2008

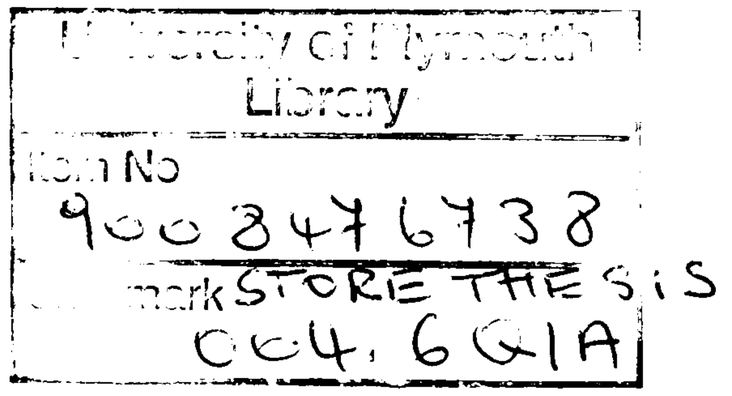
Enhancement of perceived quality of service for voice over internet protocol systems

Qiao, Zizhi

ENHANCEMENT OF PERCEIVED QUALITY OF SERVICE FOR VOICE OVER INTERNET PROTOCOL SYSTEMS

Z. Qiao

Ph.D. September 2008



ENHANCEMENT OF PERCEIVED QUALITY OF SERVICE FOR VOICE OVER INTERNET PROTOCOL SYSTEMS

by

ZIZHI QIAO

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

September 2008

Enhancement of Perceived Quality of Service for Voice over Internet Protocol Systems

Zizhi Qiao

Abstract

Voice over Internet Protocol (VoIP) applications are becoming more and more popular in the telecommunication market. Packet switched VoIP systems have many technical advantages over conventional Public Switched Telephone Network (PSTN), including its efficient and flexible use of the bandwidth, lower cost and enhanced security.

However, due to the IP network's "Best Effort" nature, voice quality are not naturally guaranteed in the VoIP services. In fact, most current VoIP services can not provide as good a voice quality as PSTN. IP Network impairments such as packet loss, delay and jitter affect perceived speech quality as do application layer impairment factors, such as codec rate and audio features. Current perceived Quality of Service (QoS) methods are mainly designed to be used in a PSTN/TDM environment and their performance in VoIP environment is unknown. It is a challenge to measure perceived speech quality correctly in VoIP system and to enhance user perceived speech quality for VoIP system.

The main goal of this project is to evaluate the accuracy of the existing ITU-T speech quality measurement method (Perceptual Evaluation of Speech Quality – PESQ) in mobile wireless systems in the context of VoIP, and to develop novel and efficient methods to enhance the user perceived speech quality for emerging VoIP services especially in mobile VoIP environment. The main contributions of the thesis are threefold:

(1) A new discovery of PESQ errors in mobile VoIP environment. A detailed investigation of PESQ performance in mobile VoIP environment was undertaken and included setting up a PESQ performance evaluation platform and testing over 1800 mobile-to-mobile and mobile-to-PSTN calls over a period of three months. The accuracy issues of PESQ algorithm was

investigated and main problems causing inaccurate PESQ score (improper time-alignment in the PESQ algorithm) were discovered . Calibration issues for a safe and proper PESQ testing in mobile environment were also discussed in the thesis.

(2) A new, simple-to-use, VoIP jitter buffer algorithm. This was developed and implemented in a commercial mobile handset. The algorithm, called "Play Late Algorithm", adaptively alters the playout delay inside a speech talkspurt without introducing unnecessary extra end-to-end delay. It can be used as a front-end to conventional static or adaptive jitter buffer algorithms to provide improved performance. Results show that the proposed algorithm can increase user perceived quality without consuming too much processing power when tested in live wireless VoIP networks.

(3) A new QoS enhancement scheme. The new scheme combines the strengths of adaptive codec bit rate (i.e. AMR 8-modes bit rate) and speech priority marking (i.e. giving high priority for the beginning of a voiced segment). The results gathered on a simulation and emulation test platform shows that the combined method provides a better user perceived speech quality than separate adaptive sender bit rate or packet priority marking 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.

Part of this study was done in cooperation of Motorola Ltd. UK

Publications:

1. Z. Qiao, T. Robinson, N. Luckcuck, L. Sun and E. Ifeachor, "Perceived QoS measurement in a live wireless mobile VoIP environment", Submitted to Wireless Personal Communications, Springer Netherlands, 2008

2. Z. Qiao, L. Sun and E. Ifeachor, "Case study of PESQ performance in live wireless mobile VoIP environment", in Proceedings of IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC 2008) - VoIP Technologies Workshop, Cannes, France, September 2008.

3. Z. Qiao, R. Venkatasubramanian, L. Sun and E. Ifeachor, "A New Buffer Algorithm for Speech Quality Improvement in VoIP Systems", Wireless Personal Communications, Volume 45, Number 2, Page 189–207, Springer Netherlands, Oct 2007.

4. 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), Volume 3, Page 1473–1477, Paris, France, June 2004.

5. Z. Qiao, and E. Ifeachor, "A New Quality of Service Control Scheme for VoIP Networks", in Proceeding of EPSRC PREP2003 conference, Volume 2, Page 23–24, Exeter, UK, April, 2003.

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

Signed ..... **Date** 25/02/2009

Word Count: 52320

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. Lingfen Sun, Dr. Kit Reeve and Dr. Paul Davey for their support and guidance.

I would also like to thank my colleagues Dr. Zhuoqun Li (Wood), Dr. Ping Hu, Dr. Brahim Hamadicharef, Mr. Nicolai Heilemann and others.

My family and friends gave me strong support during the study and I would like to thank them.

Table of Contents

List of Tables

List of Figures

List of Abbreviations and Glossary

3G	Third Generation (wireless)
3GPP	3G Partnership Project
3GPP2	3G Partnership Project 2
ACELP	Algebraic Codebook Excited Linear Prediction
ACR	Absolute Category Rating
ADPCM	Adaptive Differential Pulse Code Modulation
AMR	Adaptive Multi-Rate
ANN	Artificial Neural Networks
CBR	Constant Bit Rate
CCI	Call Clarity Index
CELP	Code Excited Linear Prediction
CODEC	COder DECoder
CoS	Class of Service
CPU	Central Processing Unit
CS-ACELP	Conjugate Structure-Algebraic Code Excited Linear Prediction
DCR	Degradation Category Rating
DiffServ	Differentiated Services
DMOS	Degradation Mean Opinion Score
DSP	Digital Signal Processing
ESTI	European Telecommunications Standards Institute
FEC	Forward Error Correction
FMC	Fix Mobile Convergence

GPS	Global Positioning System
GSM	Global System for Mobile Communications
HTTP	Hypertext Transfer Protocol
IAX	Inter-Asterisk eXchange
ICMP	Internet Control Message Protocol
IETF	Internet Engineering Task Force
iLBC	Internet Low Bit Rate Codec
IMS	IP Multimedia Subsystem
InterServ	Integrated Services
IP	Internet Protocol
ISDN	Integrated Services Digital Network
ITU	International Telecommunication Union
LAN	Local Area Network
LPC	Lost Packet Concealment
MAC	Media Access Control
MGCP	Media Gateway Control Protocol
MOS	Mean Opinion Score
MPLS	Multi Protocol Label Switching
NS-2	Network Simulator version 2
PC	Personal Computer
PCM	Pulse-code modulation
PEAQ	Perceptual Evaluation of Audio Quality
PESQ	Perceptual Evaluation of Speech Quality
PESQ-LQ	Perceptual Evaluation of Speech Quality – Listening Quality
PLA	Play Late Algorithm
POTS	Plain old telephone service

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
RSVP	Resource ReSerVation Protocol
RTCP	Real Time Transport Control Protocol
RTP	Real Time Transport Protocol
RTT	Round Trip Time
SCN	Switched Communication Network
SID	Silence Insertion Description
SIP	Session Initiation Protocol
SLA	Service Level Agreement
SNR	Signal to Noise Ratio
TCP	Transmission Control Protocol
TDM	Time-Division Multiplexing
UDP	User Datagram Protocol
UMA	Unlicensed Mobile Access
UMTS	Universal Mobile Telecommunications System
VAD	Voice Activity Detection
VoIP	Voice over Internet Protocol
WAN	Wide Area Network

Chapter 1

Introduction

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

1.1 Motivations

Voice over Internet Protocol (VoIP) is a rapid developing approach in the telecommunication industry and allows speech to be transmitted to its destination by Internet Protocol (IP) packets.

VoIP technique can provide more bandwidth efficiency because the IP trunk is shared by different users at the same time, compared with traditional circuit switching techniques. The fact that VoIP system, which includes not only the IP network operator, the VoIP service provider but also the VoIP end user, is more open to both service providers and customers means that, compared with traditional PSTN services, VoIP systems can provide more features with flexibility, more security and lower cost due to its packet switching nature. VoIP services are acting as an add-on to current infrastructure, but will convert and replace traditional telecommunication systems.

Quality of service (QoS) is one of the most important issues in this convergence process.

Only when user perceived quality is not significantly lower than the quality in a traditional telecommunication system, would the VoIP deployment take place effectively because no company would like to subject its users to poor perceived quality. However, in current VoIP system, some critical challenges need to be addressed before the perceived QoS can reach that of the traditional PSTN service. Those technical challenges including the accurate and efficient measurement of user perceived quality in the VoIP system, the reduction of the impact of delay and delay jitter to perceived quality, the enhancement of user perceived quality in specific VoIP field including the wireless mobile VoIP environment.

An accurate measurement system is necessary to improve and validate the improvement of perceived Quality of Service in the VoIP system. Current perceived speech quality measurement methods are not designed for the VoIP environment so measurement results may not reflect reality accurately enough. The development of a platform to calibrate the measurement tool and to perform VoIP QoS tests are important as this provides a basis for VoIP QoS enhancement and hence research.

Due to the packet switching nature of the VoIP technology, speech quality is likely to be affected by new impairment factors (including packet loss, delay and delay variation [1–3]), which do not exist in traditional circuit switched telecommunication networks. End-to-end packet loss may introduce noise, silence gaps or even more uncomfortable distortion in the speech, all of which will affect the user's perception. Delay and delay jitter introduced by the IP network may cause packet loss or longer "mouth to ear" delay.

When a VoIP application is used in a new environment (for instance a wireless mobile environment), the restrictions of the specific environment affect the perceived QoS. For example, the limited bandwidth of the wireless hop and the limited processing power of the mobile host need to be considered when a QoS enhancement method is applied. The limitations in the wireless mobile VoIP environment may affect perceived quality in a different way when compared with a fixed VoIP system. It is very important to consider the environmental limitations and features of the VoIP application when optimizing specific system, so that an enhanced user

perceived QoS can be delivered.

The fact that a user or application can control more elements in a VoIP system also brings perceived QoS challenges. A user can select a codec, decide audio features, adjust jitter buffer algorithm, set speech volume and have more control on a VoIP communication system, however any of these elements can cause a bad user perceived speech quality if not selected correctly or optimized.

There is a fundamental need to enhance the perceived QoS by considering different features of the VoIP system, from application layer to lower layers. It is also important to measure the perceived QoS accurately before any enhancement can be validated.

1.2 Aim and Objectives

QoS is an important issue in the VoIP environment. Multi-route, peer-to-peer, user end control instead of the network side control, are the main differences when IP network is compared with traditional PSTN system. These differences introduce more complicated QoS issues, especially in a resource restrained mobile environment.

The aim of the project on which the thesis is based is to enhance end-to-end user perceived QoS in VoIP system, with emphasis on the wireless mobile VoIP system.

This is to make it possible to achieve the same or even better perceived QoS in VoIP system, when compared with the traditional PSTN/TDM network. It will help to improve user experience in existing VoIP system and encourage further wider development and deployment of the VoIP technology.

Specific objectives of the research are presented as follows.

User perceived voice quality measures the goodness of the voice communication system. It is measured by Mean Opinion Score (MOS). MOS is measured subjectively by a group of ordinary people. But it is a very time consuming and expensive method to use because it involves a large amount of human testers. There are different measurement methods available to

measure speech quality in telecommunication systems but to use them for VoIP speech quality measurement and improvement, the performance and accuracy of the measurement method need to be investigated first.

PESQ [4] is the industrial and international standard to objectively measure user perceived speech quality and it is widely used as a cost effective alternative tool to subjective measurement method MOS [5], to measure the user perceived speech quality in a telecommunication system. The context of user perceived speech quality measurement has now being extended to VoIP system, in conjunction with the rapid development of VoIP technologies.

However, the suitability of current measurement method and their reliability in the VoIP context need to be investigated, especially in the newly developed wireless mobile VoIP system.

This brings us to the first research objective:

- To develop and implement a live-network-based VoIP QoS measurement platform including wireless mobile VoIP capability, which can be used for development and evaluation of VoIP QoS enhancement mechanisms. And to undertake a detailed investigation into possible PESQ measurement errors in the new wireless mobile VoIP environment. This will form the basis for the perceived QoS enhancement research.

At the application layer, different methods exist to improve QoS, such as Forward Error Correction (FEC) at the sender side, the jitter buffer and Packet Loss Concealment (PLC) at the receiver side. In this project, one research focus is on the jitter buffer algorithm to improve perceived speech quality.

The nature of human speech, features of speech codec and jitter buffer in the receiving end can be exploited to improve perceived QoS. The information provided by the codec about speech content can be used to indicate better operation of the adaptive jitter buffer with the awareness of end-to-end user perceived speech quality. The specific application area, for instance wireless mobile VoIP environment, need to be considered because of the specific resource restrains.

This brings us to the second research objective:

- To exploit the nature of human speech and features of codecs in the design of novel jitter buffer algorithms for specific VoIP application area such as wireless mobile VoIP application. This will provide optimized perceived QoS for dynamic network and application conditions and increase over all user satisfaction.

To improve the overall user perceived quality in complex network environments is a very challenging task. It needs to consider the trade off between individual user perceived quality and the network efficiency. There are mechanisms to support different level of QoS requirements from different applications in existing IP infrastructure.

Human speech perception has its own features, for instance different speech samples in a speech segment have different importance. The codec's voice activity detection function can be exploited to identify important parts of a speech, and priority marking can be used to protect the more important parts of a speech segment.

How the human speech's and speech codec's feature at the application layer including the adaptive codec rates and the difference importance of speech samples in a speech segment and the current VoIP systems' advanced techniques at the network layer including the QoS provisioned network management to optimize the overall end-to-end speech quality can be utilized, becomes an important research topic.

This brings us to the third research objective:

- To apply knowledge of the network and application layer QoS mechanism, features of codecs and human speech in the development of novel and efficient QoS control method for VoIP network. This objective involves not only the application layer elements but also network QoS mechanism and should provide an overall enhanced perceived quality.

The three research objectives are linked together by the main aim of this project, which is to enhance the user perceived speech quality in VoIP systems.

The first research objective addresses the platform and measurement tool calibration issue. The second research objective seeks to improve speech quality in a more real life context in the application layer. The third objective is to find and improve the link between user perceived speech quality and the VoIP system across the application layer and the lower layers including the network layer.

By finish these research objectives, the contribution to knowledge can be made not only in VoIP speech quality improvement research but in the improvement of user perceived speech quality in real VoIP systems.

1.3 Contributions of Thesis

The contributions of the thesis are the following:

1. A new discovery of PESQ errors in wireless mobile VoIP environment.

   A detailed investigation of PESQ performance in mobile VoIP environment was undertaken and a few PESQ errors are discovered in the wireless mobile VoIP environment.

   The accuracy issues of PESQ algorithm was investigated and main problems causing inaccurate PESQ score (improper time-alignment in the PESQ algorithm) were discovered. Calibration issues for a safe and proper PESQ testing in mobile environment were also discussed in the thesis.

   First, a live test platform including real wireless VoIP mobile and live IP network is developed, which allows the investigation of the suitability of objective measurement method and the impact of perceived quality improvement methods to be undertaken in a live real wireless mobile environment.

   One of the PESQ errors discovered is that it can not deal with long duration (in the order of hundreds of milliseconds) talk spurt missing situation. When a long term talk spurt is replaced by silence, the frame by frame delay calculation may not give the correct

reading on relative delay. The way to solve this PESQ error is to compare the waveforms and pick out the erroneous ones or to use a statistical method to mask out those error samples.

The other PESQ error discovered in this research is that it may give incorrect score in relative delay shifting situations, which is common in VoIP calls. An investigation of more than 1800 mobile-to-mobile and mobile-to-PSTN VoIP calls shows that in a certain test scenario, more than 5% of the PESQ score in a 60 sample group is significantly lower than subjective scores. Only 1800 samples are investigated due to time limitations. PESQ frame by frame score is not correct in those delay shifting cases due to errors caused by its time alignment algorithm. The solution to this problem is to compare the waveform of abnormal results or to use statistic tools to pick out abnormal samples. This contribution is published in [6]. Another paper related to this contribution is in the final stages of preparation. Details about this contribution is presented in Chapter 3 of this thesis.

2. A new jitter buffer algorithm

A new adaptive jitter buffer algorithm was developed and implemented in a commercial mobile handset (Motorola V560-UMA) for VoIP applications especially wireless mobile VoIP system.

The proposed new "Play Late Algorithm" extends the buffer delay inside a talkspurt if no earlier speech frames have been played, in the case of a packet containing speech frame is late. This feature increases the chance of a speech frame being played out, instead of being dropped out. This jitter buffer algorithm only adapts (increases or decreases) the jitter buffer delay in silence period of the speech to minimize the perceptual impact to speech quality. There are also mechanisms to fill gaps created by the extension of the buffer delay and to deal with extreme error situation.

