-optimised AMR rate to minimise the affect of codec rate decrease and packet loss increase. For the priority-marking scheme, the improvement over the

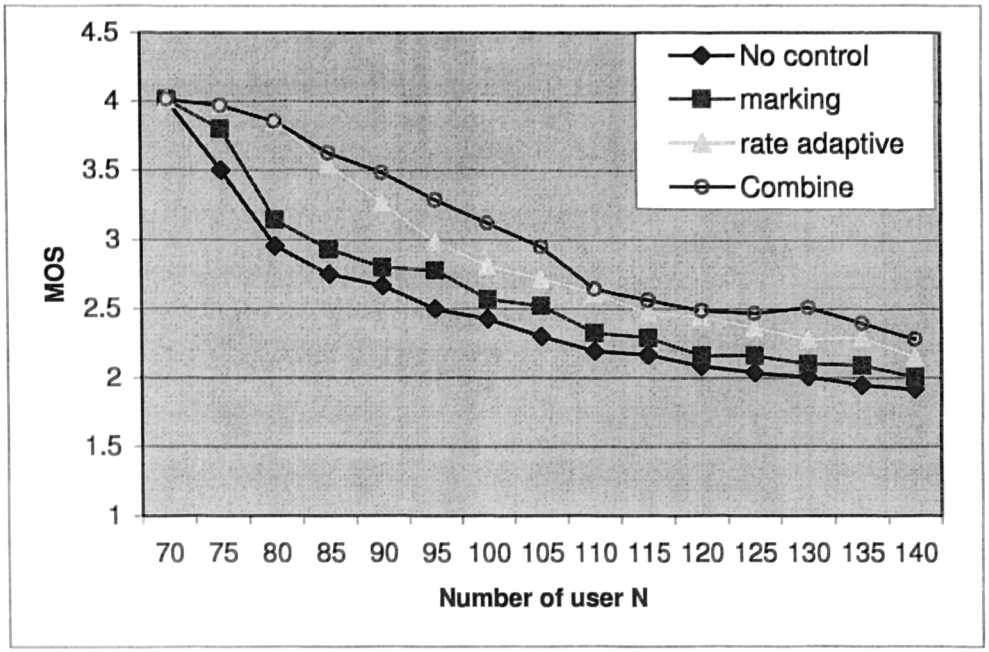


Figure 9.7: MOS vs. Number of user N for different control and non-control schemes)

non-control scheme is stable although not very significant. This is because although the Diff-Serv method can be used to treat different packets with different priority (i.e. loss rate), higher priority packets still have chance to be dropped, especially when the congestion is higher than CIR. From the figure, the performance of the new combined scheme is always better than those two different control schemes and the non-control scheme.

9.6 Summary

In this chapter, a new QoS control scheme has been proposed which combines the strengths of rate-adaptive and priority marking QoS control schemes and uses a predicted objective measure of speech quality as a control parameter. We investigated perceived speech quality for different QoS control schemes by integrating NS-2 network simulator with a real adaptive speech codec (the AMR codec) and a perceived quality evaluation system based on the ITU PESQ algorithm. We used the predicted perceived speech quality metric (measured by PESQ), instead of individual network parameters, to control the AMR codec's send bit rate. Preliminary

results show that the new control scheme achieved the best perceived speech quality compared with rate-adaptive, priority marking and no control schemes in different network congestion conditions.

In future, the investigation can be extended to include the application of the combined control scheme in a TCP/UDP mixed environment. The effects of delay in the DiffServ model and the use of conversational speech quality (instead of listening quality) as metric to control AMR rate can be studied.

This work has led to another PhD project because it is of a major research interest in its own right.

Chapter 10

Internet-based Subjective Speech Quality Measurement

10.1 Introduction/Motivation

From Chapter 5 to Chapter 9, the work is around the development of novel objective non-intrusive speech quality prediction methods/models and their applications in voice quality monitoring/prediction, buffer optimization and QoS control. As mentioned before, speech quality can be measured using either subjective or objective methods. Subjective measurement (e.g. MOS) is the benchmark for objective methods, but it is time consuming, and expensive. In this chapter, the existing subjective test methods are investigated and an efficient Internet-based subjective test methodology is proposed.

The traditional MOS test methodology has been in existence for about 20 years [133] and today its uses range from the assessment of codec quality to the assessment of VoIP network quality. The stringent test requirements for traditional tests have not changed (e.g. the use of a sound-proof room) in that time and are essential for a proper assessment of voice quality in many cases, e.g. quality assessment of codecs, as the difference between codecs may be subtle and difficult to detect. However, for VoIP applications, new impairments, such as packet loss, are much more perceptible than impairments from codecs. This has led us to investigate the possibility of conducting MOS tests under normal working/studying environments, as this is

more realistic and subjects are more relaxed. In a sound-proof room, some subjects may find it uncomfortable, psychologically, to carry out tests in the confined environments. This has led to an Internet-based subjective test methodology, which has the following advantages:

- It is closer to reality than the traditional method. Subjects remain in familiar environments, e.g. an office or a laboratory, to carry out the test. This is clearly less stressful and the test can be done at the subject's own pace.
- It is possible to organise subjective tests at more locations around the world.
- It allows easier access to a larger number of subjects (e.g. 40 - 80 subjects can be tested at the same time in one or two large rooms, e.g. a laboratory).

Overall, it has the benefits of efficiency, realism, wide access and ease of organisation. It can save money and time compared to P.800 [3]. Of course, the main disadvantage of Internet-based MOS test is the lack of a controlled testing environment (e.g. very low background noise) compared to P.800.

Two series of Internet-based MOS tests are carried out. The first one is without control (subjects did their own tests on their own computer, in their own office and at their own preferred time slot). This is extended by introducing a measure of control to reduce the impact of different working environments on the results. In the Internet-based MOS test method, all subjects sit in a large project room which they use regularly. It is not a sound-proof room, but it is quiet and has Internet access.

In this chapter, the work on these two series of Internet-based MOS tests is presented. The uncontrolled Internet-based MOS test is described in Section 10.2 and controlled Internet-based MOS test is presented in Section 10.3. The test set-up and the quality evaluation between subjective tests and objective test methods are also given. Section 10.4 concludes the chapter.

10.2 Uncontrolled Internet based MOS Tests and Quality Evaluation

10.2.1 Introduction

Packet loss is a major source of speech impairment in voice over IP (VoIP) applications. Such a loss may be caused by discarding packets in the IP networks due to congestion or by dropping packets at the gateway/terminal due to late arrival. The impact of packet loss on perceived speech quality depends on several factors, including loss pattern, codec type, packet size and loss locations. Research [42, 45] has shown that packet losses in the Internet are temporally correlated, that is, they often occur in bursts rather than in a random pattern. Our research on Internet characteristics in Section 7.3.2 has also shown that packet loss are highly correlated.

It is therefore useful to study how subjects perceive bursty losses and how objective measurement methods, such as PSQM [58], MNB [62, 63], EMBSD [64] and, in particular, the latest ITU standard, PESQ [4] correlate with subjective test results (MOS) [3] under bursty loss conditions.

The work reported here was based on the G.729B [14, 134] codec which is commonly used in VoIP applications. A 2-state Gilbert model was used in the simulation of bursty losses in IP networks. 15 different network loss conditions (a combination of different burstiness, packet size and loss locations) were chosen. 16 subjects took part in the subjective test. Four algorithms – PSQM, MNB, EMBSD and PESQ - were chosen for objective speech quality evaluation against subjective MOS.

10.2.2 Data Collection and Subjective MOS Test

The block diagram of the system that was used in the study is depicted in Figure 10.1. It is a PC-based software system that allows the simulation of key processes in voice over IP. It

enables the simulation of a variety of network conditions and objective measurement of their effects on perceived speech quality. The system includes a speech database, an encoder/decoder, a packet loss simulator and an objective quality measurement module.

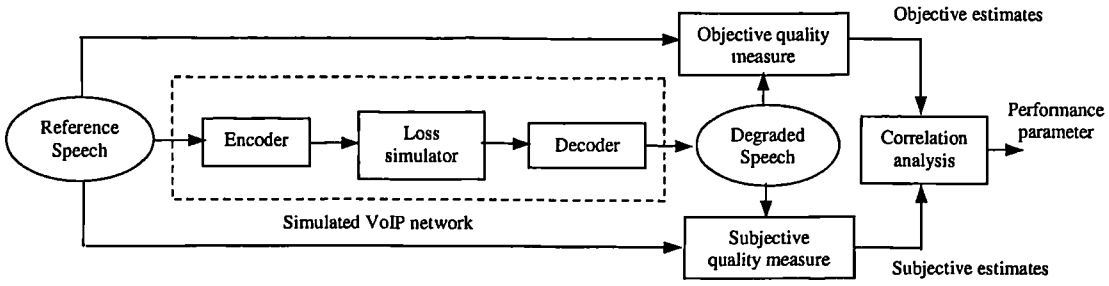


Figure 10.1: Objective and subjective speech quality evaluation system

A 2-state Gilbert model was used to simulate packet loss (see Figure 2.6). Fifteen different bursty loss conditions were chosen to cover cases of interest. They consist of combinations of unconditional loss probability (*ulp, 5%, 10% or 25%*), conditional loss probability (*clp, 20% or 60%*), packet size (2 or 4 frames/packet) as shown in Table 10.1 (the last two columns about MOS will be explained later) . Initial seeds were generated randomly to simulate different loss locations. The frame size for G.729 is 10ms. The reference speech file is about 10 seconds long and consists of four short sentences from two male and two female speakers. 15 different degraded speech files were generated for subjective and objective test.

For efficiency and to make it easier for people to participate in the listening tests, regardless of where they were, a VoIP MOS test website was created. All subjective tests were carried out via Internet by the following URL:

http://www.tech.plymouth.ac.uk/spmc/people/lfsun/mostest/

A total of 16 subjects (located on different floors within a building) participated in the MOS test. Most of them were Ph.D students with no previous MOS test experience. The participants used their own headphones to listen to the original and degraded speech files and were asked to give an opinion score between 1 to 5 (where 5 is excellent and 1 is bad) following the instructions on the web. The average score for a particular test material then gives the Mean Opinion Score (MOS).

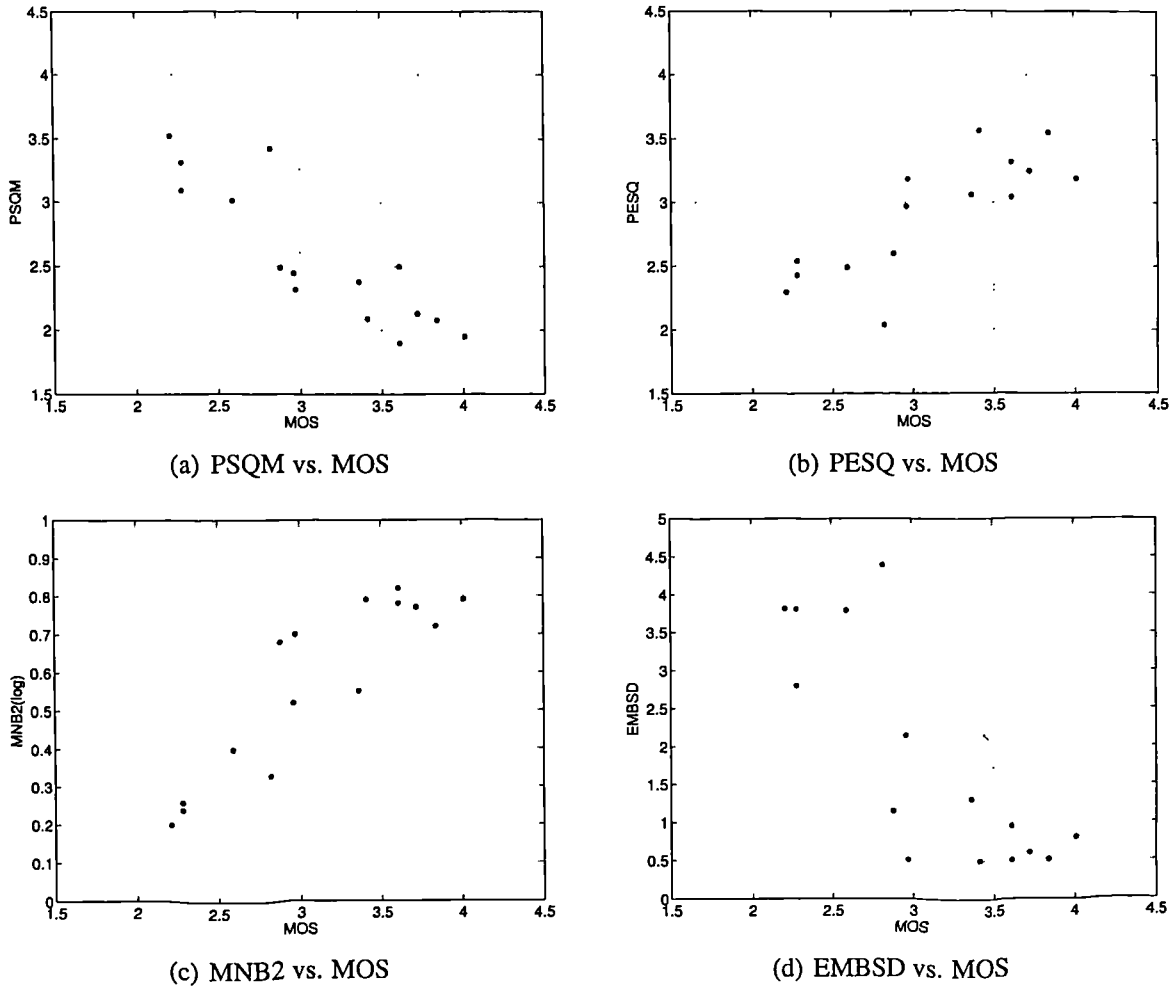
Table 10.1: 15 packet loss conditions

Conditions	*ulp* (%)	*clp* (%)	Packet size	MOS(Test)	MOS(PESQ)
1	5	20	2	3.84	3.547
2	10	60	4	4.01	3.185
3	25	60	2	2.82	2.04
4	10	20	2	3.36	3.055
5	20	60	2	2.59	2.491
6	10	20	2	2.96	2.968
7	25	20	2	2.21	2.295
8	5	60	2	3.72	3.246
9	15	60	4	2.88	2.6
10	5	60	2	3.41	3.558
11	20	20	2	2.28	2.54
12	5	60	4	3.61	3.319
13	10	60	2	3.61	3.042
14	20	20	2	2.28	2.428
15	10	60	2	2.97	3.183

10.2.3 Test Results and Analysis

For each of the 15 conditions, objective measures were obtained using each of the four measurement algorithms (PSQM, PESQ, MNB and EMBSD) separately. As the simulated network conditions were related to packet loss only (end-to-end jitter was not induced), time-alignment was not required and so all four algorithms could be used to obtain an objective measure of speech quality. The scatter diagrams for the objective test results versus subjective MOS scores are illustrated in Figure 10.2(a) to 10.2(d). The MNB produces two perceptual distances (MNB1 and MNB2) which are mapped onto a logistics value within the range [0, 1]. In our experiment, the results for MNB1 and MNB2 are similar and so only those for MNB2 are shown in Figure 10.2.

In order to assess the objective measures against subjective MOS scores, the Pearson correlation coefficient (definition see Appendix A) for each condition, after mapping the objective measures with a 3^{rd} order monotonically decreasing or increasing polynomial, was calculated. To avoid a bias by our MOS test results, the correlation coefficients before and after the 3^{rd} order polynomial mapping were both calculated and are shown in Table 10.2.



(a) PSQM vs. MOS

(b) PESQ vs. MOS

(c) MNB2 vs. MOS

(d) EMBSD vs. MOS

Figure 10.2: Scattered diagrams of objective tests vs. subjective MOS scores

Table 10.2: Correlation between subjective MOS and objective measures

Algorithms	PESQ	PSQM	PSQM+	MNB_1	MNB_2	EMBSD
Correlation before mapping	0.81	- 0.87	- 0.54	0.89	0.895	- 0.80
Correlation after mapping	0.896	0.895	0.726	0.90	0.901	0.88

From Figure 10.2 and Table 10.2, it can be seen that, surprisingly, the PESQ does not have a better performance than the other objective methods under bursty packet loss conditions and that the MNB2 performs slightly better. The results were analysed further to understand the conditions under which the PESQ differs with subjective test results. The MOS scores from the subjective and objective tests using PESQ for all 15 samples (conditions) are illustrated in Figure 10.3. From the Figure, we noticed that the subjective MOS values in samples 2 and 3 were higher than that of the objective MOS values obtained from PESQ. Sample 2 is for 10% (*ulp*), 60% (*clp*) and packet size of 4, whilst Sample 3 is for 25% (*ulp*), 60% (*clp*) and packet size of 2. Both have almost one word missing due to very heavy bursty losses. In this case, PESQ is more sensitive than the subjects. On the other hand, there were three Samples (7, 11 and 14), where subjective MOS scores were slightly lower than for objective MOS scores. It was interesting to find that all three samples belonged to high loss rate conditions (20 or 25% of *ulp*), lower burstiness (20% of *clp*) and small packet sizes (2). The loss occurs evenly and the speech sounds were annoying. In this case, PESQ is less sensitive than the subjects.

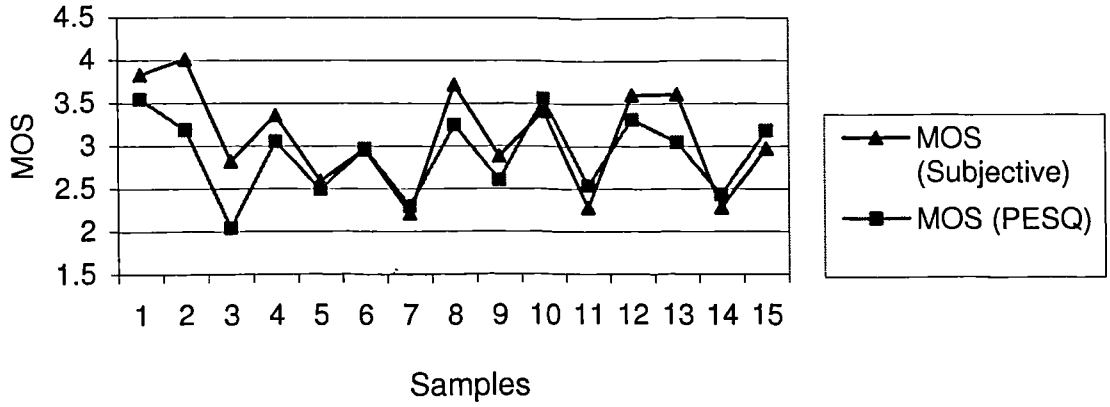


Figure 10.3: Objective (PESQ) and subjective MOS for 15 test samples

The ITU-T P.862 document [4] states that the "PESQ has *demonstrated* acceptable accuracy at factors such as packet loss and packet loss concealment with CELP codecs" and that the factors for which PESQ has *not* currently been validated include "packet loss and packet loss concealment with PCM type codecs" in which "PESQ appears to be more sensitive than subjects to front-end temporal clipping, especially in the case of missing words which *may not*

be perceived by subjects. Conversely, PESQ *may be* less sensitive than subjects to regular, short time clipping (replacement of short sections of speech by silence). "

Although the codec used in our test is G.729 8 Kb/s CS-ACELP, which belongs to the CELP type codec with internal packet loss concealment, our test results for PESQ do not show acceptable accuracy under bursty loss conditions. The results are consistent with the conditions described in the P.862 document for packet loss and loss concealment with PCM type codecs.

10.3 Controlled Internet-based Subjective Test and Quality Evaluation

10.3.1 Introduction

In the last section, we carried out the subjective tests without control (subjects did their own tests on their own computer, in their own office and at their own preferred time slot). We have extended this by introducing a measure of control to reduce the impact of different working environments on the results. In the test, all subjects sit in a large project room which they use regularly. It is not a sound-proof room, but it is quiet and has Internet access.

In this section, we will present the set-up used to evaluate voice quality using both subjective and objective methods. The preliminary test results and analysis are then presented.

10.3.2 Data Collection and Subjective Tests

Figure 10.4 depicts the set-up used for the voice quality evaluation for the experiment. It is a PC-based software system that allows the simulation of key processes in voice over IP and speech quality measurement. Objective voice quality measurements were made with the ITU PESQ and E-model to enable us to compare Internet-based MOS tests with traditional MOS tests. Reference speech files were first encoded using G.723.1 codec and then processed in accordance with network parameter values in trace data files (see later) and then decoded to

generate degraded speech signals (a fixed jitter buffer, for simplicity, was used to remove the effects of jitter. Packets that arrive too late are discarded). We selected some trace data (e.g. those with a 30 ms packet interval, consistent with G.723.1) from our trace data collection for the quality evaluation.

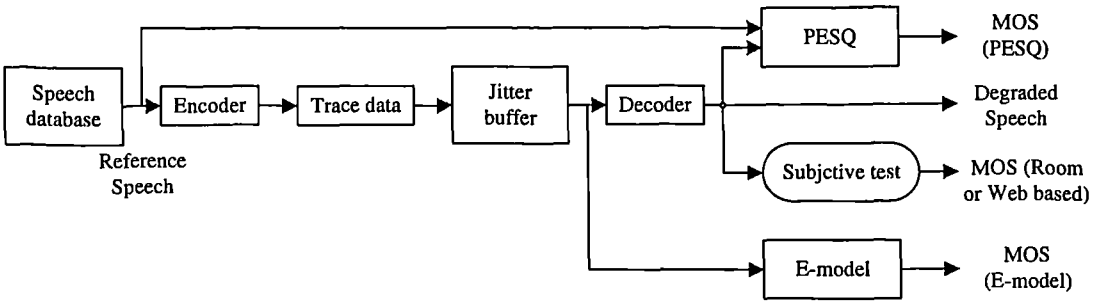


Figure 10.4: VoIP speech quality evaluation set-up

The reference speech database was taken from the TIMIT data set [92]. Each speech sample consists of four short sentences spoken by four different male and female speakers in order to keep a balanced design. Each speech sample was about 10 to 15 seconds long. A total of 10 speech samples were chosen for the VoIP quality evaluation. We also chose ten different network conditions from the Internet trace data set, covering packet loss rate (including late arrival loss due to jitter) from 0% to 30%. Ten degraded speech samples were generated and used for the quality evaluation.

Subjective Tests were carried out using two methods – Internet-based (or web-based) and P.800-based tests (Room-based). A website for the MOS test was created at the following URL:

http://www.tech.plymouth.ac.uk/spmc/people/lfsun/mos/

The 10 degraded speech samples were put on the web, together with a brief instruction about the MOS test. 15 undergraduate students were invited to attend the controlled Internet-based MOS test. The tests were carried out in the project laboratory which they use regularly. The room was quiet and similar to a normal office environment. Brief instructions were given by a supervisor before the test. The students were asked to perform the test at their own pace. The tests took about 15 minutes to complete. When all the students had submitted their opinion

scores, the MOS score were calculated and expressed as Web_MOS. The online MOS test webpage and the test results (displayed online when tests are done) are illustrated in Appendix B.

In order to compare the results of the controlled Internet-based MOS tests with similar P.800 tests, we carried out another round of MOS test in a small, quiet room (a sound-proof room was not available on-site and is another motivation for investigating the web-based approach). The room is about 8 square meters and 3.5 meters high with a desktop and a laptop PCs. A similar test procedure to the web-page was created locally. The same 10 degraded speech samples were chosen. The 15 students were invited again to carry out the tests, one by one. The test spanned over two days because of the numbers involved and their availability. The MOS score for the room-based MOS test was expressed as Room_MOS.

We also conducted MOS tests using the ITU PESQ algorithm and the E-model in order to compare the subjective test results with objective measurements. Listening-only speech quality measurements were considered in order to keep the same conditions. By comparing the reference speech and the degraded speech, an objective MOS score was obtained from the PESQ algorithm. This MOS score was referred to as MOS (PESQ) or PESQ_MOS. For the E-model, only the effects of the Equipment Impairment (I_e) were taken into account (the effects of delay, I_d, was not considered). This gives a listening-only MOS score which is referred to as MOS (E-model) or E-model_MOS.

10.3.3 Test Results and Analysis

The results for the 10 degraded speech samples for each of the four methods of voice quality measurement (PESQ, E-model, Internet-based and Room-based MOS tests) are summarised in Table 10.3. The sequences of the test speech samples from 1 to 10 are the same with that on the MOS test website. The calculated packet loss rates are included to give an indication of the impact of the network impairment.

The relationships between the MOS scores and packet loss rates for the different MOS test

Table 10.3: Objective and subjective MOS scores for different speech samples

Test samples	1	2	3	4	5	6	7	8	9	10
PESQ	3.18	2.65	2.85	3.74	2.02	1.95	2.42	2.59	2.93	2.54
E-model	2.90	2.50	2.56	3.92	1.00	1.04	1.63	2.07	2.71	2.41
Room-based	3.36	2.65	2.85	3.31	1.41	1.15	2.13	2.13	2.97	2.34
Web-based	3.37	2.32	2.52	4.00	1.22	1.11	2.19	2.04	2.90	2.35
Loss rate (%)	5.68	9.51	8.85	0.21	29.5	23.1	18.5	14.6	7.25	10.6

methods are depicted in Figure 10.5. From the figure, it can be seen that the MOS scores for all four evaluation methods decrease with increasing packet loss rate. When the packet loss rate is low, the E-model, PESQ and Web-based MOS scores are quite close. When packet loss rate is high, PESQ seems to over predict the voice quality, whilst the E-model does the opposite. The results for PESQ confirms the work in [71] (that is why PESQ-LQ is proposed for better mapping PESQ MOS score to subjective MOS score). For E-model, as it predicts voice quality directly from network parameters (e.g. packet loss rate), it does not consider factors such as packet loss location. Also as VAD (voice activity detection) was not activated in the simulation, packet loss in the silence period will not be perceived by subjects, but it was still taken into account in the E-model calculation. This is partly why the E-model gives lower MOS scores compared to the other methods when the packet loss rate is high. Room-based and Web-based MOS scores are close, except in the case when there is almost no packet loss. This is probably because the background noise (e.g. from the fan) of the computer for Room-based tests is higher than those for Web-based test.

The Pearson correlation coefficient between the results of subjective and objective methods were calculated and the results are shown in Table 10.4. From the table, it can be seen that the Internet-based MOS test (Web_MOS) compares well with the traditional MOS test (Room_MOS) (correlation coefficients of 0.95). This suggests that with the Internet-based MOS test it is possible to obtain similar results to those of traditional MOS tests for VoIP applications. For objective measurement methods (E-model and PESQ) and subjective methods (Room-based and Web-based), the correlation coefficients are between 0.93 to 0.98. This

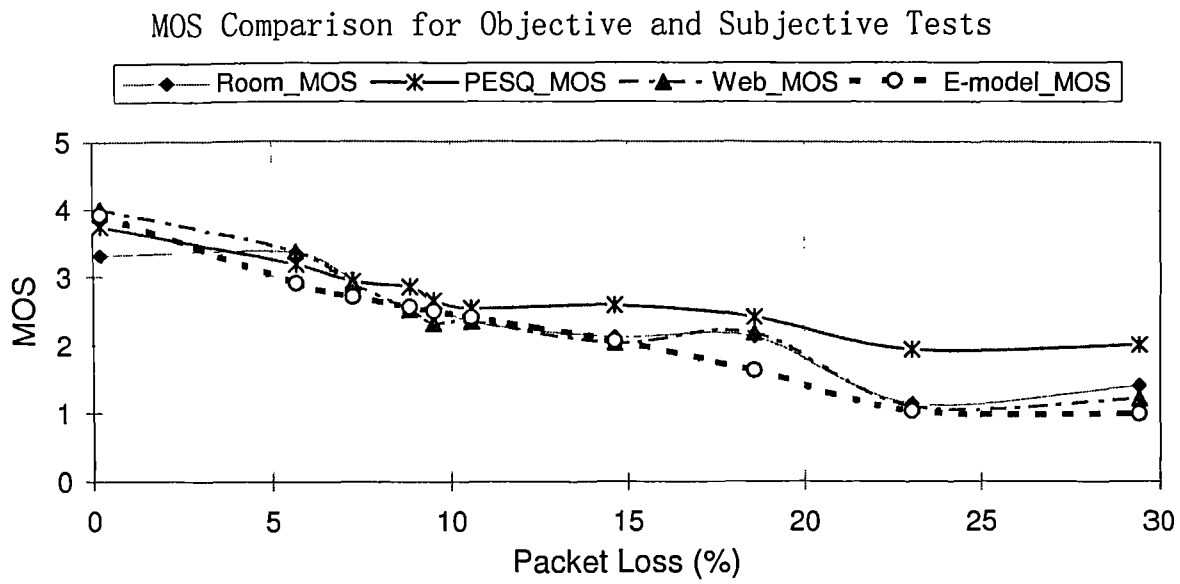


Figure 10.5: MOS comparison for objective and subjective test methods

shows that the two objective methods can both predict subject MOS score well, although both seem to predict Web-based MOS better than Room-based MOS.