The speech quality test carried out on live mobile VoIP environment proves that the new algorithm can improve overall user perceived quality in VoIP enabled mobile conversa-

tion. The advantage of this new algorithm also includes the very low usage of mobile resources when compared with other more complicated adaptive jitter buffer algorithms. This contribution is published in [7]. Details about this contribution is presented in Chapter 4 of this thesis.

3. A new combined method to enhancement perceived QoS for congested VoIP link

   A new VoIP speech quality optimization method is proposed to achieve a better overall user perceived quality in a potentially congested VoIP link.

   The new VoIP speech quality improvement is developed by combining the strength of two speech quality improvement methods. One of these two methods is the user perceived speech quality driven AMR [8] codec rate adapt mechanism, which can choose the best suitable AMR codec rate depending on the predicted user perceived speech quality based on frequent RTCP [9] feedbacks. This method also shows better performance compared with traditional loss rate driven adaptive mechanism.

   The other method is to link different parts of the speech to different priority levels in IP network QoS levels and therefore treat them with different QoS levels and achieve a better perceived speech quality. The perceptually important parts of the active speech, for instance beginning parts of voiced segments are given higher priority in packet transmission, thus get less chance of packet loss compared with unprotected packets. The overall speech quality is improved by this marking mechanism. The new combined method shows a better overall user perceived speech quality in a simulation and emulation combined test platform. This contribution is published in [10]. Details about this contribution is presented in Chpater 5.

1.4 Outline of Thesis

The outline of the thesis is shown in Figure 1.1.

In Chapter 2, a survey of VoIP and QoS issues is given. The detailed literature review includes the features of the IP communication system and its suitability for real time speech conversation, the features of IP network for supporting QoS, factors that impact end-to-end user perceived speech quality, the measurement methods, current applications of VoIP and some VoIP QoS control mechanisms. This detailed survey of current VoIP system and its QoS issues highlights problems and state-of-art solutions. This provides a basis for subsequent research and proposed improvement methods.

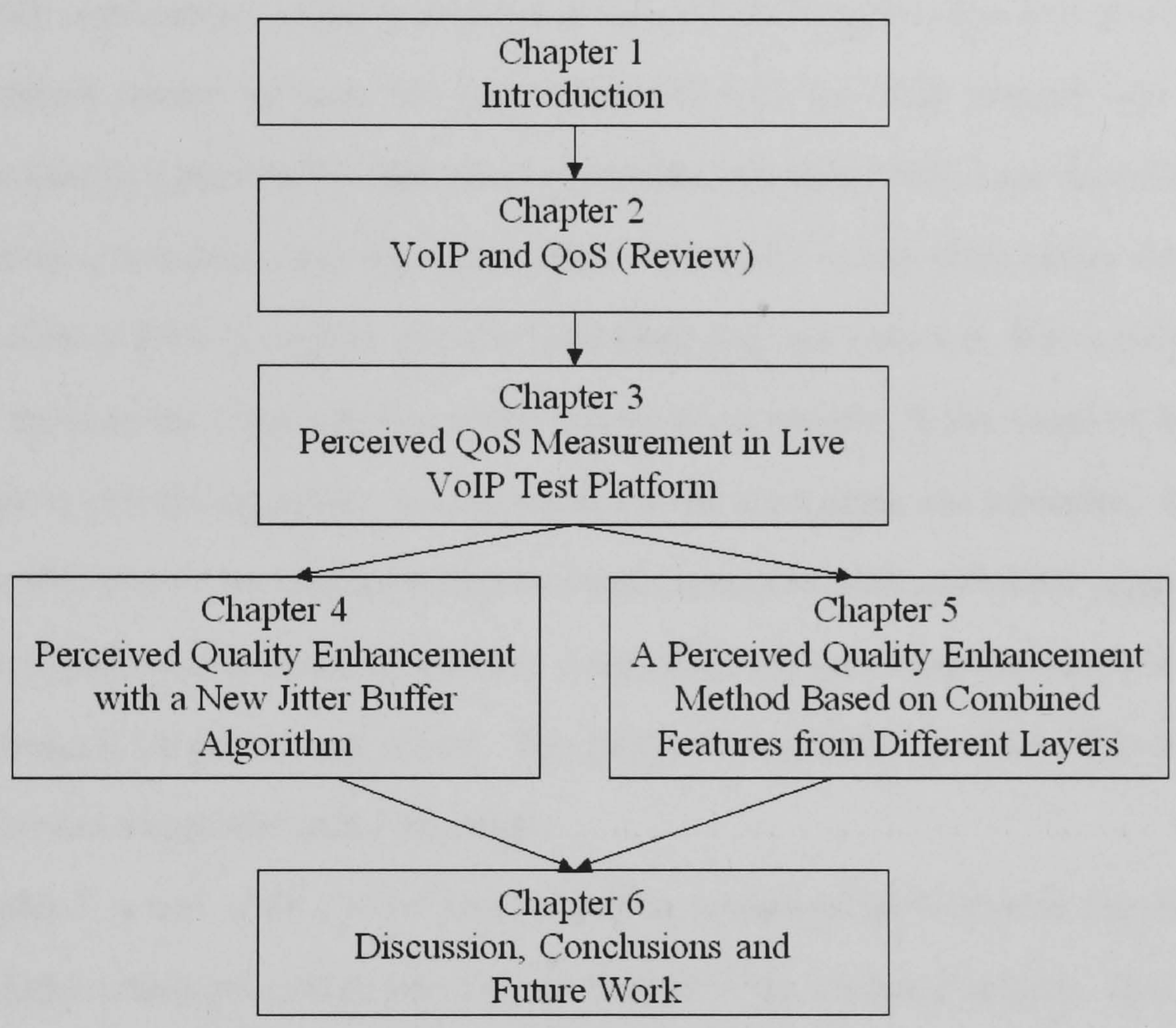


Figure 1.1: Thesis outline

In Chapter 3, the performance of the international and industrial standard for measuring user perceived speech quality, PESQ is investigated in live VoIP enabled mobile speech quality test. The live test platform is discussed in detail together with problems discussion about the calibration of the test system. The problems include those associated with gaps in play/record

speech sample, hardware and software related volume adjustment, speech quality distortion from audio features. A detailed case study of PESQ's performance in mobile VoIP environment is carried out. PESQ's incorrect relative delay calculation problem and the frame by frame PESQ error caused by relative delay shift problem are discussed in a case study format. Some solutions and discussions are given.

In Chapter 4, the new Play Late jitter buffer algorithm is presented for mobile VoIP applications. This chapter starts with an introduction on adaptive jitter buffer algorithm research and the user perceived speech quality challenge in the VoIP applications, especially in the mobile wireless VoIP application. After a detailed review of VoIP application and services' speech quality problems caused by jitter and loss introduced with the VoIP network and application and current adaptive jitter buffer algorithm evaluation, the new "Play Late Algorithm" is presented. Inside a talkspurt, this algorithm allows increases in the jitter buffer delay for late arriving packets if there is no later packets have been play out before it. But it only adapts the buffer size down in the silence period of the conversation speech. A live wireless VoIP mobile test platform is introduced before the test result for the algorithms are presented. Live mobile VoIP call results show a performance improvement compared with traditional static or adaptive jitter buffer algorithms. In terms of resource constraints, the new algorithm does not introduce excessive demand on processing power. The conclusion section concludes this chapter with some future work suggested under this topic.

In Chapter 5, a new VoIP system speech quality optimization method is introduced. This chapter includes a background introduction of the VoIP QoS control problems. This is followed by a survey of different VoIP QoS control schemes, including the rate adaptive QoS control scheme and the priority marking QoS control scheme. The new combined control mechanism is presented in detail together with the simulation system. In the simulation system section, different simulation test setups for different QoS control mechanism are presented. The test experiment results are discussed later with a performance comparison between individual QoS control methods and the new combined priority marking and user perceived speech quality

driven QoS control algorithm.

Chapter 6 concludes the thesis. The conclusion chapter includes a brief summary of contributions of this thesis and its limitations, including the methodology limitations, setup and experiments limitations. A brief suggestion for future work is also included in Chapter 6.

Chapter 2

VoIP and QoS

The purpose of this chapter is to present a background for VoIP system and its QoS issues, which underpins the work presented in this thesis.

The IP protocol, the voice applications on the IP network and current VoIP products are introduced briefly in Section 2.1. The QoS issues in VoIP systems are discussed in Section 2.2. A detailed study of QoS related elements in the VoIP system and their impacts to the perceived quality is presented in Section 2.3. QoS features and control mechanism in both network layer and other layers especially application layer are presented in Section 2.4. Section 2.5 introduced different QoS measurement methods. The user perceived speech quality measurement method PESQ and its detailed time alignment mechanism is introduced in Section 2.6.

2.1 IP and VoIP

VoIP technology is to transmit packet voice on the IP protocol based networks including the Internet. Advantages of VoIP is low cost, efficient, secure and flexible. The disadvantage includes the potential quality issue and the complexity of configuration, cooperation between user and the operator [11].

A VoIP system needs a few more protocols to operate, which include User Datagram Protocol (UDP) [12], Real-time Transport Protocol/Real-time Transport Control Protocol (RTP/RTCP) [13] for speech stream transmission and signaling protocols for instance Session Ini-

tiation Protocol (SIP) [14] and H.323 protocol group from the ITU [15].

2.1.1 IP Protocol and the Internet

Internet Protocol [16] is the fundamental protocol for Internet. The original idea of Internet is to build an unbreakable network, which can survive even under nuclear weapon attack. Therefore, the policy of Internet is packet switching and so called "Best Effort". Based on the packet switch network, many applications developed to support different applications including email, www browsing, ftp, telnet etc. With the development of multimedia techniques, multimedia applications also start to be wildly deployed over the Internet. VoIP is one of the most popular applications.

IP protocol is commonly introduced with a serious of associating protocols, called TCP/IP protocol family. The relations of those protocols are illustrated in the following Figure 2.1.

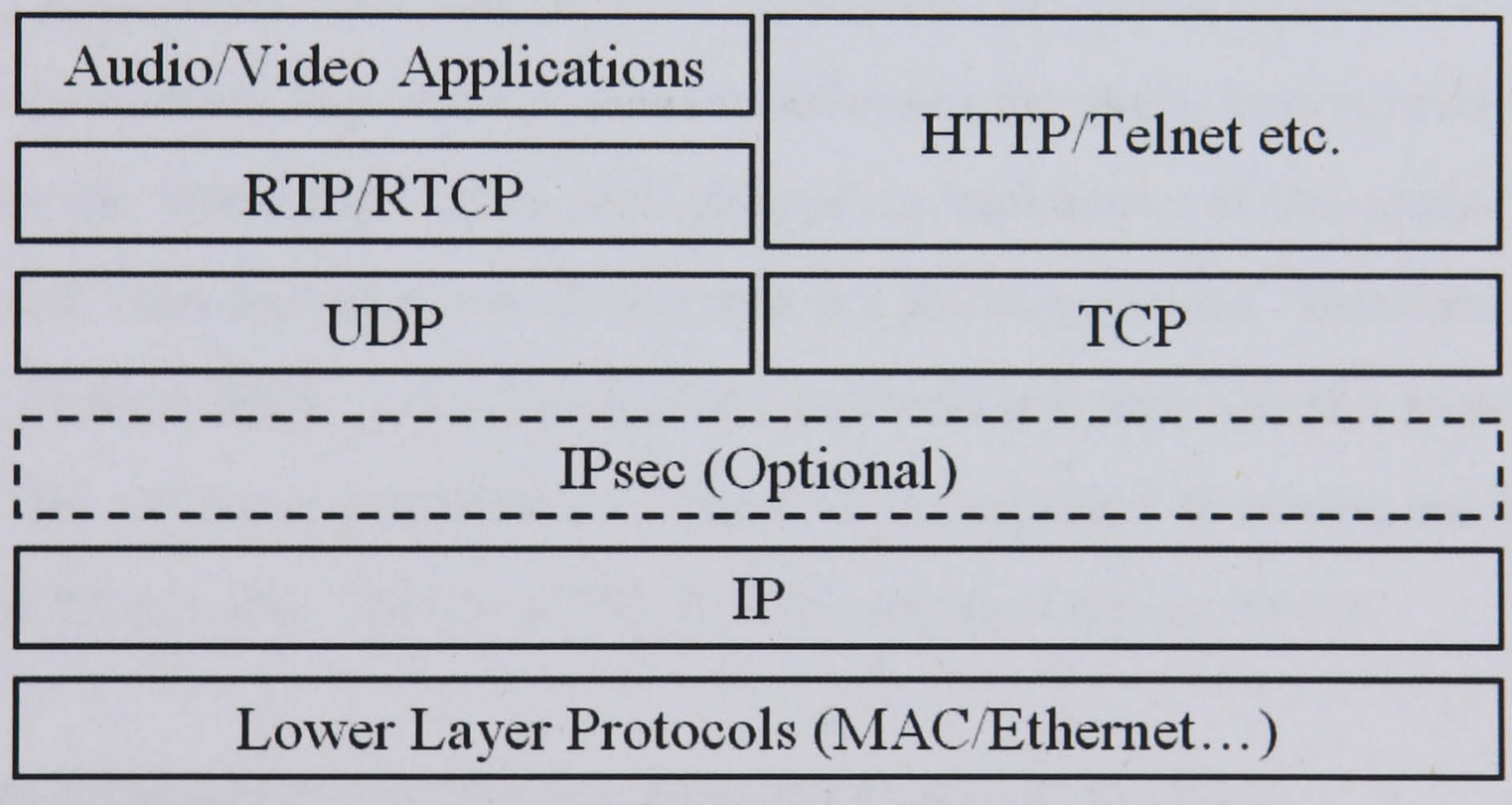


Figure 2.1: IP protocol family

The basic idea of IP based networks is the idea of "Best Effort". And this is the fundamental different between an IP based communication system and a circuit switching communication system, in which the QoS is guaranteed by real or virtual circuits established between communication parties. "Best Effort" means the network does not guarantee a certain level of quality

to the user but only promises "to do it's very best". If a network receives a packet from a user, it tries to find one path to the packet's destination and deliver the packet along this path. There is no guarantee that this delivery works or how much time it will take.

The capacities of the Internet have been considered as sufficient for a long period of time. Thus the "Best Effort" principle has satisfied most users because most of the packets have been delivered in a reasonable time and in most of the cases the percentage of lost packets has been low [17]. And in most of the cases, lost packets can be resend and cause no harm to the application.

With the increasing load in the networks the level of dissatisfaction also increases due to higher network impairments such as higher loss rate and higher delay and jitter. Applications such as VoIP with higher requirements on network impairments add further challenges.

The idea of increasing the network capacities to keep the load low and the "Best Effort" principle working is hardly economical because new applications may come up and take the bandwidth aggressively. This kind of "over engineering" the system is also restricted by the technical opportunities. Hence new systems for delivering the quality have to be developed.

Because the "Best Effort" nature of IP protocol, to implement real time applications such as multimedia communication over IP networks is a challenging work. Therefore, Real-time Transport Protocol (RTP) [13] is developed to deal with real time transport problems in IP networks. The following subsection 2.1.2 discusses the suitability of running real time applications over IP networks. The role of RTP in VoIP network is also introduced.

2.1.2 Realtime Related Protocols in IP Protocol Family

In real time application including voice conversation applications, timely interaction between two parties is required. A delay of minutes or even seconds cannot be accepted. Thus reordering parts which have arrived out of order or are lost would consume too much time. Consequently it is not appropriate to use TCP as transport protocol for this kind of traffic. User Datagram Protocol (UDP) is used instead which adds a smaller header to the packet to speed

up the transmission. It also does not use sequence numbers and acknowledgements, leaving the detection of packet loss to higher level protocols. Therefore 100% correctness cannot be guaranteed when using UDP.

Real-time data is mainly transported through a network using the Real-time Transport Protocol (RTP). This protocol adds a sequence number to the packet header that allows the receiver to detect lost packets. Further it adds information about the type of real-time data sent in the packet and a time stamp that can be used for jitter calculations. Finally media stream information needed for multi-party conferences is given [13].

As a part of the RTP protocol, a protocol named Real-time Transport Control Protocol (RTCP) is introduced. This protocol enables the exchange of further information between the parties of a call. As specified in RFC3550 [9], the RTCP protocol can be used to provide out of band control information for a RTP stream. For example a receiver can report packet loss and jitter to the sender of a transmission, or the parties can exchange their email-addresses and phone numbers.

RTCP packets are periodically sent to transmit control information to both sides of the RTP steam. An important function of the RTCP is to provide QoS report from the receiver side so the sender can perform some adaptation if necessary.

Information contains in a typical RTCP report includes statistics of a RTP multimedia stream such as bytes sent, packets sent, lost packets, round trip delay, jitter and other feedback information. There are few types of RTCP report packets including Sender Report, Receiver Report, Source Description, Goodbye and other custom specified RTCP report packets.

Security is an important feature of a VoIP system. IPsec [18] is developed to enable this feature. It is optional in a VoIP system but important to a commercial VoIP system. IPsec operates between the IP protocol and the UDP/TCP protocol. IPsec header process requires extra processing power and processing time hence have some impact to perceived quality.

UDP and RTP/RTCP are the basic protocols for multimedia applications over IP networks because the connectionless feature of UDP and sequence and time stamp features in RTP/RTCP

make the real time transmission possible. There are still some QoS problems in VoIP applications compare with PSTN system. The following section describes QoS issues in IP based networks.

2.1.3 VoIP

Figure 2.2 shows a typical VoIP communication system, where user can make calls to other user via IP network. To make VoIP calls, a VoIP phone hardware or VoIP software i.e. a soft VoIP phone is necessary. The signalling messages are exchanged before the speech media stream can be transmitted. As shown in Figure 2.2, signalling and speech stream are transmitted inside the IP cloud in a VoIP telephony system, but they may use different protocols and take different routes in the IP network. In a commercial VoIP service, some signalling message may need to be send to a registration server but the voice stream may be send directly to the receiver IP after the signalling process established the call. If the media stream are send through a central server, the central server may introduce extra impairment.

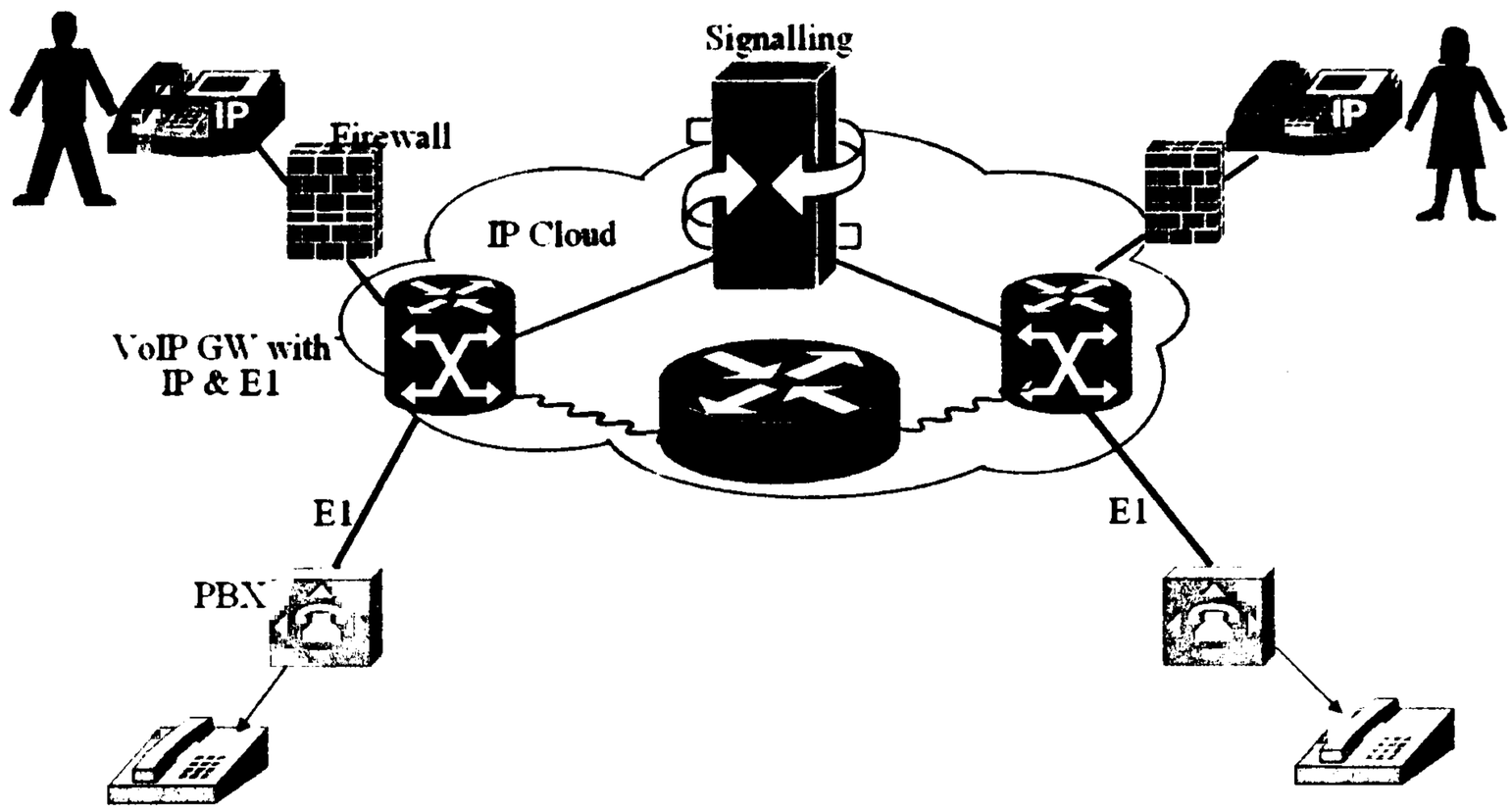


Figure 2.2: A typical VoIP system

There are a few popular VoIP signalling methods include the H.323 protocol group from the

ITU [15], the Session Initiation Protocol (SIP) [14] from the IETF. Other VoIP signalling protocols also include the Inter-Asterisk eXchange protocol (IAX) for software PBX Asterisk [19] and Media Gateway Control Protocol (MGCP) [20].

The methods to transmit speech streams between calling parties are mostly UDP and RTP/RTCP. Some security protocols defined in IPsec [18] may also be involved in the transmission of encoded speech packets.

VoIP Signaling

Besides the exchange of real-time data more information has to be transmitted between the participants of a phone call. This kind of data is called signaling data and is used to control the session. The first controlling data has to be exchanged when the call session is established. The caller has to request the start of a conversation and the called party has to accept or reject this request. If the start of the conversation is accepted the used protocols and codecs have to be negotiated between the telephone sets of the two parties to allow the receiver to decode the transmitted speech. Each change of the settings of the session (for example more parties join the conference or protocols are changed) causes further signaling. Finally the termination of the call needs more information exchange to ensure every party has realized that the session is finished.

SIP protocol is a text based signalling protocol for multimedia services. It is a client-server protocol transported on either TCP or UDP. Currently SIP traffic is mainly transported over UDP due to latency constraints. For reliability, SIP employs its own retransmission mechanisms [21].

H.323 protocol group is a group of recommended protocols introduced by the ITU for implementing packet based multimedia services over IP based networks. It is been developed firstly for local area networks (LAN) environment and then extended to Internet usage. In the protocol group, H.245 is the logical channel signalling protocol, H.225.0 is the call signalling and the terminal to Gatekeeper signalling, H.261 and H.263 are the media codec protocol,

together with G series of voice codecs. H.323 uses TCP to carry most of the signalling messages and has Resource ReSerVation Protocol (RSVP) [22] to ensure predictable QoS.

Details about signalling procedures are not introduced here because the main focus of this project is on perceived speech quality in a VoIP call. There are a few quality related issues including call set up time, resource based admission control, quality feedback in signalling and so on but they are not in focus of this project. In this project, there is a default assumption that different VoIP signalling methods will not introduce speech quality difference into to the conversation if the media transport is done by the same method. All tests and evaluations in this project are performed without considering the impact of different signalling method. This assumption is true if no different network QoS mechanisms are defined or negotiated in the signalling, such as RSVP in H.323. There is no such difference in the experiments in this project.

End-to-End Speech Stream Transfer

In VoIP communications, the frame of encoded speech samples are packetized into different packets and transmitted via the IP network. The speech transport is implemented with the real time transport protocol (RTP). RTP itself does not provide any QoS mechanisms but relies on the signalling protocol to setup the connection, negotiate the media format and if necessary, provide QoS support. RTP runs mostly on top of UDP instead of TCP, for the simplicity and speed requirement.

There is no flow control mechanism in the UDP protocol as in the TCP. This feature makes UDP simple and quick, very suitable for carrying speech streams. There is no error correction mechanism in RTP payload so if there is any error happens in the RTP payload the error is not been treated until it reaches the application layer. Because the real-time nature of speech conversation, residual errors is normally treated in the codec and the wrong samples are not retransmitted [23].

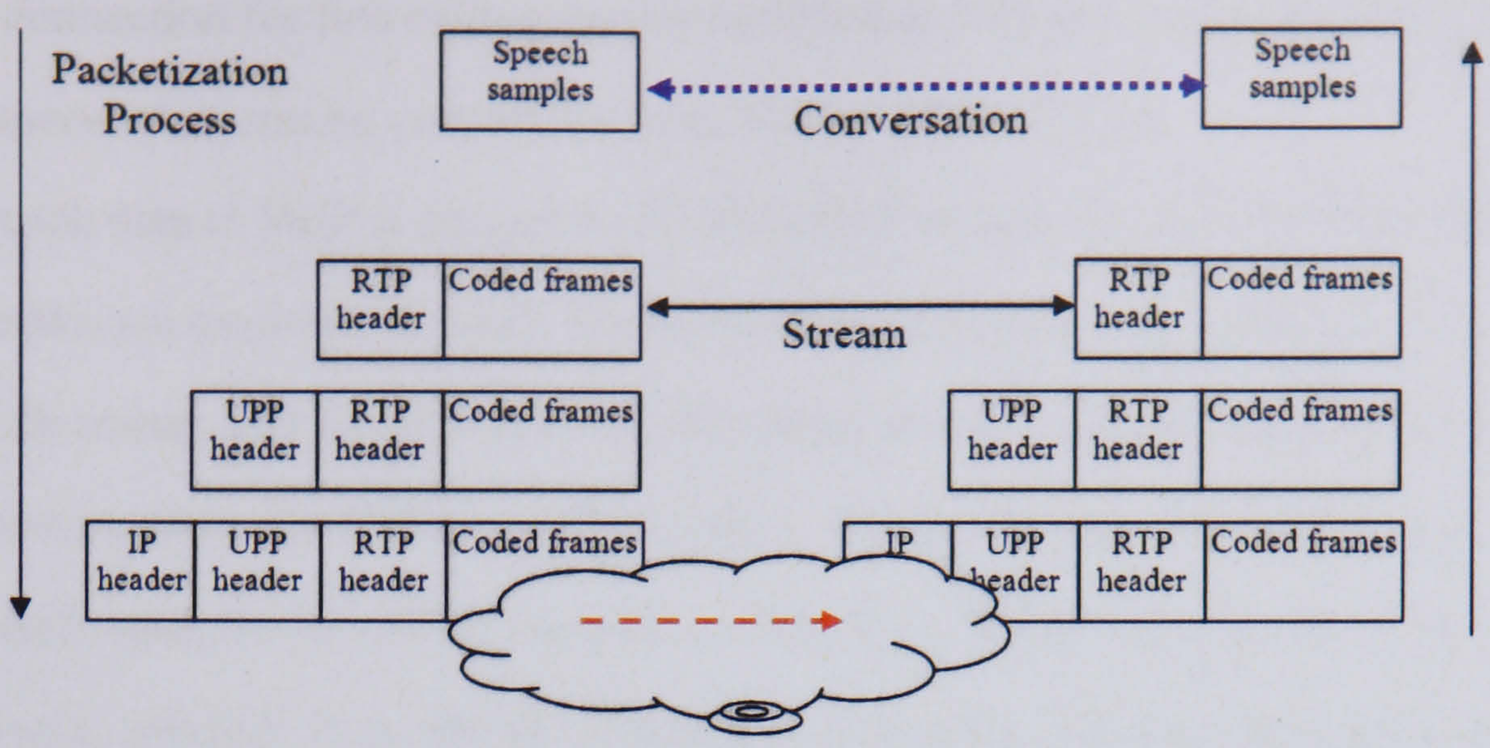


Figure 2.3: Packetization and transmission in IP network

Figure 2.3 shows the packetization process and the packet transmission in IP network. Speech samples in a voice call are coded and packetized in each layer and send to next layer. Then the IP network carries the IP packets to the other side of the call. The reverse depacketization process removes headers in each layer and reassembly the speech samples to voice. Only in RTP layer there is the concept of stream. Other lower layers using UDP and IP do not keep the status of the stream. The advantage of this structure is that the whole IP network is more efficient with less bandwidth usage and more flexibility on the route a packet can take. The disadvantage is that the packets are not guaranteed to be delivered on time, or even if been delivered at all [24].

VoIP Connectivity to PSTN or TDM network

VoIP have not fully replace the PSTN network yet so they need to co-exist for the time been. The connectivity between the VoIP system and the PSTN or TDM system becomes an important issue in the telecommunications industry.

The equipment to connect a VoIP call to a PSTN or TDM based device for instance a POTS land line phone or a mobile is a VoIP gateway. The VoIP gateway should be able to handle the

signalling connection for two calling parties in different network, and it also should be able to deal with speech data stream conversion from VoIP to PSTN/TDM.

The speech data in VoIP is carried by RTP/UDP/IP streams but in the PSTN/TDM domain, speech samples are transmitted inside physical circuit or logical circuit. The circuit is dedicated to the speech stream and synchronized but IP routes are not. The codec used in the VoIP call maybe different from the PSTN or TDM calls. To convert the unsynchronized IP packets carried speech samples to circuit samples in PSTN or TDM network, the gateway need to retrieve speech samples from the IP packet by performing the depacketization process and decoding process. Then the speech samples need to be reassembled into PCM or relevant format which can be carried by PSTN or TDM system. The depacketization process in the VoIP gateway is the same as a VoIP terminal, also involves the process to remove of headers from each packets and to decode them. There also need to be a jitter buffer to eliminate the delay jitter introduced by the packet switched IP network. The time sequence of the speech samples need to be rebuilt in the PSTN or TDM side of the gateway so the call can be connected to a PSTN or TDM terminal.

There are other possible speech quality distortions could be introduced by the gateway and the translation of VoIP to PSTN or TDM, including echo, noise, or gaps in speech sample due to packet loss in IP side and so on. These distortions can affect the end-to-end user perceived speech quality and need to be considered in a VoIP to PSTN or TDM end-to-end speech quality test.

2.1.4 Current Services and Products

Examples of current popular VoIP services and products are introduced in this section. A more detailed user perceived speech quality test on some of them is discussed in Chapter 4. The UMA wireless VoIP mobile introduced here is an important part of the live wireless mobile VoIP test platform.

SIP or H.323 VoIP Phones

SIP and H.323 are the most popular VoIP signalling system and there are many soft phones using these protocols to perform VoIP calls from computer. When software in the computer can make VoIP calls to another VoIP peer, it is called a soft phone. This kind of soft VoIP phone can be used to call another VoIP phone or a PSTN phone, if the VoIP provider supports it in the gateway to PSTN network. Those soft phones include Netmeeting [25], Twinkle [26], xLite [27], MSN messenger [28], GoogleTalk [29] and many more. There are some hardware WiFi based wireless phone products available on the market for instance Zyxel wireless phone [30] and so on.

VoIP Calling Card Service and VoIP Trunk

VoIP can also be used to carry long distance or international calls to reduce cost. When the VoIP applications like the soft phone or wireless VoIP mobile is used to call a PSTN land line, only the IP terminal to VoIP gateway part is IP based network. After the VoIP gateway converted IP packet carried speech frames to PSTN speech samples, they are carried by the circuit switched PSTN network to the PSTN terminal.

The traditional trunk service for instance T1 or OC3 are divided into virtual circuits for circuit switched network but they can be converted into VoIP trunks as well. That means the advantage of packet switching can be used in the long distance or international trunks as well. Some international calling card service is using this technology to save circuit rent.

VoIP trunk can also be used to link distance branch offices. It is more flexible and secure than the PSTN and TDM network. Many network equipment providers can provide solutions to this kind of IP trunk application.

New Development in IP Multimedia Subsystem (IMS) and Unlicensed Media Access (UMA) - Mobile VoIP

There are many new features in the ever evolving VoIP systems. To use VoIP in mobile handset is an important one not only because of the huge business potential but also because of the technical challenge involved in the convergence system. It is obvious that to introduce new features like VoIP into a mobile handset will attract more users. With the facility of VoIP, not only cheaper calls can be made when available, but also lots of other features like online messaging service can be build on top of it.

Most of the VoIP in mobile implementations are working towards the Fixed Mobile Convergence (FMC). IP Multimedia Subsystem (IMS) is a new architectural framework for delivering IP based multimedia feature to mobile users. The aim of the IMS is to help a fixed mobile convergence and bring the content higher mobility and accessibility [31].

Unlicensed Mobile Access (UMA) as shown in Figure 2.4 [32], is another form of fix mobile convergence. It allows seamless roaming and handover between Local Area Networks (LAN) and Wide Area Networks (WAN) using multi-mode mobiles. The mobile can switch between VoIP mode to GSM mode based on availability. It is normally an add-on function to a GSM mobile to allow the mobile to take advantage of wireless form of VoIP service in wireless access point coverage. The difference from IMS is that the UMA mobiles can only be one mode at one given time although it can switch between them.

The physical layer wireless connection method for UMA mobiles can be bluetooth or WiFi, which are widely used in home wireless access point devices. Those wireless access points are connected to IP network with broadband connection on the back. In the VoIP mode, the UMA mobile can connect to the wireless AP and use the broadband connection to send VoIP packets to a gateway in the IP network. The gateway then transfers these VoIP packets into TDM format and send to the other end of the call.