Table 10.4: Correlation coefficients (ρ) for MOS comparison

Name	PESQ vs Room_MOS	PESQ vs Web_MOS	Emodel vs Room_MOS	Emodel vs Web_MOS	Web_MOS vs Room_MoS	Emodel vs PESQ
ρ	0.933	0.984	0.935	0.964	0.952	0.975

10.4 Summary

In this Chapter, a new subjective, Internet-based MOS test methodology which allows rapid assessment of voice quality has been proposed. Two series of subjective tests are carried out. The first one is uncontrolled Internet-based MOS tests focusing on voice quality under bursty packet loss. The preliminary results show that the four objective algorithms (PSQM, PESQ, MNB and EMBSD) do not have a sufficiently high correlation with subjective scores under network bursty loss conditions. The MNB2 has a slightly higher correlation than the other three algorithms. The PESQ only predicts the MOS score well for the average bursty loss

cases. When loss burstiness is higher/lower, PESQ shows an obvious sensitivity to these conditions than human subjects. This highlights the need for a further modification/improvement of the latest ITU objective speech quality measurement algorithm to make it suitable for a variety of network bursty loss conditions. The second test is the controlled Internet-based MOS test focusing on comparing between objective tests (e.g. PESQ and E-model) and subjective tests (e.g. Internet-based and Room-based). Results show that the Internet-based MOS test compares well with traditional MOS test (correlation coefficient of 0.95). In general, the two ITU objective test methods (PESQ and E-model) can predict subjective MOS scores well.

Chapter 11

Discussion, Future Work and Conclusions

11.1 Introduction

The need to evaluate voice quality in Voice over IP applications is an important requirement for technical and commercial reasons. Speech quality can be measured by subjective tests or by objective methods. The Mean Opinion Score (MOS) is the most widely used subjective measure of voice quality and is recommended by the ITU [3]. The MOS score is the internationally accepted metric as it provides a direct link to voice quality as perceived by the end user. The inherent problem in subjective MOS measurement is that it is time consuming, expensive, lack of repeatability and cannot be used for long-term or large scale voice quality monitoring in an operational network infrastructure. This has made objective methods very attractive for meeting the demand for voice quality measurement in communications networks.

Objective measurement of voice quality can be intrusive or non-intrusive. Intrusive methods (e.g. ITU PESQ [4]) are more accurate, but normally are unsuitable for monitoring live traffic because of the need for a reference data and to utilise the network. Non-intrusive methods (e.g. ITU E-model [7]) are appropriate for monitoring voice quality directly from IP network and/or non-network parameters. However, current non-intrusive methods (e.g. statistical E-model or neural network models [11]) rely on subjective tests to derive model parameters (e.g. for the E-model) or to create the training sets (for neural network models). Unfortunately subjective tests are costly and time-consuming and as a result the derived models are limited and cannot

cover all the possible scenarios in dynamic and evolving networks, such as the Internet.

The main aims of this project are (1) to undertake a fundamental investigation to quantify the impact of network impairments and speech related parameters on perceived speech quality in IP networks, (2) to apply the results to develop novel and efficient models for non-intrusive speech quality measurement and prediction for VoIP applications, and (3) to apply the developed models to new and emerging applications in voice quality prediction/monitoring, voice quality optimization (e.g. jitter buffer optimization) and QoS control.

This Chapter discusses the main contributions of this work and highlights the novelty, future work and the conclusions.

11.2 Contributions to Knowledge

To summarise, this thesis presents research that has achieved the following:

(1). A detailed understanding of the relationships between voice quality (in terms of MOS score), and IP network impairments (e.g. packet loss, delay and delay variation) and non-network parameters (e.g. codec type, voiced/unvoiced, gender and language)

An understanding of the perceptual effects of these key parameters on voice quality is important as it provides a basis for the development of non-intrusive voice quality prediction models. A fundamental investigation of the impact of these parameters on perceived voice quality is undertaken using the latest ITU intrusive voice quality measurement algorithm, PESQ, and a combined structure of PESQ and E-model. The results show that packet loss burstiness, loss locations/patterns, packet size, delay, codec, gender of talkers and language all have an impact on perceived speech quality. The work is described in Chapter 4.

(2). The development of a new methodology and models to predict voice quality non-intrusively (including non-linear regression and neural network models)

The method is based on the latest intrusive ITU algorithm, PESQ, and a combined structure of PESQ/E-model and allows a perceptually accurate prediction of both listening and conversational voice quality non-intrusively to be obtained. This avoids time-consuming subjective tests and so removes one of the major obstacles in the development of models for voice quality prediction. The method is generic and as such has wide applicability in multimedia applications (e.g. objective, non-intrusive, prediction of end-to-end voice, audio image or video quality; optimization of quality of the multimedia services) and to other packet networks (e.g. ATM or managed IP networks). Both efficient non-linear regression models and robust artificial neural network learning models are developed for predicting voice quality non-intrusively. Nonlinear regression models are efficient and static, but not flexible, whereas the neural network models are more complicated, but have the learning ability which can adapt to the dynamic environment of IP networks. The new methodology and regression models are presented in Chapter 5 and the neural network models are described in Chapter 6.

(3). Demonstrated the application of the new non-intrusive voice quality prediction models to three areas

The three applications are voice quality prediction for the current Internet, perceived quality driven playout buffer optimization and perceived quality driven QoS control. The neural network and non-linear regression models are both used for predicting voice quality in the current Internet based on international links between UK and USA, UK and China, and UK and Germany. Results show that the neural network models have accuracy close to the ITU PESQ/E-model (correlation coefficient of 0.94) and the regression models with correlation coefficient of 0.98. The work is presented in Chapter 7.

A new adaptive playout buffer algorithm and a novel perceptual optimization playout buffer algorithm are presented in Chapter 8. The adaptive buffer algorithm can automatically adapt to the most suitable buffer algorithm according to network delay and delay variation. The perceptual optimization buffer algorithm is based on the regression models for voice quality

prediction, a new minimum impairment criterion and the Weibull delay distribution model. Results show that perceptual based optimum buffer algorithm can achieve the optimum perceived voice quality, and the adaptive algorithm can achieve sub-optimum perceived voice quality with low complexity, when compared with four existing buffer algorithms.

A novel QoS control scheme that combines the strengths of rate-adaptive and priority marking control schemes to provide a superior QoS control in terms of measured perceived voice quality is described in Chapter 9. The perceived speech quality (e.g. MOS score), instead of individual network parameters (e.g. packet loss or delay), has been used for adaptive control of sender bit rate. This has formed the basis for another PhD project because it is of a major research interest in its own right.

(4). The development of a novel subjective MOS test methodology

A novel methodology for Internet-based subjective speech quality measurement is presented in Chapter 10. This allows rapid assessment of voice quality for VoIP applications. Both uncontrolled and controlled Internet-based MOS tests are carried out and the results are compared with various of objective test methods and traditional subjective MOS tests. The results show that Internet-based MOS tests compare well with traditional MOS tests (correlation coefficient of 0.95).

11.3 Limitations of the Current Work and Discussions

There are a number of limitations in the project which should be addressed.

(1). The accuracy of PESQ

At present, ITU PESQ represents the most accurate objective, intrusive speech quality measurement method. The models we developed for predicting speech quality non-intrusively are highly correlated with PESQ algorithm (with correlation coefficients over 0.94). However, the

subjective measurement remains the most accurate benchmark for all objective measurement methods. As a result, PESQ accuracy [135] may be continuously enhanced by future generations of intrusive speech quality measurement methods. The speech quality prediction models we developed can be readily updated with the future generations of PESQ algorithm.

(2). The additivity of impairments

The new regression and neural network models for speech quality prediction use a combined structure of PESQ/E-model. This, as in the E-model, is still based on the assumption that the impairment factors from packet loss (I_e) and delay (I_d) are additive (or independent of each other) at a psychological (quality) scale (or transmission rating scale, R scale). This may not be entirely correct and needs further validation [77, 136].

(3). Limited consideration of end-to-end impairments

In the thesis, the impairments considered are mainly IP network impairments (e.g. packet loss, delay and jitter), which, of course, are the most important impairments for Voice over IP networks, and speech-related impairments (e.g. codec, gender and language). The impairments relevant in telecommunication networks (e.g. echo, sidetone, background noise, cross-talk, and too low/high volume) [37] have not been taken into account in the thesis. Obviously these impairments have an impact on end-to-end perceived speech quality.

(4). Limited validation of the work

The thesis has presented several new methods for predicting speech quality for VoIP networks, objectively or subjectively. Although individual elements of the work have been validated where possible, large scale validation or cross-validation are still needed. For example, the neural network training set has to be refined according to real Internet trace data and neural network models have to be tested/validated in large scale networks for VoIP applications. The Internet-based subjective tests should also be compared with real P.800 results (from sound-proof room).

(5). Limited trace data collection

In the project, only limited Internet trace data was collected and analyzed. This should be expanded to cover more local, national and international links with various of transmission mediums (e.g. modem, ADSL links and different ISPs). Trace data should be collected periodically, for example, every three months, new network information can be collected and analysed to determine whether there are significant changes.

11.4 Future Work

There are four main areas for future work.

(1). Adaptive, self-turning learning models

In the thesis, it has been demonstrated that neural network learning models are feasible for predicting both listening and conversational voice quality, non-intrusively. Changes in IP networks may lead to new sources of impairments or changes in the character of the new impairments which may in turn lead to a deterioration in the performance of the learning models. Thus, it would be desirable to exploit the key attribute of neural networks, namely the ability to learn, to adapt dynamically to significant new situation in the communication networks over time.

The objective here can be twofold. First, to provide new knowledge of the behaviour of the learning models in situations in which their predictions may not be accurate. Knowledge of their limitations can be used to improve the performance of the models and make them more robust across new classes of impairments. Second, to provide a basis for research into self-turning, learning models for voice quality prediction. For example, whenever the performance of a learning model falls below a predefined threshold (e.g. because the network has been modified or a new VoIP application significantly alters the voice quality/impairment relationships), the learning model is re-trained, automatically, to adapt to the new situation. Clearly, this will

require automated training of the learning models and a means of determining when re-training is necessary.

(2). Perceived quality prediction for multimedia services over IP networks

The methodology presented in the thesis for predicting voice quality non-intrusively is generic and based on end-to-end, intrusive measurement (in this case, using PESQ) which avoids time-consuming subjective tests. Thus, it can be easily applied to other applications, such as audio (e.g. using ITU-T PEAQ [86]) and video (e.g. using ANSI objective video quality standard T1.801.03 [98]), provided the neural networks models are retrained and appropriate parameters for the regression models determined.

Also the differences between speech, audio and video mean that new parameters that impact on quality have to be taken into account when developing models for predicting audio or video quality non-intrusively. For example, parameters such as the bit rate (BR), the frame rate (FR), the ratio of the encoded intra macro-blocks to inter macro-blocks [13] have to be considered for video quality prediction for video over IP networks.

(3). *Perceived speech quality prediction for voice over other packet networks*

Although the thesis has focused on IP networks (mainly best-effort IP networks), the approach of the end-to-end quality consideration can be applied to managed IP networks (e.g. DiffServ), other packet networks (e.g. ATM), or to wireless networks as well. The end-to-end intrusive measurement (e.g. using PESQ) is suitable for any networks, as the method is based on a comparison of the reference speech signal and the degraded speech signal transmitted through the network. An important requirement for applying intrusive technique for non-intrusive application is to understand and obtain the relevant parameters which affect the corresponding end-to-end perceived speech quality. These parameters or the range of the values of the parameters are network-dependent(e.g. depending on wired or wireless, IP or ATM networks).

For example, additional impairment parameters due to wireless links (e.g. bit error rate) have to be taken into account when predicting voice quality for wireless networks.

For managed IP networks (e.g. DiffServ), the network performance (e.g. packet loss, jitter and delay) will differ with that of the best effort IP networks. For example, the range of packet loss rate or end-to-end delay may be much smaller than that from the best effort networks for certain quality of service classes. Thus, the neural network training set has to be refined and neural network models retrained.

Similarly, for a network deployed with different packet loss recovery strategies (e.g. using FEC [137,138]), the influence from these loss recovery strategies have to be taken into account.

(4). QoS performance optimization and control

The perceived speech quality metric can be used to optimise the quality of VoIP services in accordance with changing network conditions and to control the QoS and manage the utilisation of available resources. It is a better metric than the traditional use of only individual network impairments (e.g. packet loss, jitter and delay) as it provides a direct link to end user's perception. Perceived speech quality driven control or optimization can have variety of potential applications. The following are some examples.

In 3G wireless networks, it can help to control retransmission mechanisms to reduce packet loss and errors in wireless VoIP and hence optimise the quality of voice [139].

For media streaming (voice, audio or video) over IP networks, perceived voice/audio/video quality metric can be used for server selection. For example, it can be used to search for an audio/video server which can provide an optimum end-to-end perceived audio/video quality, instead of traditionally obtaining an optimum individual network parameters (e.g. minimum end-to-end delay, jitter or packet loss). The media streaming applications may include eHealthcare, tele-lcarning and home entertainment.

11.5 Conclusions

An important conclusion of this project is that it is possible to exploit perceptually more accurate intrusive speech quality measurement (e.g. PESQ) for non-intrusive applications. This is an important development as it avoids time-consuming subjective tests and removes a major obstacle in the development of models for non-intrusive prediction of voice quality. This is applicable to audio, image and video applications over packet networks.

Non-linear regression models and neural network models have been developed based on the new methodology and further applied in three main applications, namely, voice quality prediction for VoIP in Internet, perceived voice quality driven adaptive buffer optimization and perceived voice quality driven QoS control. A detailed understanding of the relationships between voice quality, IP network impairments and non-network impairments has been provided in order to derive the models properly. An Internet-based subjective test methodology has also been presented for more efficient subjective tests for VoIP applications.

The novelty in this work is in a new methodology to predict voice quality non-intrusively, nonlinear regression and neural network models for speech quality prediction, adaptive and perceived quality optimized jitter buffer algorithms, a new QoS control scheme that combines the strengths of adaptive rate and speech priority marking QoS control techniques, and Internet-based subjective test methodology.

Much of the work and effort over the course of this programme of work has gone into understanding the problems for speech quality prediction for Voice over IP networks, objective and subjective speech quality measurement, end-to-end speech quality prediction, optimization and control. Several new models, methodology, and algorithms have been proposed for speech quality prediction, jitter buffer algorithm optimization and QoS control. The models and algorithms need to be further validated in a large scale networks and for other applications before they are made available for commercial applications.

Bibliography

[1] European Telecommunications Standards Institute, "Specification and Measurement of Speech Transmission Quality; Part 1: Introduction to Objective Comparison Measurement Methods for One-way Speech Quality Across Networks," *ETSI Guide, EG 201 377-1 V1.1.1*, April 1999.

[2] L. Yamamoto and J.G.Beerends, "Impact of Network Performance Parameters on the End-to-end Perceived Speech Quality," in *Proceedings of Expert ATM Traffic Symposium*, (Mykonos, Greece), Sep. 1997.

[3] International Telecommunicaion Union, "Methods for Subjective Determination of Transmission Quality," *ITU Recommendation P.800*, August 1996.

[4] International Telecommunication Union, "Perceptual Evaluation of Speech Quality (PESQ), An Objective Method for End-to-end Speech Quality Assessment of Narrow-band Telephone Networks and Speech Codecs," *ITU-T Recommendation P.862*, Feb. 2001.

[5] A. W. Rix, M. P. Hollier, A. P. Hekstra, and J. G. Beerends, "Perceptual Evaluation of Speech Quality (PESQ): The New ITU Standard for End-to-End Speech Quality Assessment, Part I – Time-Delay Compensation," *Journal of the Audio Engineering Society*, vol. 50, pp. 755–764, October 2002.

[6] J. G. Beerends, A. P. Hekstra, A. W. Rix, and M. P. Hollier, "Perceptual Evaluation of Speech Quality (PESQ): The New ITU Standard for End-to-End Speech Quality Assessment Part II – Psychoacoustic Model," *Journal of the Audio Engineering Society*, vol. 50, pp. 765–778, October 2002.

[7] International Telecommunicaion Union, "The E-model, A Computational Model for Use in Transmission Planning," *ITU-T Recommendation G.107*, July 2000.

[8] N. O. Johannesson, "The ETSI Computation Model: A Tool for Transmission Planning of Telephone Networks," *IEEE Communications Magazine*, pp. 70–79, Jan. 1997.

[9] R. G. Cole and J. Rosenbluth, "Voice over IP Performance Monitoring," *Journal on Computer Communications Review*, vol. 31, April 2001.

[10] A. P. Markopoulou, F. A. Tobagi, and M. Karam, "Assessment of VoIP Quality over Internet Backbones," in *Proc. of IEEE Infocom*, vol. 1, (New York, USA), pp. 150–159, June 2002.

[11] S. Mohamed, F. Cervantes-Pérez, and H. Afifi, "Real-Time Audio Quality Assessment in Packet Networks," *Network and Information Systems Journal*, pp. 595–609, 2000.

[12] S. Mohamed, F. Cervantes-Pérez, and H. Afifi, "Integrating Networks Measurements and Speech Quality Subjective Scores for Control Purposes," in *Proc. IEEE INFOCOM'01*, vol. 2, (Anchorage, Alaska), pp. 641–649, April 2001.

[13] S. Mohamed and G. Rubino, "A Study of Real-Time Packet Video Quality Using Random Neural Networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, pp. 1071–1083, Dec. 2002.

[14] International Telecommunicaion Union, "Coding of Speech at 8 kbit/s Using Conjugate-Structure Algobraic-Code-Excited Linear-Prediction (CS-ACELP)," *ITU-T Recommendation G.729*, March 1996.

[15] International Telecommunicaion Union, "Dual Rate Speech Coder for Multimedia Communication Transmitting at 5.3 and 6.3 kbit/s," *ITU-T Recommendation G.723.1*, March 1996.

[16] European Telecommunications Standards Institute, "Digital Cellular Tlecommunications System (Phase 2+); Adaptive Multi-Rate (AMR) Speech Transcoding," *ETSI-EN-301-704 V7.2.1*, April 2000.

[17] S. V. Andersen, W. B. Kleijn, R. Hagen, J. Linden, M. N. Murthi, and J. Skoglund, "iLBC - A Linear Predictive Coder with Robustness to Packet Losses," in *Proceedings of IEEE 2002 Workshop on Speech Coding*, (Tsukuba Ibaraki, Janpan), pp. 23–25, Oct 2002.

[18] L. Sun, G. Wade, B. Lines, and E. Ifeachor, "Impact of Packet Loss Location on Perceived Speech Quality," in *Proc. of IPTEL'01*, (New York, USA), pp. 114–122, April 2001.

[19] L. Sun and E. Ifeachor, "Perceived Speech Quality Prediction for Voice over IP-based Networks," in *Proceedings of IEEE International Conference on Communications ICC'02*, (New York, USA), pp. 2573–2577, April 2002.

[20] L. Sun and E. Ifeachor, "New Models for Perceived Voice Quality Prediction and their Applications in Playout Buffer Optimization for VoIP Networks," in *Proceedings of IEEE International Conference on Communications ICC 2004*, (Paris, France), June 2004.

[21] L. Sun and E. Ifeachor, "Learning Models for Non-intrusive Prediction of Voice Quality for IP Networks," *Submitted to IEEE Transactions on Neural Networks*, 2004.

[22] L. Sun and E. Ifeachor, "Prediction of Perceived Conversational Speech Quality and Effects of Playout Buffer Algorithms," in *Proceedings of IEEE International Conference on Communications ICC'03*, (Anchorage, USA), pp. 1–6, May 2003.

[23] Z. Qiao, L. Sun, N. Heilemann, and E. Ifeachor, "A New Method for VoIP Quality of Service Control based on Combined Adaptive Sender Rate and Priority Marking," in

Proceedings of IEEE International Conference on Communications ICC 2004, (Paris, France), June 2004.

[24] L. Sun and E. Ifeachor, "Subjective and Objective Speech Quality Evaluation under Bursty Losses," in *Proceedings of On-line Workshop Measurement of Speech and Audio Quality in Networks (MESAQIN 2002)*, (Prague, Czech Republic), pp. 25–29, Jan 2002.

[25] L. Sun and E. Ifeachor, "New Methods for Voice Quality Evaluation for IP Networks," in *Proceedings of 18th International Teletraffic Congress (ITC-18)*, (Berlin, Germany), pp. 1201–1210, Sep 2003.

[26] D. Minoli and E. Minoli, *Delivering Voice over IP Networks, 2nd Edition.* John Wiley and Sons, 2002. ISBN 0-471-38606-5.

[27] H. Sanneck, "Adaptive Loss Concealment for Internet Telephony Applications," in *Proceedings INET'98*, (Geneva, Switzerland), July 1998.

[28] J. Rosenberg, "G.729 Error Recovery for Internet Telephony," *Columbia University Computer Science, Technical Report CUCS-016-01*, Dec 2001.

[29] J. Rosenberg, L. Qiu, and H. Schulzrinne, "Integrating Packet FEC into Adaptive Voice Playout Buffer Algorithms on the Internet," in *Proceedings of IEEE Infocom 2000*, vol. 3, (Tel Aviv, Israel), pp. 1705–1714, March 2000.

[30] P. L. Tien and M. C. Yuang, "Intelligent Voice Smoother for Silence-Suppressed Voice over Internet," *IEEE Journal on Selected Areas in Communications*, vol. 17, pp. 29–41, Jan. 1999.

[31] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An Architecture for Differentiated Services," *RFC 2475, IETF*, 1998.

[32] J. Wroclawski, "The Use of RSVP with IETF Integrated Services," *RFC2210 IETF*, Sep. 1997.

[33] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson., "RTP: a Transport Protocol for Real-time Applications," *RFC 1889, IETF*, Jan. 1996. `ftp://ftp.ietf.org/rfc/rfc1889.txt`.

[34] International Telecommunicaion Union, "H.323 Visual Telephone Systems and Equipment for Local Area Networks Which Provide a Non-guaranteed Quality of Service," *ITU-T Recommendation H.323*, May 1996.

[35] M. Handley, H. Schulzrine, E. Schooler, and J. Rosenberg, "SIP: Session Initiation Protocol, RFC 2543," *IETF*, March 1999.

[36] J.-Y. Monfort, "Basic Requirements to Quality of Service (IP centric)," in *Workshop on Standardization in E-health*, (Geneva, Swissland), May 2003.

[37] W. C. Hardy, *QoS Measurement and Evaluation of Telecommunications Quality of Service*. John Wiley & Sons, 2001. ISBN 0-471-49957-9.

[38] S. Andersen and A. Duric, "Internet Low Bit Rate Codec (iLBC), IETF Draft," Feb 2002.

[39] A. Duric and S. Andersen, "RTP Payload Format for iLBC Speech," Oct 2003. `http://www.ietf.org/internet-drafts/draft-ietf-avt-rtp-ilbc-03.txt`.

[40] J. F. Kurose and K. W. Ross, *Computer Networking A Top-Down Approach Featuring the Internet*. Pearson Addison Wesley, 2001. ISBN 0-201-47711-4.

[41] H. Schulzrinne, "IP Networks." `http://citeseer.nj.nec.com/schulzrinne00ip.html`.

[42] J. C. Bolot, "Characterizing End-to-end Packet Delay and Loss in the Internet," *Journal of High-Speed Networks*, vol. 2, pp. 305–323, Dec. 1993.

[43] M. S. Borella, "Measurement and Interpretation of Internet Packet Loss," *Journal of Communication and Networking*, vol. 2, pp. 93–102, June 2000.

[44] V. Paxson, "End-to-end Internet Packet Dynamics," *IEEE/ACM Transactions on Networking*, vol. 7, no. 3, pp. 277–292, 1999.

[45] M. Yajnik, S. Moon, J. Kurose, and D. Towsley, "Measurement and Modelling of the Temporal Dependence in Packet Loss," in *Proceedings of IEEE INFOCOM 99*, vol. 1, (New York, USA), pp. 345–352, March 1999.

[46] W. Jiang and H. Schulzrinne, "QoS Measurement of Internet Real-Time Multimedia Services," *Technical Report, CUCS-015-99, Columbia University*, Dec. 1999.

[47] H. Sanneck, "Packet Loss Recovery and Control for Voice Transmission over the Internet," *Ph.D Dissertation, Technical University of Berlin*, Oct. 2000.

[48] H. Sanneck and N. T. L. Le, "Speech Property-Based FEC for Internet Telephony Applications," in *Proceedings of the SPIE/ACM SIGMM Multimedia Computing and Networking Conference*, (San Jose, CA, USA), pp. 38–51, Jan. 2000.

[49] W. Jiang, "QoS Measurement and Management for Internet Real-time Multimedia Services," *Ph.D Dissertation, Columbia University*, 2003.

[50] EURESCOM Project P905-PF, "AQUAVIT - Assessment of QUality for Audio-Visual signals over Internet and UMTS – Deliverable 1: Dscription of testbeds," August 2000. `http://www.eurescom.de/~pub-deliverables/p900-series/p905/d1/p905d1.pdf`.

[51] European Telecommunications Standards Institute, "TIPHON Release 3; Technology Compliance Specification; Part 5: Quality of Service (QoS) measurement methodologies," *ETSI TS 101 329-5 V1.1.1*, Nov. 2000.

[52] A. D. Clark, "Modeling the Effects of Burst Packet Loss and Recency on Subjective Voice Quality," in *Proc. of IPTEL'2001*, (New York, USA), pp. 123–127, April 2001.

[53] M. Yajnik, J. Kurose, and D. Towsley, "Packet Loss Correlation in the MBone Multicast Networ Experimental Measurements and Markov Chain Models," Tech. Rep. University of Massachusetts, UM-CS-1995-115, , 1995. http://citeseer.nj.nec.com/yajnik96packet.html.

[54] International Telecommunicaion Union, "One-way Transmission Time," *ITU-T Recommendation G.114*, May 2003.

[55] P. Gray, M. P. Hollier, and R. E. Massara, "Non-intrusive Speech Quality Assessment using Vocal-tract Models," *IEE Proceedings - Vision, Image and Signal Processing*, vol. 147, pp. 493–501, Dec. 2000.

[56] EURESCOM Project P905-PF, "AQUAVIT - Assessment of QUality for Audio-Visual signals over Internet and UMTS – Deliverable 2: Methodology for subjective audio-visual quality evaluation in mobile and IP networks," August 2000. http://www.eurescom.de/~pub-deliverables/p900-series/p905/d2/p905d2.pdf.

[57] A. Watson and M. A. Sasse, "Measuring Perceived Quality of Speech and Video in Multimedia Conferencing Applications," in *Proceedings of ACM Multimedia '98*, (Bristol, England), pp. 55–60, Sep. 1998.

[58] International Telecommunicaion Union, "Objective Quality Measurement of Telephone-band (300-3400 Hz) Speech Codecs," *ITU-T Recommendation P.861*, Feb. 1998.

[59] J. G. Beerends and J. A. Stemerdink, "A Perceptual Speech Quality Measure Based on a Psychoacoustic Sound Representation," *J. Audio Eng. Soc.*, vol. 42, no. 3, pp. 115–123, 1994.

[60] M. P. Hollier, M. O. Hawksford, and D. R. Guard, "Algorithms for Assessing the Subjectivity of Perceptually Weighted Audible Errors," *J. AES*, vol. 43, pp. 1041–1045, Dec. 1995.

[61] A. Rix, R. Reynolds, and M. Hollier, "Perceptual Measurement of End-to-end Speech Quality over Audio and Packet-based Networks," in *AES 106th Convention*, (Munich, Germany), May 1999. Preprint 4873.

[62] S. Voran, "Objective Estimation of Perceived Speech Quality - Part I: Development of the Measuring Normalizing Block Technique," *IEEE Trans. on Speech and Audio Processing*, vol. 7, pp. 371–382, July 1999.

[63] S. Voran, "Objective Estimation of Perceived Speech Quality - Part II: Evaluation of the Measuring Normalizing Block Technique," *IEEE Trans. on Speech and Audio Processing*, vol. 7, pp. 383–390, July 1999.

[64] W. Yang, "Enhanced Modified Bark Spectral Distortion (EMBSD): An Objective Speech Quality Measure Based on Audible Distortion and Cognition Model," *Ph.D Dissertation, Temple University*, May 1999.

[65] W. Yang, M.Benhouchta, and R. Yantorno, "Performance of a Modified Bark Spectral Distortion Measure as An Objective Speech Quality Measure," in *Proc. of IEEE ICASSP*, pp. 541–544, 1998.

[66] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual Evaluation of Speech Quality (PESQ) - A New Method for Speech Quality Assessment of Telephone Networks and Codecs," in *Proceedings of 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2001.

[67] International Telecommunicaion Union, "Improvement of the P.861 Perceptual Speech Quality Measure," *ITU-T Contribution Com12-20*, Dec. 1997.

[68] S.Wang, A.Sekey, and A.Gersho, "An Objective Measure for Predicting Subjective Quality of Speech Coders," *IEEE J. Selected Areas on Commun.*, vol. 10, pp. 819–829, June 1992.

[69] International Telecommunicaion Union, "Performance of the Integrated KPN/BT Objective Speech Quality Assessment Model," *ITU-T Contribution COM12 - D.136*, May 2000.

[70] International Telecommunicaion Union, "Report of the Question 13/12 Rapporteus's Meeting," *ITU-T Contribution Com12-117*, March 2000.

[71] A. W. Rix, "Comparison between Subjective Listening Quality and P.862 PESQ Score," in *Proceedings of Online Workshop Measurement of Speech and Audio Quality in Networks*, (Czech Republic), pp. 17–25, May 2003.

[72] International Telecommunicaion Union, "In-service, Non-intrusive Measurement Device - Voice Service Measurements," *ITU-T Recommendation P.561*, Feb. 1996.

[73] International Telecommunicaion Union, "Analysis and Interpretation of INMD Voice-service Measurements," *ITU-T Recommendation P.562*, May 2000.

[74] W. Jiang and H. Schulzrinne, "Speech Recognition Performance as an Effective Perceived Quality Predictor," in *Proceedings of International Workshop on Quality of Service (IWQOS)*, (Miami, FL, USA), May 2002.

[75] European Telecommunications Standards Institute, "Speech Communication Quality from Mouth to Ear of 3.1 kHz Handset Telephony across Networks," *Tech. Report. ETR 250*, 1996.

[76] J. Allnatt, "Subjective Rating and Apparent Magnitude," *International Journal Man - Machine Studies*, vol. 7, pp. 801–816, 1975.

[77] S. Möller and J. Berger, "Describing Telephone Speech Codec Quality Degradations by Means of Impairment Factors," *J. Audio Eng. Soc.*, vol. 50, pp. 667–680, Sep. 2002.

[78] International Telecommunicaion Union, "Definition of Categories of Speech Transmission Quality," *ITU-T Recommendation G.109*, Sep. 1998.

[79] E. Nordström, J. Carlström, O. Gällmo, and L. Asplund, "Neural Networks for Adaptive Traffic Control in ATM Networks," *IEEE Communications Magazine*, pp. 43–49, October 1995.

[80] J. E. Neves, M. J. Leitao, and L. B. Almeida, "Neural Networks in B-ISDN Flow Control: ATM Traffic Prediction or Network Modeling," *IEEE Communications Magazine*, pp. 50–56, October 1995.

[81] A. D. Doulamis, N. D. Doulamis, and S. D. Kollias, "An Adaptable Neural-Network Model for Recursive Nonlinear Traffic Prediction and Modeling of MPEQ Video Sources," *IEEE Transactions on Neural Networks*, vol. 14, pp. 150–166, Jan 2003.

[82] A. Bhattacharya, A. G. Parlos, and A. F. Atiya, "Prediction of MPEQ-Coded Video Source Traffic Using Recurrent Neural Networks," *IEEE Transactions on Signal Processing*, vol. 51, pp. 2177–2190, August 2003.

[83] H. Yousefi'zadeh, "A Neural-Based Technique for Estimating Self-Similar Traffic Average Queueing Delay," *IEEE Communications Letters*, vol. 6, pp. 419–421, Oct 2002.

[84] H. Yousefi'zadeh, E. A. Jonckheere, and J. A. Silvester, "Utilizing Neural Networks to Reduce Packet Loss in Self-Similar Teletraffic Patterns," in *Proceedings of IEEE International Conference on Communications ICC'03*, vol. 3, (Anchorage USA), pp. 1942–1946, May 2003.

[85] M. M. Meky and T. N. Saadawi, "Prediction of Speech Quality Using Radial Basis Functions Neural Networks," in *Proceeding of Second IEEE Symposium on Computers and Communications*, (Alexandria, Egypt), pp. 174–178, July 1997.

[86] International Telecommunicaion Union, "Method for Objective Measurement of Perceived Audio Quality," *ITU-R Recommendation BS.1387*, Nov. 2001.

[87] R. Koodli and R. Ravikanth, "One-way Loss Pattern Sample Metrics," *Internet Draft, Internet Engineering Task Force, draft-ietf-ippm-loss-pattern-03*, July 2000.

[88] International Telecommunicaion Union, "Objective measuring apparatus, Appendix 1: Test signals," *ITU-T Recommendation P.50*, Feb. 1998.

[89] W. Jiang and H. Schulzrinne, "Comparison and Optimization of Packet Loss Repair Methods on VoIP Perceived Quality under Bursty Loss," in *Proceedings of the 12th International Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV)*, (Miami, Florida, USA), pp. 73–81, 2002.

[90] R. Cox and R. Perkinds, "Results of a Subjective Listening Test for G.711 with Frame Erasure Concealmen," *Committee T1*, May 1999.

[91] International Telecommunicaion Union, "The Effect of Packet Losses on Speech Quality," *ITU-T Contribution COM12-D.15*, Feb. 2001.

[92] J. S. Garofolo, L. Lamel, and W. M. Fisher, "TIMIT Acoustic-Phonetic Continuous Speech Corpus," 1993.

[93] International Telecommunicaion Union, "Objective Measurement of Active Speech Level," *ITU-T Recommendation P.56*, 1993.

[94] P. A. Barrent, R. M. Voelcker, and A. V. Lewis, "Speech Transmission over Digital Mobile Radio Channels," *BT Technol J*, vol. 14, pp. 45–56, Jan. 1996.

[95] International Telecommunicaion Union, "Methodology for Derivation of Equipment Impairment Factors from Subjective Listening-only Tests," *ITU-T Recommendation P.833*, Feb. 2001.

[96] B. V. Ghita, S. M. Furnell, B. Lines, D. L. Foll, and E. Ifeachor, "Network Quality of Service Monitoring for IP Telephony ," *Internet Research*, vol. 11, no. 1, pp. 26–34, 2001.

[97] Z. Wang and A. C. Bovik, "A Universal Image Quality Index," *IEEE Signal Processing Letters*, vol. 9, pp. 81–84, March 2002.

[98] American National Standards Institute, "American National Standard for Telecommunications - Digital Transport of One-way Video Signals-parameters for Objective Performance Assessment," *ANSI T1.801.03*, 2003.

[99] C. Hoene, B. Rathke, and A. Wolisz, "On the Importance of a VoIP Packet," in *Proc. of ISCA Tutorial and Research Workshop on the Auditory Quality of Systems*, April 2003.

[100] iLBC website, 2003. `http://www.ilbcfreeware.org/`.

[101] C. Boutremans and J. Y. Le Boudec, "Adaptive Joint Playout Buffer and FEC Adjustment for Internet Telephony," in *Proceedings of IEEE INFOCOM'2003*, (San-Francisco, CA), pp. 652–662, April 2003.

[102] SNNS, "Stuttgart Neural Network Simulator," 2002. `http://www-ra.informatik.uni-tuebingen.de/SNNS/`.

[103] S. Haykin, *Neural Networks: A Comprehensive Foundation, 2nd edition.* Prentice Hall, 1999. ISBN 0-13-273350-1.

[104] S. Lawrence, C. L. Giles, and A. Tsoi, "What Size Neural Network Gives Optimal Generalization? Convergence Properties of Backpropagation," Tech. Rep. UMIACS-TR-96-22 and CS-TR-3617, University of Maryland, April 1996. `http://www.neci.nec.com/~lawrence/papers/minima-tr96/`.

[105] T. Masters, *Practical Neural Network Recipes in C++.* Academic Press, 1993. ISBN 0-12-479040-2.

[106] V. Paxson, "End-to-end Routing Behavior in the Internet," *IEEE/ACM Transactions on Networking*, vol. 5, no. 5, pp. 601–615, 1997.

[107] A. Mukherjee, "On the Dynamics and Significance of Low Frequency Components of Internet Load," *Technical Report MS-CIS-92-83/DSL-12, University of Pennsylvania*, Dec. 1992.

[108] O. Hagsand, K. Hanson, and I. Marsh, "Measuring Internet Telephony Quality: Where Are We Today," in *Proceedings of IEEE Globecom: Global Internet*, (Rio De Janeiro, Brazil), Dec 1999.

[109] J. Rosenberg, *Distributed Algorithms and Protocols for Scalable Internet.* PhD thesis, Columbia University, PhD thesis, 2001.

[110] K. Fujimoto, S. Ata, and M. Murata, "Statistical Analysis of Packet Delays in the Internet and its Application to Playout Control for Streaming Applications," *IEICE Trans. on Communications*, vol. E00-B, June 2001.

[111] W. Jiang 2002. `http://www.cs.columbia.edu/~wenyu/research/ott_mon-v1.3b.tar.gz`.

[112] D. Sanghi, A. K. Agrawala, O. Gudmundsson, and B. N. Jain, "Experimental Assessment of End-to-end Behavior on Internet," in *Proceedings of IEEE Infocom*, pp. 867–874, March 1993.

[113] S. B. Moon, P. Skelly, and D. Towsley, "Estimation and Removal of Clock Skew from Network Delay Measurements," in *Proc. of IEEE Infocom*, (New York, USA), March 1999.

[114] P. Brady, "A Technique for Investigating on/off Patterns of Speech," *Bell Labs Tech. Journal*, vol. 44, pp. 1–22, Jan 1965.

[115] W. Jiang and H. Schulzrinne, "Analysis of On-off Patterns in VoIP and Their Effect on Voice Traffic Aggregation," in *Proceedings of International Conference on Computer Communications and Networks (ICCCN)*, (Las Vegas, USA), pp. 82–87, Oct 2000.

[116] K. Fujimoto, S. Ata, and M. Murata, "Playout Control for Streaming Applications by Statistical Delay Analysis," in *Proceedings of IEEE International Conference on Communications (ICC)*, vol. 8, pp. 2337–2342, June 2001.

[117] R. Ramachandran, J. Kurose, D. Towsley, and H. Schulzrinne, "Adaptive Playout Mechanisms for Packetized Audio Applications in Wide-area Networks," *Proc. of IEEE Infocom*, vol. 2, pp. 680–688, 1994.

[118] S. B. Moon, J. Kurose, and D. Towsley, "Packet Audio Playout Delay Adjustment: Performance Bounds and Algorithms," *Multimedia Systems*, vol. 6, pp. 17–28, 1998.

[119] V. Ramos, C. Barakat, and E. Altman, "A Moving Average Predictor for Playout Delay Control in VoIP," in *Proc. Quality of Service - IWQoS 2003, 11th International Workshop*, (Berkeley, CA, USA), pp. 155–173, June 2003.

[120] K. Fujimoto, S. Ata, and M. Murata, "Adaptive Playout Buffer Algorithm for Enhancing Perceived Quality of Streaming Applications," in *Proceedings of IEEE Globecom2002*, vol. 3, pp. 2451–2457, Nov 2002.

[121] M. Gardner, V. S. Frost, and D. W. Petr, "Using Optimization to Achieve Efficient Quality of Service in Voice over IP Networks," in *Proceedings of IPCCC 2003-The 22nd International Performance, Computing, and Communications Conference*, (Phoenix,Arizona), April 2003.

[122] Telecommunications Industry Association, "Voice quality recommendations for IP telephony," *TIA/EIA Telecommunications Systems Bulletin, TSB116*, March 2001.

[123] A. Feldmann, "Characteristics of TCP Connection Arrivals," *Technical report, AT & T Labs Research*, 1998. `http://citeseer.nj.nec.com/feldmann98characteristics.html`.

[124] Y. J. Liang, N. Frber, and B. Girod, "Adaptive Playout Scheduling and Loss Concealment for Voice Communication over IP Networks," *IEEE Trans. on Multimedia*, vol. 5, pp. 532–543, Dec. 2003.

[125] R. Eejaie, M. Handley, and D. Estrin, "RAP: An End-to-end Rate-based Congestion Control Mechanism for Realtime Streams in the Internet," in *Proc. IEEE INFOCOM'99*, pp. 21–25, March 1999.

[126] F. Beritelli, G. Ruggeri, and G. Schembra, "TCP-Friendly Transmission of Voice over IP," in *Proceedings of IEEE International Conference on Communications*, vol. 2, (New York USA), pp. 1204–1208, April 2002.

[127] A. Barberis, C. Casetti, J. D. Martin, and M. Meo, "A Simulation Study of Adaptive Voice Communications on IP Networks," *Computer Communications*, vol. 24, pp. 757–767, 2001.

[128] H. Sanneck, N. T. L. Le, M. Haardt, and W. Mohr, "Selective Packet Prioritization for Wireless Voice over IP," in *Proc. of Fourth International Symposium on Wireless Personal Multimedia Communication*, (Aalborg, Denmark), Sep. 2001.

[129] J. C. D. Martin, "Source-driven Packet Marking for Speech Transmission over Differentiated-Services Networks," in *Proceedings of IEEE ICASSP*, (Salt Lake City, Utah, USA), pp. 753–756, May 2001.

[130] "The Network Simulator - NS-2," `http://www.isi.edu/nsnam/ns`.

[131] K. Nichols, V. Jacobson, and L. Zhang, "A Two-bit Differentiated Services Architecture for the Internet," *RFC 2638, IETF*, July 1999.

[132] P. Pieda, J. Ethridge, M. Baines, and F. Shallwani, "A Network Simulator, Differentiated Services Implementation," July 2000. Open IP, Nortel Networks.

[133] International Telecommunicaion Union, "Subjective Performance Assessment of Telephone-band and Wideband Digital Codecs," *ITU-T Recommendation P.830*, Feb. 1996.

[134] International Telecommunicaion Union, "A Silence compression scheme for G.729 optimized for terminals conforming to Recommendation V.70," *ITU-T Recommendation G.729 – Annex B*, March 1996.

[135] S. Pennock, "Accuracy of the Perceptual Evaluation of Speech Quality (PESQ) algorithm," in *Proceedings of Online Workshop on Measurement of Speech and Audio Quality in Networks*, (Prague, Czech Republic), Jan. 2002.

[136] A. Radke, "Speech Quality of Heterogeneous Networks Involving VoIP: Are Time-Varying Impairments Additive to Classical Stationary Ones," in *Proc. 1st ISCA Tutorial and Research Workshop on Auditory Quality of Systems*, (Akademie Mont-Cenis, D-Herne), pp. 63–70, April 2003.

[137] C. Perkins, O. Hodson, and V. Hardman, "A Survey of Packet-loss Recovery Techniques for Streaming Audio," *IEEE Network Magazine*, vol. 12, pp. 40–48, Sept./Oct. 1998.

[138] J.-C. Bolot, S. Fosse-Parisis, and D. F. Towseley, "Adaptive FEC-Based Error Control for Internet Telephony," in *Proc. IEEE Infocom'99*, (New York, USA), pp. 1453–1460, March 1999.

[139] Z. Li, L. Sun, Z. Qiao, and E. Ifeachor, "Perceived Speech Quality Driven Retransmission Mechanism for Wireless VoIP," in *Proceedings of IEE Fourth International Conference on 3G Mobile Communication Technologies*, (London, UK), pp. 395–399, June 2003.

[140] G. C. Canavos, *Applied Probability and Statistical Methods*. Little, Brown and Company, 1984. ISBN 0-316-12778-7.

Appendix A

Statistical Analysis

1. Correlation Coefficient

If $\{(x_i, y_i);\ i = 1, 2, \ldots, n\}$ are the pairs of values X and Y, then the Pearson correlation coefficient is calculated as below:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\left[\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2\right]^{1/2}} \qquad -1 \leq r \leq 1 \tag{A.1}$$

where $\overline{x}$ and $\overline{y}$ are the mean of x and y.

2. Confidence Interval (CI)

Standard Deviation:

$$Stdev = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}} \tag{A.2}$$

Confidence Interval (CI):

$$CI = t \cdot \frac{Stdev}{\sqrt{n}} \tag{A.3}$$

The Confidence Interval is between $\overline{x} - t \cdot \frac{Stdev}{\sqrt{n}}$ to $\overline{x} + t \cdot \frac{Stdev}{\sqrt{n}}$

For 90% confidence interval of T distribution, $t = 1.652$ when sample numbers are 300 [140].

Appendix B

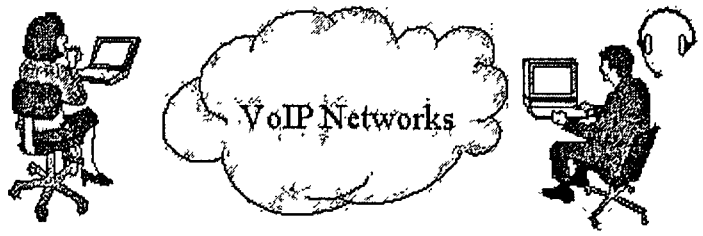
Online VoIP Mean Opinion Score (MOS) Test

1. Online VoIP Mean Opinion Score (MOS) Test

   (from website: http://www.tech.plymouth.ac.uk/spmc/people/lfsun/mos/)

2. Online MOS Test Result (from website, for a test carried out on May 8th, 2003 by 39 students)

Online VoIP Mean Opinion Score Test



Instructions:

Thanks for your time to attend VoIP Mean Opinion Score (MOS) test at Signal Processing & Multimedia Communications Group (SPMC), Department of Communication and Electronic Engineering (DCEE), University of Plymouth.

During this test, you will be asked to listen several test speech clips and rate the overall speech quality according to the following opinion scale.

Rating	Definition	Description
5	Excellent	e.g. a perfect AM radio reception
4	Good	e.g. long distance telephone quality, such as PCM (PSTN)
3	Fair	e.g. communication quality, such as GSM (requires some hearing effort)
2	Poor	e.g. low bit rate vocoder, such as LPC (hard to understand the speech)
1	Bad	communications breakdown

1. Please listen carefully the test speech clips (about 10-15 seconds each) with a head phone and give your opinion score from 5 to 1, or more detail with one digit after decimal point, such as 3.4.

2. Please do your best to make the scores fair and comparable during the whole test. The score should reflect the overall speech quality or your overall impression based on clarity and degree of distortion. You can listen many times before you give a

score.

3. There are totally 10 degraded speech clips within 2 groups. After you finish all the test, please press submit button to submit your opinion scores.

4. The whole test will take about 10 - 15 minutes.

For record-keeping purpose, could you please leave your full name:

Full Name:

Where do you do this test (please choose one)?

○At home ○At office ◉At lab ○At library ○At others

Please listen the following test speech clip and **adjust** your headphone volume to comfortable level.

Test speech clip

Now you can start MOS test!

Group	Test Speech	MOS score
1	Speech Clip 1	
	Speech Clip 2	
	Speech Clip 3	
	Speech Clip 4	
	Speech Clip 5	
2	Speech Clip 1	
	Speech Clip 2	
	Speech Clip 3	
	Speech Clip 4	
	Speech Clip 5	

Submit Test Reset

(Make sure you have entered your name, before you press Submit. Do not submit twice.)

MOS Test Results

Thanks for your contribution to the MOS test.

The current MOS test results are as following.

Groups	Conditions	MOS	Min	Max	Stdev	Count
1	1	3.25	2	4	0.50	39
	2	2.37	1	3.1	0.49	39
	3	2.16	1	4	0.70	39
	4	3.57	2	5	0.80	39
	5	1.30	1	2	0.43	39
2	1	1.15	1	3	0.38	39
	2	2.04	1	3	0.55	39
	3	1.81	1	4	0.85	39
	4	2.63	1.3	4	0.59	39
	5	2.12	1	3	0.60	39

Thank You! Have a Good Day!

Last updated 7/5/2003 by Lingfen Sun, SPMC, DCEE, University of Plymouth

Appendix C

Selected Published Papers

1. L Sun and E Ifeachor, New Models for Perceived Voice Quality Prediction and their Applications in Playout Buffer Optimization for VoIP Networks, in Proceedings of IEEE International Conference on Communications (IEEE ICC 2004), Paris, France, June 2004.

2. L Sun and E Ifeachor, New Methods for Voice Quality Evaluation for IP Networks, in Proceedings of 18^{th} International Teletraffic Congress (ITC-18), Berlin, Germany, 31 August - 5 September 2003, pp. 1201 – 1210.

3. L Sun and E Ifeachor, Prediction of Perceived Conversational Speech Quality and Effects of Playout Buffer Algorithms, in Proceedings of IEEE International Conference on Communications (IEEE ICC 2003), Anchorage, USA, May 2003, pp. 1 – 6.

4. L Sun and E Ifeachor, Perceived Speech Quality Prediction for Voice over IP-based Networks, in Proceedings of IEEE International Conference on Communications (IEEE ICC 2002) New York, USA, April 2002, pp.2573 – 2577.

5. L Sun, G Wade, B Lines and E Ifeachor, Impact of Packet Loss Location on Perceived Speech Quality, in Proceedings of 2nd IP-Telephony Workshop (IPTEL '01), Columbia University, New York, April 2001, pp.114 – 122.

New Models for Perceived Voice Quality Prediction and their Applications in Playout Buffer Optimization for VoIP Networks

Lingfen Sun and Emmanuel Ifeachor
Centre for Signal Processing & Multimedia Communication
School of Computing, Communications and Electronics
University of Plymouth
Plymouth PL4 8AA, U.K.
Email: L.Sun@plymouth.ac.uk; E.Ifeachor@plymouth.ac.uk

Abstract—**Perceived voice quality is an important metric in VoIP applications. The quality is mainly affected by network impairments such as delay, jitter and packet loss. Playout buffer at the receiving side can be used to compensate for the effects of jitter based on a tradeoff between delay and loss. The main aim in this paper is to find an efficient perceived quality prediction method for perceptual optimization of playout buffer. The contributions of the paper are three-fold. First, we propose an efficient new method for predicting voice quality for buffer design/optimization. The method can also be used for voice quality monitoring and for QoS control. In the method, non-linear regression models are derived for a variety of codecs (e.g. G.723.1/G.729/AMR/iLBC) with the aid of ITU PESQ and the E-model. Second, we propose the use of minimum overall impairment as a criterion for buffer optimization. This criterion is more efficient than using traditional maximum Mean Opinion Score (MOS). Third, we show that the delay characteristics of Voice over IP traffic is better characterized by a Weibull distribution than a Pareto or an Exponential distribution. Based on the new voice quality prediction model, the Weibull delay distribution model and the minimum impairment criterion, we propose a perceptual optimization buffer algorithm. Preliminary results show that the proposed algorithm can achieve the optimum perceived voice quality compared with other algorithms under all network conditions considered.**

I. INTRODUCTION

In Voice over IP (VoIP) applications, delay, jitter and packet loss are the main network impairments that affect perceived voice quality. Jitter can be partially compensated for by using a playout buffer at the receiving end, but this introduces further delay and additional packet loss. A tradeoff is necessary between increased packet loss and buffer delay to achieve satisfactory results for any playout buffer algorithm.

In the past, the choice/design of buffer algorithms was largely based on buffer delay and loss performance (e.g. a design objective could be to achieve a minimum average delay for a specified packet loss rate [1]–[3] or minimum late arrival loss [1]. This approach is inappropriate as it does not provide a direct link to perceived speech quality. From QoS perspective, the choice of the best buffer algorithm for a given situation should be determined by the likely perceived speech quality. The importance of this is now starting to be recognised [4]–[6]. For example, in [5], perceived voice quality is used to control the playout buffer in order to maximise the MOS values in terms of delay and loss. The concept of perceptual optimization has also been extended to other QoS control problems, such as joint playout buffer/FEC control [7] to maximise MOS values in terms of delay, loss and rate.

However, current methods of perceptual optimization are based on assumptions about perceived voice quality which are inappropriate. In [5], the method is based on the assumption that the effects of packet loss and delay on voice quality are linearly additive on the MOS scale which is doubtful. A further assumption is that the relationship between MOS and packet loss for codecs is linear which is not correct for most codecs. It has also been suggested in [7] that one equation may be used to represent the impairments due to packet loss for all codecs. This may not be appropriate, especially for newer codecs.

In all perceptual-based buffer design/optimisation and QoS control for VoIP, voice quality is used as the key metric because it provides a direct link to user perceived QoS. However, this requires an efficient and accurate objective way to measure perceived voice quality. Most current methods [7] [8] use the E-model [9] to predict voice quality, but the E-model requires subjective tests to derive model parameters which is time-consuming and often impractical. As a result, the E-model is only applicable to a limited number of codecs and network conditions. It is also inevitable that discontinuities exist in subjective results [10] because only a limited range of scenarios can be tested for. PESQ [11] gives a good measure of voice quality, but it is not appropriate for optimisation because of the overhead involved in its use in real-time.

In this paper, we have extended the method and developed new models which can be used for voice quality monitoring, buffer design/optimisation and for QoS control applications. As the method is based on end-to-end objective measurement instead of subjective tests, it can be easily applied to new codecs and network conditions.

For perceived buffer design, it is important to understand the delay distribution modeling as it is directly related to buffer loss. The characteristics of packet transmission delay

over Internet can be represented by statistical models which follow Normal, Exponential, Pareto and Weibull distributions depending on applications. For example, the delay distribution for Internet packets (for a UDP traffic) has been shown to be consistent with an Exponential distribution [12], whereas, Pareto distribution may be the most appropriate one to represent the tail delay characteristics for streaming media [13]. As delay characteristics may change with networks and applications, it is unclear what the appropriate delay distribution modelling is the best fit for current VoIP traffic. This motivated us to investigate the delay distribution modelling for VoIP trace data collected internationlly.

The contributions of the paper are three-fold:

(1) A new method for predicting voice quality for VoIP. In the method, a non-linear regression model is derived for each codec with the aid of the PESQ and the E-model. We illustrate the method for four modern codecs - G.729, G.723.1, AMR and iLBC. (2) Second, we propose the use of minimum impairment as a criterion for buffer optimization. This criterion is more efficient than using traditional maximum MOS score. (3) Third, we show that the delay characteristics of VoIP traffic is better characterized by a Weibull distribution than a Pareto or an Exponential distribution. Based on the new voice quality prediction model, the Weibull distribution model and the minimum impairment criterion, we propose a perceptual optimization buffer algorithm. Preliminary results show that the proposed algorithm can obtain the best voice quality when compared with other algorithms under the network conditions considered.

The remainder of the paper is structured as follows. In Section II, a new method for predicting voice quality is presented. In Section III, the perceptual optimization and minimum impairment criterion, and the delay distribution are discussed. In Section IV, a perceptual optimization buffer algorithm is proposed and the performance is compared with other algorithms. Section V concludes the paper.

II. New Models for Predicting Voice Quality

Fig 1a illustrates how the E-model may be used to predict voice quality in VoIP applications. Information about the codec, packet loss rate and delay is suitably transformed by the I_e and I_d models and then processed by the E-model to produce a MOS value. The MOS value is a prediction of what the perceived voice quality would be under these conditions. However, the I_e model is codec dependent and as indicated above, the derivation of the model parameters for each codec requires subjective tests which is impractical.

An important aim of our work is to develop an objective method which can be used to derive the I_e model for any codec without the need for subjective tests. The proposed method is depicted in Fig 1b and is based on the PESQ [11] (and the new PESQ-LQ [14]). The reference speech files are first encoded and then processed in accordance with the network impairments parameter values and then decoded to generate the degraded speech. The degraded speech and the reference speech are then processed by PESQ (or PESQ-LQ) to provide a MOS value. The MOS values can then be suitably transformed to give measured I_e values. As shown later, given a set of measured I_e values for a codec we can then derive an I_e model for the codec using regression techniques without the need for subjective tests.

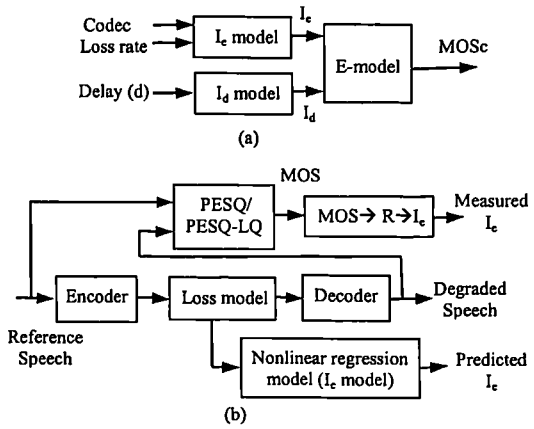


Fig. 1. (a) An illustration of how to predict voice quality using the E-model, (b) Prediction of I_e model using the PESQ

We will illustrate this for four modern codecs which are relevant for VoIP - G.729 (8 Kb/s), G.723.1 (6.3 Kb/s), AMR (the highest mode, 12.2 Kb/s and the lowest, 4.75 Kb/s) and iLBC (15.2 Kb/s). In the study, the reference speech database was taken from the ITU-T data set [15]. Packet loss was generated from 0% to 30%, in an incremental step of 3% and Bernoulli loss model was used for simplicity. PESQ-LQ (Listening Quality), the latest improvements on PESQ algorithm, is also included for comparison.

For each speech sample in the ITU-T data set, a MOS (PESQ or PESQ-LQ) score is obtained by averaging over 30 different packet loss locations (via different random seed setting) in order to remove the influence of loss location. Further, the MOS score for one loss rate is obtained by averaging over all speech samples (a total of 16 samples, consisting of 8 males and 8 females), so that the influence of gender is removed. The relationships between the average MOS and packet loss rate (expressed as ρ) for each of the four codecs are shown in Fig 2.

From Fig 2, it can be seen that PESQ-LQ has a much lower MOS score when the loss rate is high. iLBC shows the best voice quality when ρ is high (over 4%). AMR (H, 12.2 Kb/s) has the highest MOS score when ρ is zero. AMR (L, 4.75 Kb/s) has the lowest quality no matter with or without loss.