New technical challenges come with the wireless VoIP features. As described in Figure 2.4,

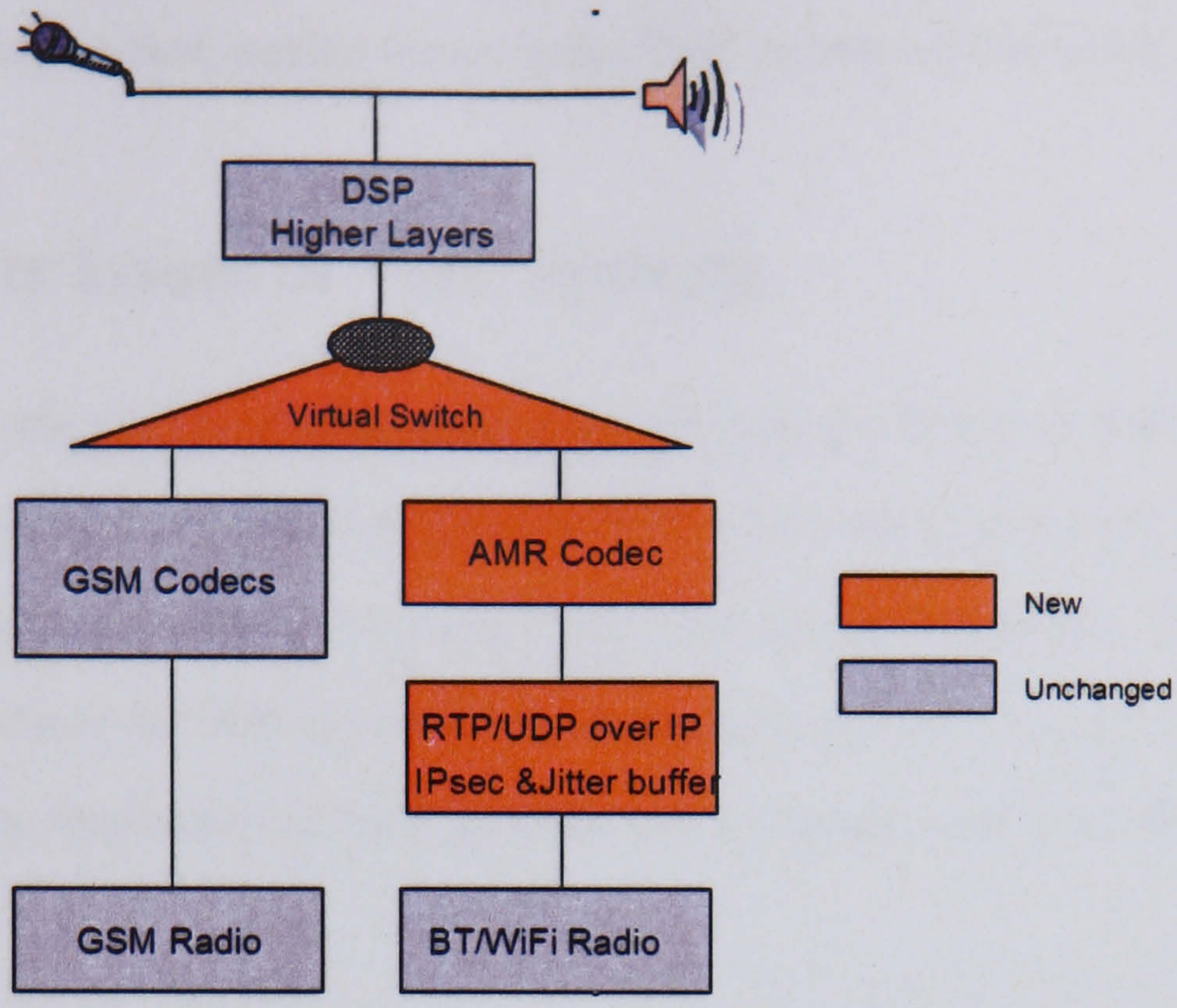


Figure 2.4: UMA VoIP modification to a normal GSM mobile handset

new function blocks need to be added into the existing mobile architecture. Those new function blocks include whole IP stack with UDP and RTP protocols, plus the optional IPsec protocol to enable security of the speech content.

As mentioned before, IP network is known as "Best Effort" so there is no quality guarantee as in TDM based networks, which becomes a challenge because an important design issue is that the user perceived speech quality in VoIP mode can not be lower than speech quality in the GSM mode, otherwise the user will complain about it and lead to business failure. Delay jitter introduced in the IP network and the mobile itself will affect speech quality so a jitter buffer function block is needed to improve the speech quality in the VoIP mode.

These VoIP protocol stack, jitter buffer and the virtual switch, which takes care of the handover and mobility issues, are just add-ons to the already resource-restrained mobile handset. The processing power of the mobiles is not as high as in the PC yet, which restricted the usage of lots of speech quality optimization methods. The memory space, internal communication bus and other resources are also limited in a mobile, when compared with a PC, where VoIP software can run as an application and enjoy much flexible resources.

In the following section, quality issues in the VoIP system are discussed.

2.2 Quality Issues in VoIP Systems

Coming with the benefit of bandwidth efficiency and flexibility of VoIP techniques is the QoS issues in the IP system. The QoS issues in VoIP system are different from traditional PSTN or TDM networks because the architecture of the VoIP system is different. There are challenge to speech quality from the VoIP network side for example priority management algorithms and challenge from the application or VoIP services side including codec selection and jitter buffer algorithm.

2.2.1 Issues in the IP Network

User perceived speech quality may not reach the same level as PSTN services. This is because in the PSTN network, speech samples are transmitted in dedicated circuits with no interference from other stream of speeches in other circuits. The speech sample are synchronized as well, not like in the IP network. As discussed in previous sections, VoIP service us IP packets to carrier speech frames via the packet switched IP network. To rebuild the speech samples' time sequence is not guaranteed as in the PSTN network, because IP packets can get delayed or lost in the packet switched IP network, where IP packets carrying one speech frame is sharing the IP route with other traffics and could be interrupted by others. The time of transmitting speech frames is different for each IP packet. The difference of delay is called delay jitter and it will introduce distortion to speech quality. The jitter buffer is introduced to reduce the impact of jitter.

When the IP packet for a speech stream is lost, concealment method need to be involved to fill the gaps generated in the speech stream. The filling speech samples are different from the original missing speech samples so the user perceived speech quality is impaired. If no concealment method introduced, the relative delay is moved and the user perceived speech

quality is impaired as well. The same situation happens if the delay jitter is introduced to speech samples.

Because the traffic condition in a packet switched IP network can change from time to time, the QoS of a voice call carried over IP network may change dramatically during a call. The changes in the IP network may include the packet loss rate change, the delay change, the delay jitter change, the bit error rate change and so on.

2.2.2 Issues in the VoIP Applications

The end-to-end user perceived speech quality is the ultimate measure for goodness of the speech quality in any communication system includes the VoIP system. In the central controlled PSTN or TDM network architecture, the terminal or end user have very limited access to the signalling and control plane of the system. Therefore the network introduced QoS impairment is almost as the same as end-to-end QoS to the user. Because the VoIP terminals or end users have more control of the IP based communication system, application layer elements of the VoIP service including codec rate and jitter buffer implementation have more importance to the over all end-to-end speech quality. The network QoS only contributes to a part of the end-to-end user perceived speech quality. It is necessary to consider end-to-end user perceived quality instead of network QoS. Figure 2.5 shows the concept of network QoS and the end-to-end user perceived QoS.

There are new techniques developed in the VoIP system to prevent the end-to-end speech quality degradation. Codec with redundant information like iLBC [33] and lost packet concealment (LPC) mechanism as in [34, 35] can help to reduce the impact of packet loss in the VoIP network. But their performance is different between different implementation or algorithm approaches.

The same situation happens with jitter buffer such as in [7, 36, 37], which is used to smooth out the delay difference i.e. jitter in the received speech samples. The performance of these new

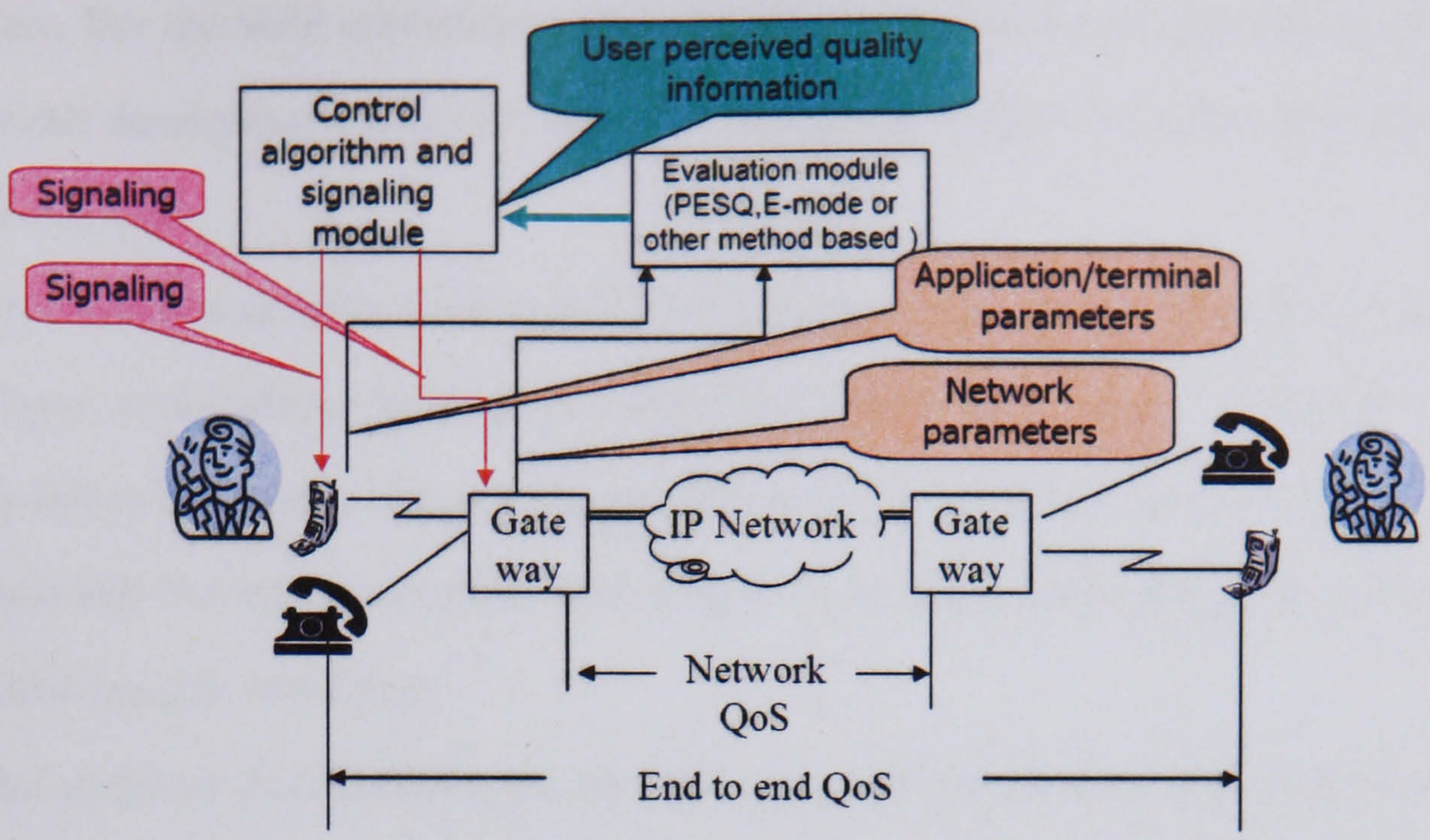


Figure 2.5: End-to-end user perceived speech quality in VoIP system

elements in the VoIP system is different and they are contributing to the overall performance of the VoIP system in terms of user perceived speech quality.

The sender side of the VoIP call may change codec or codec rate during a call so the adaptation to dynamic sender condition and IP network condition needs to be quick and efficient.

2.2.3 Specific QoS Issues in Wireless Mobile VoIP System

VoIP application can be used in wireless mobile environment as well a fixed land line environment. The wireless mobile VoIP as a part of UMA or IMS is a specify type of VoIP application. The mobile and wireless environment brings more specific wireless environment related challenges than ordinary VoIP services.

In a wireless mobile, part of the IP route is not provided by cable connection but wireless radio frequency link. Normally the wireless link is just one hop of the IP link from the mobile to the base station or an access point. Then the IP packets are transmitted through a fixed IP infrastructure as in a fixed VoIP system.

The most common QoS issues with the wireless link is the limited bandwidth and the higher

bit error rate. For the VoIP application, these wireless link introduced QoS issues are affecting the bandwidth throughput and packet loss rate, which could limit the codec rate and affect the retransmission time.

From the wireless mobile VoIP application's point of view, every IP packet is transmitted by the IP layer, which means lower physical and MAC layer mechanism is transparent to it. But the quality issues in wireless link are affecting the IP layer QoS parameters so these issues need to be address and the wireless mobile VoIP system needs to be optimized while considering the wireless links specific condition.

Another common performance consideration in a wireless mobile VoIP system is the VoIP application host's processing power limitation. Different from a typical soft VoIP phone application in a PC platform, the mobile VoIP application is located in an embedded mobile system. The processing power of the handhold device is much less than the PC platform. This processing power restriction limited the complexity of the VoIP application in the mobile platform. This brings another challenge to the design and optimization of VoIP services in the wireless mobile environment.

It is necessary to highlight these wireless mobile VoIP related QoS issues because QoS enhancement method developed in Chapter 4 are optimized for this environment.

In the next section, impairment factors and QoS related elements in VoIP services and products are introduced.

2.3 Elements and Impairment Factors in the VoIP System

Elements in the VoIP system and their impact to end-to-end user perceived quality are studied in this section. These elements include codecs, DSP features, IP network elements, security elements, jitter buffer, packet loss concealment element and so on. These elements can affect other elements' performance so the relationship and their overall impact to end-to-end user perceived speech quality is discussed later in this section.

2.3.1 Codecs and DSP Features

Codec is an important part in a telecommunication system. Analogue voice sample cannot be transferred directly through IP networks. The voice source (analogue signals) has to be coded (converted) into a digital format. On the receiving end, the digital format signal has to be decoded (reconverted) back into an analogue format in order to be intelligible to the human ear, as shown in Figure 2.6. Other DSP features work closely with the codec element to provided good voice.

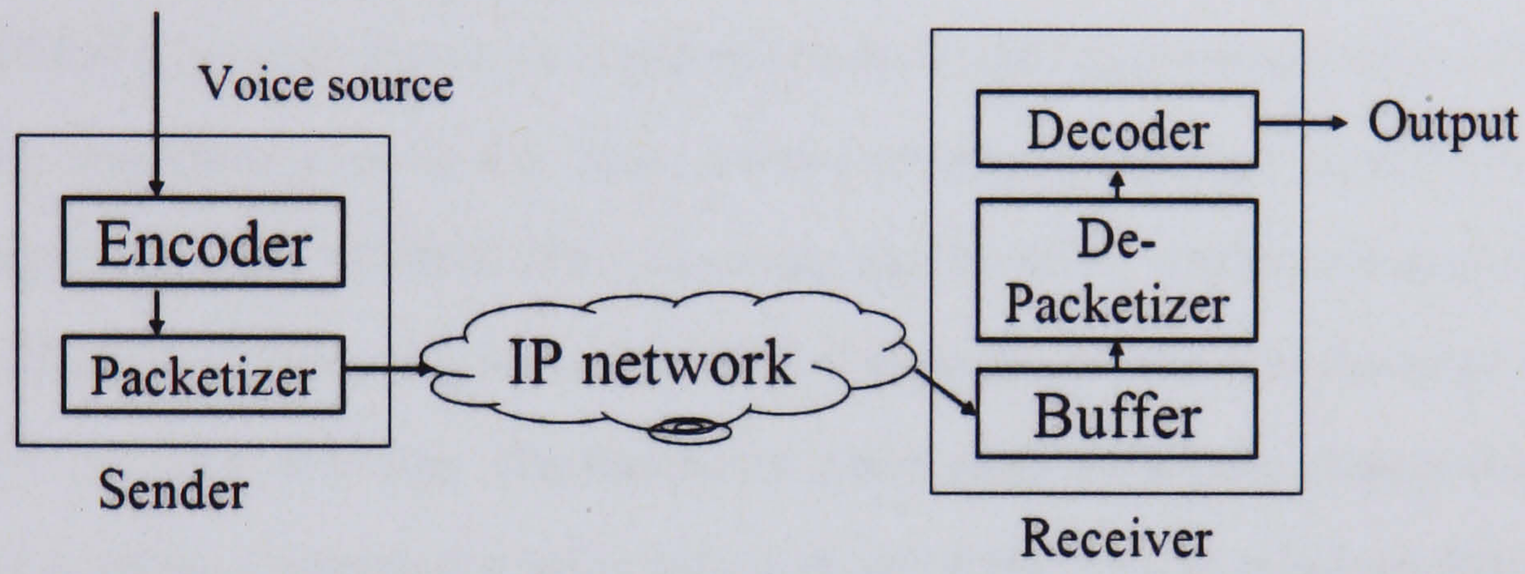


Figure 2.6: Codec and packetization process in the network

Codecs

This conversion process is done by a matching pair of process named COder and DECoder (CODEC). Traditional codec in PSTN system is PCM (Pulse Code Modulation), standardized by the ITU-T as G.711.

VoIP system has higher requirements for codecs because of the dynamic nature of IP networks. Several low bandwidth, efficient, robust and more network impairment tolerant codecs have been developed to meet the more critical requirement in VoIP system.

PCM is based on the Nyquist Theorem, developed by Harry Nyquist of Bell Telephone Laboratories in 1928 [38]. As a basic coding technique used in traditional PSTNs worldwide,

PCM requires bandwidth of 64 kbps. PCM specifies that the sampling process take place every 125 ms, which is exactly 1/8000th of a second. Each 16 bits sample is coded into an 8 bits word. It is a waveform-based codec.

ADPCM (Adaptive Differential Pulse Code Modulation) [39], a technique also used in some PSTN networks, offers voice coding at 40, 32, 24 and 16 kbps. At the most common implementation rate of 32 kbps, the compression rate is 2:1, which exactly halves the bandwidth required for voice. At 32 kbps, ADPCM yields voice quality at a 4.2 MOS (Mean Opinion Score), which is very close to that of PCM. At the higher compression rates of 24 kbps and 16 kbps, there is a corresponding drop in quality.

CS-ACELP (Conjugate Structure-Algebraic Code Excited Linear Prediction) [40] runs at 8 kbps with a compression rate of 8:1. There are two versions, which vary in terms of computational complexity, either of which offers raw voice quality that is similar to that of ADPCM at 32 kbps. CS-ACELP rates as high as a 4.2 MOS. Compression delay is in the range of 10 ms.

G.723.1 [41] runs at 8 kbps. The frame size it used is 20 ms and the compressed rates are 5.3 kbps or 6.3 kbps. Compression delay fot it is about 30 ms. Two bit rate have different MOS score. 6.3 kbps achieved MOS 3.9 and 5.3 kbps achieved MOS 3.5.

LD-CELP (Low Delay-Code Excited Linear Prediction) [42] runs at 16 kbps with a compression rate of 4:1. Compression delay is 3.0 ms - 5.0 ms, and raw voice quality is similar to that of ADPCM at 32 kbps.

AMR (Adaptive Multi Rate) [34] is developed by ETSI and has been standardized for GSM. It has been chosen by 3GPP as the mandatory codec. AMR is a multi-mode codec with eight modes (MR475 to MR122) with bit rates between 4.75 kbps to 12.2 kbps. Mode switching can occur at any time (frame-based).

Table 2.1 shows comparison of different codec's performance in terms of their average PESQ score. In the table MOS score is obtained by PESQ. MOS and PESQ is discussed in later sections.

Method	ITU-T Standard	Data Rate	MOS score
PCM	G.711	64 kbps	4.4
MPMLQ	G.723.1	3.9 kbps	3.9
ADPCM	G.726	32 kbps	4.2
LD-CELP	G.728	16 kbps	4.2
CS-ACELP	G.729	8 kbps	4.2
AMR	3GPP standard	12.2 – 4.75 kbps	4.1-3.1 *

(* MOS score vary with AMR rate)

Table 2.1: Parameters comparison for different codecs

The following subsection introduced some end point or codec related mechanisms that will affect voice quality.

Speech Quality Related Features

FEC (Forward Error Correction) mechanisms are open-loop mechanism, based on the transmission of redundant information along with original information. Previous research including [43] and [44] have concluded that FEC can provide good performance and generally reliable.

Transcoder is a converter or interpreter to translate one codec into another codec. It is introduced to link two or more different users with different codecs. Voice coding and transcoding algorithm causes loss in quality as the compression process losses some information and the voice signal cannot be perfectly reconstructed. The flexibility of VoIP end users brings higher importance to the transcoder element in the network because users could use different codecs in their end points.

A few other audio features such as Noise suppression and Echo Cancelation are discussed in later chapters.

2.3.2 Network Impairment

The user perceived quality of a VoIP system strongly depends on the network performance (packet loss rate, delay/jitter, amount of bandwidth available etc.). Although VoIP requires a

low bandwidth, it still needs a constant available bandwidth even for nonlinear codecs and low-bit-rate codecs. These requirements are hard to be guaranteed in the IP networks so it is hard to predict quality.

Several factors influence perceived QoS including network packet loss, network jitter or network delay, which are the main parameters determine IP network QoS as well.

Delay

Delay does not cause any reduction in voice quality but affects the interactive nature of conversations. Codecs, packetizer and network transmission all introduce delay. All these add up towards the overall transmission delay.

Delay also causes echo and talker overlap in two-way transmission. There are two types of echo in VoIP, first is the usual far-end echo caused by the 4-to-2 wire hybrid conversion. The user will hear her/his own voice reflected back from the remote central office or gateway's line-card hybrid. Second type of echo is called acoustic feedback echo, occurs when free-air microphone and speakers are used, as is the case for most PC endpoint. It happens when the remote user's voice signal produced by the speakers is picked up by the microphone and echoed back to the remote user. Talker overlap is caused by one caller stepping on the other talker's speech. Talker overlap becomes significant if the one-way delay more than 250 ms. To avoid such impairment, the end-to-end delay need to be controlled under certain limit, which is different for different environment.

Packet Loss

Packet loss is a common problem in IP networks. It could happen in many cases such as router overflow, which caused by queue full or almost full in the router, and bottleneck links experiencing congestion. It is a major impairment factor because if voice packets are lost, or can not be played due to corruption or excessively delayed, gaps will occur in the compressed voice stream. A corrupted speech frame may bring unacceptable noise if forced to be played

because the corrupted bits in the payload frame may be interpreted into noise. Packet loss causes more noticeable degradation in voice quality than other network impairments. There is several packet loss concealment strategies used to fill in the gap, such as to insert comfort noise or silence into the gap. Other packet loss compensations include local repair (interpolation of missing data using previous or later packets), or interleaving. To reduce the impact of packet loss, FEC can be used by sending redundant information in packets.

Jitter

Jitter (delay variation), is the variation in inter-packet arrival time as introduced by the variable transmission delay over the network. The transmission time of packets through an IP network will vary due to queuing effects. Removing jitter requires collecting packets and holding them long enough to allow the slowest packets to arrive in time to be played and re-sequence if necessary, which causes additional delay. Jitter in network does not necessary introduce information loss but jitter will still cause quality degradation.

Jitter buffer adds extra delay to incoming packets so that variations in transmission time can be removed. By increasing the size of the jitter buffer, more network jitter will be compensated, but this will also introduce more delay. On the opposite side, by decreasing the size of the jitter buffer, more packet will be lost due to late arrival. To find the right balance between delay and loss is the main goal of a jitter buffer.

2.3.3 Other Impairment Factors

Other than application related codec and audio features impairment and network related impairments, there are a few more other impairments need to be considered, which include the security elements, voice activity detection, play out buffer, packet loss concealment mechanism.

Security Element

Security element play an important role in the VoIP system [45]. One of the important reason users convert from traditional circuit switch based telecommunication system to IP based system is that security features such as user authentication and encryption on speech data can be implemented with more flexibility. The security elements are normally located either on the VoIP application side or on the entrance point to the core network. Both these two places have some resource restrains. In the VoIP application side, authentication or encryption may uses extra processing power and bring extra delay to the call setup time or end-to-end conversational delay. In the entering point to the core network, encryption process may cause higher jitter and congestions due to heavy demand of processing power.

Voice Activity Detector

VAD (Voice Activity Detector) is a component of a voice gateway or terminal that suppresses packet transmission when voice signals are not present (silence period). There is no or very litter signal encoded during silence period thus a save of bandwidth is possible. This element sometimes is implemented within the codec. It can affect the QoS in several ways. If the threshold of silence or active speech is not set correctly, it could cause clipping to active speech. The switching between silence and active speech need to be very accurate, otherwise the speech could be impaired.

Play Out Buffer

Play out buffer is an important QoS element in the receiver end. It can be used to minimize the effect of jitter and sort frames into sequence if necessary. Typical play out buffer stores the arriving packets until a later play out time in order to ensure there are enough packets buffered to be played out continuously and any packet arrival later this time will be simply discarded. The play out buffer size depends on the network delay. It is important to avoid unnecessary

delay and dropping of packets. A fixed buffer size normally causes a constant delay, but no end-to-end delay jitter. More details about jitter buffer is discussed in Chapter 4.

Packet Loss Concealment

Packet Loss Concealment (PLC) is a QoS mechanism to reduce the effects of packet losses. Most modern codecs such as G.729, G.723.1 and AMR have build-in PLC mechanism in the decoder. External PLC mechanism are also possible. PLC function can produce a replacement of loss packets by inserting comfort noise, silence or waveform substitution. It is also possible to reuse part of the earlier packets or insert replacement speech generated by complicated models, depends on the PLC strategies used.

2.3.4 Relationship Between the Impairment Factors

Figure 2.7 shows inter-relationship between these factors.

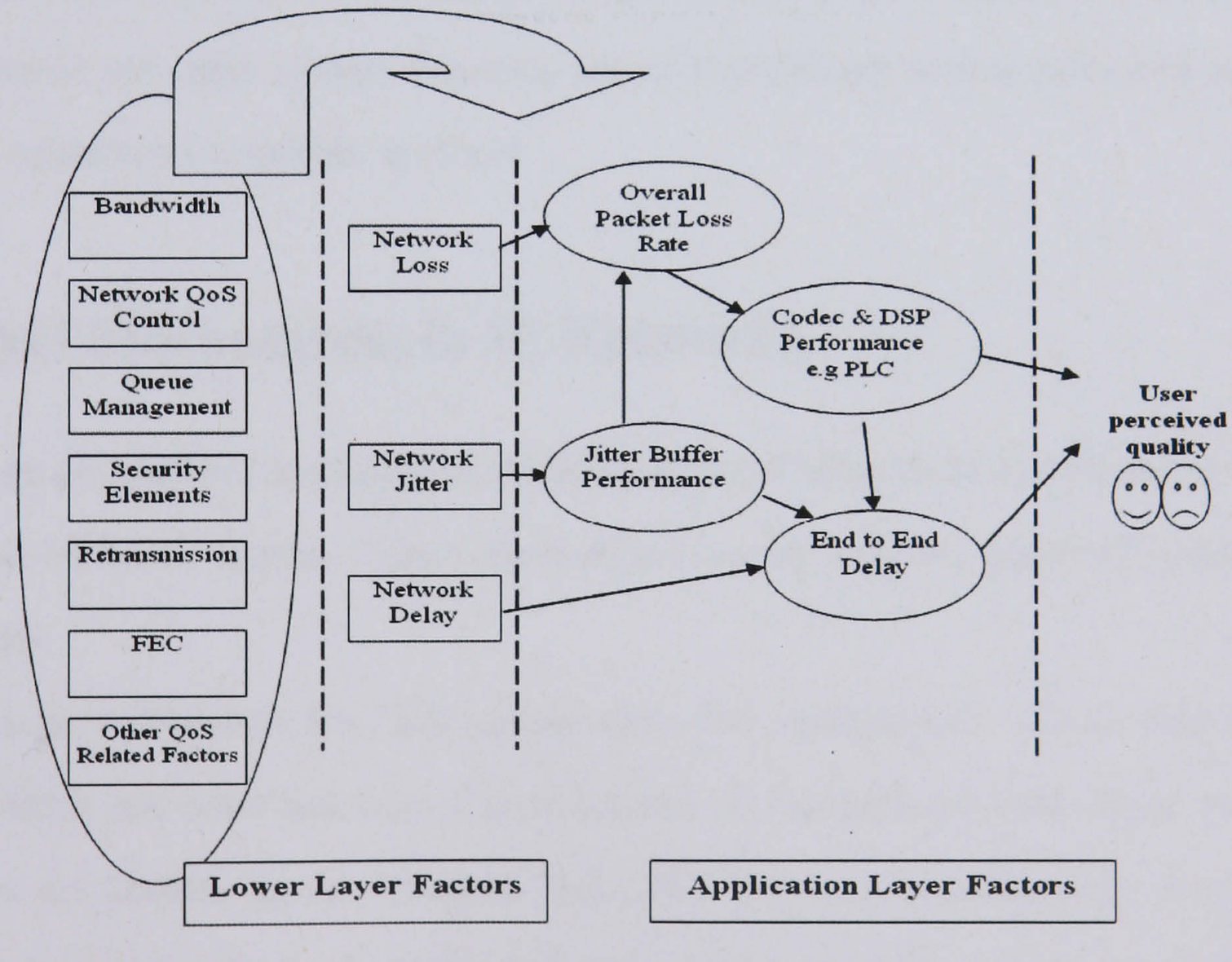


Figure 2.7: Relationship between impairment factors

As the figure shows, the user perceived quality is the ultimate judgement of the goodness of the telecommunication system, in our case, the end-to-end VoIP system.

Codec and DSP output provides the content (WHAT) of the speech to the user and end-to-end delay represents it on time (WHEN). They are the two main quality factors of the speech quality.

The performance of codec/DSP are linked to end-to-end delay because they can trade content quality with latency. The overall Packet Loss Rate and Jitter Buffer Performance contribute to the Codec/DSP performance and end-to-end Delay as well. The jitter buffer can trade packet loss with delay. To find the best balance between loss and delay is the main aim of the jitter buffer element.

The three main IP network quality factors are the network loss, the network jitter and network delay. They contribute to overall loss, jitter buffer performance and the end-to-end delay.

Other related QoS factors, including bandwidth limitation, network QoS control, queue management algorithms, security elements, retransmission of error frames or packets, forward error correction and other related elements, are all contributing to user perceived quality and need to be optimized for specific scenario.

2.4 QoS Mechanisms in IP Networks

There are several QoS mechanisms available in the IP network to control and optimize the overall QoS of the IP system. These mechanisms can be used for perceived speech quality enhancement.

IP network architecture has QoS requirements and mechanisms. In the data link layer, there are MPLS and other link related QoS method. In the network layer, there are DiffServ and IntServ mechanism. In the Transport layer, TCP has its own mechanism to control flow, bandwidth, throughput etc. In the application layer, every application has their application layer method to measure and control QoS. A few such mechanisms are introduced in this section.

Best Effort Nature of the IP System

IP protocol is natively a "Best Effort" system. It is coming with the fundamental advantage of multi-pathing, multiplexing, flexibility and bandwidth efficiency. Applications on top of it have their own mechanism like retransmission and forward error check to achieve better quality. This situation is satisfactory until the raise of time critical applications including VoIP. More systematic improvement need to be considered to achieve the requirements for real time applications. Those methods are introduced in the following subsections.

Integrated Services (IntServ)

The Internet Engineering Task Force (IETF) proposed a model called Integrated Services (IntServ) in 1994 [46]. This model proposed the classification of applications depends on their sensitivity to delay. If network resources are reserved for time critical applications, a certain level of quality (in terms of delay, jitter and packet loss) can be guaranteed by those reserved resources.

In IntServ architecture, flow represents stream of packets with common Source Address, Destination Address and port number. IntServ mechanism requires router to maintain state information on each flow; router determines what flows get what resources based on available capacity. The architecture has some limitations such as the reservation path need to be establish all the way through the route and any changes to the route change would trigger a difficult dynamic resource update.

IntServ components include Traffic classes, Traffic control, setup protocol and reservation protocol (RSVP) [22] etc. The traffic classes includes: best effort, controlled load ("best-effort like" without congestion), guaranteed service (real-time with delay bounds). Traffic control includes: admission control, packet classifier, packet scheduler etc. RSVP is used to setup, reserve and release bandwidth allocation for flows. RSVP is a good approach to solve the dynamic resource update problem but it has some limitations as well. For example it needs to

make sure that each application applies for a suitable class, otherwise if everyone is using the highest class, the whole system would end up with another "Best Effort" situation.

IntServ's limitations are lack of policy control mechanisms and lack of large scale flexibility. Some external mechanism such as class related charging could be used to solve some of those limitations but there are out of scope of this thesis.

Differentiated Service (DiffServ)

Compare with IntServ maintaining individual flows on all routers, DiffServ [47] flows are aggregated into an aggregate flow that receives "treatment" (per class or per service state). In DiffServ, service classes are identified, packet is marked as belonging to a particular service or class, sent on its way and routers in path examine header to determine treatment.

The basic functions in DiffServ architecture are: admission control, which is the ability of network to refuse customers when demand exceeds capacity, packet scheduling which is the method for treating different customers' data differently as needed, traffic classification which is the ability to sort streams into "substreams" that receive different treatments and policies/rules for allocating the network's resources including bandwidth.

DiffServ have several proposed service levels. For example Premium service is a "virtual leased line" level low delay, loss and jitter service, Assured service emulates a lightly loaded network or "drop me last", and default level support only usual "best effort" service.

The problems in DiffServ are: edge router is lack of information to mark the packets with correct service level; fixed policy is not working well in a dynamic mixed application environment.

Multi Protocol Label Switching (MPLS)

MPLS works in the data link layer, lower than the IP protocol's layer i.e. the network layer, over the physical layer. The IP network is a connection-less network and each router searches for the next hop of each received packet based on its destination address and forwards

the packet. The search to find the destination IP address in the packet header means all bytes before the destination IP address need to be read before the forwarding process can begin. This process makes it difficult to perform high-speed packet forwarding.

The Multi-Protocol Label Switching (MPLS) system uses a frame format having the fixed-length label wrapped on to each packet, and it forwards packets based on their label values. As the destination is determined by fixed-length label search instead of variable length as in original IP packet, the MPLS can perform high-speed packet forwarding. The packet-forwarding path determined by this label is called the Label Switched Path (LSP). The MPLS controls the LSP explicitly and it can calculate an optimum edge-to-edge path based on the QoS required for the traffic. Also, it can provide traffic engineering with load distribution of each path within the network.

By selecting and utilizing the above network QoS mechanisms, IP network can provide better QoS and an enhanced user perceived speech quality in the VoIP system running on top of the IP network can be achieved.

Above QoS mechanisms are more focused on network parameters such as packet loss, delay/jitter, priority, bandwidth, throughput but less aware of end user's perception.

It is important to utilizing them together with end user controllable mechanisms including codec rate, jitter buffer algorithm to achieve an enhanced overall end-to-end perceived quality. The methods of such enhancement method are discussed in later chapters.

2.5 User Perceived QoS Measurement Methods

To measure the goodness of a telecommunication system, there are many measurement methods. This thesis interests in the speech quality measurement area because accurate and reliable measurement of speech quality is the basis for further speech quality enhancement. The focus of this research is mainly on end-to-end speech quality on live telecommunication network, especially on wireless mobile VoIP environment. The measurement aspect is on live

system evaluation and performance evaluation. It is necessary to compare different speech quality measurement method and choose the suitable one to serve the aspect of this research.

2.5.1 Subjective Speech Quality Measurement

Subjective speech quality measurement is to use human testers to subjectively measure the speech quality. There are a few methods developed includes Mean Opinion Score (MOS), which uses Absolute Category Rating (ACR) and Degradation Mean Opinion Score (DMOS), which uses Degradation Category Rating (DCR).

MOS and ACR

When assessing the quality of speech coding and transmission systems, the subjective quality measurement plays an important role and works as benchmark for evaluating objective measures. ITU P.800 [5] describes several methods and procedures for conducting subjective evaluations of transmission quality [48]. The most commonly used method is Absolute Category Rating (ACR) test giving Mean Opinion Score (MOS).

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 voting on the quality uses an opinion scale such as the following Table 2.2.

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

Table 2.2: Subjective speech quality measurement categories

DMOS and DCR

Degradation Category Rating (DCR) is also used in some occasions [49], which gives Degradation Mean Opinion Score(DMOS) [2]. 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 or so called reference speech. The rating scales or the degradation levels are shown in the following Table 2.3.

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

Table 2.3: Degradation quality categories for DCR

The average of the opinion scores of subjects in DCR is called Degradation Mean Opinion Score (DMOS) [2]. There is a strict requirement for preparation of test speech sentences and procedures of the whole subject listening test. It makes subjective test more time-consuming, costly and stringent.