For the same data, the relationships between packet loss rate, ρ and the equipment impairment factor, I_e, for the four codecs are shown in Fig 3. The relationship between the MOS vs. ρ in Fig 2 can be converted to the Equipment impairment I_e, (measured I_e in Fig 1b) vs. ρ via Equations 1 and 2 [6].

$$R = 3.026MOS^3 - 25.314MOS^2 + 87.060MOS - 57.336 \quad (1)$$

$$I_e = R_0 - R \quad (2)$$

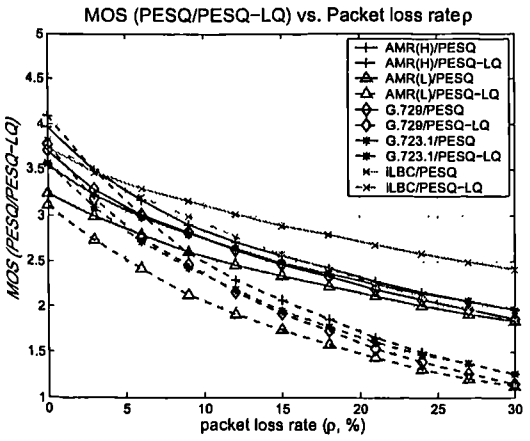


Fig. 2. MOS vs. Packet loss rate ρ

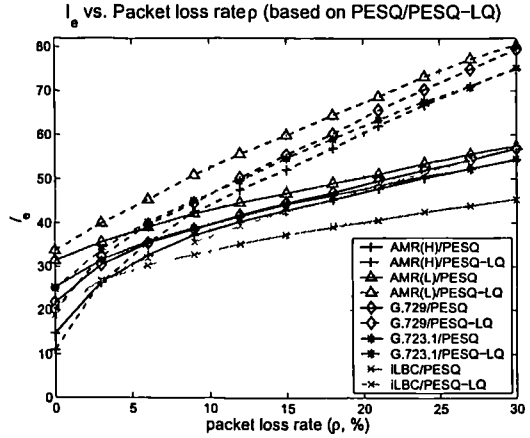


Fig. 3. I_e vs. Packet loss rate ρ

From Fig 3, a non-linear regression model (similar to the logarithm fitting function in [10]) can be derived for each codec based on the PESQ or PESQ-LQ by the least squares method and curve fitting. The derived I_e model has the following form:

$$I_e = a\ln(1 + b\rho) + c \tag{3}$$

where ρ is the packet loss rate in percentage. The parameters (a, b and c) for different codecs under PESQ and PESQ-LQ are shown in Table I and Table II, respectively.

TABLE I
PARAMETERS FOR DIFFERENT CODECS (PESQ)

Parameters	AMR (H)	AMR (L)	G.729	G.723.1	iLBC
a	16.68	30.86	21.14	20.06	12.59
b*100	30.11	4.26	12.73	10.24	9.45
c	14.96	31.66	22.45	25.63	20.42

In Fig 3, the I_e value for zero packet loss represent the codec impairment itself. The AMR (L, 4.75 Kb/s) has the largest codec impairment (the largest I_e), whereas, the AMR (H, 12.2 Kb/s) has the lowest I_e value. G.729 and iLBC codecs have similar I_e values at zero packet loss, but iLBC has the lowest I_e of all four codecs when loss rate is over 3%.

TABLE II
PARAMETERS FOR DIFFERENT CODECS (PESQ-LQ)

Parameters	AMR (H)	AMR (L)	G.729	G.723.1	iLBC
a	40.0	93.66	63.20	60.09	31.72
b*100	12.11	2.16	4.84	4.17	7.22
c	12.2	33.82	21.71	25.79	19.65

Considering that the effect of codec impairment (without loss) is fixed for any codec, I_e can be viewed as consisting of two main components: $I_e = I_{ec} + I_{e\rho}$, where I_{ec} is the impairment without loss and $I_{e\rho}$ the impairment with loss. The $I_{e\rho}$ vs. ρ for PESQ-LQ is shown in Fig 4.

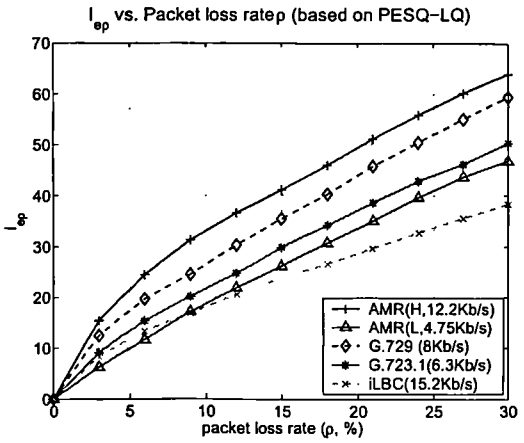


Fig. 4. $I_{e\rho}$ vs. packet loss rate ρ

Fig 4 illustrates the ability of a codec to cope with network packet loss. From the curves, the iLBC has the lowest slope, whereas, the AMR (H) has the highest. This further shows that the iLBC has an obvious high robustness to packet loss. AMR (H) has the highest MOS score under zero packet loss condition (as shown in Fig 2), but it has the least ability to cope with packet loss (quality decreases sharply as packet loss increases). From Fig 4, it is clear that to use only one curve (or model) as suggested in [7] to represent all codecs is inappropriate. Obviously with emerging new network codecs (with even higher robustness to loss), the diversity in the ability of codecs to cope with packet loss will be even larger. Thus, we recommend to use different models for each codec for accurate parameter optimization or quality control.

Unlike I_e which is codec dependent, the delay impairment factor, I_d, is common to all codecs. I_d can be derived by a simplified fitting process in [10] with Eq 4 as below.

$$I_d = 0.024d + 0.11(d - 177.3)H(d - 177.3)$$
$$where \begin{cases} H(x) = 0 \text{ if } x < 0 \\ H(x) = 1 \text{ if } x \geq 0 \end{cases} \tag{4}$$

By using I_d (Eq 4) and I_e (Eq 3), voice quality can be predicted using the E-model as shown in Fig 1(a). These

models can be used for buffer optimization as described in the following section or for voice quality monitoring/control.

III. Perceptual Optimization of Playout Delay and Delay Distribution Modelling

A. Optimum voice quality and minimum impairment criterion

For perceptual buffer optimization, the aim is to achieve an optimum end-to-end voice quality (e.g. in the term of MOS score). Considering the relationship of voice quality and impairments (e.g. packet loss and delay), the problem of an optimum voice quality can be converted to an issue of minimum impairment.

We define an overall impairment function I_m which is a function of delay d and packet loss ρ, with $I_m = f(d, \rho) = I_d + I_{e\rho}$. If ignoring other impairments such as echo, R factor can be further simplified as Eq 5.

$$R = 93.2 - I_d - I_e = (93.2 - I_{ec}) - I_m \quad (5)$$

As MOS increases monotonously with R (see Eq 1), a maximum R value corresponds to a maximum MOS score. Further when maximum R is obtained, it corresponds to a minimum impairment function, I_m.

Using Eqs 3 and 4, I_m can be further expressed as:

$$I_m = I_{e\rho} + I_d = a\ln(1 + b\rho) + 0.024d + 0.11(d - 177.3)H(d - 177.3) \quad (6)$$

where a and b are codec related constants. d is the playout delay, including network delay (d_n) and buffer delay (d_b). ρ consists of network packet loss (ρ_n) and buffer loss (ρ_b).

It is a trade-off between delay and packet loss for any buffer algorithm. When playout delay $d \uparrow$ ($I_d \uparrow$), then buffer loss $\rho_b \downarrow$ ($I_{e\rho} \downarrow$). When $d \downarrow$ ($I_d \downarrow$), then $\rho_b \uparrow$ ($I_{e\rho} \uparrow$). An optimum playout delay d can be obtained when minimum impairment I_m is reached. A minimum impairment criterion for buffer optimization is set and defined in Table III.

TABLE III
DEFINITION OF A MINIMUM IMPAIRMENT CRITERION

Given:	delay d_n, loss ρ_n and codec type
Required to estimate:	an optimized playout delay d_{opt}
Such that:	minimum I_m can be reached

Obviously seeking for a minimum I_m is more efficient than for traditionally seeking for a maximum MOS, as it is not necessary to convert I_m to R and then to MOS (a 3^{rd} order polynomial) for each buffer adaptation/calculation.

In order to find the best tradeoff of delay d and packet loss ρ, we now look at the relationship between d and ρ (or buffer loss ρ_b) which is described in the next section.

B. Playout delay and delay distribution funtion

The relationship between d and ρ_b can be described by delay Cumulative Distribution Function (CDF) which is defined as $F(x) = P(X \leq x)$. For a playout delay d, the buffer loss ρ_b can be calculated as $\rho_b = P(X \geq d) = 1 - F(d)$.

To understand the delay distribution for current VoIP traffic, we investigated the delay distribution for the VoIP trace data which were collected from Internet connections between Uni. of Plymouth (UoP), UK to Beijing Uni. of Posts & Telecomm. (BUPT) China, UoP to Columbia Uni.(CU), USA, UoP to Darmstadt Uni. of Tech.(DUT), Germany, and UoP to Nanchang (NC) China. A detailed description of trace data collection is in our previous paper [6]. We experimented with Exponential, Pareto and Weibull distributions. The definition of CDF for three distributions are listed in Table IV. The RMSE (Root Mean Square Error) for the five selected traces for different approximation models are tabulated in Table V. The empirical and fitted CDF for trace 1 is illustrated in Fig 5.

TABLE IV
DEFINITION OF SEVERAL CUMULATIVE PROBABILITY DISTRIBUTIONS

Distribution	Exponential	Pareto	Weibull
CDF: F(x)	$1 - e^{-(x-\mu/\beta)}$	$1 - (k/x)^{\alpha}$	$1 - e^{-((x-\mu)/\alpha)^{\gamma}}$

TABLE V
RMSE OF DIFFERENT DISTRIBUTION FUNCTIONS FOR DIFFERENT TRACES

Traces	Exp.	Pareto	Weibull
1 (BUPT → UoP, 7/6/02)	0.04467	0.03916	0.005607
2 (UoP → CU, 3/04/02)	0.0007858	0.0007389	0.0007233
3 (UoP → BUPT, 11/06/02)	0.05228	0.03398	0.01064
4 (UoP → DUT, 10/06/02)	0.01926	0.02029	0.004269
5 (UoP → NCT, 30/05/02)	0.01376	0.01366	0.003032

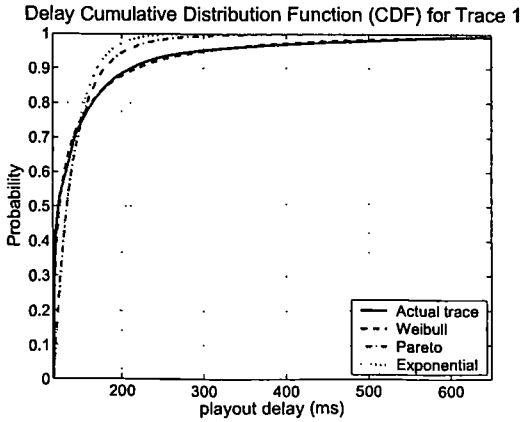


Fig. 5. Empirical and fitted CDF for trace 1 (Weibull: μ = 116, α = 15.9, γ = 0.4451; Pareto: k = 116, α = 5.277; Exp: μ = 116, β = 23.47)

From Table V, Fig 5, it can be seen that Weibull distribution achieved the best fit for all five traces (with the lowest RMSE) when compared with Pareto and Exponential distribution. As a result, we use Weibull distribution to represent delay distribution in the perceptual-based buffer design.

C. Perceptual Optimization of Playout Delay

Given network packet loss ρ_n (in percentage) and playout delay d, the buffer loss (ρ_b) for a Weibull Distribution can be

calculated in the following Equation.

$$\rho_b = (1 - \rho_n/100)P(X \geq d) = (1 - \rho_n/100)e^{-((d-\mu)/\alpha)^\gamma} \quad (7)$$

Replacing ρ_b of Eq 7 into Eq 6, overall impairment factor, I_m, can be depicted as follows:

$$I_m = 0.024d + 0.11(d - 177.3)H(d - 177.3) + a \ln\left[1 + b[\rho_n + (100 - \rho_n)e^{-((d-\mu)/\alpha)^\gamma}]\right] \quad (8)$$

For a given trace segment, the Weibull Distribution location parameter μ equals to the minimum network delay d_n, the scale parameter α and shape parameter γ can be estimated using maximum-likelihood-estimator (MLE) method [16] (we use Matlab's *weibfit* function directly in the simulation for simplicity). The optimum playout delay (d_{opt}) can be obtained by searching for a playout delay d which meets the minimum impairment criterion. Fig 6 shows an example of I_m vs. d for a trace segment (with 1000 packets) selected from trace #1. In order to see how different codecs and objective measurement methods (e.g. PESQ/PESQ-LQ) affect playout delay optimization, Fig 6 also shows I_m vs. d for AMR122 and iLBC using PESQ and PESQ-LQ. It is obviously that the optimum playout delay differs according to which codec and which objective quality method are used. The iLBC/PESQ has the smallest optimum playout delay (d_1) and AMR122/PESQ-LQ has the largest one (d_4). The minimum impairment values obtained also differ for different codecs, with iLBC (PESQ) the lowest I_m and AMR122 (PESQ-LQ) the highest I_m.

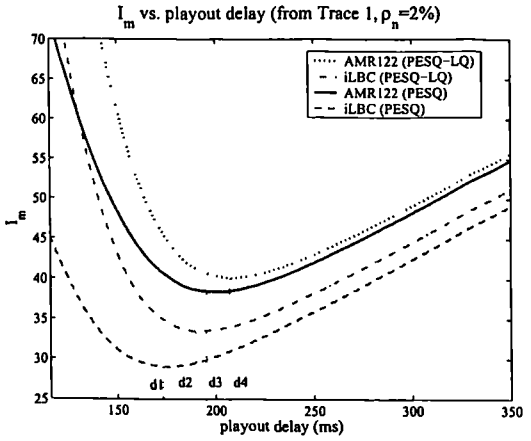


Fig. 6. Optimization of playout delay

IV. Perceptual Optimization Buffer Algorithm

A. Perceptual Optimization Buffer Algorithm (P-optimum)

In Section III, we have derived Eq 6 which relates impairment (I_m) with playout delay (d) and network packet loss (ρ_n) for a given trace. This can be used directly for perceived jitter buffer algorithm optimization. For simplicity, we only use the equation for G.723.1 codec to show the concept of perceptual optimization buffer design.

As network traces show high possibility of "spike" which is defined as a number of packets that have significantly higher delays than the rest. The "spike" state can be regarded as an exceptional state in the trace data (seen as a short-term delay characteristics) and the remaining "non-spike" state can be analysed in long-term delay distribution. Several algorithms exist for spike detection. For example, Ramachandran et al [1] proposed to use $(n_i - n_{i-1}) > threshold$ as the detection of a start of a spike (n_i is the network delay for i^{th} packet). This accounts for the spike with a sudden increase of delay. However through the analysis of our collected Internet trace data, we notice that large amounts of spike is with gradual increase which cannot be detected by the above algorithm. Considering spikes with sudden or gradual increase, we follow the spike detection based on ($n_i > threshold$) as in [2]. The proposed perceptual optimum buffer algorithm (P-optimum) is shown in Algorithm 1.

Algorithm 1 Perceptual Optimum Buffer Algorithm

```
For every packet i received, calculate the network delay n_i
if mode == SPIKE then
  if n_i ≤ tail × old_d then
    mode = NORMAL /* the end of a spike */
  end if
else if n_i > head × d_i then
  mode = SPIKE /* the beginning of a spike */
  /* save d_i to detect the end of a spike later */
  old_d = d_i
else
  /* normal model*/
  - update delay records for the past W packets
end if

At the beginning of a talkspurt
if mode == SPIKE then
  d_i = n_i /* estimated playout delay d_i */
else
  - obtain (μ, α, γ) in Weibull distribution
  - search playout delay d for d_i = d_opt which meets → min(I_m)
end if
```

Depending on the current mode, the playout delay for the next talkspurt is estimated differently in each mode as shown in Algorithm 1. In spike-detection mode, the delay of the first packet of a talkspurt becomes the estimated playout delay for the talkspurt. Otherwise, the perceptually optimized playout delay based on the delay distribution of the last W packets (in *NORMAL* mode) is used. The large the W value, the less responsive the scheme to adapt. The *head* and *tail* parameters are used to set the threshold for spike detection.

B. Performance Analysis and Comparison

In order to compare with other buffer algorithms, we also implemented "exp-avg", "fast-exp", "min-delay", "spk-delay" and "adaptive" algorithms (detail see [6]). The results are shown in Table VI for the above five traces. The window size W is set to 1000. The *head* is 4 and the *tail* is 2,

as suggested in [2]. During the experiment, we changed the window size W from 100 packets (3sec) to 10,000 packets (300 sec, as suggested by [2] and [5]), we noticed that the performance (the overall MOS score) does not show a big difference within the range. We chose W of 1000 (30 sec), as it is an appropriate duration for the I_m or MOS calculation and has higher computation efficiency than the longer window length.

From Table VI, it can be seen that "P-optimum" obtained the optimum MOS scores among all the five traces. Our previous proposed "adaptive" algorithm achieved sub-optimum results. The remaining buffer algorithms can achieve good results only in some traces, but not for all. It has to be mentioned that P-optimum has the highest complexity, whereas the others including "adaptive" have the similar low complexity.

TABLE VI
PERFORMANCE COMPARISON FOR DIFFERENT BUFFER ALGORITHMS

Trace	Buffer algorithms	Loss p (%)	Delay d (ms)	MOS
Trace 1	Exp-avg	4.9	298.5	2.01
	Fast-exp	1.5	750.8	1.00
	Min-delay	9.4	208.8	2.34
	Spk-delay	10.4	225.0	2.18
	Adaptive	9.0	208.1	2.37
	P-optimum	10.5	188.2	2.43
Trace 2	Exp-avg	1.8	27.3	3.28
	Fast-exp	0	35.9	3.44
	Min-delay	1.7	27.3	3.29
	Spk-delay	3.4	24.9	3.15
	Adaptive	0	35.9	3.44
	P-optimum	0.1	44.5	3.42
Trace 3	Exp-avg	18.2	432.4	1.01
	Fast-exp	14.3	1408.6	1.00
	Min-delay	22.1	312.7	1.30
	Spk-delay	23.8	325.4	1.22
	Adaptive	22.1	299.8	1.35
	P-optimum	32.0	171.1	1.80
Trace 4	Exp-avg	5.9	24.0	2.97
	Fast-exp	4.3	94.4	2.99
	Min-delay	5.3	23.0	3.01
	Spk-delay	7.6	21.9	2.86
	Adaptive	4.3	72.8	3.02
	P-optimum	5.1	34.4	3.02
Trace 5	Exp-avg	3.5	150.9	2.98
	Fast-exp	0.5	176.1	3.22
	Min-delay	4.5	148.8	2.91
	Spk-delay	6.3	144.3	2.79
	Adaptive	0.5	170.3	3.22
	P-optimum	0.5	169.8	3.22

V. CONCLUSIONS

In this paper, we have proposed a non-linear regression model to predict perceived voice quality based on PESQ/PESQ-LQ and E-model. We derived new models for variety of codecs for VoIP applications. These models can be efficiently used for voice quality monitoring, perceptual buffer design/optimization and other QoS control purposes. As the method is based on objective tests instead of subjective tests, it can be easily extended to other new codecs or network conditions. We proposed the use of minimum overall impairment as a criterion for quality control and buffer optimization. This is more efficient than traditional maximum MOS score criterion. We investigated delay distribution characteristics based on VoIP trace data collected. We find that a Weibull distribution is a better fit than a Pareto and Exponential distribution. Based on the derived voice quality prediction models, the Weibull delay distribution model and the minimum impairment criterion, we proposed a perceptual optimization playout buffer algorithm. Preliminary results show that the proposed algorithm can achieve the optimum perceived voice quality compared with other algorithms under all network conditions considered.

As the work is based on the buffer adaptation at the beginning of each talkspurt, it cannot adapt to any delay changes during a talkspurt. Future work will extend the idea to consider buffer adaptation during a talkspurt in order to achieve a best trade-off among delay, loss and end-to-end jitter.

ACKNOWLEDGEMENT

The work is supported in part by an EU grant under the Sixth Framework Programme (BIOPATTERN Project, No. 508803) and by Acterna.

REFERENCES

[1] R. Ramachandran, J. Kurose, D. Towsley, and H. Schulzrinne, "Adaptive playout mechanisms for packetized audio applications in wide-area networks," *Proc. of IEEE Infocom*, vol. 2, pp. 680–688, 1994.

[2] S. B. Moon, J. Kurose, and D. Towsley, "Packet audio playout delay adjustment: performance bounds and algorithms," *Multimedia Systems*, vol. 6, pp. 17–28, 1998.

[3] V. Ramos, C. Barakat, and E. Altman, "A moving average predictor for playout delay control in VoIP," in *Proc. Quality of Service - IWQoS 2003, 11th International Workshop*, pp. 155–173, June 2003.

[4] A. P. Markopoulou, F. A. Tobagi, and M. Karam, "Assessment of VoIP quality over Internet backbones," *Proc. of IEEE Infocom*, 2002.

[5] K. Fujimoto, S. Ata, and M. Murata, "Adaptive playout buffer algorithm for enhancing perceived quality of streaming applications," *Proceedings of IEEE Globecom2002*, Nov 2002.

[6] L. Sun and E. Ifeachor, "Prediction of perceived conversational speech quality and effects of playout buffer algorithms," in *Proceedings of IEEE ICC'03*, pp. 1–6, 2003.

[7] C. Boutremans and J. Y. Le Boudec, "Adaptive joint playout buffer and FEC adjustment for Internet telephony," in *Proceedings of IEEE INFOCOM'2003*, pp. 652–662, April 2003.

[8] M. Gardner, V. S. Frost, and D. W. Petr, "Using optimization to achieve efficient quality of service in voice over IP networks," in *Proceedings of IPCCC 2003*, April 2003.

[9] International Telecommunicaion Union, "The E-model, a computational model for use in transmission planning," *ITU-T Recommendation G.107*, July 2000.

[10] R. G. Cole and J. Rosenbluth, "Voice over IP performance monitoring," *Journal on Computer Communications Review*, vol. 31, April 2001.

[11] International Telecommunication Union, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *ITU-T Recommendation P.862*, Feb 2001.

[12] J. C. Bolot, "Characterizing end-to-end packet delay and loss in the internet," *Journal of High-Speed Networks*, vol. 2, pp. 305–323, Dec 1993.

[13] K. Fujimoto, S. Ata, and M. Murata, "Playout control for streaming applications by statistical delay analysis," *Proceedings of IEEE ICC*, vol. 8, pp. 2337–2342, June 2001.

[14] A. W. Rix, "Comparison between subjective listening quality and p.862 PESQ score," *Proceedings of Online Workshop Measurement of Speech and Audio Quality in Networks*, pp. 17–25, May 2003.

[15] International Telecommunicaion Union, "Objective measuring apparatus, Appendix 1: Test signals," *ITU-T Recommendation P.50*, Feb 1998.

[16] A. Feldmann, "Characteristics of TCP connection arrivals," *Technical report, AT&T Labs Research*, 1998. `http://citeseer.nj.nec.com/feldmann98characteristics.html`.

New Methods for Voice Quality Evaluation for IP Networks

Lingfen Sun and Emmanuel Ifeachor

Department of Communication and Electronic Engineering
University of Plymouth
Plymouth PL4 8AA, United Kingdom

The need to evaluate voice quality in VoIP (Voice over IP) applications is an important requirement for technical and commercial reasons. This may involve subjective and/or objective voice quality measurements, but existing methods may not always be appropriate for VoIP applications. The aims of the study reported in the paper are to investigate new subjective and objective measurement methods for VoIP applications. The contributions of the paper are two-fold. First, we present a new subjective, Internet-based MOS (Mean Opinion Score) test methodology which allows rapid assessment of voice quality. We conducted MOS tests using the new method as well as traditional MOS tests under different VoIP network conditions and compared the results using objective measurement methods. Preliminary results show that the Internet-based MOS test compares well with traditional MOS test (correlation coefficients of 0.95). Second, we propose novel conversational intrusive and non-intrusive speech quality measurement methods, based on the ITU PESQ and E-model to extend the applicability of existing methods. We illustrate the application of the novel approach to the derivation of model parameters for a new codec for VoIP applications (the AMR codec).

1. INTRODUCTION

The convergence of communications and computer networks has led to a rapid growth in real-time applications such as Voice over IP (VoIP). However, IP networks are not designed to support real-time applications and factors such as network delay, jitter and packet loss lead to unpredictable deterioration in perceived voice quality. A major challenge that faces network and service providers is how to measure or predict voice quality accurately and efficiently for Quality of Service (QoS) monitoring and/or control purposes to meet technical and commercial requirements.

Voice quality measurement can be carried out using either subjective or objective methods. The Mean Opinion Score (MOS) is the most widely used subjective measure of voice quality and is recommended by the ITU [1]. A MOS value is normally obtained as an average opinion of quality based on asking people to grade the quality of speech signals on a five-point scale (Excellent, Good, Fair, Poor and Bad) under controlled conditions. In voice communication systems, MOS is the internationally accepted metric as it provides a direct link to voice quality as perceived by the end user. The inherent problem in subjective MOS measurement is that it is slow, time consuming, expensive and cannot be used for long-term or large scale

voice quality monitoring in an operational network infrastructure. This has made objective methods very attractive for meeting the demand for voice quality measurement in communication networks.

Objective measurement of voice quality in modern communication networks can be intrusive or non-intrusive. Intrusive methods are more accurate, but normally are unsuitable for monitoring live traffic because of the need for a reference data and to utilise the network. A typical intrusive method is based on the latest ITU standard, P.862 Perceived Evaluation of Speech Quality (PESQ) Measurement Algorithm [2]. This involves a comparison of the reference and the degraded speech signals to predict the listening-only one-way MOS score.

Non-intrusive methods do not need a reference signal and are appropriate for monitoring live traffic. ITU-T E-model [3] is the most widely used non-intrusive voice quality measurement method and may be used to predict conversational MOS score directly from IP network and/or terminal parameters [4,5].

Subjective methods are crucial for benchmarking objective methods. The need remains for an efficient method for subjective MOS tests. In the paper, we introduce a new subjective, Internet-based MOS test methodology, intended to simplify MOS tests for VoIP applications. We conducted MOS tests using the new method as well as traditional MOS tests for speech samples under different VoIP network conditions, and compared the performance with the latest ITU-T objective measurement methods (e.g. PESQ and E-model). Preliminary results show that the Internet-based MOS test compares well with traditional MOS test method.

The PESQ algorithm provides a more accurate measure of quality, but it is intrusive and can only predict one-way listening speech quality. In practice, there is a need for objective measure of conversational speech quality to account for interactivity in voice communication. In this paper, we present a novel conversational, intrusive speech quality measurement method, based on a combination of PESQ and E-model.

The current E-model [3] and extended E-models [4,5] rely on subjective tests for the derivation of model parameters when they are used for VoIP applications. This is obviously time consuming, impractical and hinders the use of the E-model in new and emerging applications. Non-subjective derivation of model parameters has recently been proposed [6], but this is limited to only codec impairments and is unsuitable for VoIP applications. In this paper, we introduce a new objective method for deriving model parameters for VoIP applications. This should extend the applicability of the E-model to meet the needs of new and emerging applications.

The remainder of the paper is structured as follows. In Section 2, a new Internet-based MOS test and results of its use to assess voice quality under different VoIP network conditions are presented. New intrusive and non-intrusive conversational speech quality measurement methods are presented in Sections 3 and 4, respectively. Section 5 concludes the paper.

2. INTERNET-BASED MOS TEST

The traditional MOS test methodology has been in existence for about 20 years [7] and today its use range from the assessment of codec's quality to the assessment of VoIP network quality. The stringent test requirements for traditional tests have not changed (e.g. the use of a sound-proof room) in that time and are essential for a proper assessment of voice quality in many cases, e.g. quality assessment of codecs, as the difference between codecs may be subtle and difficult to detect. However, for VoIP applications, new impairments, such as packet loss,

are much more perceptible than impairments from codecs. This has led us to investigate the possibility of conducting MOS tests under normal working/studying environments, as this is more realistic and subjects are more relaxed. In a sound-proof room, some subjects may find it uncomfortable, psychologically, to carry out tests in the confined environments. This has led to an Internet-based subjective test methodology, which has the following advantages:

- It is closer to reality than the traditional method. Subjects remain in familiar environments, e.g. an office or a laboratory, to carry out the test. This is clearly less stressful and the test can be done at the subject's own pace.
- It is possible to organise subjective tests at more locations around the world.
- It allows easier access to a larger number of subjects (e.g. 40 - 80 subjects can be tested at the same time in one or two large rooms, e.g. a laboratory).

Overall, it has the benefits of efficiency, realism, wide access and ease of organisation. It can save money and time compared to P.800. Of course, the main disadvantage of Internet-based MOS test is the lack of a controlled testing environment (e.g. very low background noise) compared to P.800.

In a previous Internet-based MOS study [8], we carried out the tests without control (subjects did their own tests on their own computer, in their own office and at their own preferred time slot). We have extended this by introducing a measure of control to reduce the impact of different working environments on the results. In the Internet-based MOS test method, all subjects sit in a large project room which they use regularly. It is not a sound-proof room, but it is quiet and has Internet access.