Other Subjective Methods

There are some other subjective methods developed for different specific measurement requirements. Continuous scaled MOS tests [50] were developed to specifically measure continuous multimedia services. In this test, a slider was used for each signal to indicate subject's

opinion of the voice quality. Similarly, a Quality Assessment Slide (QUASS) [51] was used to continuously rate perceived quality along a specified dimension for audio-visual applications.

2.5.2 Intrusive Objective Voice Quality Measurement

Intrusive objective speech quality measurement systems, as shown in Figure 2.8, 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 to utilize the network. They are more accurate to measure end-to-end perceived speech quality and are not very suitable for monitoring live traffic if used frequently because the injected traffic will affect other live traffic.

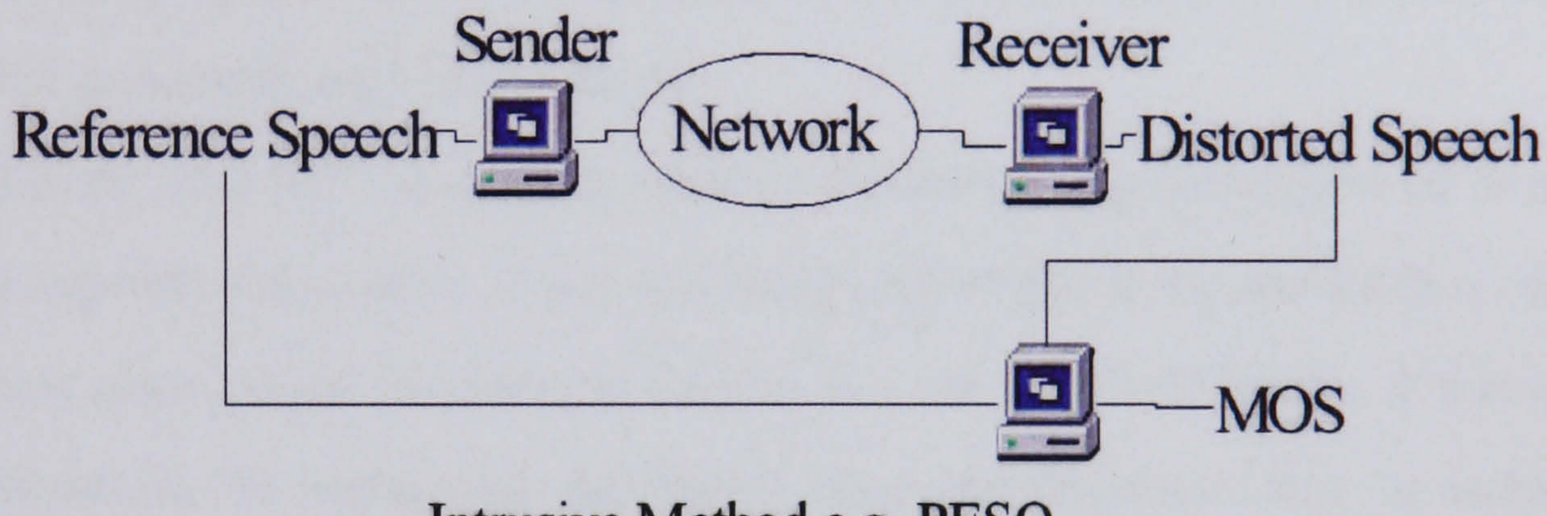


Figure 2.8: Intrusive speech quality measurement method

There are three groups of objective speech quality measurement methods developed [52].

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) [2] 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.

The third group of measurement method is perceptual domain measurement. These group of measurement methods are based on models of human auditory perception. They have been shown to be the most successful objective speech quality measures so far. These measurement methods transform speech signal into a perceptually relevant domain such as bark spectrum or loudness domain, and incorporate human auditory models [52].

The typical perceptual measure methods are Perceptual Speech Quality Measure (PSQM) [53, 54], Perceptual Analysis Measurement System (PAMS) [55,56], Measuring Normalizing Blocks (MNB) and Enhanced Modified Bark Spectral Distortion (EMBSD) [49, 57–59].

The Perceptual Evaluation of Speech Quality (PESQ) [4, 60] method is the latest ITU standard for measure speech quality for communication systems and networks and it has been widely used in research and industry fields.

PESQ is the new 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. It is the most popular method to measure end-to-end voice quality and still improving.

PESQ-LQ (PESQ Listening Quality) [61] is later invented to compensate the impact of high packet loss rate which is out of the designer's consideration when PESQ was first invented. And because 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 PESQ score to a ITU-T P.800 standard defined ACR listening quality average MOS scale, in the range of 1 to 4.5. PESQ-LQ is defined as follows, where x is the ITU-T P.862 standard defined PESQ score and y is the corresponding PESQ-LQ score:

$$y = \begin{cases} 1.0 & \text{for } x \leq 1.7 \\ -0.157268x^3 + 1.386609x^2 - 2.504699x + 2.023345 & \text{for } x > 1.7 \end{cases} \quad (2.1)$$

The convertion curve from P.862 to P.862.1 is shown in Figure 2.9 [52].

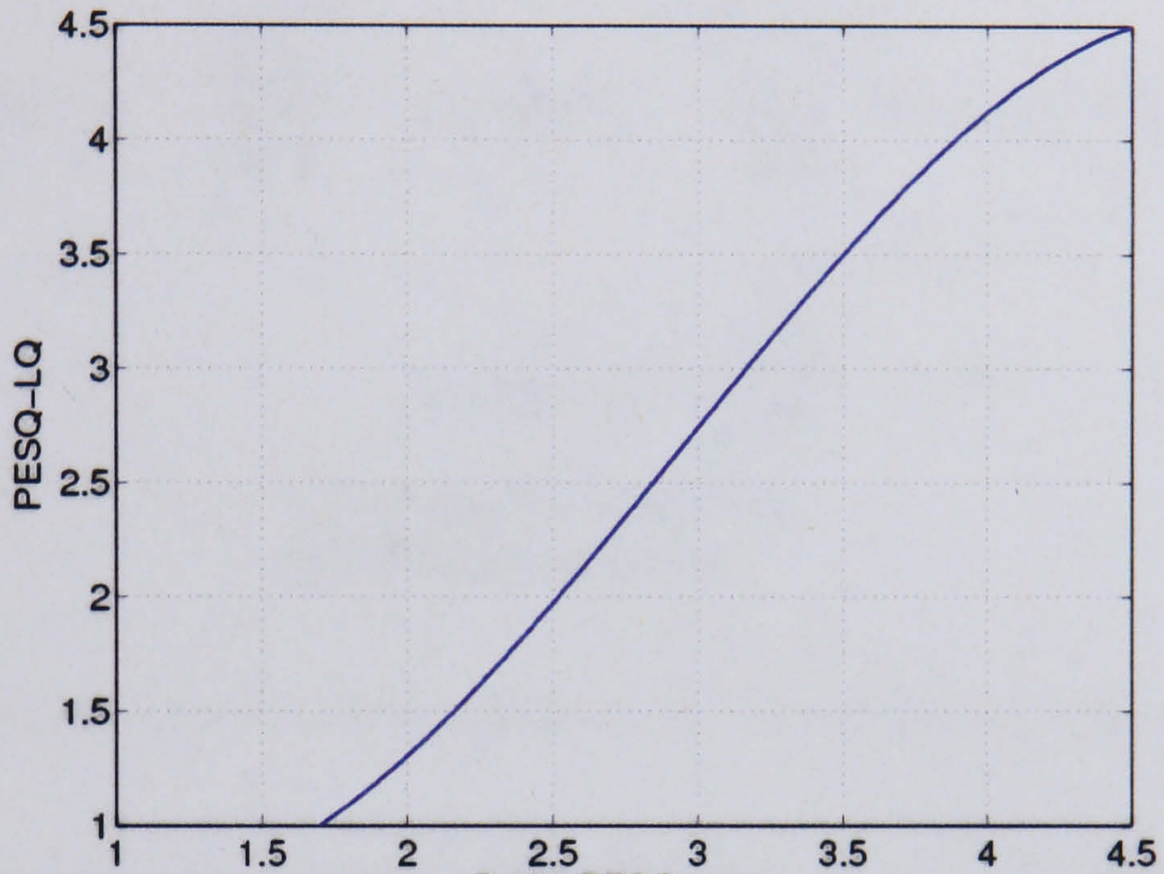


Figure 2.9: Convert from PESQ score to PESQ-LQ

PESQ is the most important tool uses in the evaluation of end-to-end speech quality in this project so the details of PESQ are discussed in a later chapter, where the accuracy of PESQ in a VoIP environment is evaluated and discussed.

2.5.3 Non-intrusive Objective Voice Quality Measurement

Intrusive methods inject a reference or original test signal into the tested system and compare the degraded signal with the reference. Thus the live traffic has to be interrupted during the test. The non-intrusive speech quality measurement or prediction techniques, as shown in Figure 2.10, aims for monitoring live traffic and no reference signal is needed.

There are two general types of non-intrusive speech quality prediction system.

One is to predict speech quality directly from varying IP network impairment parameters (for instance packet loss, jitter and delay) and non-IP network parameters (for instance codec, echo, language and/or talker issues). Typical methods are E-model based and artificial neural network (ANN) prediction method, which are presented in the following section.

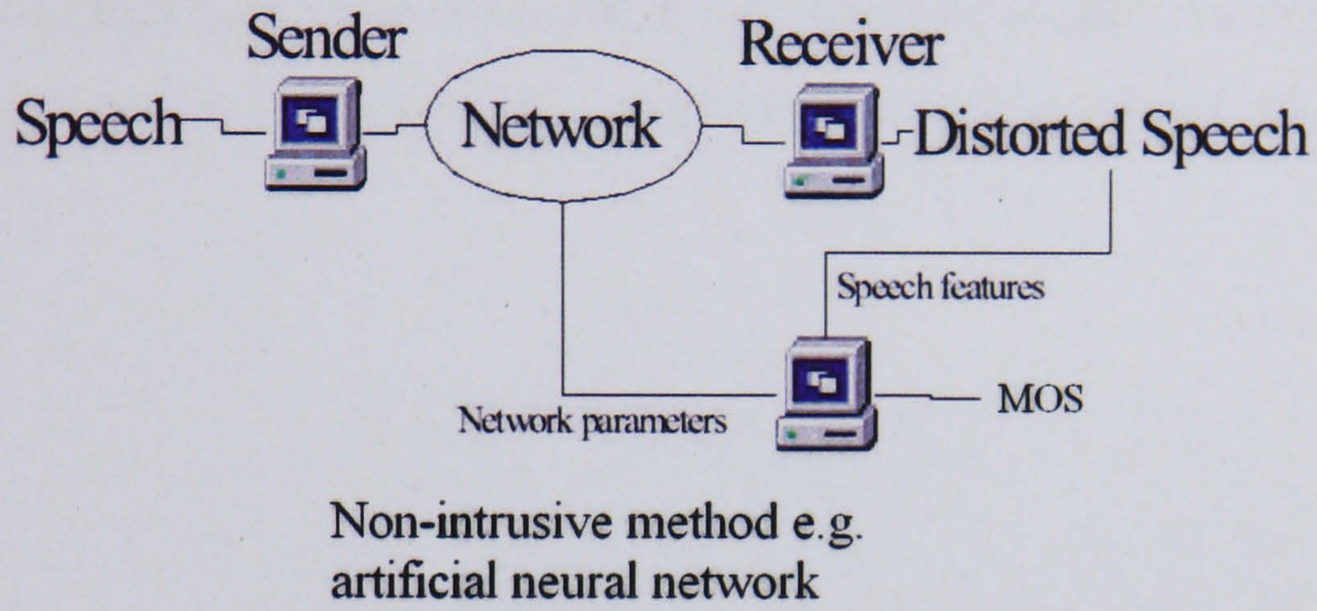


Figure 2.10: Intrusive speech quality measurement method

Another type of non-intrusive speech prediction system is to predict speech quality directly from degraded speech signal using signal-processing methods, such as INMD/CCI [62, 63], which is currently used for PSTN networks. There is another non-intrusive method called speech recognition performance as an effective perceived quality predictor, which is also a non-intrusive measurement method [64]. Their applications in VoIP networks are still not very clear at this moment and the topic is beyond the focus of this study. Some combination of the above methods are also developed.

E-model Based Non-intrusive Objective QoS Measurement Method

The E-model is developed by a working group within ETSI during the work on ETSI Technical Report ETR 250 [65]. It is a computational tool originally developed for network planning [2, 66, 67], but it is able to be used as a non-intrusive voice quality prediction tool for

VoIP applications [68–70]. It is also been used to provide speech quality indications for QoS enhancement mechanisms [71].

The E-model is based on the concept developed by J. Alnatt [72] that "Psychological factors on the psychological scale are additive". ITU G.107 [67] defines the transmission Rating factor (R). On the transmission rating scale of speech quality, i.e. R [73] scale, all the impairments are additive (by assumption of E-model) and 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:

$$R = R_0 - I_s - I_d - I_e + A \tag{2.2}$$

Where

R_0: Basic signal-to-noise ratio, including noise sources such as circuit noise and room noise.

I_s: Impairments that occur simultaneously with speech (including quantization noise, received speech level and sidetone level)

I_d: Impairments that are delayed with respect to speech (including talker/listener echo and absolute delay)

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

A: Advantage factor or expectation factor (for instance 0 for wireline and 10 for GSM)

ITU G.109 [74] defines the speech quality classes with the Rating (R), as shown in Figure 2.11. A rating below 50 indicates the speech quality is unacceptable to users.

MOS score can be converted from R value by using the equation 2.3 in ITU G.107 [67].

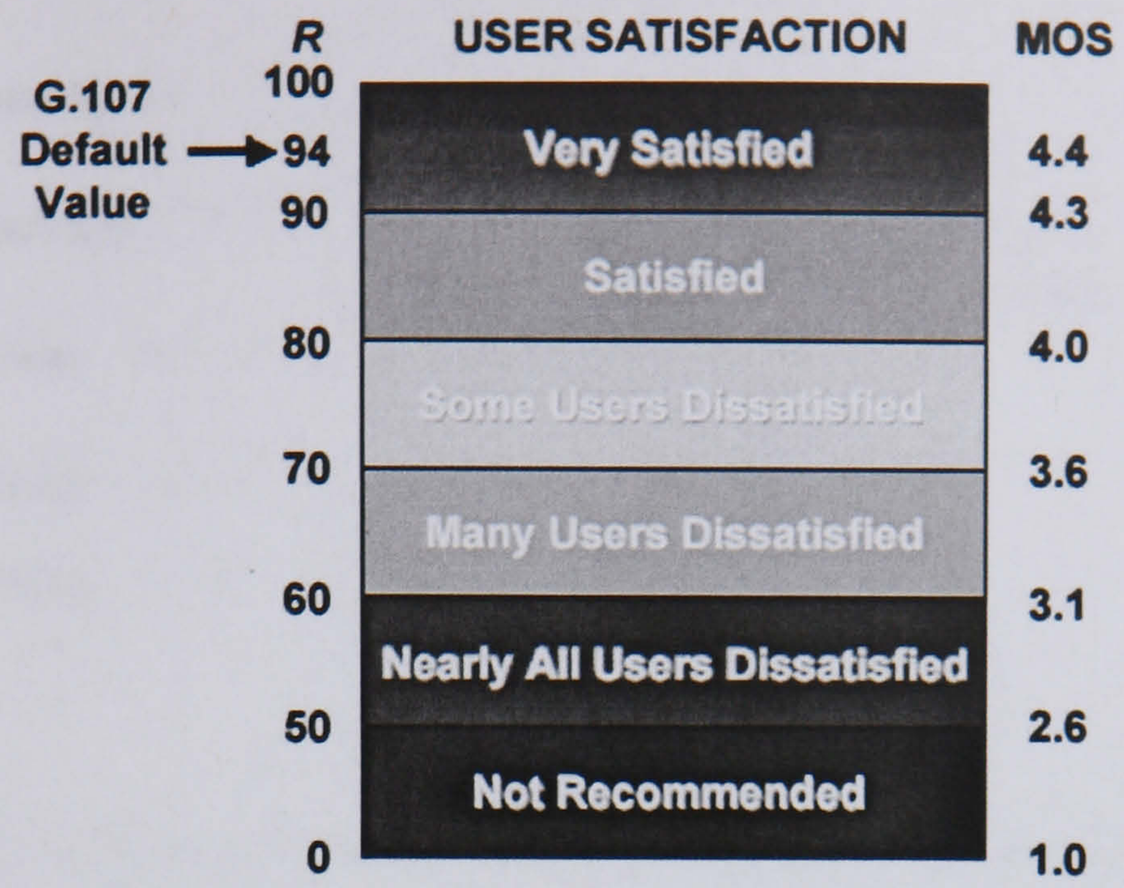


Figure 2.11: Speech quality classes according to E-model

$$MOS = \begin{cases} 1 & \text{for } R \leq 0 \\ 1 + 0.035R + R(R-60)(100-R) * 7 * 10^{-6} & \text{for } 0 < R < 100 \\ 4.5 & \text{for } R \geq 100 \end{cases} \tag{2.3}$$

It is possible to convert MOS into R value as well if necessary.

ANN Based Non-intrusive Objective QoS Measurement Method

ANN (Artificial Neural Networks), is widely used in solving engineering problems such as speech and image recognition, adaptive control, robotics and estimation. ANN has been successfully used in objective speech quality measurement as well, such as in Perceptual Evaluation of Audio Quality (PEAQ) [75]. Unlike other non-intrusive measurements such static, computational method of E-model, ANN model can adapt to the dynamic environment of IP networks, because of its ability to learn. Figure 2.12 shows the basic structure of the artificial neural network based non-intrusive objective speech quality measurement method.

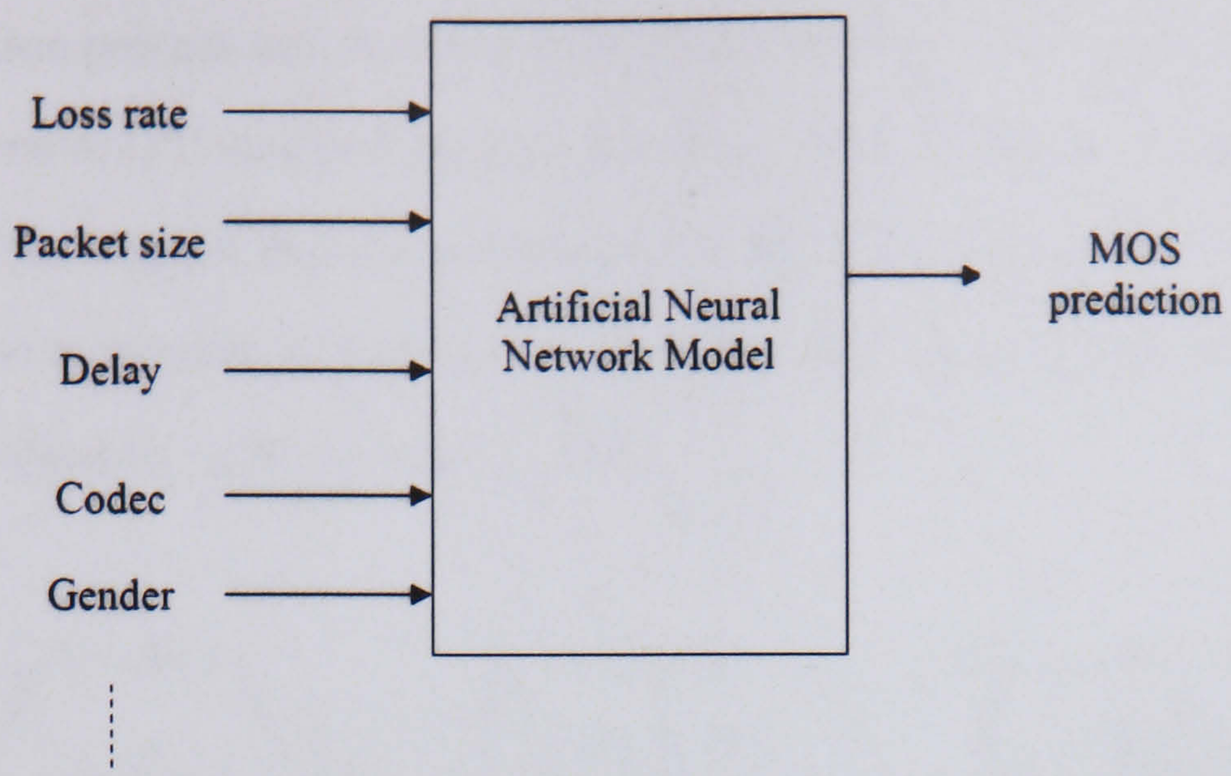


Figure 2.12: ANN based non-intrusive objective QoS measurement method

The input parameters to ANN model can be packet loss, delay, codec type, echo noise and so on. A three-layer feed-forward ANN is used to create a nonlinear model that associates the network impairment factors with MOS [52].

2.5.4 Some Mixed Usage of Above Methods

Mixed use of above method can bring benefit of each method and reduce their weak points.

Local Simulation Based Non-intrusive Method

A method has been developed to use intrusive voice quality measurement in a non-intrusive way. There is an initial idea about how to combine the advantage of intrusive measurement method and the advantage of passive prediction method developed [76].

The idea is shown in the Figure 2.13. Packets are passively captured from the network, then stamped with time stamp and filtered to individual RTP stream. The filer can be set in front of the capture or after. Then the RTP payloads of the packets are replaced with local generated payload. Those payloads are generated locally from sample voice signal, encoded and packetized exactly as the format of the original RTP payload.

The payload replaced RTP packets with time stamps are sending to a simulate decoding

and depacketization process and decoded voice signal retrieve. The original local voice signal sample is then send to ITU standard intrusive objective voice quality measurement method (i.e. P.862 PESQ) to be compare with the retrieved decoded voice signal. The result from the ITU intrusive objective measurement method will give a objective quality level which is easily to be transformed to subjective quality score i.e. MOS.

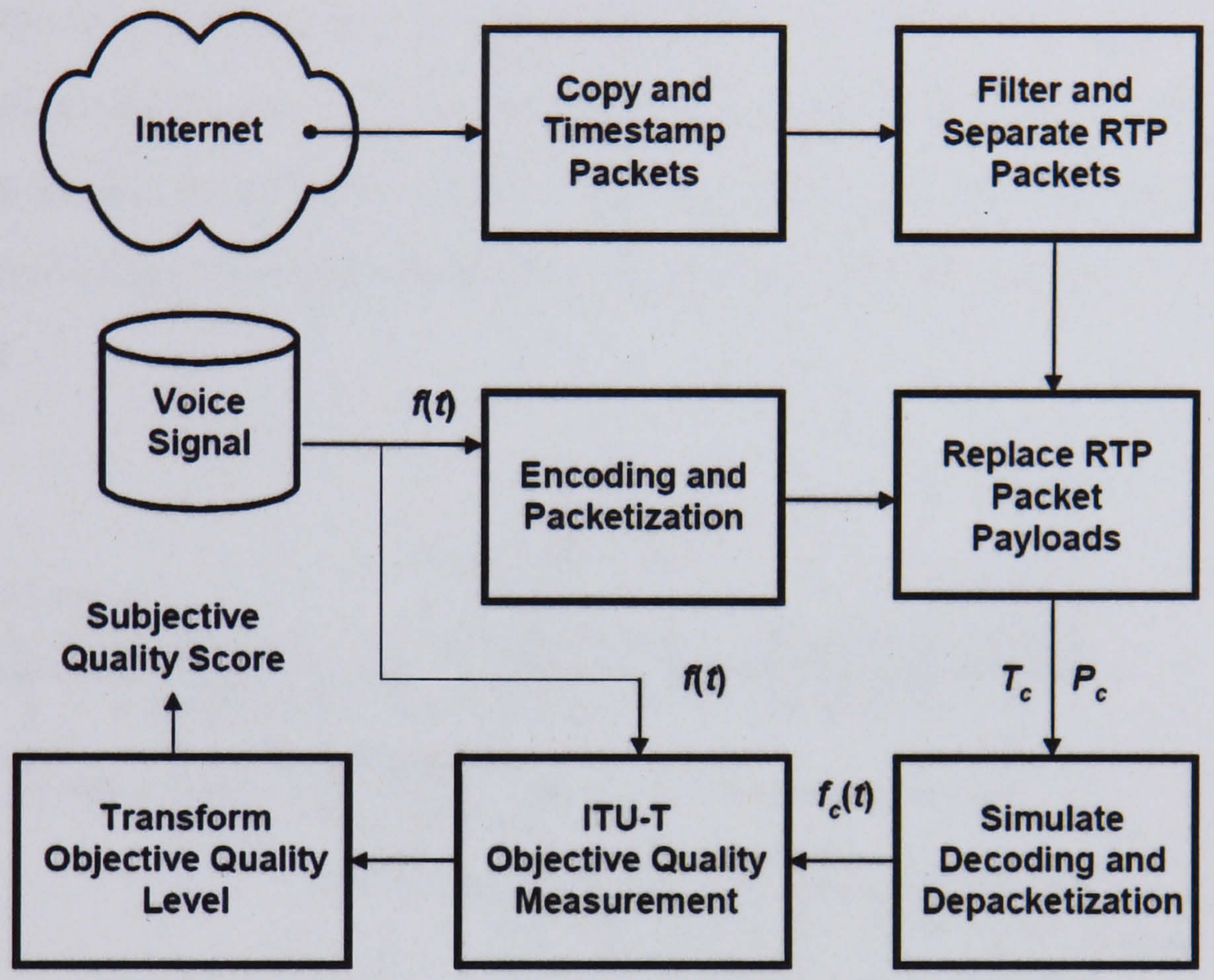


Figure 2.13: Local simulation based measurement

This method used the advantage of ITU standard intrusive measurement method, which is not network or codec related and purely voice comparison and combined with the advantage of non-intrusive measurement method, which is convenient and passive to the traffic.

There are some disadvantage for this method as well. The local voice signal is not as complicated as the real conversation and it will not suitable for all conversation scenarios and this will cause the inaccurate of the measurement. The ITU standard objective quality measurement method requires a high calculation power, which is hardly achievable from a multiplexed

network node or voice gateway.

Combine PESQ and E-model

A method has been developed [77] to combine PESQ and E-Model and thus get the advantage of conversational non-intrusive speech quality prediction. The main aim of this method is to get the measurement and prediction for conversational QoS from a non-intrusive measurement way. The idea combined the strongpoint of PESQ, which is suitable for a one-way speech quality measurement because it will not compare the time delay difference between two input speech samples and the strongpoint of ITU E-model, which is suitable to measure time delay effect to perceived speech quality. The detail of the method from [77] is shown in the following Figure 2.14.

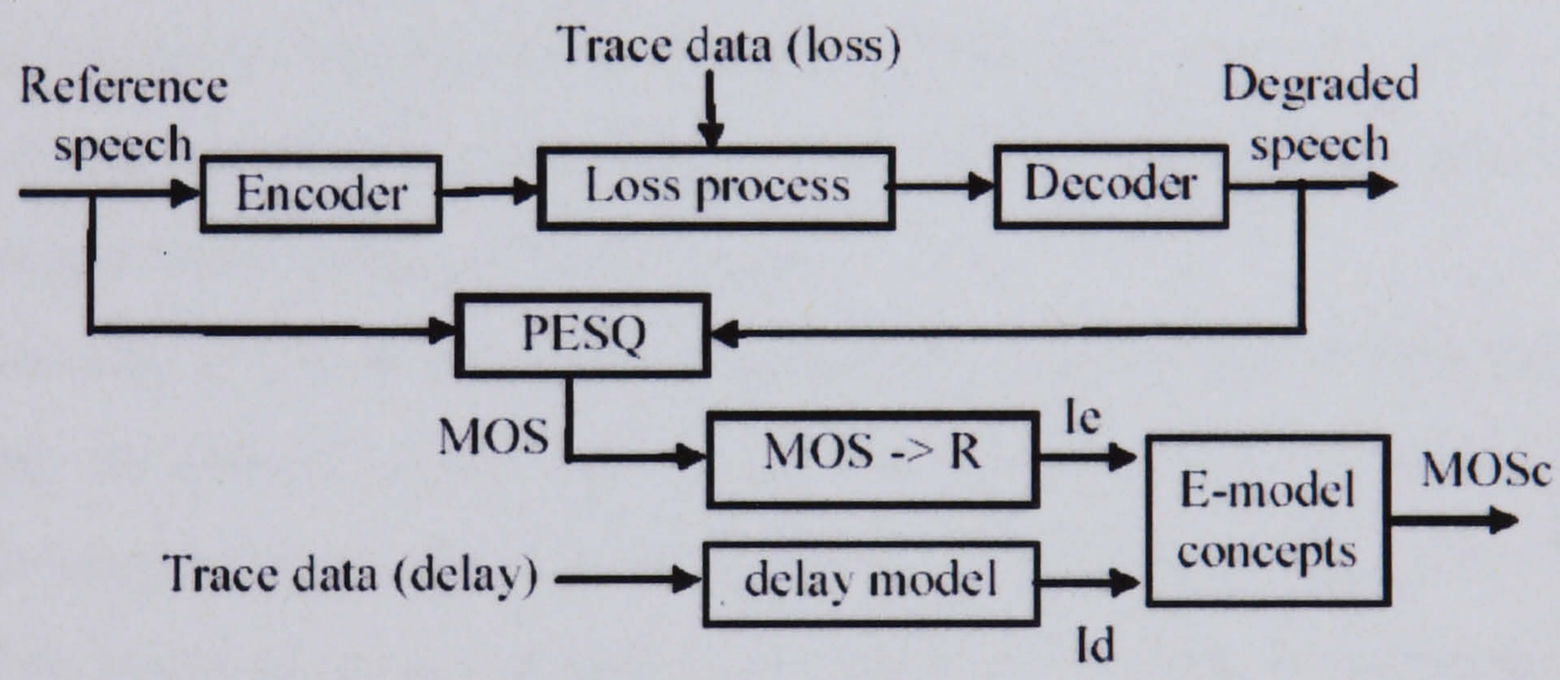


Figure 2.14: Combination of PESQ and E-Model for speech quality measurement

The advantage of this method is to bring the ability to non-intrusively measure conversational speech quality by introduce E-Model to measure the delay affects. The disadvantage of this method is similar to Conway's method, i.e. the ITU standard objective quality measurement method requires a high calculation power, which is hardly achievable from a multiplexed network node or voice gateway. Conway suggested another client/server structure approach to solve the calculation-power-demanding problem in later papers.

2.6 PESQ and Its Time Alignment Method

Different measurement method have different measurement aspects, including network planing, codec performance evaluation, live system performance evaluation, performance enhancement evaluation and so on. Subjective measurement is not suitable for performance evaluation on live network because the high cost, time and resource restrains.

This thesis is focus on end-to-end speech quality on live telecommunication network, especially on wireless mobile VoIP environment. The measurement aspect is on live system evaluation and performance evaluation. PESQ is the most popular perceived speech quality measurement tool and it is selected to benchmark most of the works in this thesis.

PESQ is the international standard for user perceived speech quality measurement. One of the important reasons PESQ can replace the old P.861 PSQM measurement method is that it considers the effect of delay in its algorithm and this makes it suitable for end-to-end perceived speech measurement. The time alignment feature, which is used to provide a correct alignment of time between the reference speech and the degraded speech, is an important part of the PESQ algorithm apart from the psychoacoustic model.

As described in [78], in the process of compare the degraded speech file to the reference speech file, the speech file need to be divided in to smaller parts i.e. so called frames and compared frame by frame. Because the time of each frame is only 32ms as in PSQM and PESQ, if the frames are not time aligned very accurately, it is possible to compare the degraded speech frame to the wrong part of the reference frame and result to a wrong measurement.

PESQ has developed a two stage time alignment method to estimate the delay of a speech path. The first stage is the crude envelop-based delay estimation, which uses cross-correlation to decide delay time roughly. The second stage is "the fine-scale delay identification using a weighted histogram of the frame by frame delay" [78].

Variable delay is also considered in PESQ. A group of methods including utterance delay estimation, utterance splitting, bad frame identification and re-alignment method is developed

in PESQ. This group of mechanism can help to deal with IP network introduced variable delay such as packet loss introduced gap and adaptive jitter buffer adaptation caused delay movement.

The delay alignment feature is very useful for IP network environment because the jitter introduced in IP network could be identified by user and affect the perceived quality.

However, PESQ's performance needs to be analyzed in detail before it can be reliably used in live network performance evaluation and apply it to algorithm performance enhancement evaluation.

To calibrate and optimize PESQ algorithm's performance in live network environment, is the main topic of next chapter.

2.7 Summary

A detailed review of the VoIP system is presented in this chapter. The QoS challenges and state of the art QoS mechanisms are also reviewed in this chapter. QoS measurement methods are reviewed in this chapter as well.

The review of current VoIP system and the QoS problems forms the basis for the performance enhancement research presented in this thesis. The background knowledge gained from the study of current technologies presented in this chapter can be used to improve the current VoIP system and reach better performance of user perceived quality.

Chapter 3

Perceived QoS Measurement in Live VoIP Test Platform

User perceived speech quality is very important for telecommunication service providers and equipment developers. To measure it correctly forms the basis for speech quality enhancement research.

Because speech quality could be affected in many points in the end-to-end conversation, it is getting more and more important to have a reliable speech quality test tool to measure telecommunication products such as a VoIP mobile's performance in the live network environment. A speech quality test platform is necessary for real tests including wireless VoIP mobile test in the live network.

PESQ is designed to objectively measure end-to-end user perceived speech quality by an intrusive comparison method, which needs both the original speech and the degraded speech to give a speech quality assessment result. The live performance of PESQ measurement in the new wireless mobile VoIP environment needs to be investigated and calibrated.

The objective of this chapter is to investigate some live speech quality test examples of using PESQ in the field and present some discoveries about PESQ's error in such cases. This will help us to understand PESQ better and form a solid base for the quality enhancement objective. Setting up a proper VoIP speech quality measurement and evaluation platform including PESQ and live VoIP mobiles, and calibrate it for optimized speech quality test result in a real VoIP

mobile environment is presented before the PESQ error cases because those PESQ error cases are discovered in such environment.

The main contribution of this chapter is the discovery of PESQ performance issues and errors in the live wireless mobile speech quality test environment. A few performance enhancement solutions are also proposed for more accurate speech quality test and better PESQ performance.

From the calibration test cases in Section 3.3 and the field test cases discussed in Section 3.4, it is concluded that PESQ's algorithm limitations may bring certain errors in the new wireless VoIP environment. Those limitations should be considered and well prevented in real cases. It is also necessary to calibrate the test platform and prevent some setup or configuration errors to affect the accuracy of PESQ measurement.

This chapter is structured as following. Section 3.1 introduces the background of user perceived Quality of Service measurement in live test platforms. Detailed speech quality test platform setup, including features and connections, is introduced in Section 3.2. The issues about correct integration of PESQ into the speech quality test platform is introduced in Section 3.3. This section also discusses the platform calibration issues for PESQ. In Section 3.4, a few real test cases are presented and discussed in detail. The suitability and calibration discussion of using PESQ for mobile VoIP speech quality test is carried out in this section as case study. Section 3.5 summarize this chapter with a few future improvement points proposed.