In the next section, we will present the set-up used to evaluate voice quality using both subjective and objective methods. The preliminary test results and analysis are then presented.

2.1. Voice Quality Evaluation

Figure 1 depicts the set-up used for the voice quality evaluation. It is a PC-based software system that allows the simulation of key processes in voice over IP and speech quality measurement. Objective voice quality measurements were made with the ITU PESQ and E-model to enable us to compare Internet-based MOS tests with traditional MOS tests. Reference speech files were first encoded using G.723.1 codec [9] and then processed in accordance with network parameter values in trace data files (see later) and then decoded to generate degraded speech signals (a fixed jitter buffer, for simplicity, was used to remove the effects of jitter. Packets that arrive too late are discarded).

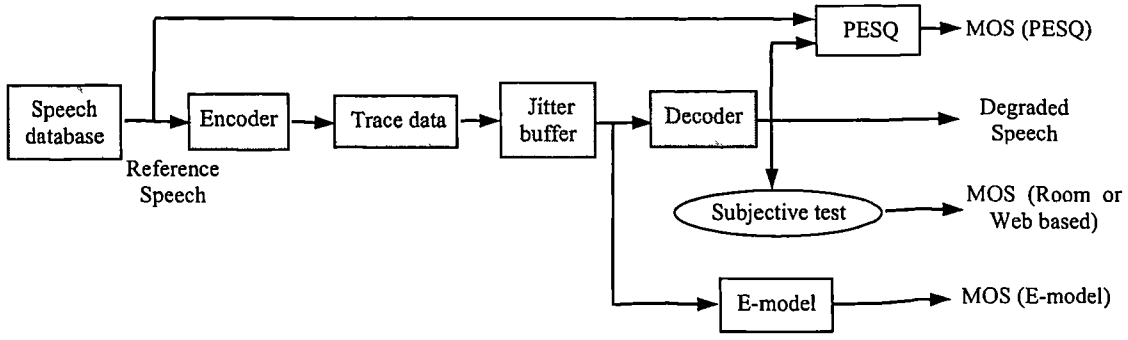


Figure 1. VoIP Speech Quality Evaluation Set-up

We collected Internet trace data between the UK and USA, UK and China and UK and Germany using a UDP/IP probe tool in the past year. A detailed description of the trace data collection and the effects of jitter buffer can be found in our paper [10]. Some trace data (e.g. those with a 30 ms packet interval, consistent with G.723.1) was selected for the quality evaluation.

The reference speech database was taken from the TIMIT data set [11]. Each speech sample consists of four short sentences spoken by four different male and female speakers in order to keep a balanced design. Each speech sample was about 10 to 15 seconds long. A total of 10 speech samples were chosen for the VoIP quality evaluation. We also chose ten different network conditions from the Internet trace data set, covering packet loss rate (including late arrival loss due to jitter) from 0% to 30%. Ten degraded speech samples were generated and used for the quality evaluation.

Subjective Tests were carried out using two methods -- Internet-based (or web-based) and P.800-based tests (Room-based). A website for the MOS test was created at the following URL:
http://www.tech.plymouth.ac.uk/spmc/people/lfsun/mos/

The 10 degraded speech samples were put on the web, together with a brief instruction about the MOS test. 15 undergraduate students were invited to attend the controlled Internet-based MOS test. The tests were carried out in the project laboratory which they use regularly. The room was quiet and similar to a normal office environment. Brief instructions were given by a supervisor before the test. The students were asked to perform the test at their own pace. The tests took about 15 minutes to complete. When all the students had submitted their opinion scores, the MOS score were calculated and expressed as Web_MOS.

In order to compare the results of the controlled Internet-based MOS tests with similar P.800 tests, we carried out another round of MOS test in a small, quiet room (a sound-proof room was not available on-site and is another motivation for investigating the web-based approach). The room is about 8 square meters and 3.5 meters high with a desktop and a laptop PCs. A similar test procedure to the web-page was created locally. The same 10 degraded speech samples were chosen. The 15 students were invited again to carry out the tests, one by one. The test spanned over two days because of the numbers involved and their availability. The MOS score for the room-based MOS test was expressed as Room_MOS.

We also conducted MOS tests using the ITU PESQ algorithm [2] and the E-model [3] in order to compare the subjective test results with objective measurements. Listening-only speech quality measurements were considered in order to keep the same conditions. By comparing the reference speech and the degraded speech, an objective MOS score was obtained from the PESQ algorithm. This MOS score was referred to as MOS (PESQ) or PESQ_MOS. For the E-model, only the effects of the Equipment Impairment (I_e) were taken into account (the effects of delay, I_d, was not considered). This gives a listening-only MOS score which is referred to as MOS (E-model) or E-model_MOS.

Table 1. Objective and subjective MOS scores for different speech samples

Test samples	1	2	3	4	5	6	7	8	9	10
PESQ	3.18	2.65	2.85	3.74	2.02	1.95	2.42	2.59	2.93	2.54
E-model	2.90	2.50	2.56	3.92	1.00	1.04	1.63	2.07	2.71	2.41
Room-based	3.36	2.65	2.85	3.31	1.41	1.15	2.13	2.13	2.97	2.34
Web-based	3.37	2.32	2.52	4.00	1.22	1.11	2.19	2.04	2.90	2.35
Loss rate (%)	5.68	9.51	8.85	0.21	29.5	23.1	18.5	14.6	7.25	10.6

2.2. Test Results and Analysis

The results for the 10 degraded speech samples for each of the four methods of voice quality measurement (PESQ, E-model, Internet-based and Room-based MOS tests) are summarised in Table 1. The sequences of the test speech samples from 1 to 10 are the same with that on the MOS test website. The calculated packet loss rates are included to give an indication of the impact of the network impairment.

The relationships between the MOS scores and packet loss rates for the different MOS test methods are depicted in Figure 2. From the figure, it can be seen that the MOS scores for all four evaluation methods decrease with increasing packet loss rate. When the packet loss rate is low, the E-model, PESQ and Web-based MOS scores are quite close. When packet loss rate is high, PESQ seems to over predict the voice quality, whilst the E-model does the opposite. As the E-model predicts voice quality directly from network parameters (e.g. packet loss rate), it does not consider factors such as packet loss location. Also as VAD (voice activity detection) was not activated in the simulation, packet loss in the silence period will not be perceived by subjects, but it was still taken into account in the E-model calculation. This is partly why the E-model gives lower MOS scores compared to the other methods when the packet loss rate is high. Room-based and Web-based MOS scores are close, except in the case when there is almost no packet loss. This is probably because the background noise (e.g. from the fan) of the computer for Room-based tests is higher than those for Web-based test.

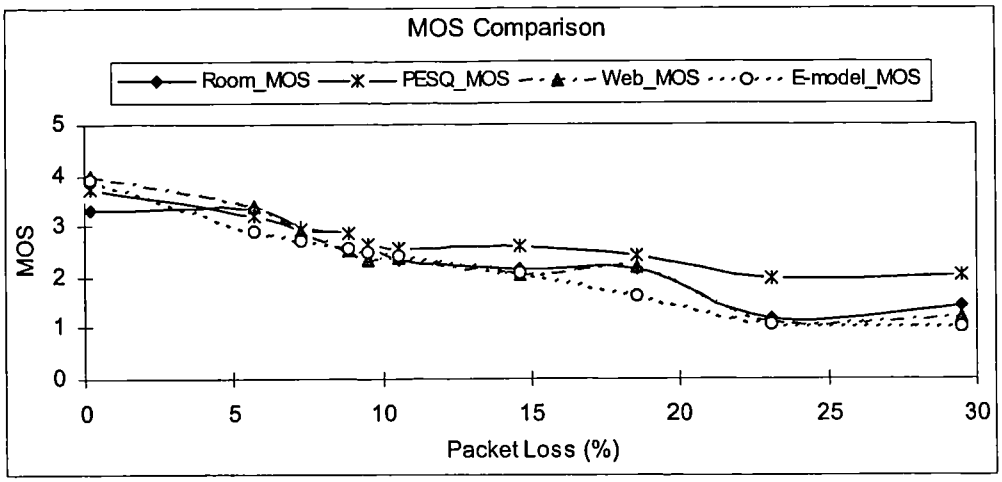


Figure 2. MOS comparison for objective and subjective test methods

Table 2. Correlation coefficients for MOS comparison

Name	PESQ vs Room_MOS	PESQ vs Web_MOS	E-model vs Room_MOS	E-model vs Web_MOS	Web_MOS vs Room_MoS	E-model vs PESQ
Correlation Coefficients	0.933	0.984	0.935	0.964	0.952	0.975

The Pearson correlation coefficient between the results of subjective and objective methods were calculated and the results are shown in Table 2. From the table, it can be seen that the Internet-based MOS test (Web_MOS) compares well with the traditional MOS test (Room_MOS) (correlation coefficients of 0.95). This suggests that with the Internet-based MOS test it is possible to obtain similar results to those of traditional MOS tests for VoIP applications. For objective measurement methods (E-model and PESQ) and subjective methods (Room-based and Web-based), the correlation coefficients are between 0.93 to 0.98. This shows that the two objective methods can both predict subject MOS score well, although both seem to predict Web-based MOS better than Room-based MOS.

3. INTRUSIVE CONVERSATIONAL SPEECH QUALITY MEASUREMENT

PESQ is an intrusive method and can only predict one-way listening speech quality. It does not consider the impact of end-to-end delay which is important for interactivity in communications. We propose a conversational quality measurement method which exploits the accuracy of the PESQ algorithm and the delay model of the E-model, see Figure 3.

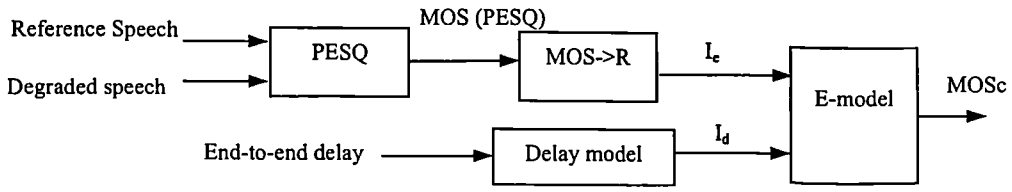


Figure 3. An intrusive conversational speech quality measurement

As shown in Figure 3, the listening *MOS* score is first obtained using PESQ (referred to as *MOS (PESQ)*). The *MOS* score is then converted to a rating factor (the *R* factor) and then to an equipment impairment value (the I_e value). The conversational MOS score, *MOSc*, is obtained by combining the I_e value and the effects of end-to-end delay (I_d values). The detailed procedures are given in the following steps.

The ITU-T G.107 [3] defines the relationships between the *R* to *MOS* as in (1).

$$\begin{aligned} MOS &= 1 && for \quad R \le 0 \\ MOS &= 1 + 0.035R + R(R-60)(100-R)7\times10^{-6} && for \quad 0 < R < 100 \\ MOS &= 4.5 && for \quad R \ge 100 \end{aligned} \tag{1}$$

However, Equation (1) cannot be inverted directly to obtain the R values because it covers the R-values between 0 and 6.5, which maps to MOS scores below 1. Thus, the R-values are normally restricted to the range [6.5, 100] , with R-values below 6.5 assigned a $MOS = 1$

before inversion. Candono's Formula [15] can be used to obtain the R-values from the MOS, but the equations are very complicated. Thus, we propose to use a simplified 3rd order polynomial fitting (Equation 2) to obtain the equation for mapping from MOS to R values. The fitting curve and original curve from G.107 are shown in Figure 4.

$$R = 3.026MOS^3 - 25.314MOS^2 + 87.060MOS - 57.336 \quad (2)$$

If we consider only the equipment impairment, R values can be converted to I_e using Equation (3) (a default R value of 93.2 is used [3]).

$$I_e = 93.2 - R \quad (3)$$

The delay impairment factor, I_d, represents all impairments due to delay of voice signals, and includes impairments due to Listener Echo, Talker Echo and Absolute delay. I_d can be calculated by a series of complex equations [3]. The relationships between I_d and one-way delay can be expressed by a simplified equation (4) according to [5] . The corresponding fitting curve and the curve from G.107 [3] are shown in Figure 5.

$$I_d = 0.024T_a + 0.11(T_a - 177.3)H(T_a - 177.3)$$
$$where \begin{cases} H(x) = 0 & \text{if } x < 0 \\ H(x) = 1 & \text{if } x \geq 0 \end{cases} \quad (4)$$

Considering I_d and I_e, E-model R factor can be simplified as in (5).

$$R = 93.2 - I_d - I_e \quad (5)$$

From R, the conversational *MOS* score (*MOSc*) can be calculated using Equation (1). Overall, the conversational *MOS* score can be obtained from comparing the reference and degraded speech samples and taking into account the effects of end-to-end delay.

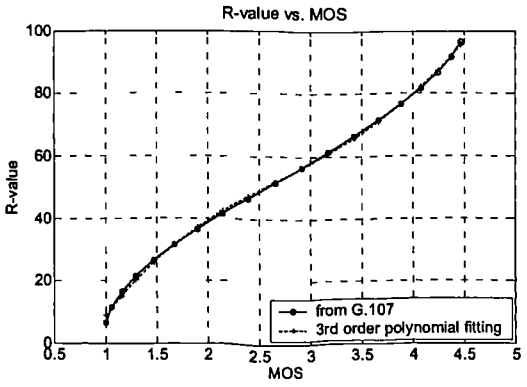


Figure 4. R-value vs. MOS score

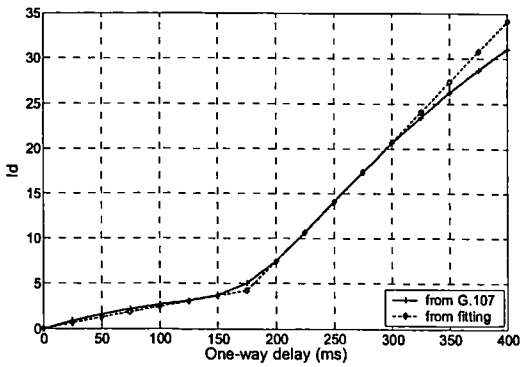


Figure 5. I_d vs. one-way delay

4. NON-INTRUSIVE OBJECTIVE MEASUREMENT

The basic E-model has been extended in several directions (e.g. extended or simplified models) and used for non-intrusive voice quality monitoring for VoIP applications. However, the equations for equipment impairment (I_e) are still based on subjective tests. Such tests are required in order to derive the model parameters for I_e for each codec (e.g. new emerging codecs), application (e.g. with/without *VAD*) and network condition (e.g. packet size, network packet loss such as random or burst packet loss). This is obviously time consuming and impractical and hinders the development of the E-model for future applications.

To improve the applicability of the E-model, we have extended it in two directions:

First, we have proposed the use of the Internet-based MOS test methodology to increase the efficiency of subjective MOS tests. As described above, the use of the Internet-based MOS test method to derive model parameters is more efficient. Second, we have proposed the use of an objective method, such as PESQ, to replace the subjective tests which are currently required for deriving the model parameters. As in Figure 3, the Equipment Impairment factor (I_e) can be derived directly from an objective measurement method (e.g. PESQ). Of course, the accuracy of the resulting model parameters will depend on the accuracy of the objective measurement method used, but this will improve as new objective measurement algorithms (e.g. the next generation of PESQ) become available.

As an example, we derived the I_e value for a new codec for VoIP applications using PESQ (the AMR [12] at the highest mode of 12.2 Kb/s. I_e model does not exists for AMR codecs at present in public domain). The procedures are as follows:

Step 1: Obtain MOS (PESQ) vs. packet loss rate for the AMR codec (Figure 6). Following the approach in Figure 1, we have used a random packet loss generator instead of real trace data. The packet size is set to 1 frame/packet (20ms). We obtained MOS (PESQ) at different packet loss rates (from 0 to 30% in steps of 3%) for the AMR Codec. Each MOS value was calculated by averaging over 25 different random seeds for both male and female speech samples from ITU-T dataset [13] to avoid the influence from packet loss location and gender.

Step 2: Convert the MOS vs. packet loss rate to I_e vs. packet loss rate as shown in Figure 7 (the curve from PESQ) using Equations (2) and (3). A logarithm fitting function, similar as in [5], can be derived as Equation (6). The fitting curve is also shown in Figure 7 (from fitting).

$$I_e = 13.2 + 15.84 * \ln(1 + 0.38 * loss) \qquad (6)$$

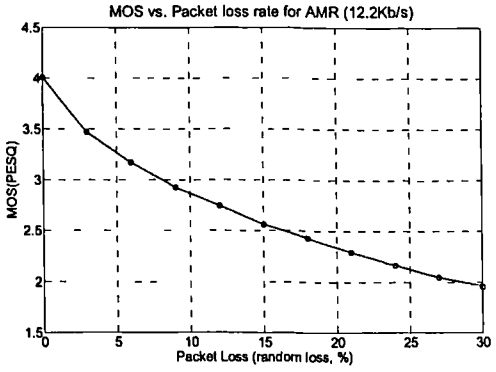


Figure 6. MOS vs. Packet loss for AMR

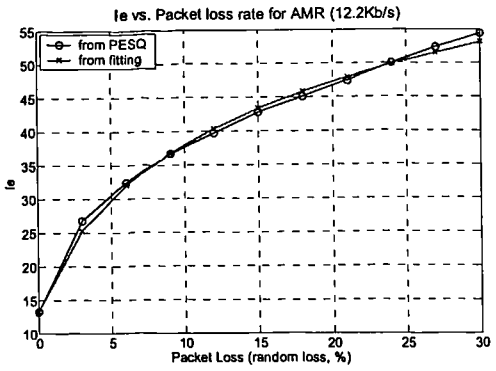


Figure 7. I_e vs. Packet loss for AMR

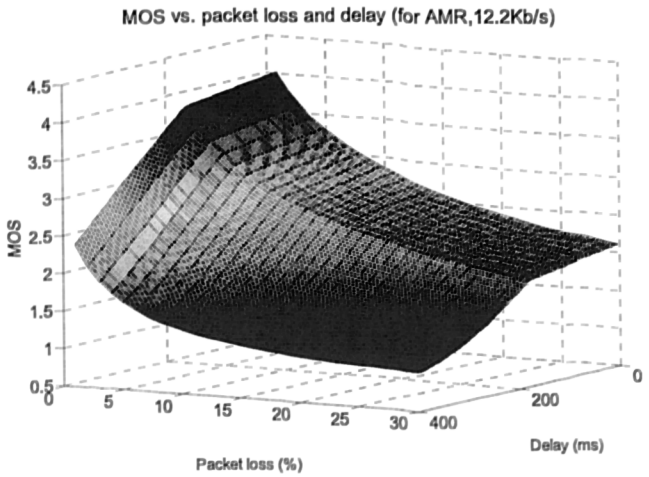


Figure 8. MOS vs. Packet loss rate and delay for AMR (12.2.Kb/s)

Step 3: Calculate the MOS for AMR codec (12.2 Kb/s mode) using Equations (6), (4), (5), and (1) for a given random packet loss rate and end-to-end delay. The MOS vs. packet loss rate and delay is shown in Figure 8.

This method can be extended to other speech codecs, current or new ones (above 4.8Kb/s [2]) and packet loss patterns (e.g. burst packet loss). It can be easily used to monitor/predict conversational speech quality from network impairments (e.g. packet loss rate and end-to-end delay) non-intrusively [5, 14].

5. CONCLUSIONS AND FUTURE WORK

In this paper, we have investigated novel subjective and objective speech quality evaluation methods for VoIP applications. We have proposed a new Internet-based subjective MOS test and carried out informal subjective MOS tests. Preliminary results show that the Internet–based MOS test compares well to the Room-based MOS test (correlation coefficients of 0.95). In general, the two ITU objective test methods (PESQ and E-model) can predict subjective MOS scores well. We have introduced improved intrusive and non-intrusive conversational speech quality measurement methods which exploit the capabilities of PESQ and E-model.

Future work will focus in two directions. First, we will investigate further the Internet based MOS test methodology by undertaking a more extensive MOS tests. We wish to establish, for example, how the test environment affects the results, what the differences between controlled Internet-based test and uncontrolled Internet-based tests and between Internet-based MOS test and formal P.800-based MOS tests (in sound-proof room) really are. Secondly, we will investigate further the new intrusive and non-intrusive measurement methods and how to use them in new applications (e.g. in perceived quality driven QoS control systems).

ACKNOWLEDGEMENT

We would like to thank Mr. Chunshui Liu for his contribution during his MSc study at the university.

We are grateful to Acterna for part sponsorship of our work.

REFERENCES

1. ITU-T Rec. P. 800, Methods for subjective determination of transmission quality, August 1996.
2. 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, February 2001.
3. ITU-T Recommendation G.107, The E-model, a computational model for use in transmission planning, 2000.
4. A. D. Clark, Modeling the Effects of Burst Packet Loss and Recency on Subjective Voice Quality, IPTEL'2001, pp. 123-127, April, New York, 2001.
5. R. G. Cole and J.H. Rosenbluth, Voice over IP Performance Monitoring, Journal on Computer Communications Review, vol. 31, no.2, April 2001.
6. S Möller and J Berger, Describing Telephone Speech Codec Quality Degradations by Means of Impairment Factors, J. Audio Eng. Soc., Vol. 50, No. 9, September 2002, pp. 667-680.
7. ITU-T P.830, Subjective Performance Assessment of Telephone-band and Wideband Digital Codecs, February 1996.
8. L Sun and E Ifeachor, Subjective and Objective Speech Quality Evaluation under Bursty Losses, Proceedings of On-line Workshop Measurement of Speech and Audio Quality in Networks (MESAQIN 2002), Prague, Czech Republic, Jan. 2002, pp.25 - 29.
9. ITU-T Recommendation G.723.1, Dual Rate Speech Coder for Multimedia Communication Transmitting at 5.3 and 6.3 kbit/s, March 1996.
10. L Sun and E Ifeachor, Prediction of Perceived Conversational Speech Quality and Effects of Playout Buffer Algorithms, to appear in the Proceedings of IEEE International Conference on Communications (ICC), Anchorage, USA, May 2003.
11. TIMIT data set, J. S. Garofolo, L. F. Lamel, W. M. Fisher, et al. TIMIT Acoustic-Phonetic Continuous Speech Corpus.
12. ETSI EN 301 704 V7.2.1 (2000-04), Digital cellular telecommunications system (Phase 2+); Adaptive Multi-Rate (AMR) speech transcoding.
13. ITU-T Recommendation P.50, Appendix 1, Test signals, 1999.
14. A. P. Markopoulou, F. A. Tobagi, M.J. Karam, Assessment of VoIP Quality over Internet Backbones, Proc. of IEEE Infocom, June 2002.
15. C. Hoene, B. Rathke and A Wolisz, On the Importance of a VoIP Packet, In Proc. of ISCA Tutorial and Research Workshop on the Auditory Quality of Systems, Germany, April 2003 (http://www.tkn.tu-berlin.de/publications/papers/paper10.pdf).

Prediction of Perceived Conversational Speech Quality and Effects of Playout Buffer Algorithms

Lingfen Sun and Emmanuel C. Ifeachor

Department of Communication and Electronic Engineering
University of Plymouth
Plymouth PL4 8AA, U.K.

Abstract— **Perceived conversational speech quality is a key quality of service (QoS) metric for voice over IP (VoIP) applications. Speech quality is mainly affected by network impairments, such as delay, jitter and packet loss. Playout buffer algorithms are used to compensate for jitter, based on a tradeoff between delay and loss, but can have a significant effect on perceived quality. The main aim in this paper is to assess how buffer algorithms affect perceived speech quality and how to choose the best algorithm and its parameters to obtain optimum perceived speech quality (in terms of an objective Mean Opinion Score). The contributions of the paper are three-fold. First, we introduce a new methodology for predicting conversational speech quality (conversational Mean Opinion Score or MOSc) which combines the latest ITU-T speech quality measurement algorithm (PESQ) and the concepts of the E-model. Second, we assess different playout buffer algorithms using the new MOSc metric on Internet trace data. Our findings indicate that, in general, end-to-end delay has a major effect on the selection of a buffer algorithm and its parameters. For small end-to-end delays, an algorithm that seeks to minimise loss is preferred, whereas for large end-to-end delays, an algorithm that aims at a minimum buffer delay is best. Third, we propose a modified buffer algorithm together with an adaptive parameter adjustment scheme. Preliminary results show that this can achieve an "optimum" perceived speech quality for all the traces considered. The results are based on Internet trace data measurements between UK and USA, UK and China, and UK and Germany.**

Keywords- Voice over IP; Conversational Speech Quality; Playout Buffer Algorithm; Jitter; Packet Loss; Perceived Quality

I. INTRODUCTION

IP networks are on a steep slope of innovation that will make them the long-term carriers of different types of traffic including speech, but they are not designed to support real-time voice communication. In voice over IP (VoIP) applications, delay, jitter (i.e. delay variation) and packet loss are the main network impairments that affect perceived speech quality. Jitter can be partially compensated for by using a playout buffer at the receiving end, but this introduces further delay (buffer delay) and additional packet loss (packets arriving after their playout times will be dropped by the receiver). A tradeoff is necessary between increased packet loss and buffer delay to achieve a satisfactory result for any playout buffer algorithm. For example, the longer the buffer delay, the lower the late arrival loss and vice versa.

In the past, the choice of a buffer algorithm was purely based on buffer delay and loss performance (e.g. minimum end-to-end delay for a given packet loss rate [1,2,3] or minimum late arrival loss [1]). Given that the ultimate purpose of a buffer algorithm is to obtain a better perceived speech quality, this approach is inappropriate, as it does not provide a direct link to perceived speech quality. From QoS perspective, the choice of the best buffer algorithm for a given situation should be determined by the likely perceived speech quality. This issue is now recognized [8], but the work so far has been limited to one adaptive buffer algorithm and a fixed parameter. It is still unclear how different buffer algorithms and parameters affect perceived quality and how to determine the buffer algorithm/parameters to achieve the optimum perceived speech quality (in terms of an objective MOS).

Perceived speech quality during a VoIP communication can be expressed as a conversational Mean Opinion Score (MOSc). MOSc values may be obtained by subjective listening tests [4] or by objective measurement methods, such as the ITU E-model [5]. The E-model (or the Extended E-model) has been widely used for measuring and assessing conversational speech quality for VoIP applications [6,7,8]. It is based on the principle that the perceptual effects of different impairments are additive on a psychophysical scale. As the E-model consists of very complicated equations and is only applicable to a limited number of codecs at present, we have developed a more general method to predict MOSc. The new method combines the latest ITU-T perceived speech quality measurement algorithm (PESQ) [9] and the concepts of the E-model. The method is suitable for any codec (from 64Kb/s to 4Kb/s) [9] and may also be used to monitor/predict conversation speech quality in practice. The accuracy of the method is limited mainly by the accuracy of the PESQ algorithm, but this can be readily replaced by, for example, the next generation PESQ algorithm, if necessary.

The main contributions of this paper are threefold. First, we introduce a new methodology to predict MOSc score, based on a combination of PESQ and E-model. Second, we assess different buffer algorithms/parameters using the new MOSc instead of existing packet loss/delay metric. Third, we propose a modified buffer algorithm and an adaptive parameter adjustment scheme which can achieve an "optimum" perceived quality for different categories of traces. To assess the quality of current VoIP networks and to evaluate the performance of

The work is sponsored in part by Acterna.

0-7803-7802-4/03/$17.00 (C) 2003 IEEE

buffer algorithms, we collected Internet trace data between UK and USA, UK and China, and UK and Germany.

The remainder of the paper is structured as follows. Section II presents the method used to collect the Internet trace data, the data, and the conversational speech quality measurement method. Section III compares and analyses the performance of different buffer algorithms and parameters using the new MOSc metric. In Section IV, we present a modified buffer algorithm and an adaptive parameter adjustment scheme. Section V concludes the paper.

II. DATA COLLECTION AND MEASUREMENT

A. Internet trace data collection

We use a UDP/IP probe tool [10] to collect and measure the main network parameters that affect voice quality. It consists of a client/server program, which runs in a local host and at a remote host. It transmits a stream of UDP/IP packets over the network to emulate VoIP traffic, and at the remote host the packets received are echoed back to the local host. Each packet has a sequence number, which indicates the order the packets were sent and can be used to deduce packets that have been lost in the network. The timestamps can be used to deduce network delay and delay variations. Similar tools have been used for experimental assessment of end-to-end behavior of Internet in the past [12,13,14] and more recently for speech quality prediction [8]. In our experiments, the size of the probe packets is set to 32 bytes. The interval between successive packets is 30 ms, which is similar to G.723.1.

In determining one-way delay, the collected trace data is first processed to remove the differences between the clocks at the two hosts and clock drift (or clock skew) [11]. Further, the data is processed to contain talkspurts and silences using a well-known on/off model with an exponential distribution [15]. A mean of 1.5sec for both talkspurts and silences is selected as in [8,16]. For about a week, we collected trace data from Internet connections between the University of Plymouth (UoP), U.K and Columbia University (CU), USA; between UoP and Beijing University of Posts & Telecommunications (BUPT), China and between UoP and Darmstadt University of Technology (DUT), Germany. These sites were selected because they are international connections with different delay characteristics. The basic information for 4 selected traces with a duration of 30 min (1800 sec) is listed in Table 1. Examples of traces (after synchronisation, deskewing and talkspurt/silence processing) are shown in Fig. 1 (a) to (d).

TABLE I. BASIC INFORMATION FOR TRACES #1 TO #4

Trace Num. #	Trace Path	Start Time (Sender)	Average Network Delay (ms)	Average Packet Loss (%)
1	BUPT → UoP	16:50pm, 0706/02, Fri	255	1.8
2	UoP → CU	13:22pm, 13/04/02, Sat	46	0.3
3	UoP → BUPT	9:11am, 11/06/02, Tue	186	14.2
4	UoP → DUT	17:44pm, 10/06/02, Mon	16	4.2

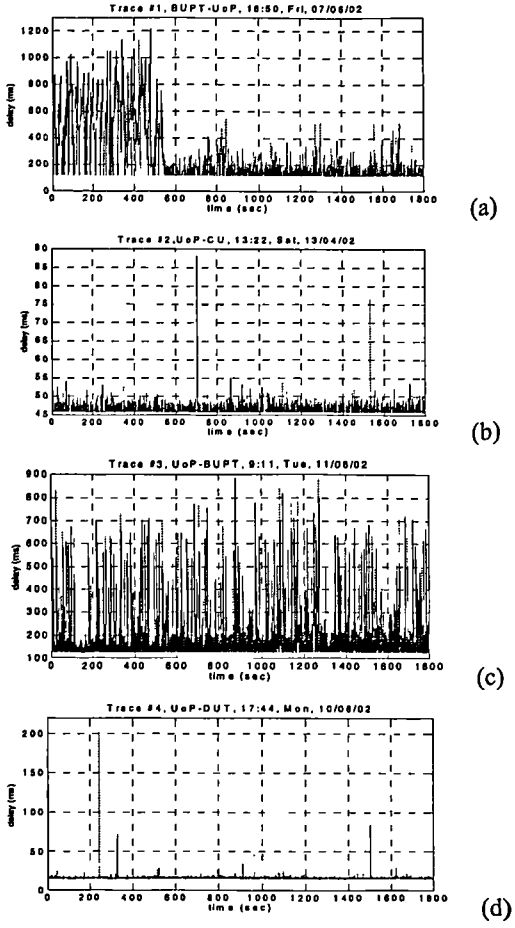


Figure 1. Trace data #1 to #4 ((a) to (d))

We classify traces into two major categories. The first category are traces with small end-to-end delay/jitter, such as traces from UoP to CU and UoP to DUP (Fig 1 (b) and (d)). The other are traces with large/medium delay/jitter, such as those between UK and China (Fig 1 (a) and (c)).

B. Conversational Speech Quality (MOSc) Measurement

The methodology for conversational speech quality measurement is based on the PESQ and the E-model (see Fig. 2). The reference speech signal is first encoded using a suitable codec (e.g. G.723.1) and then processed in accordance with the loss characteristics of the trace data to generate the degraded speech (equivalent to IP impaired speech). In practice, the relevant parameters (i.e. end-to-end delay, delay variation and packet losses) can be obtained from analysis of the RTP header and RTCP report. The ITU voice test signals [18] are used as the reference speech data in our study. The reference speech and degraded speech are then fed to the PESQ to obtain a measure of speech quality due to loss and codec. PESQ is designed for one-way listening-only perceived quality measurement and does not consider the effects of delay, which is required for conversational speech quality. The E-model concepts are used to combine the effects of loss and delay to obtain an overall quality score, MOSc.

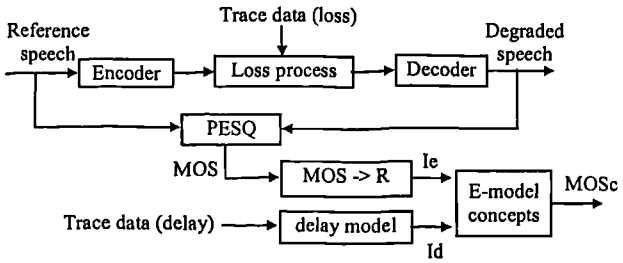


Figure 2. Schematic diagram for MOSc measurement

The PESQ is an intrusive, end-to-end measurement algorithm and requires a reference signal. However, provided a suitable local reference is available, it can be used for non-intrusive measurement [17] to exploit its greater accuracy and applicability to a wide range of codecs. We have followed this approach and extended it to account for delay in the study of the effects of buffer algorithms on IP speech quality.

Ignoring the effects of other impairments (e.g. echo), the rating scale for the E-model, R, may be simplified as follows:

$$R = R_0 - I_d - I_e \tag{1}$$

where R_0 is the optimum quality value (the default value for R_0 is 93.2 [5] which is used in the study). I_e is known as the equipment impairment factor and accounts for impairments due to non-linear codec and packet loss. I_d accounts for echo and delay. Under perfect echo cancellation conditions, I_d can be calculated by (2) [5].

$$I_d = 0 \quad \text{for} \quad T_a < 100ms \tag{2a}$$

$$I_d = 25\left\{(1+x^6)^{1/6} - 3\left[1+(x/3)^6\right]^{1/6} + 2\right\} \text{ for } T_a \geq 100\text{ms} \tag{2b}$$

where $x = (\lg[Ta/100])/\lg 2$ and Ta represents absolute delay (or end-to-end delay).

Given the R value, the corresponding MOS score can be obtained using the following relationship: [5].

$$MOS = 1 \quad \text{for} \quad R \leq 0 \tag{3a}$$

$$MOS = 1 + 0.035R + R(R-60)(100-R)7\times10^{-6} \quad \text{for } 0 < R < 100 \tag{3b}$$

$$MOS = 4.5 \quad \text{for} \quad R \geq 100 \tag{3c}$$

Using a similar 3rd order polynomial the expression for transforming MOS to R is given by (4).

$$R = 3.026x^3 - 25.314x^2 + 87.060x - 57.336 \tag{4}$$

where x represents MOS value.

Values of MOS obtained from PESQ are first transformed to R using (4) and then to I_e ($I_e = R_0 - R$). The overall score, MOSc, is obtained from (1) and (3).

For every 9 sec trace data (9 sec is chosen because it is within the recommended length for PESQ algorithm [9]), the overall packet loss (including late arrival loss) and overall end-to-end delay (including buffer delay) are calculated based on the playout buffer algorithm used (see next section for details of buffer algorithms). An average end-to-end delay (i.e. real delay) for the 9 sec trace data is also calculated and sent to delay model to get delay impairment I_d. PESQ MOS score is also transformed to I_e value. From I_e and I_d values, the conversational speech quality (MOSc) is calculated. The average MOSc score at the end of the selected trace data (30min) is calculated as the overall MOSc score (recency effect was not considered).

III. BUFFER ALGORITHMS AND PERFORMANCE ANALYSIS

Playout buffer can be fixed or adaptive. A fixed buffer cannot adapt to changing network delay conditions and this may result in poor speech quality. Thus, we have focused on adaptive buffer algorithms and adjust the buffer at the beginning of each talkspurt [1][2][3].

The notations used to describe buffer algorithms are defined in Fig. 3. For packet i, we define t_i as the send time; a_i and p_i as the arriving and playout times, respectively. n_i represents network delay and d_i is the actual end-to-end delay or "playout delay". b_i is the buffer delay.

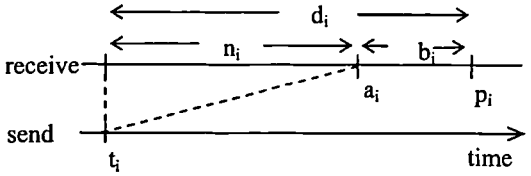


Figure 3. Timing associated with packet i

We first implemented four algorithms proposed by Ramjee et al. [1]. These four algorithms maintain running estimates of the mean and variation of network delay, i.e. $\hat{d}_i$ and $\hat{v}_i$, seen up to the arrival of the i[th] packet. If packet i is the first packet of a talkspurt, its playout time p_i is computed as:

$$p_i = t_i + \hat{d}_i + \mu * \hat{v}_i \tag{5}$$

where μ is a constant and $\hat{v}_i$ is given by:

$$\hat{v}_i = \alpha\hat{v}_{i-1} + (1-\alpha)\left|\hat{d}_i - n_i\right| \tag{6}$$

n_i is the network delay of the i[th] packet.

The playout delay for subsequent packets (e.g. packet j) in a talkspurt is kept the same as $d_j = d_i$.

The four algorithms differ only in the computation of $\hat{d}_i$.

1) Algorithm 1 ("exp-avg"): This algorithm estimates the mean delay through an exponentially weighted average.

$$\hat{d}_i = \alpha\hat{d}_{i-1} + (1-\alpha)n_i \quad (\text{with } \alpha = 0.998002) \tag{7}$$

2) Algorithm 2 ("fast-exp"):This algorithm is similar to the first, except it adapts more quickly to increases in delays by using a smaller weighting factor as delays increase:

$$\hat{d}_i = \begin{cases} \beta\hat{d}_{i-1} + (1-\beta)n_i & n_i > \hat{d}_{i-1} \\ \alpha\hat{d}_{i-1} + (1-\alpha)n_i & n_i \le \hat{d}_{i-1} \end{cases} \quad (8)$$

with $\beta = 0.75$ and $\alpha = 0.998002$ as before.

3) Algorithm 3 ("min –delay"): This is more aggressive in minimizing delays. It uses the minimum delay of all packets received in the current talkspurt. Let S_i be this set of delays.

$$\hat{d}_i = \min_{j \in s_i} \{n_j\} \quad (9)$$

4) Algorithm 4 ("spk-delay")

This algorithm contains a spike detection algorithm. During a spike, the delay estimate tracks the delays closely, but after a spike, it is similar to Algorithm 1 (with α = 0.875 under Normal mode). We avoid a detailed description here and refer the reader to [1] for details.

There are other more complicated algorithms which can achieve better spike detection than Algorithm 4, such as those mentioned in [2]. As our purpose here is not to find a better algorithm for spike detection, those algorithms are not covered.

In the first experiment, we investigated how the buffer algorithm parameters affect speech quality using MOSc metric. We assume no limitations in buffer size and adapt μ in (5) from 1 to 20, as in [2]. In comparing with the existing performance metrics, we also include the performance of average playout delay (or real delay) and average loss rate (or real loss).

The real delay and loss vs. μ for traces #1 and #2 are shown in Fig 4 (a) to (d), respectively. It is clear that the "fast-exp" has the lowest loss rate but the highest delay for both traces, as it adapts more quickly to increase in delay. The "min-delay" has the lower delay and higher loss for both traces, as it targets at minimum delay. The results for other two algorithms are between that of the "fast-exp" and the "min-delay".

Four buffer algorithms show similar trends at real delay and loss metrics for traces #1 and #2 (similar results obtained for traces #3 and #4). However, the combined effect on perceived quality shows a big difference for two categories of traces (see Figures 4 (e) to (h)). There is an obvious similarity within the same category of traces (e.g. trace #1 and #3, trace #2 and #4). This suggests that the perceived performance of the four buffer algorithms for different parameters is mainly affected by the end-to-end delay/jitter of the trace data.

For small delay/jitter traces, the MOSc score can achieve its "optimum" value when μ is set within a proper range for a certain algorithm (e.g. any μ within 1 to 20 for the "fast-exp" algorithm, and $\mu > 10$ for other three algorithms). The reason behind this is that the end-to-end delay for these two traces does not affect MOSc score, as the overall end-to-end delay is near or less than 100 ms, with the I_d in (2) near to zero. In this case, MOSc is only affected by packet loss and codec.

The performance of the four algorithms differs slightly for traces #1 and #3 (see Fig 4 (e) and (g)). It seems that "min-delay" algorithm can reach the maximum MOSc value for both traces #1 and #3 at different μ values (e.g. μ=6 for trace #1 and μ=2 for trace #3). This maximum MOSc score represents the best overall tradeoff between delay and loss for the selected traces. As the two traces have both large end-to-end delay (over 100ms), delay has a major effect on the perceived speech quality. The "min-delay" algorithm can achieve its good performance as it induces lower buffer delay among the four algorithms. For the "exp-avg" and the "spk-delay" algorithms, there also exists a μ value to achieve a maximum MOSc score, although this maximum value is lower than that of the "min-delay" algorithm. For the "fast-exp" algorithm, MOSc scores just decrease monotonously with μ increasing. This suggests that the impact on speech quality due to buffer delay induced by this algorithm is much higher than the benefits due to lower late arrival loss.

The curves of MOS (from PESQ) and MOSc vs. time for traces #1 and #2 are shown in Fig 5 (a) and (b). Fig 5 (a) is for the "min-delay" algorithm with μ of 6, while Fig 5 (b) is for the "fast-exp" with μ of 2 (both under the best MOSc scenarios). It is almost the same for MOS and MOSc for trace #2, as there is no direct impact from delay, whereas, MOSc is obviously lower than MOS for trace #1 due to the impact from delay.

IV. A MODIFIED PLAYOUT BUFFER ALGORITHM – PERCEIVED QUALITY OPTIMIZATION

From the performance analysis on these two categories of traces, we find that there is no 'best' algorithm/parameter, which can always achieve the 'best' MOSc value for all the traces. However, there is a best algorithm among the four, which is most suitable for each category of traces. For example, the "fast-exp" algorithm is preferred for low delay path/trace within a wide range of μ value (μ within 1 to 20), whereas, the "min-delay" algorithm seems the best for a longer delay trace/path under a certain μ value (μ=6 for trace #1 and μ=2 for trace #3). It suggests that a different algorithm or μ value should be chosen for different traces to achieve an "optimum" perceived quality. Based on this, we propose a modified buffer algorithm which can adapt to the preferred algorithm (e.g. "fast-exp" or "min-delay") automatically according to the running estimate of mean network delay $\hat{d}_i$. The algorithm (abbreviated as "adaptive") is as follows:

if $(\hat{d}_i \ge delay\_threshold)$

$\quad \hat{d}_i = \min_{j \in s_i} \{n_j\}$

else { if $(n_i > \hat{d}_{i-1})$

$\quad\quad \hat{d}_i = \beta\hat{d}_{i-1} + (1-\beta)n_i$

$\quad$ else $\hat{d}_i = \alpha\hat{d}_{i-1} + (1-\alpha)n_i$

}

Considering the impact of delay on MOSc (imperceptible when delay is under 150ms [6]), we first set the *delay_threshold* (mean delay) to 150ms and calculate the MOSc score under different μ values (as before). The "adaptive" algorithm can adapt to the "fast-exp" for traces #2 and #4 and to the "min-delay" for traces #1 and #3 (in most cases). The result is the same as that of their adapted algorithms in Fig 4 (e) to (h)).

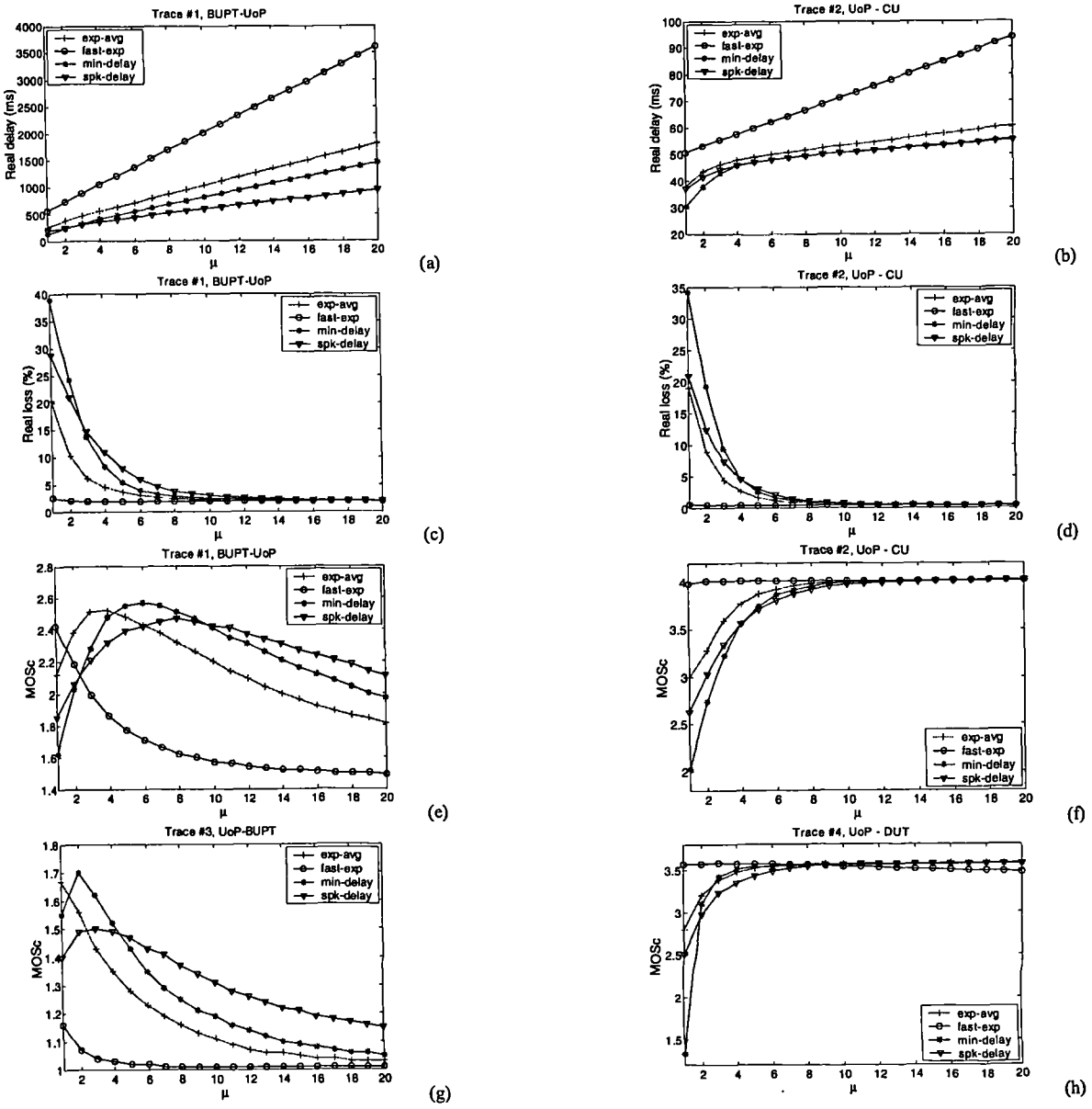


Figure 4. Performance comparison of playout buffer algorithms for Traces #1 to #4

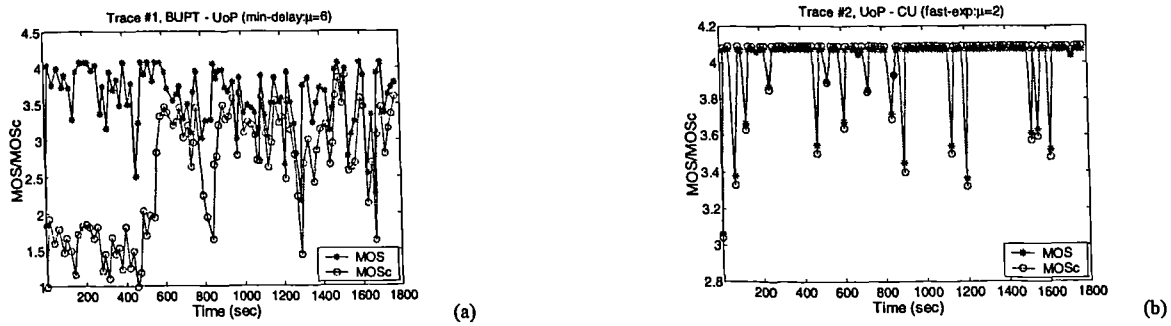


Figure 5. MOS (PESQ) and MOSc vs. time for traces #1 and #2

In order to see how delay threshold affects MOSc, we also set *delay_threshold* to 170, 190, 210 and 250ms and calculate the MOSc score for trace #3 (its average network delay is 186ms as in Table 1). The "adaptive" algorithm swaps between the "fast-exp" and the "min-delay" algorithms according to the change of end-to-end delay. The results for the "adaptive", the "min-delay" and the "fast-exp" algorithms are shown in Fig 6. When the threshold is 170ms, the result of "adaptive" algorithm is similar to that of the "min-delay". With the increase of threshold, the results move towards the direction of the "fast-exp" algorithm and the maximum MOSc score becomes lower.

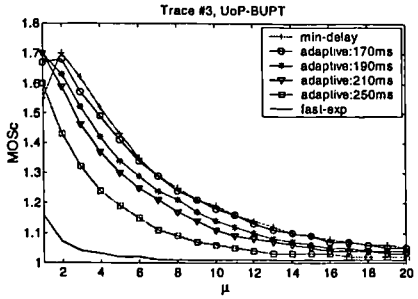


Figure 6. Performance comparison for Trace #3

The results suggest that the "adaptive" algorithm can adapt to the best algorithm for four traces to achieve the best perceived speech quality under the selected delay threshold.

We further investigated how to choose and adapt the parameters (e.g. μ value) of the buffer algorithm to achieve the "optimum" perceived quality all the time. The stages in the adaptation strategy are as follows:

(1). The best μ value (corresponding to the maximum MOSc score) is searched for each test segment (e.g. 9 sec), and this best μ value is used in next segment for the calculation of playout time (p_i) in (5).

(2). For each segment, also calculate MOS_{max}, the maximum PESQ MOS score with only network packet loss.

(3). Search μ ($\mu_{i+1} = \mu_i + 1$, $\mu_{i-1} = \mu_i - 1$), until

$$(((MOSc_{u_i} \geq MOSc_{u_{i-1}}) \wedge (MOSc_{u_i} \geq MOSc_{u_{i+1}})) \vee (MOSc_{u_i} = MOS_{\max}))$$

, then, μ_i is the best one for the segment. For the first segment, the search starts from μ =1, for other segment, the search starts from the best μ of the previous segment. If $(MOSc_{u_i} = MOS_{\max})$, the lowest μ that meets these criteria is selected, as this suggests an "optimum" MOSc score with the lowest end-to-end delay.

We implemented this parameter adjustment scheme on the four traces. The preliminary results show that the MOSc score increased obviously for traces #1 and #3. For traces #2 and #4, the MOSc scores always remained at MOS_{max}.

V. CONCLUSIONS

In this paper, we have proposed a new methodology for predicting conversation speech quality (MOSc). We investigated the performance of different buffer algorithms and parameters using the new MOSc metric based on newly collected Internet trace data. Results show that end-to-end delay, in general, has a major effect on the selection of buffer algorithms/parameters. For large to medium end-to-end delay, a buffer algorithm that aims for a minimum delay is preferred, whereas, for small end-to-end delay, an algorithm that targets minimum loss is best. Based on this, we proposed a modified buffer algorithm and an adaptive parameter adjustment scheme. Results show that it can achieve an "optimum" perceived quality for all the traces.

Future work will focus on the analysis of the impact of other parameters (e.g. buffer size) and the impact of parameter adjustment rate on perceived speech quality.

ACKNOWLEDGMENT

We would like to thank Mr. Wenyu Jiang from Columbia University, Mr. Michael Zink from Darmstadt University of Technology, Prof Wendong Wang and Mr. Lunyong Zhang from Beijing University of Posts & Telecommunications for their cooperation in trace data collection.

REFERENCES

[1] R. Ramachandran, J. Kurose, D. Towsley and H. Schulzrinne, "Adaptive playout mechanisms for packetized audio applications in wide-area networks, Proc. of IEEE Infocom, 1994, vol.2, pp.680 - 688.

[2] S. B. Moon, J. Kurose, D. Towsley, Packet audio playout delay adjustment: performance bounds and algorithms, Multimedia Systems, 1998, vol.6, pp.17 – 28.

[3] J. Rosenberg, L.Qiu and H. Schulzrinne, Integrating Packet FEC into Adaptive Voice Playout Buffer Algorithms on the Internet, Proc. of IEEE Infocom 2000, vol.3, pp.1705 – 1714.

[4] ITU-T P.800, Methods for subjective determination of transmission quality.

[5] ITU-T Recommendation G.107 (05/2000), The E-model, a computational model for use in transmission planning.

[6] TIA/EIA Telecommunications Systems Bulletin, Voice Quality *Recommendations for IP Telephony, TSB116, March 2001.*

[7] A. Clark, Modeling the Effects of Burst Packet Loss and Recency on Subjective Voice Quality, 2nd IPTel Workshop, 2001, pp.123 – 127.

[8] A. P. Markopoulou, F. A. Tobagi, M.J. Karam, Assessment of VoIP Quality over Internet Backbones, Proc. of IEEE Infocom, 2002.

[9] ITU-T Recommendation P.862, Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs.

[10] http://www.cs.columbia.edu/~wenyu/

[11] W. Jiang and H. Schulzrinne, QoS Measurement of Internet Real-Time Multimedia Services, Technical Report, CUCS-015-99, Columbia University, Dec. 1999.

[12] D.Sanghi, A.K. Agrawala, O. Gudmundsson, Experimental Assessment of End-to-end Behavior on Internet, Proc. of IEEE Infocom, 1993.

[13] J. –C. Bolot, Characterizing end-to-end packet delay and loss in the Internet, Jour. of High Speech Networks, vol.2, pp. 305-323, 1993.

[14] M. S. Borella, Measurement and Interpretation of Internet Packet Loss, Journal of Communications & Networking, 2(2), June 2000.

[15] P. Brady, A Technique for Investigating on/off Patterns of Speech, Bell Labs Tech. Journal, 44(1):1-22, January 1965.

[16] W. Jiang, H.Schulzrinne, Analysis of on-off patterns in VoIP and their effect on voice traffic aggregation, Proc. of ICCCN 2000.

[17] A. E. Conway, A Passive Method for Monitoring Voice-over-IP Call Quality with ITU-T Objective Speech Quality Measurement Methods, Proc. of IEEE ICC, 2002.

[18] ITU-T Recommendation P.50, Appendix 1, Test signals, 1999.

Perceived Speech Quality Prediction for Voice over IP-based Networks

Lingfen Sun and Emmanuel C. Ifeachor
Department of Communication and Electronic Engineering,
University of Plymouth, Plymouth PL4 8AA, U.K.

Abstract – **Perceived speech quality is the key metric for QoS for VoIP applications. The primary aims of the study reported in the paper are to carry out a fundamental investigation of the impact of packet loss and talkers on perceived speech quality using an objective method to provide the basis for developing an artificial neural network (ANN) model to predict speech quality for VoIP. The impact of packet loss (e.g. loss burstiness, loss patterns and packet size) and different talkers on speech quality was investigated for three modern codecs (G.729, G.723.1 and AMR) using the new ITU PESQ algorithm. Results show that packet loss burstiness, loss locations/patterns and the gender of talkers have an impact on perceived speech quality. Packet size has, in general, no obvious influence on perceived speech quality for the same network conditions, but the deviation in speech quality depends on packet size and codec. Based on the investigation, we used talkspurt-based conditional and unconditional packet loss rates (instead of network packet loss rates because they are perceptually more relevant), codec type and the gender of the talker (extracted from decoder) as inputs to an ANN model to predict speech quality directly from network parameters. Results show that high prediction accuracy was obtained from the ANN model (correlation coefficients for the test and validation datasets were 0.952 and 0.946 respectively). This work should help to develop efficient, non-intrusive QoS monitoring and control strategies for VoIP applications.**

Keywords – **Voice over IP, Speech Quality, Artificial Neural Network, Packet Loss, Codecs, Talker Dependency**

I. INTRODUCTION

In real-time voice communication, perceived speech quality, expressed as a Mean Opinion Score (MOS), is the key metric for Quality of Service (QoS) as it provides a direct link to quality as perceived by the end user. MOS values may be obtained by subjective listening tests [1] or by objective perceptual measurement methods, such as the new ITU algorithm, the Perceptual Evaluation of Speech Quality (PESQ) [2].

In voice over IP applications, statistical and artificial intelligence methods are being developed to predict speech quality directly from IP network parameters for QoS monitoring and control purposes [3][4][5][6]. The E-model as well as artificial neural networks (ANN) models have recently been used to predict speech quality from network parameters [4][5][6][7]. Unlike the E-model which is static, artificial neural networks models can adapt to the dynamic environment of IP networks, such as the Internet, because of its ability to learn. However, the success of ANN approach in voice over IP depends on the ability of the models to fully learn the non-linear relationships between IP networks impairments (e.g. packet loss and jitter) and the perceived speech quality.

At present, both the E-model and ANN based methods rely on databases obtained by subjective tests. Unfortunately, subjective listening tests are costly and time-consuming and as a result the databases are limited and do not cover all the possible scenarios and network conditions. The impact of a variety of network parameters (e.g. loss rate, burstiness, loss pattern and packet size) on perceived speech quality remains unclear. Further, little attention has been paid to talker dependency and the development of current ANN models are based on a limited number of codecs. The assumptions about the behaviour of network losses do not reflect reality. For example, only the numbers of consecutively lost packets (e.g. 1 to 5) were used to represent different bursty losses.

The aims of the study reported in this paper are three fold: (1). to undertake a fundamental investigation of the impact of packet loss (e.g. loss rate and loss pattern) on perceived speech quality using an *objective* measurement algorithm (the new ITU PESQ algorithm), (2) to investigate the impact of different talkers on perceived speech quality, and (3) to develop a robust ANN model that exploits perceptually relevant information for speech quality prediction.

The remainder of the paper is organised as follows. In Section II, the experimental system used in the study is introduced. In Section III, a fundamental study of the impact of packet loss and different talkers on speech quality is presented. The study provides a basis for the development of the ANN model for speech quality prediction which is presented in Section IV. Section V concludes the paper.

II. SIMULATION SYSTEM

A block diagram of the system that was used in the study is depicted in Figure 1. It is a PC-based software system that allows the simulation of key processes in voice over IP. It enables the simulation of a variety of network conditions and objective measurement of the effects on perceived speech quality. The system includes a speech database, an encoder/decoder, a packet loss simulator, a speech quality measurement module, a parameter extraction and an ANN model. The speech database is taken from the TIMIT data set [15] and ITU dataset [2]. Speech files from different male and female talkers are chosen for talker dependency analysis and to generate a data base for ANN model development

Three modern codecs were chosen for the study. These are G.729 CS-ACELP (8 Kbps) [9], G.723.1 MP-MLQ/ACELP (5.3/6.3 Kbps) [10] and Adaptive Multi-Rate (AMR) codecs with eight modes (4.75 to 12.2 Kbps) [11].

0-7803-7400-2/02/$17.00 (C) 2002 IEEE

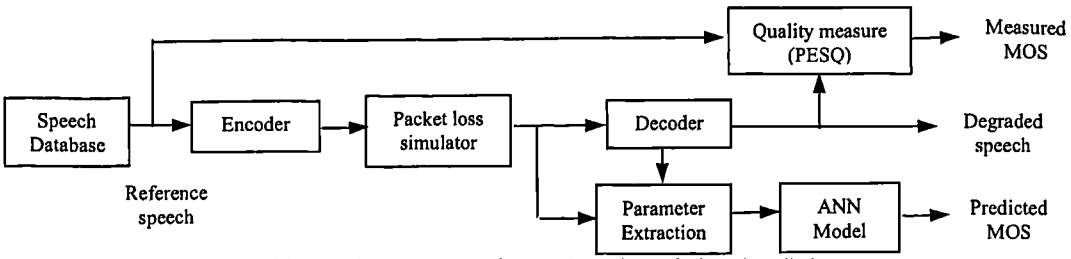


Figure 1. System structure for speech quality analysis and prediction

A 2-state Gilbert model was used to simulate packet loss (see Figure 2). The Gilbert model is well known to represent the packet loss behaviour of a real network, even after the late arrival loss due to jitter is taken into account (if a packet arrives too late, it will be discarded by jitter buffer) [8]. In the figure, State 0 is for a packet received (no loss) and State 1 is for a packet dropped (loss). p is the probability that a packet will be dropped given that the previous packet was received. q is the probability that a packet will be dropped given that the previous packet was dropped. q is also referred to as the conditional loss probability (*clp*). The probability of being in State 1 is referred to as unconditional loss probability (*ulp*). The *ulp* provides a measure of the average packet loss rate. It is given by:

$$ulp = p/(p+1-q)$$

The conditional loss probability (*clp*) and unconditional loss probability (*ulp*) are used in the paper to characterise the loss behaviour of the network.

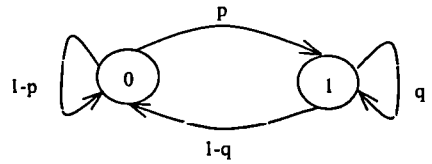


Figure 2. Gilbert model

In our system, the new ITU perceptual measurement algorithm, the Perceptual Evaluation of Speech Quality (PESQ), is used to measure the perceived speech quality under different network conditions and for different talkers. The PESQ compares the degraded speech with the reference speech and computes an objective MOS value in a 5-point scale. In the study, the MOS score obtained from the PESQ is referred to as the 'measured MOS' to differentiate it from the 'predicted MOS' obtained from the ANN model. The Parameter Extraction module is used to extract salient information from the IP network and the decoder (including the codec type and network packet loss). In real VoIP applications, codec type and packet loss would be parsed from the RTP header. After processing, the information is fed to the ANN model to predict speech quality.

To provide a basis for the development of a robust ANN model, a fundamental study of the impact of packet loss and gender on perceived speech quality was undertaken. This enabled us to determine the relevant parameters to be used as input to the neural networks model to predict speech quality. The study is based on three modern codecs described above.

III. Perceived Speech Quality Analysis

A. The impact of packet loss on perceived speech quality

We first investigated how packet loss burstiness affects perceived speech quality. A fixed packet size was set for different codec. Different network *ulp* and *clp* were chosen and the corresponding MOS score was calculated. To account for a wide range of possible type of packet loss patterns and locations, 300 different initial seeds for random number generation were chosen for each pair of *ulp* and *clp*. The average MOS score and 90% Confidence Internal (CI) were calculated. The results for G.729 and G.723.1 (6.3 Kb/s mode) are shown in Figures 3 and 4. The length of the test speech sentence was about 12 seconds. The packet size for G.729 and G.723.1 was 2 and 1 frames/packet, respectively. No VAD was activated.

From Figures 3 and 4, it can be seen that the *clp* has an obvious impact on the perceived speech quality even for the same average loss rate (*ulp*). When burst loss increases (*clp* increasing), the MOS score decreases and the variation of the MOS score (shown in CI) also increases. This is because losses may occur more concentrated with high burst losses and this results in large variation in the MOS scores due to the locations of the losses, whereas it may occur evenly in low burst loss cases which results in small deviations. There is only a small difference between the results for G.729 and G.723.1, when *ulp* is 10%, and *clp* is from 40% to 70%.

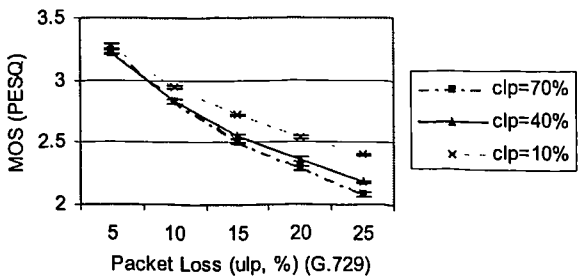


Figure 3. MOS vs packet loss for G.729

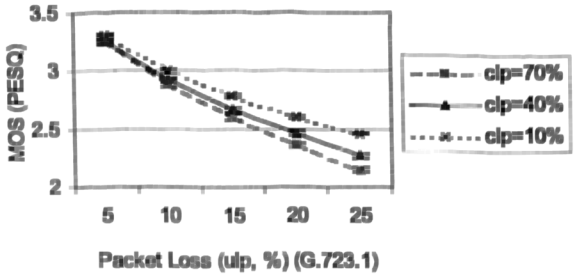


Figure 4. MOS vs packet loss for G.723.1

We then investigated how packet size affects perceived speech quality. A fixed *clp* (40%) was set and *ulp* was changed from 0% to 40% in 5% increment. The packet size was changed from 1 to 6 frames/packet. As before, 300 different initial seeds were generated. The average and the standard deviation of MOS scores for G.729 are shown in Figure 5 (a) and (b). The standard deviations for the MOS scores for AMR (12.2 Kb/s mode) are shown in Figure 6.

From Figure 5 (a), it can be seen that the packet size has in general no obvious influence on perceived speech quality for a given packet loss rate. Similar results were obtained for G.723.1 and AMR. However, the variation in speech quality for the same network loss rate depends on packet size and codec, as shown in Figures 5 (b) and 6.

When packet loss rate is lower and packet size is larger, the higher values of the standard deviation of MOS score means larger variation in speech quality for the same network conditions. The variation in quality is the main obstacle in the prediction of speech quality directly from network parameters. When packet loss (e.g. *ulp* and *clp*) was calculated from the Gilbert model, the loss is perceptual irrelevant as some losses may occur during a silent period which is imperceptible [13]. As a solution, we proposed to calculate losses only during talkspurts.

A network packet may include a speech talkspurt frame or a silence frame. The number of silence frames depends on whether VAD (Voice Activity Detection) is activated at the encoder side. If VAD is activated, silence frame only represents SID (Silence Insertion Description) frame. Here we combined the information from decoder's VAD indicator and network packet loss, and calculated the *ulp* and *clp* according to Gilbert model only during the speech talkspurt. In this case, State 1 in Figure 2 represents loss during a talkspurt, and State 0 represents no loss or loss during a silence period. We use *ulp(VAD)/clp(VAD)* to differentiate them from network *ulp/clp*. As the calculation of *ulp(VAD)/clp(VAD)* was based on speech frame, the loss pattern and the factor of packet size have both been taken into account. The codec type, *ulp(VAD) and clp(VAD)* were identified as inputs for neural network analysis.

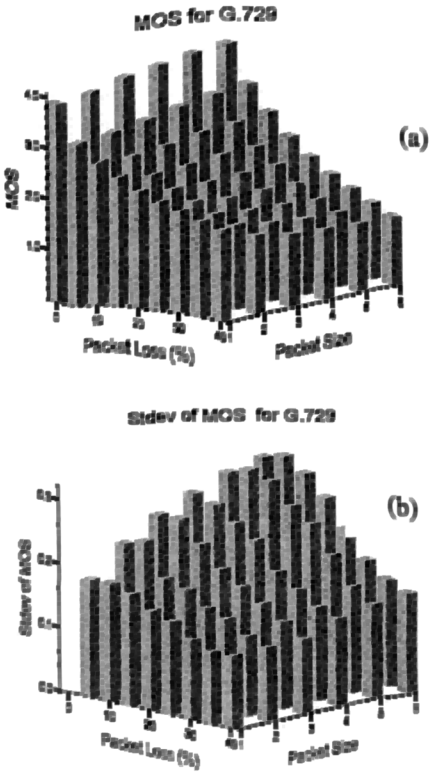


Figure 5. (a) Average MOS and (b) Standard Deviation of MOS for G.729

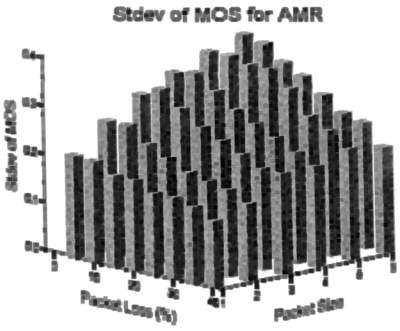


Figure 6. Standard Deviation of MOS for AMR

B. The impact of talker on perceived speech quality

This experiment was to investigate whether difference in talker (male or female) has an effect on perceived speech quality for the same network conditions. We first chose 6 English speakers (3 male and 3 female) from the TIMIT [15] Data Set (dialects 1 and 2). Speech files from the same talker were grouped to form a longer file (about 10s). The activity factor [16] was about 0.82 for all files.

We altered *ulp* from 0 to 30 % in 5% increment, set *clp* to 10% and packet size to 2 for G.729 (no VAD). As before, 300 different initial seeds were chosen. The average MOS

scores for the six talkers are shown in Figure 7. The speech file name starts with letter "f" for female and "m" for male.

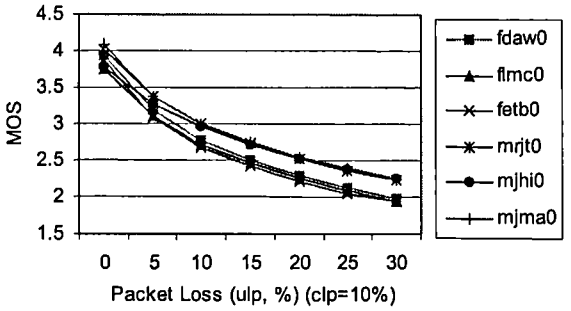


Figure 7. MOS vs Loss Rate for English speakers

We further tested another four speech files (2 male and 2 female of Dutch speakers) from an ITU data base [2]. Each speech file was about 8s, with about 45% to 49% activity. The results for the four speech files are illustrated in Figure 8.

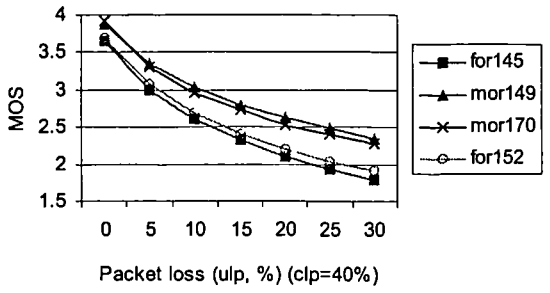


Figure 8. MOS vs Loss Rate for Dutch speakers

From inspection Figures 7 and 8, it can be seen that the impact of different talkers on perceived speech quality appears to depend mainly on the gender of the talker (male or female), irrespective of language and accent. The quality for the female talker tends to be worse than that of the male talker for the same network impairments. This effect is more obvious when loss increases.

The reason for talker dependency is due to the codec algorithm. As the G.729, G.723.1 and AMR are all CELP-based codecs, the use of linear predictive model of speech production can lead to variations in codec performance with different talkers and languages [14]. In this paper, we focused on gender issue, and identified gender as one of the input parameters for neural network analysis. The gender can be decided according to pitch delay derived from the decoder.

IV. Prediction of Perceived Speech Quality using Artificial Neural Network (ANN)

In order to model the relationships between network impairments and perceived speech quality, a neural network model was developed to learn the non-linear mapping from network parameters to MOS score.

Four variables were identified as inputs to the neural network model, namely: codec type, gender, *ulp(VAD)* and *clp(VAD)*. The predicted MOS score was the only output (see Figure 1). Stuttgart Neural Network Simulator (SNNS) package [12] was used for neural network training and testing. A three-layer feed-forward neural net architecture and the Standard Backpropagation learning algorithm were selected for simplicity.

In order to train and test the neural network, a database was generated from 2 talkers (1 male and 1 female) and three codecs, G.729, G.723.1 (6.3Kb/s) and AMR (12.2 Kb/s). For dual–mode G.723.1 and eight-mode AMR, only one mode was chosen for simplicity. The network loss *ulp* was set to 0, 10, 20, 30 and 40% and *clp* was set to 10, 50 and 90%. The packet size was set to 1, 2, 3, 4 and 5 frames/packet. For each case, an initial seed was chosen randomly to cater for a range of possible loss patterns. The state transitions were counted according to the Gilbert model (see Figure 2). In order to compare the results to real network loss and talkspurt-based network loss, the real loss rate at the end of the test sentence, *ulp(Real)/clp(Real)* and loss rate during talkspurt, *ulp(VAD)/clp(VAD)*, were calculated at the same time. The difference between *ulp(Real)/clp(Real)* and *ulp/clp* is due to pseudo-random number generation and initial seeds selection. For each loss condition, the perceived speech quality between the reference and degraded speech files was calculated using PESQ. A total of 362 samples (patterns) were generated. 70% of the samples were chosen randomly as the training set and the remaining 30% as the testing set.

Different network structures (e.g. the number of neurons in the hidden layer and the parameters of learning algorithm) were investigated to determine a suitable architecture for ANN model. Comparing the predicted MOS score from the ANN model and the measured MOS, we obtained a maximum Correlation Coefficient (ρ) of 0.967 and an average error of 0.12 for the training set. For the testing set, ρ was 0.952 and the average error was 0.15. The learning rate (η) was 0.4 and the maximum difference (d_{max}) was 0.01 for a 4-5-1 net. The scatter diagrams of the predicted versus the measured MOS scores for the training and test data sets are illustrated in Figure 9 (a) and (b). Increasing the number of neurons in the hidden layer did not improve the prediction accuracy. However, when *ulp(Real)/clp(Real)* was used instead of *ulp(VAD)/clp(VAD)*, the Correlation Coefficients for the training and testing datasets both dropped by 2-3 percent. This suggested that *ulp(VAD)/clp(VAD)* are better for speech quality prediction than *ulp(Real)/clp(Real)*. We also investigated the effect of including packet size as an input to the neural net (i.e. 5 inputs) and obtained similar results. This suggested that packet size may not be necessary as an input to the neural network.

As the training and test data sets were from the same talkers, we further generated a validation data set from another male and female talkers and set the different network loss conditions (*ulp:* 5, 15, 25, 35%, and *clp:* 30, 70%). A total of 210 new patterns were generated and used to validate the trained ANN model. We obtained ρ of 0.946 and an average error of 0.19. This suggested that the neural network model works well for speech quality prediction in general.

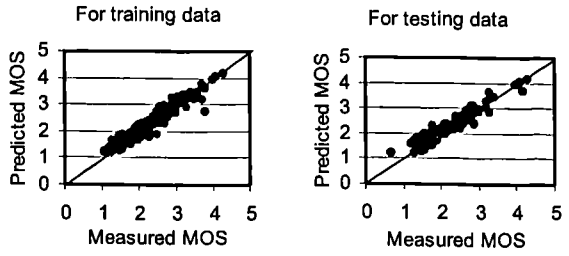


Figure 9. Predicted MOS vs. Measured MOS for (a) training data and (b) test data

The correlation coefficients obtained from training, test and validation datasets are between 0.946 to 0.967. It seems difficult to improve the performance further from neural network side. We think this is mainly due to the following two reasons. (1). *ulp(VAD)/clp(VAD)* is still not accurate enough to express perceptually relevant loss information for some loss patterns/locations; (2). Objective MOS scores from PESQ may not be as accurate as subjective MOS scores for some loss conditions. Our subjective test results have also confirmed that PESQ shows higher sensitivity than human subjects in high bursty conditions, especially in the case of missing words, whereas, it shows lower sensitivity than human subjects in lower bursty cases for G.729.

V. CONCLUSIONS

We have investigated the impact of packet loss, codec and talker on perceived speech quality based on the new ITU PESQ measurement algorithm and developed an ANN model for speech quality prediction. Results show that the loss pattern, loss burstiness and the gender of the talker have an impact on perceived speech quality. Packet size has in general no obvious influence on perceived speech quality for a given packet loss rate, but the deviation in speech quality depends on packet size and codec. The quality for the female talker tends to be worse than that of the male talker for the same network impairments. Based on the investigation, we used talkspurt-based conditional and unconditional packet loss rates (instead of the network packet loss rates because they are perceptually more relevant), codec type and the gender of the talker (extracted from decoder) as inputs to an ANN model to predict speech quality directly from the network parameters. Results show that high prediction accuracy was obtained from the ANN model (correlation coefficients of the test and validation datasets are 0.952 and 0.946 respectively). This work should help to develop efficient, non-intrusive QoS monitoring and control strategies for VoIP applications.

Future work will focus on further analysis of the loss pattern in order to incorporate more information from speech content (e.g. signal energy, voiced/unvoiced information) and to obtain more accurate perceptually relevant loss information. The neural networks based model will be optimised using real Internet VoIP trace data. More speech data will be investigated for the analysis of talker dependency.

ACKNOWLEDGEMENT

We are grateful to Acterna for sponsorship.

REFERENCES

[1] ITU-T Recommendatiaon P.800, Methods for subjective determinatio of transmission quality

[2] ITU-T Recommendation P.862, Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs.

[3] L.A.R. Yamamoto, J.G.Beerends, Impact of network performance parameters on the end-to-end perceived speech quality, Expert ATM Traffic Symposium, Mykonos, Greece, Sep. 1997

[4] S. Mohamed, F. Cervantes-Perez and H. Afifi, Integrating Networks Measurements and Speech Quality Subjective Scores for Control Purposes, IEEE Infocom 2001

[5] S. Mohamed, F. Cervantes-Perez and H. Afifi, Audio Quality Assessment in Packet Switched Networks: an "Inter-Subjective" Neural Network Model, Proc. International Conference on Information Networks, Japan, Jan. 2001.

[6] A. D. Clark, Modeling the Effects of Burst Packet Loss and Recency on Subjective Voice Quality, IPTEL'2001, pp. 123-127, April, New York, 2001

[7] ITU-T Recommendation G.107, The E-model, a computational model for use in transmission planning

[8] W. Jiang and H. Schulzrinne, QoS Measurement of Internet Real-Time Multimedia Services, Technical Report, CUCS-015-99, Columbia University, Dec. 1999, http://www.cs.columbia.edu/~wenyu

[9] ITU-T Recommendation G.729, Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP), March 1996

[10] ITU-T Recommendation G.723.1, Dual Rate Speech Coder for Multimedia Communication Transmitting at 5.3 and 6.3 kbit/s, March 1996

[11] ETSI EN 301 704 V7.2.1, Digital cellular telecommunications system; Adaptive Multi-Rate (AMR) speech transcoding

[12] http://www-ra.informatik.uni-tuebingen.de/SNNS/

[13] L. F. Sun, G. Wade, B. Lines and E. Ifeachor. Impact of Packet Loss Location on Perceived Speech Quality, IPTEL'2001, April, New York, 2001

[14] P. A. Barrent, R. M. Voelcker and A. V. Lewis, Speech transmission over digital mobile radio channels, BT Technol J Vol 14 No. 1 January 1996, pp.45-56

[15] J. S. Garofolo, L.F. Lamel, W. M. Fisher, et al. TIMIT Acoustic-Phonetic Continuous Speech Corpus, 1993

[16] ITU-T Recommendation P.56 - Objective measurement of active speech level

Impact of Packet Loss Location on Perceived Speech Quality

L. F. Sun, G. Wade, B. M. Lines, E. C. Ifeachor
Department of Communication and Electronic Engineering,
University of Plymouth,
Drake Circus, Plymouth PL4 8AA, United Kingdom,
{ L.F.Sun@jack.see.plym.ac.uk, j.wade@plymouth.ac.uk, B.Lines@plymouth.ac.uk, E.Ifeachor@plymouth.ac.uk}

Abstract – In VoIP applications, packet loss can have a major impact on perceived speech quality. The impact is affected by factors such as packet loss size, loss pattern and loss locations. In this paper, we report an investigation into the impact of loss location on perceived speech quality and the relationships between convergence time and loss location for three different codecs (G.729, G.723.1 and AMR) using perceptual-based objective measurement methods (PSQM+, MNB and EMBSD). Our results show that loss location has a severe effect on perceived speech quality. The loss at unvoiced speech segments has little impact on perceived speech quality for all codecs. However, the loss at the beginning of voiced segments has the most severe impact on perceived speech quality. The convergence time depends on the speech content (voiced/unvoiced). For unvoiced segments, the convergence time is stable whereas for voiced segments it varies but has an upper bound at the end of the segment. Our method allows a more accurate measurement of the exact effect of packet loss on perceived speech quality. This could help in the development of a perceptually relevant packet loss metric, which could be valuable in non-intrusive VoIP measurements.

Keywords -- Voice over IP, Packet loss, Speech quality, Objective perceptual measurement, Codecs, Concealment performance

I. INTRODUCTION

Packet loss is a major source of speech impairment in voice over IP (VoIP) applications. Such a loss could be caused by discarding packets in the IP networks due to congestion or by dropping packets at the gateway/terminal due to late arrival. The impact of packet loss on perceived speech quality depends on several factors, including loss pattern, codec type, and packet loss size [1][2]. It may also depend on the location of loss within the speech.

In modern codecs (e.g. G.729, G.723.1 and Adaptive Multi-Rate, AMR codec), internal concealment algorithms are used to alleviate the effects of packet loss on perceived speech quality [3][4][5]. When a loss occurs the decoder derives the parameters for the lost frame from the parameters of previous frames to conceal the loss. The loss also affects subsequent frames because the decoder takes a finite time (the convergence time) to resynchronise its state to that of the encoder. Recent research has shown that for some codecs (e.g. G.729) concealment works well for a single frame loss, but not for consecutive or burst losses [1], and that the convergence times are dependent on speech content. Further, the effectiveness of a concealment algorithm is affected by which part of speech is lost (e.g. voiced or unvoiced). For example, it has been shown that concealment for G.729 works well for unvoiced frames, but for voiced frames it only works well after the decoder has obtained sufficient information [6]. Further, the decoder fails to conceal the loss of voiced frames at an unvoiced/voiced transition. Thus, the location of packet loss in relation to different parts of speech is important.

In most studies [1][6], the analysis of concealment performance and convergence times is based on the mean square error (MSE) and signal-to-noise ratio (SNR) criteria (with subjective or perceptual-based objective methods only used to assess overall quality under stochastic loss simulations). The perceptual impact of concealment algorithms or convergence times for different loss locations is still unknown. It is important to understand the effects of loss location and loss pattern on perceived speech quality, for different types of codec, to allow a more accurate measurement of voice quality. This requires the use of perceptual-based objective methods in the analysis. This could be helpful in setting up more efficient speech recovery system and for the development of perceptually relevant packet loss metrics which could be valuable in non-intrusive VoIP measurement.

The IETF has recently proposed a set of new metrics for packet loss [2]. This includes loss constraint distance (i.e. distance threshold between two losses) and "noticeable" loss rate (i.e. percentage of lost packets with loss distances smaller than loss constraint distance). For the same loss rate, different loss patterns may have different effects on perceived speech. In VoIP applications, the loss constraint is related to the convergence times of the decoder. However, it is still unclear how to determine the loss constraint threshold and whether (or how) the threshold is related to codec type, burst size or speech.

The aims of the study reported in this paper are two fold: (1) to investigate the impact of loss location on perceived speech quality and hence the concealment performance of codecs, and (2) to investigate the relationships between convergence times and loss locations/speech content, codec type or loss size.

The work reported here is based on three codecs – two existing codecs (G.729B [13] and G.723.1) and a new codec (AMR [7][14]) for VoIP. Three major perceptual distance measurement algorithms (PSQM/PSQM+ [8][9], MNB [10][11] and EMBSD [12]) are used for perceptual performance analysis for different loss location. Each

algorithm quantifies perceptual quality, but has a different range of perceptual distance.

The results show that the loss location has a severe effect on perceived speech quality. The loss at unvoiced speech segments has little impact on perceived speech quality for all three codecs. However, the loss at the beginning of voiced segments has the most severe impact on perceived speech quality. The extent of the impact depends on the size of the burst loss and codec type. The convergence time depends on the speech content. For unvoiced segments the convergence time is stable whereas for voiced segments it varies but constrained by the duration of the segment.

The remaining sections of the paper are structured as follows: Section II presents a brief overview of the codecs used and their concealment algorithms. The perceptual distance measurement algorithms (PSQM/PSQM+, EMBSD and MNB) are summarised briefly in Section III. The simulation system is described in Section IV, the experiments, results and their analysis are given in Section V. Section VI concludes the paper.

II. CODECS AND THEIR INTERNAL CONCEALMENT

A. *Codec types - G.729, G.723.1 and AMR*

The G.729 CS-ACELP (Conjugate Structure Algebraic Codebook Excited Linear Prediction, 8 Kbps) and G.723.1 (MP-MLQ/ACELP: Multipulse excitation with a maximum-likelihood-quantizer/Algebraic Codebook Excited Linear Prediction, Dual rate: 5.3/6.3 Kbps) are both standardized by the ITU and have been used in VoIP applications. The AMR (Adaptive Multi-Rate, ACELP) speech codec was developed by ETSI and has been standardized for GSM. It has been chosen by 3GPP as the mandatory codec. The AMR is a multi-mode codec with 8 narrow band modes with bit rates between 4.75 to 12.2 Kb/s. Mode switching can occur at any time (frame-based). AMR speech codec represents a new generation of coding algorithms which are developed to work with inaccurate transport channels. The flexibility on bandwidth requirements and the tolerance in bit errors of AMR codecs are not only beneficial for wireless links , but are also desirable for VoIP applications.

The three codec types belong to CELP analysis-by-synthesis hybrid codec. At each speech analysis frame, the speech signal is analysed to extract the parameters of the CELP model (Linear Prediction, or LP filter coefficients, adaptive and fixed codebooks' indices and gains). For stability and efficiency, LP filter coefficients are transformed into Line Spectral Frequencies, or LSF's for transmission. These parameters are then encoded and transmitted. At the decoder, the parameters are decoded and speech is synthesized by filtering the reconstructed excitation signal throuth the LP synthesis filter.

The major differences between the three codecs lie in the excitation signals, the partitioning of the excitation space (the algebraic codebook), delay and the way in which the coefficients of the filter are represented. For example, the G.729 uses two stage codebook structures for LSP parameters and gets the name "conjugate structure".

The frame sizes for the three codecs are 10 ms (80 samples at 8 kHz sampling) for G.729, 20 ms (160 samples) for AMR and 30 ms (240 samples) for G.723.1. They all have voice activity detection and silence suppression processing. The frames are classified as normal speech frame, SID (Silence Insertion Description) frame and null frame (non-transmitted frame).

B. *Codec Internal Concealment*

All three codecs have built-in concealment algorithms, which can interpolate the parameters for the loss frames from the parameters of the previous frames. For example, for the G.729 the concealment algorithm works in accordance to the following steps:

- The line spectral pair coefficients of the last good frame are repeated
- The adaptive and fixed codebook gain are taken from the previous frame but are damped to gradually reduce their impact.
- If the last reconstructed frame was classified as voiced, the fixed codebook contribution is set to zero. The pitch delay is taken from the previous frame and is repeated for each following frame. If the last reconstructed frame was classified as unvoiced, the adaptive codebook contribution is set to zero and the fixed codebook vector is randomly chosen.

III. PERCEPTUAL SPEECH QUALITY MEASURE – PERCEPTUAL DISTANCE

Perceptual distance is used to measure the perceptual difference between a reference speech signal and a degraded speech signal. It normally includes a perceptual model and a cognition model to mimic the process in the human's hearing perceptual process. Various perceptual speech quality measurement algorithms exist with different perceptual or cognition models.

PSQM (Perceptual Speech Quality Measurement) developed by KPN has been adopted as ITU-T Recommendation P.861 for assessing the speech quality for codecs [8]. PSQM+ was proposed by KPN to improve the performance of PSQM for loud distortions and temporal clipping [9]. PSQM/PSQM+ can generate a perceptual distortion value for each frame (32 ms for 8 kHz sampling, with 50% overlapping) and the overall PSQM/PSQM+ value is calculated for the whole test sentence via different weighting factors for silence or non-silence frames. As PSQM+ provides a more accurate measure of perceived speech quality under frame loss situations, we have chosen it

for overall perceived speech quality and perceptual distance calculation for each frame.

The MNB (Measuring Normalizing Blocks) developed by the US department of Commerce [10][11], is included as an Appendix in ITU-T P.861 Recommendation. The MNB does not generate a distortion value for each frame since each MNB is integrated over frequency or time internals.

EMBSD (Enhanced Modified Bark Spectral Distortion) was developed by Temple University in USA [12]. It estimates speech distortion in the loudness domain taking into account the noise masking threshold in order to include only audible distortions in the calculation of the distortion measure. As EMBSD only takes into account the non-silence frame for the final perceptual distortion calculation, the setting of the threshold of silence or non-silence will affect the final result.

In the paper, MNB and EMBSD are used for the overall quality measurement.

IV. SIMULATION SYSTEM

In order to investigate the impact of packet loss location on perceived speech quality, and the relationships between convergence time and loss location, we set up a simulation system. This includes speech encoder/decoder, loss simulation, perceptual quality measure and convergence time analysis, as shown in Figure 1. For codecs, we have a choice of G.729, G.723.1 and AMR. The standard 16 bit, 8 kHz sampled speech signal is processed by the encoder first. Then the parameter-based bit stream is sent to the decoder without frame losses (speech quality degradation in this case is only due to codec). The bitstream is also sent to the loss simulation module where the loss position and frame loss size can be selected. After loss simulation the bit stream is processed by the decoder to obtain the degraded speech signal with loss. The overall perceptual speech quality is measured between the reference speech signal and the degraded speech signal with loss by calculating the perceptual distance values using the PSQM+, MNB and EMBSD algorithms. The perceptual distance for each frame is also measured between the degraded speech without loss and the degraded speech with loss using PSQM+ for the analysis of convergence time. This eliminates coding impairment from the computation. The convergence time is also calculated using the normal Mean Square Error (MSE) method [1].

Loss simulation for each codec differs from the loss specification in the codecs. For G.729, if a parameter byte in the bit stream is set to zero, the frame is treated as a loss by the decoder and concealment is initiated automatically. For AMR, there is an extra byte for the transmit/receive frame type. For a lost frame, there is only a need to set the type as a BAD/ERASED frame. For G.723.1, a loss location mark file is created and serves as the input to the decoder.

V. EXPERIMENTS AND ANALYSIS OF RESULTS

A. Loss location and perceived speech quality

In the first experiment, the impact of loss position on the overall perceptual speech quality or the performance of concealment under different loss locations is investigated. The PSQM+, MNB and EMBSD perceptual distance values are calculated for the whole test speech sentence (about 6 seconds), while only one loss is produced each time and the loss position moves smoothly from left to right. The move is one frame each time and the frame size is decided by the codec chosen. At each loss location , the frame loss size can change by one, two, three or four frames to simulate different packet size or burst loss size.

The waveform for the first talkspurt for the test sentence "Each decision show (s)" is shown in Figure 2. It consists of four voiced segments - V (1) to V (4) corresponding to the vowels 'i', 'i', 'ə' and 'au'. The voiced segments are separated by unvoiced segments.

The overall perceptual distance values for PSQM+, MNB and EMBSD for G.729 are shown in Figures 3, 4 and 5, respectively. The values (using PSQM+) for G.723.1 (6.3 Kb/s) and AMR (12.2 Kb/s and 4.75 Kb/s mode) are shown in Figures 6, 7 and 8. In all the figures, the horizontal scales are in the unit of frames. As the frame sizes are 10, 20 and 30ms for G.729, AMR and G.723.1, respectively, the total number of frames for the test segments shown are 134, 67 and 45.

Examination of Figure 3 shows that the perceptual distance value varies between 1.4 and 2.4 as the loss location moves from left to right. In the PSQM+, a change in perceptual distance indicates a change in perceptual speech quality (the smaller the distance, the better the perceived quality). Similar changes in perceived speech quality can also be seen for the MNB (Figure 4) and EMBSD (Figure 5), as well as for the different codecs (Figure 6, 7 and 8). It is evident that the same loss condition (one packet loss for the whole test speech segment) causes an obvious variation in overall perceived speech quality, but the variation is dependent on speech content. A loss at unvoiced speech segments shows little impact on perceived speech quality (almost the same perceptual distance values as for no-loss cases). However, a loss at voiced segments has different effects on perceived speech quality depending on its location within the voiced segment. At the beginning of a voiced segment, it has the most severe impact (the peaks in the figures). At the end of voiced segments, the impact is small. In the middle voiced segments, perceptual distances change depending on the codec and frame loss size. For example, for the G.729 one-frame loss (Figure 3), the perceptual distance value reaches its peak when the loss is at the beginning of voiced segments. Then, as the loss position moves to the right (for each voiced segment), the perceptual

distance rapidly returns to the minimum value, showing a good convergence performance for voiced segments 1, 2 and 3. For voiced segment 4, the value varies depending on the speech content. As the frame loss size increases, the perceptual distance increases.

We explain this phenomenon from two perspectives:
(i). From the perspective of the codec or concealment algorithms

In the case of a loss at the beginning of voiced segment, as the previous frame is clearly an unvoiced frame or an unvoiced/voiced transition frame. The concealment algorithm will conceal the loss using the filter coefficients and the excitation for an unvoiced sound. It causes the lost frame to be concealed using the unvoiced features. In other words, during the unvoiced to voiced transition period, the shape of the vocal tract is in transition (not stable), and the LP filter coefficients will change rapidly for each frame. The excitation signal is also changing from unvoiced to voiced. The concealment algorithm can not conceal properly for the loss at this transition stage.

For a loss during the stationary part of a voiced segment, the concealment algorithm will conceal the current frame with the gain further reduced from the previous frame (adaptive codebook gain). The line spectral pair coefficients (or LP filter coefficients) of the last good frame are repeated. In other words, the vocal tract is at a stable stage (after the transition) and keeps the same shape. The LP filter coefficients are very stable during this stage. If the pitch delay does not change much within a short time period, a small loss can be concealed perfectly using the parameters of the previous frames. However, when there is an increase in burst loss size or frame size, it is difficult to conceal the losses adequately. The concealment performance degrades depending on the features in the voiced segments.

(ii). From the perspective of the perceptual quality measurement algorithms

The signal energy is very important for the overall perceived speech quality for all the perceptual algorithms. If a reference signal frame has a large signal energy (e.g. the beginning of a voiced segment), and the degraded signal has a very small energy (due to improper concealment), this will cause a significant increase in the perceptual distance. For a loss during the voiced segment, the degraded signal will normally have a rather large energy. Perceptual distance will vary for different loss size and loss location.

For different codecs (G.729, G.723.1 and AMR), the perceived speech quality shows large variations due to differences in the frame sizes. The perceptual distances using PSQM+ for the three codecs for a loss at the beginning of voiced segment 4 is summarized in Table 1 (including perceptual distances for no-loss cases).

Table 1: Perceptual distance using PSQM+

Codec Type	No-loss	1-frame	2-frame	3-frame	4-frame
G.729 (8 Kb/s)	1.36	1.62	1.83	2.11	2.42
G.723.1 (6.3 Kb/s)	1.51	1.79	2.84	3.54	4.03
AMR (12.2Kb/s)	0.98	1.35	1.6	2.06	2.45
AMR (4.75Kb/s)	1.92	2.17	2.42	2.81	3.34

From Table 1, it can be seen that the AMR (12.2 Kb/s) has the best perceptual quality and the AMR (4.75 Kb/s) the worst for no-loss cases. For a one-frame loss, the quality sequences remain the same. For a two-frame loss, the G.723.1 has the worst quality while AMR (12.2 Kb/s) remains the best. For three-frame and four-frame loss, G.729 and AMR (12.2 Kb/s) have similar perceptual quality, while G.723.1 remains the worst.

Of the three perceptual measurement methods (PSQM+, MNB and EMBSD), the PSQM+ provides perceptual distance values for most parts of the speech segment. The EMBSD and MNB only show the variations in perceived speech quality for frames with high energy. A loss at the unvoiced or voiced segments with small energy (see Figure 2) has no impact on perceived speech quality (flat line area in Figures 4 and 5). This is due to the different processing methods for silence and non-silence frames in the perceptual quality measurement algorithms. For EMBSD, the perceptual distance for an entire test speech segment is obtained by averaging over all non-silence frames (which are defined as the frames with the energy of the reference speech and the degraded speech both above their preset thresholds). For a loss at short and small energy voiced segments (e.g. voiced segment 1), the degraded speech with a loss has a limited energy. This is not taken into account by the EMBSD in the overall perceptual distance calculation and causes a flat area in Figure 5 (e.g. for voiced segments 1 and 3). A similar phenomenon exists for the MNB. The PSQM+ also classifies the frames as silence or non-silence. But it calculates all perceptual distances for silence or non-silence frames and uses different weighting factors for the overall perceptual distance calculation. Thus PSQM+ (Figure 3) also gives the perceptual distance value for a loss during small energy.

B. Convergence time with loss location

The second experiment was carried out to analyze the convergence time and its relationship to speech content or loss position. The convergence time is calculated by comparing the difference between the degraded signal without loss and the degraded signal with loss (as shown in Figure 1). First the MSE method [1] is used to calculate the convergence time for each loss position for a speech waveform such as that shown in Figure 2. Here the convergence time is defined as the first good frame received

after a burst of lost frames until the frame with its MSE value below a threshold (1% of the maximum MSE value seen so far). The convergence time for G.729 is shown in Figures 9, in units of frames (10ms/frame). From the figure, we can see that the convergence times are almost the same for different loss sizes. It shows a good linear relationship for loss at the voiced segments. It is at a maximum at the beginning of the voiced segments and decreases gradually to a minimum at the end of the voiced segments. The convergence time for a loss at the unvoiced segments appears stable. Similar results were also obtained for the AMR and G.723.1 codecs. It seems that the convergence time is only related to the speech content and not to codec and frame loss size.

We analyze further the convergence time based on perceptual distance. We measured the frame-based PSQM+ values between degraded speech without loss and degraded speech with loss. We choose two voiced segments in Figure 2. One with only voiced part (V(2) in Figure 2) and another one with the adjacent unvoiced part (V(4) in Figure 2). We change loss positions from the beginning to the end of the waveforms. The perceptual distance variation curves for selected loss positions are shown in Figure 10 and 11, in the unit of frames (here it is the frame of PSQM+ calculation, which is 32ms frame size with 50% overlapping resulting in 16 ms real frame size). Curves 1 to 5 (Figure 10) and 1 to 12 (Figure 11) correspond to the loss position from left to right. The loss position for each curve corresponds to the first non-zero point in the curve. The duration of the frames with non-zero (or over a threshold) perceptual distance is related to the convergence time.

From Figures 10 and 11, we can see that if a loss occurs during a voiced segment, then the convergence time is almost the remainder of the length of that voiced segment from the loss point (curve 1 to 5 in Figure 10 and curve 6 to 12 in Figure 11). The perceptual distance itself changes significantly with changes in the location of loss while the influence of the loss seems only limited to the voiced segment. The convergence times are almost the same as for a loss at unvoiced parts (curves 1 to 5 in Figure 11). The PSQM+ curves vary in a similar way. This explains the linear relationship of the convergence time during the voiced segments and flat variation during the unvoiced segments as shown in Figure 9. PSQM+ variation curves also show the overall PSQM+ values for the different loss position. We also tested other voiced segments and obtained similar results. The convergence time is more closely related to speech content and less affected by frame loss size and codec type. The convergence time is constrained by the duration of the voiced segments.

VI. CONCLUSIONS

We have investigated the impact of loss positions on perceived speech quality and the relationships between the convergence time and loss locations. Preliminary results show that a loss at unvoiced speech segment has almost no obvious impact on perceived speech quality. However, a loss at the beginning of voiced segments has the most severe impact on perceived speech quality. We have explained this effect from both the perspectives of the concealment and objective perceptual measurement algorithms. The impact of loss position on perceived speech or the concealment performance of three modern codecs (G.729, G.723.1 and AMR) have also been compared and analyzed. Three different perceptual speech quality measurement algorithms (PSQM+, MNB and EMBSD) are compared for the purpose of loss location analysis. We have analyzed the convergence times for different loss locations and different codecs by taking into account the normal MSE and perceptual PSQM+ measure. The results show that the convergence time is affected mainly by speech content (e.g. it is very stable within unvoiced segment whereas it varies but constrained by the duration of the voiced segments).

This work should help to fully understand the real impact of packet loss on perceived speech quality and the features of the convergence time in order to set the real loss constraint distance between the losses. This could be help for the development of a perceptually relevant packet loss metric, which could be valuable in non-intrusive VoIP measurements or to set up more efficient speech recovery systems.

Further research will focus on a more extensive analysis of the impact of packet loss on speech content.

ACKNOWLEDGEMENT

We are grateful to the Speech Processing Lab of the Electrical and Computer Engineering Department at Temple University, especially Dr. Wonho Yang and Prof. Robert Yantorno, for providing us with the Enhanced Modified Bark Spectral Distortion (EMBSD) software to evaluate the performance of concealment in this paper.

We are grateful to WWG/Acterna for sponsorship.

REFERENCES

[1] J. Rosenberg. G.729 Error Recovery for Internet Telephony. Project Report, Columbia University, May 1997

[2] R. Koodli and R. Ravikanth, One-way Loss Pattern Sample Metrics <draft-ietf-ippm-loss-pattern-03.txt>, Internet Draft, Internet Engineering Task Force, July 2000

[3] ITU-T Recommendation G.729, Coding of Speech at 8 kbit/s Using Conjugate-Structure Algobraic-Code-Excited Linear-Prediction (CS-ACELP), March 1996

[4] ITU-T Recommendation G.723.1, Dual Rate Speech Coder for Multimedia Communication Transmitting at 5.3 and 6.3 kbit/s, March 1996

[5] 3G TS 26.091, AMR Speech Codec; Error Concealment of Lost Frames

[6] H. Sanneck and N. Tuong Long Le, Speech Property-Based FEC for Internet Telephony Applications, Proceedings of the SPIE/ACM SIGMM Multimedia Computing and Networking Conference 2000, San Jose, CA, January 2000

[7] Q. Xie, S. Gupta, Error Tolerant RTP Payload Format for AMR <draft-xie-avt-et-rtp-amr-00.txt>, Internet Draft, Internet Engineering Tasks Force, October 2000

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

[9] ITU-T Contribution COM 12-20-E, Improvement of the P.861 Perceptual Speech Quality Measure, KPN Research, Netherlands, Dec. 1997

[10] S. Voran, Objective Estimation of Perceived Speech Quality – Part I: Development of the Measuring Normalizing Block Technique, IEEE Trans. on Speech and Audio Processing, Vol. 7, No.4. July 1999, pp. 371-382

[11] S. Voran, Objective Estimation of Perceived Speech Quality – Part II: Evaluation of the Measuring Normalizing Block Technique, IEEE Trans. on Speech and Audio Processing, Vol. 7, No.4. July 1999, pp. 383 -390

[12] W. Yang, Enhanced Modified Bark Spectral Distortion (EMBSD): An Objective Speech Quality Measurement Based on Audible Distortion and Cognition Model, Ph.D Dissertation, May 1999, Temple University, USA

[13] ITU-T Recommendation G.729 Annex B, A silence compression scheme for G.729 optimized for terminals conforming to Recommendation V.70, November 1996

[14] ETSI EN 301 704 V7.2.1 (2000-04), Digital cellular telecommunications system (Phase 2+); Adaptive Multi-Rate (AMR) speech transcoding (GSM 06.90 version 7.2.1 Release 1998)

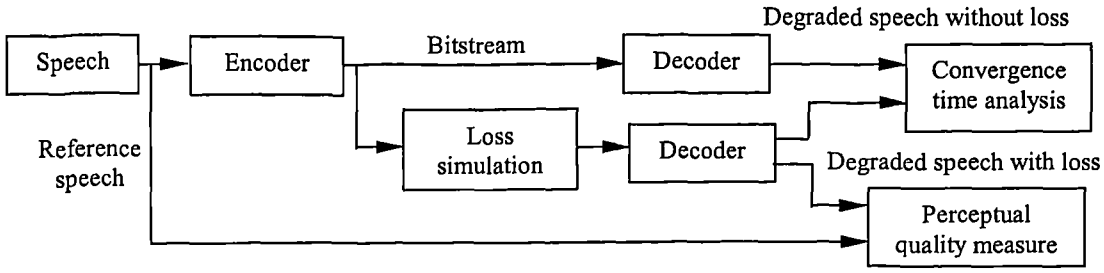


Figure 1: Structure of the simulation system

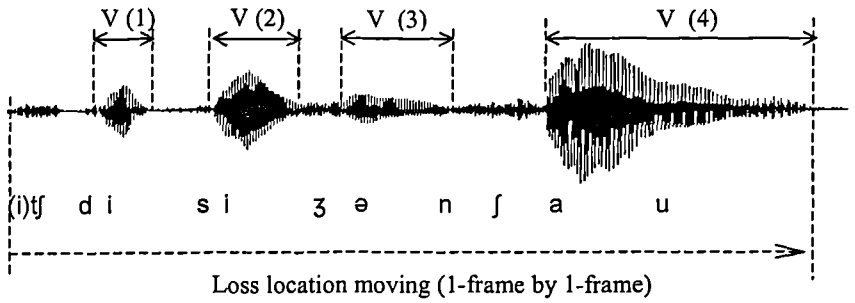


Figure 2: Speech waveform for the 1st talkspurt of test sentence
(The sentence is "_each decision show(s)_". V(1) to V(4) corresponds to 4 voiced segments)

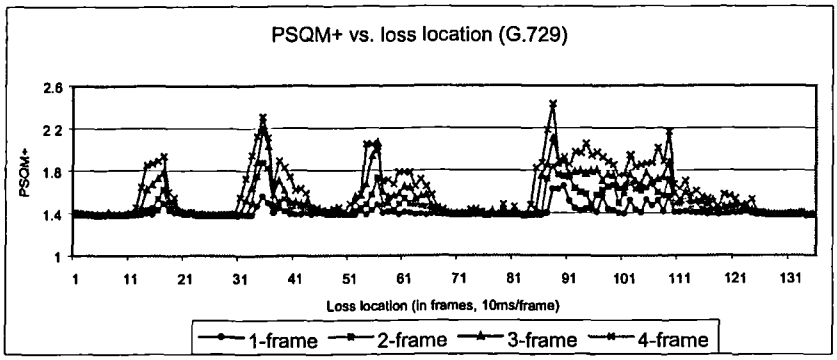


Figure 3: Overall PSQM+ values vs. loss location for G.729

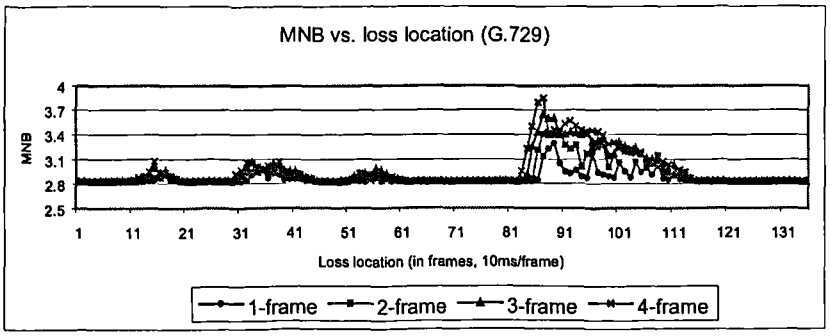


Figure 4: Overall MNB value vs. loss location for G.729

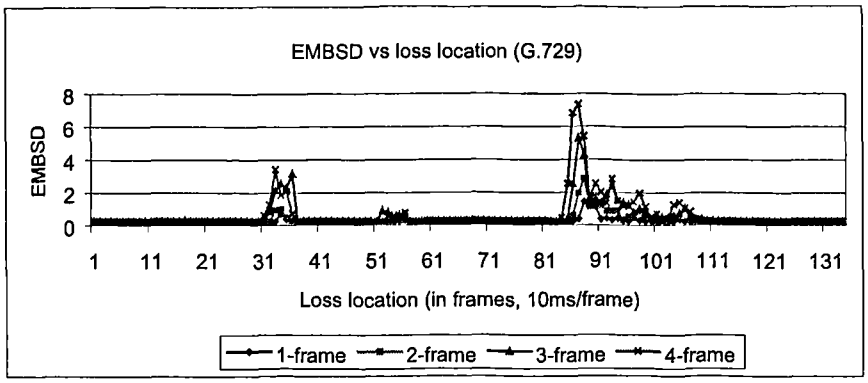


Figure 5: Overall EMBSD value vs. loss location for G.729

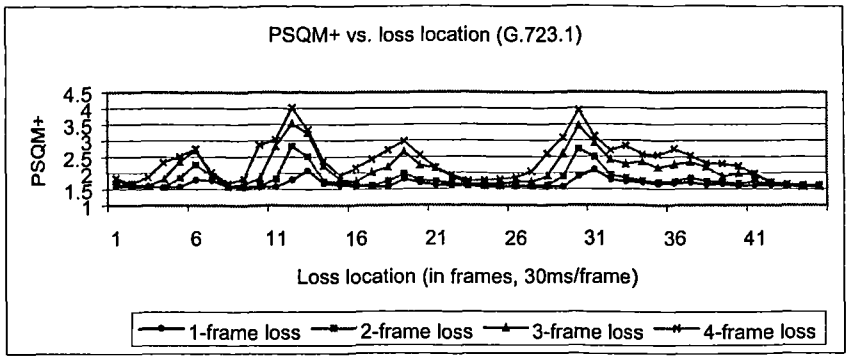


Figure 6: Overall PSQM+ value vs. loss location for G.723.1 (6.3 Kb/s)

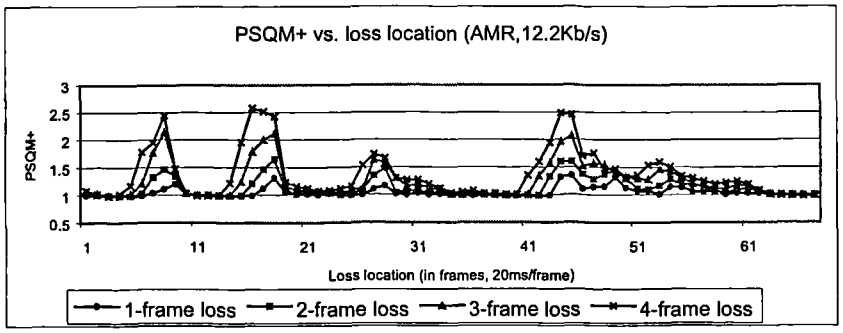


Figure 7: Overall PSQM+ value vs. loss location for AMR (12.2 Kb/s)

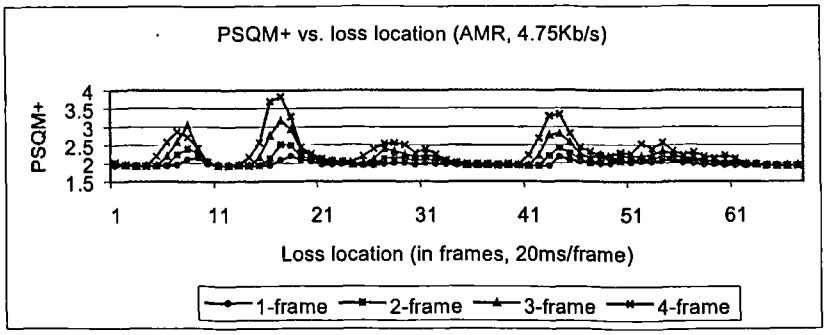


Figure 8: Overall PSQM+ values vs. loss location for AMR (4.75Kb/s)

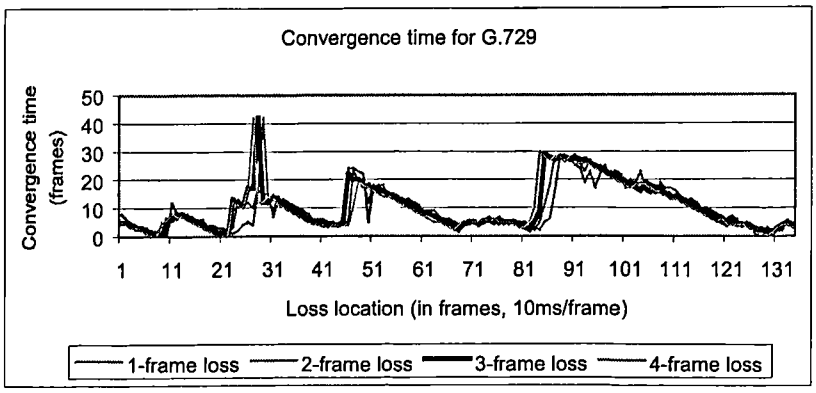


Figure 9: Convergence time vs. loss location for G.729

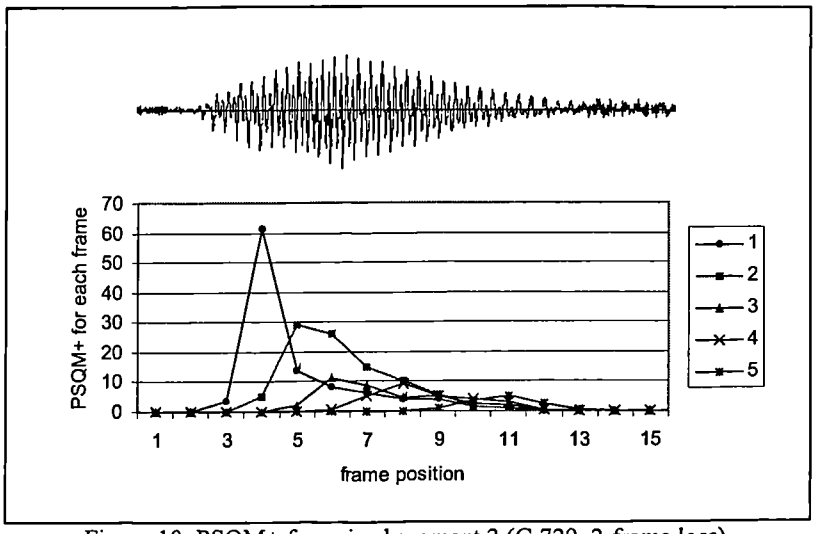


Figure 10: PSQM+ for voiced segment 2 (G.729, 2-frame loss)
(Curves 1 – 5 correspond to 5 loss locations from left to right)

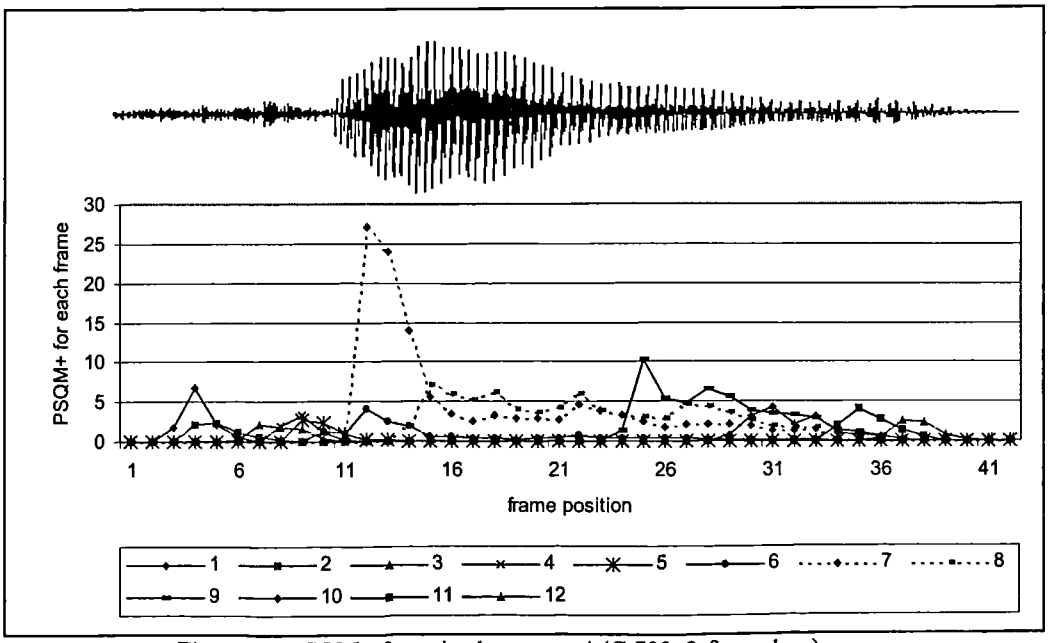


Figure 11: PSQM+ for voiced segment 4 (G.729, 2-frame loss)
(Curves 1 to 12 correspond to 12 loss locations from left to right)