3.1 Introduction

Users are always demanding better speech quality from a telecommunication service. So the telecommunication equipment such as mobile producers and network carrier are consistently asking the question:" How good is the user perceived speech quality in this product when its used in the field?" They need to setup requirements for their products, measure the user perceived speech quality accurately and if possible, improve it.

This measurement ability demand is getting more and more important as VoIP features start to emerge in some mobile handsets.

For example, in a mobile producer's product requirement, it could specify the following requirement for a VoIP based UMA product :

In the UMA product with AMR codec highest rate (AMR122), the mobile to mobie speech quality requirement is user perceived speech quality score at 3.6 or higher; end-to-end delay requirement is at 400ms or lower from conventional GSM based channels to IP based channels. Test cases should reflect the two aspects through out the whole test and should not allow compensation on one aspect to achieve higher results on the other side.

Please note, the above paragraph is taken from a Motorola internal document which can not be revealed publicly but it can provide an idea about test requirements.

To fulfil the test requirements, a speech quality measurement test platform needs to be established. Before it can be used to provide accurate measurement results, it needs to be calibrated. The test platform is the basic tool for further VoIP QoS research.

However, current VoIP speech quality research including [79–81] are mainly using simulation or trace data collected form network log to replace network elements including real VoIP mobiles.

PESQ is the international standard for objective speech quality measurement and has been widely used in industry for voice quality assessment for VoIP products and systems. However, current research efforts involving PESQ are mainly focused on simulated IP networks. The need remains to establish how accurate PESQ is when used in real wireless systems, especially in new wireless VoIP environment.

PESQ can handle relative delay shift and measure VoIP quality as discussed in [78], but its performance needs to be investigated and validated in new wireless mobile VoIP context where packet loss concealment and Adaptive Jitter Buffer (AJB) are commonly employed to provide better quality.

PESQ's accuracy issues have been discussed in [82] in a statistic manner but this lacks de-

tailed reason for PESQ's inaccuracy. Other research efforts, such as those in [83–85] discussed the performance of PESQ in VoIP environments, but are mainly focusing on packet loss conditions in simulated IP networks. An ITU-T Study Group 12 contribution [86, 87] investigated the performance issue of PESQ for specific codec (EVRC) and audio features. P.862.1 [88] suggested a better correlation from raw PESQ score to MOS but it is a general calibration of PESQ score and does not cover specific PESQ error issue.

More than 1800 calls were made under different network and jitter buffer conditions in live mobile VoIP environment. In this chapter, VoIP mobile to PSTN calls and VoIP mobile to mobile calls are focused. It's the observation that in some wireless VoIP speech quality tests, PESQ scores were significantly lower than the user's opinion (obtained by a simplified MOS [5] test with 6-7 listeners) led to the discovery of PESQ error in the new environment. It highlights the importance of test platform calibration for different scenarios and the importance of investigating PESQ's performance in new wireless mobile VoIP environment.

The main objective of this chapter is to present the discovery of PESQ errors found in the live wireless VoIP mobile tests. The objective speech quality measurement platform involving PESQ, live network and real VoIP mobile is introduced in section 3.2. Calibration of the test platform in specific wireless mobile VoIP scenarios is also presented in section 3.3 because of its importance to the whole test and measurement research. Cases about PESQ performance issues are presented in section 3.4. In this section, the discovery of PESQ errors are presented in case studies.

3.2 The Test Platform's Elements, Functions and Connections

A voice quality test platform is established to perform live tests on a telecommunication carrier's live network in order to represent the real field speech quality measurement situation.

The aim of the speech quality test platform is to objectively measure user perceived speech

quality in a telecommunication conversation. In this study the test platform has been setup specifically for wireless mobile VoIP test, but it can also be used to test other forms of telecommunication applications including mobile to PSTN calls, mobile to mobile calls, VoIP calls and so on.

Figure 3.1 shows the over all architecture of the live wireless VoIP mobile speech quality test platform, which is build for this research.

Part A of Figure 3.1 shows the subjective end-to-end user perceived speech quality measurement. The subjective speech quality test i.e. MOS test is expensive, time consuming and difficult to repeat so objective speech quality measurement method is developed to objectively measure speech quality in telecommunication systems.

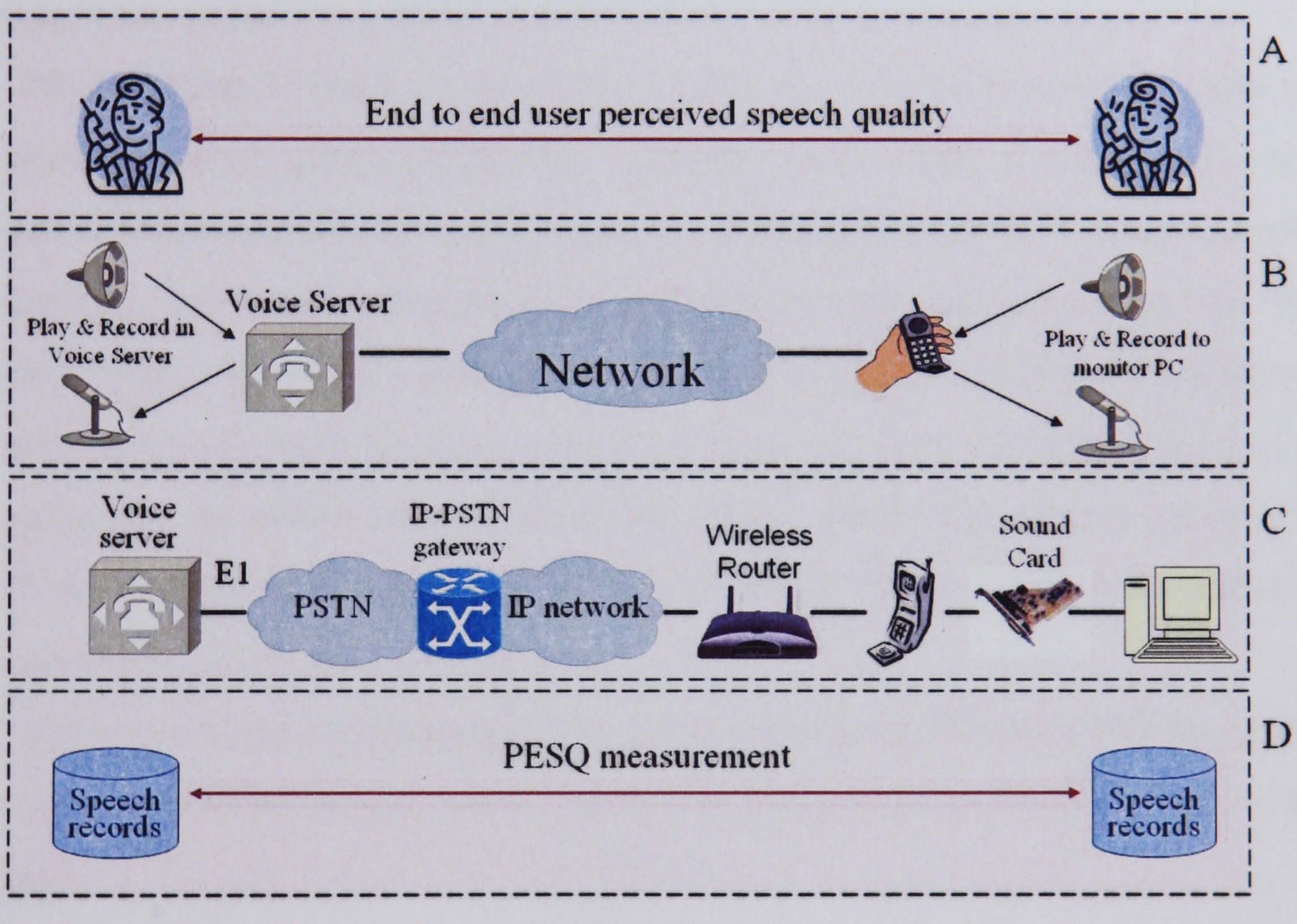


Figure 3.1: The structure of objective speech quality test platform

The goal of the speech quality test platform is to objectively measure user perceived speech quality using PESQ, as illustrated in part D of Figure 3.1. Then PESQ score can be interpreted

into MOS following the recommendations specified in P.862.1 if necessary.

To objectively measure the end-to-end user perceived speech quality, a set of speech transmission and receiving device needs to be introduced to replace the human speaker and listener. As shown in the part B of Figure3.1, microphone and line-out of a sound card device plays the role of human peers in the conversation. Between them, the conversation takes place in a network, including the mobile device and the carrier network.

One side of the network is a mobile handset with wireless VoIP feature and the other side of the network is a voice server, plays the counter party roll of the conversation.

The detailed technology infrastructure is transparent to the measurement method PESQ but should be well considered and optimized in the design and development of the test platform because these technical details will affect the accuracy of the test system.

Part C of Figure 3.1 shows a more detailed connectivity structure figure of the live wireless VoIP mobile speech quality test platform. Left side of the system shows the voice server, which can play/record and store speech samples. It is connected to the PSTN network via a E1 connection, with 30 speech channels. So it can handle up to 30 tests on the same time. The PSTN network is connected to the IP network via a PSTN/IP gateway, which can convert packet based VoIP speech stream to circuit switched PSTN speech stream. The IP speech stream is introduced by the wireless access point and the wireless VoIP mobile, which is connected to a PC through a sound card. On the right side of the figure, the PC can play/record and store speech samples for PESQ measurement.

Different test scenarios including mobile to PSTN speech test (UPLINK), PSTN to mobile speech test (DOWNLINK) and mobile to mobile test are discussed in different sections of this chapter.

3.2.1 The Voice Server

The voice server is a feature enhanced soft PBX platform Asterisk [89] and it has functions to make and receive calls. Based on the basic functions of make and receive calls, other

functions including speech quality measurement are developed. The live VoIP mobile speech quality evaluation platform uses the voice server as communication peer in the network to carry out the speech quality test. In order to provide accurate speech quality measurement results, the Asterisk PBX platform needs to be modified and calibrated. The calibration process is discussed in Section 3.3.

To be able to receive and place voice calls is the fundamental function of the voice server. This function is performed in the software PBX Asterisk platform. The PBX can take calls from both VoIP and TDM side depends on the connection method. After answering the call, the voice server can take further actions to perform lots of tasks including the speech quality test or even speech quality measurement because there is a PESQ module integrated into the test platform. It can place calls to different channels as well. The call may follow with specified actions such as play speech records or echo back the receiver speech.

3.2.2 Codecs and Audio Features

This subsection introduces the play, record and codec functions in the test platform. These are the basic functions used to perform the objective speech quality tests.

In the voice server, there are Asterisk functions available for play and record speech sample files. In voice server play process, the speech file taken from the file system of the voice server can be played to the mobile via the carrier network and through the air interface from the base station to the mobile. Then the mobile can play it out through its speaker and the connected monitor can record the downlink speech. The recorded speech is the degraded version of the played speech and it contains network and mobile introduced distortions for the downlink direction. The record process is the reverse direction of the same process and the voice server recorded speech contains distortion from the network and the voice server side distortion.

The Asterisk platform is able to play almost everything including GSM, G.729, G.726 and PCM, if related format and codec available. The play function can read the Microsoft format wave files (.wav files) if the format of the file is in 8,000 Hz and coded at 16 bits liner. It

can also take raw GSM coded speech files as well if necessary and play parameters are set correctly. Some times there are other codec used in Microsoft .wav format files, for example a GSM codec coded wave file may also have a .wav filename. In this case, file format should be specified in Asterisk, same as other formats, before it can be played out properly.

In the test platform context, because almost all speech analysis is done in a Microsoft Windows based platform and it only takes wav encapsulated speech file, the .wav file format is used and every speech samples and record speech records are coded at 8,000 Hz and 16 bits with a .wav header. It is worth to note that all our speech sample and recorded degraded speech segments are mono, which means it has only one channel.

The play speech function, which as the name suggests, is able to play a speech file to the channel and thus the user end will hear the speech from the test server. The volume of the speech is able to be amplified in the voice server is necessary. However, due to the distortion caused by amplification, the volume is not changed when playout the speech samples but adjusted by the speech volume level in the speech sample preparation stage.

The record function in Asterisk allows the uplink speech to be recorded in the test platform and stored in the specified format. In this group of tests, all records are take at 8,000 Hz and 16 bits liner and wrapped into Microsoft .wav format. Once the speech files are recorded, they can be stored into file system with specific names.

Asterisk the open source PBX supports most common codecs such as Mu-law, A-law, G.723.1, G.726, G.729, GSM, iLBC and Speex. Some of those codecs need license but there are some sample demo codes for educational purpose available. The voice server inherited them and integrated a few more codecs including the much needed AMR codec. There is an Motorola internal version of AMR tested but the 3GPP TS 26.104 demo code is chosen because this is most popular code used in the research community.

Transcoder is necessary when two channels using different codecs are joining each other. In a conference call, transcoder is also necessary for mixing different speech streams. In the mobile speech quality test cases, there is no need for transcoding in the voice server side be-

cause the PCM to AMR transcoding is done in the wireless to TDM gateway or the base station side. The conversion from A-Law or Mu-Law to 8 KHz sampling and 16bits per sample record stream is a standard task and well handled by Asterisk's play and record function. The record and play of wave file functions of Asterisk is discussed in the next subsection.

3.2.3 Connection of the Test Platform to Carriers

Figure 3.2 shows a typical connection configuration for an Asterisk box. It can be connected to the carrier via different connection methods. In the voice server platform build for this research, it is connected to the carrier via a E1-PRI link with 30 digital circuits, which means the carrier trunk is a E1 Primary Rate Interface (PRI) trunk and the time slot are assigned to 30 digital telephone lines. Other methods can be used if the carrier side requires. In our test platform, digital connections are preferred as it does not introduce extra echo and noise into the test as analog line would.

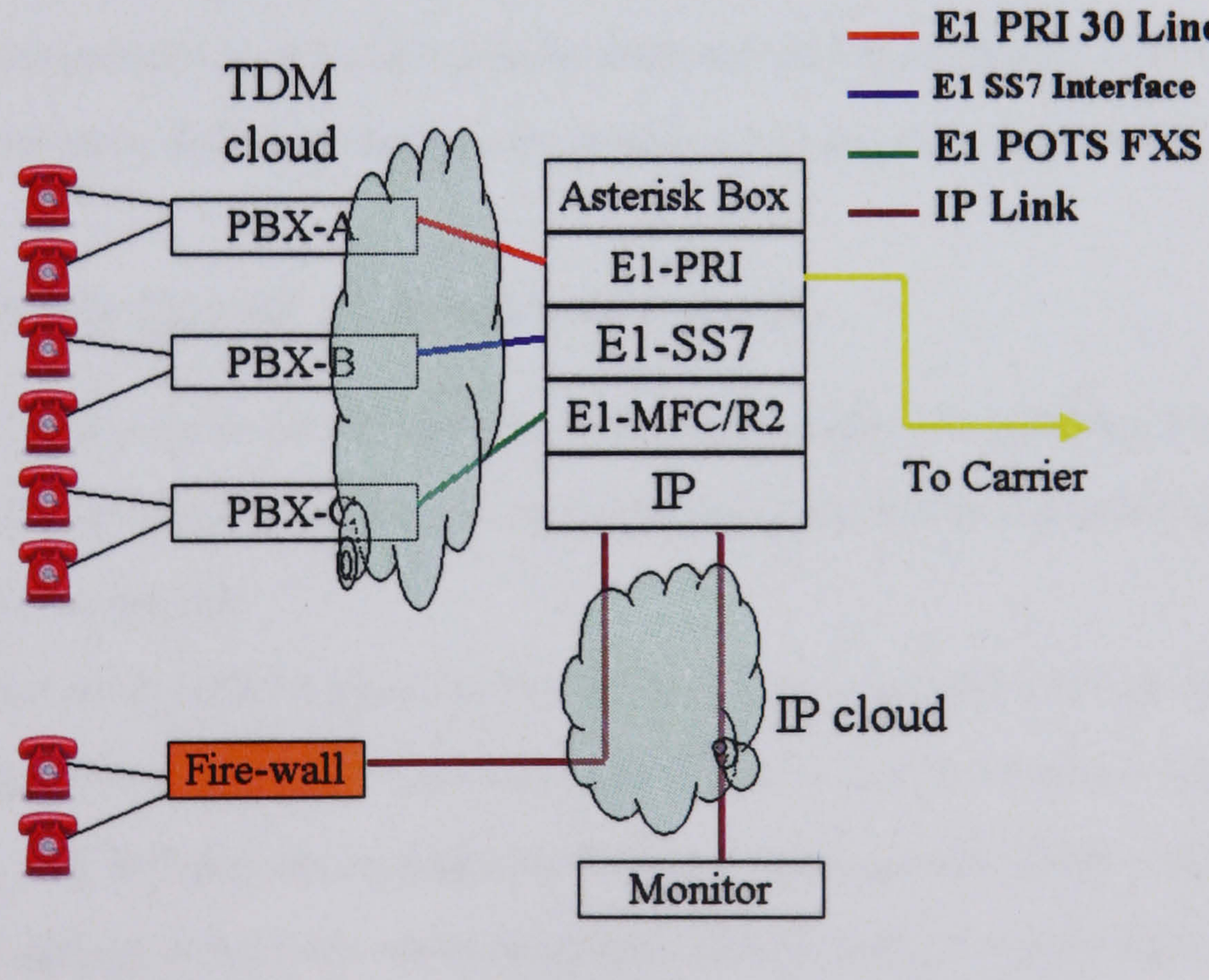


Figure 3.2: Asterisk connection to TDM and IP interface

To use digital circuit of the communication network is necessary to reduce echo in the

system. The test platform is located in the UK, where most of the PBX connections are using E1. Therefore E1 port is selected to be the main connection method to the carrier. IAX (Inter-Asterisk eXchange Protocol) or SIP [14] connection via IP network will introduce unpredictable distortions so they are not selected. The management monitor can be connected via an IP link together with some IP based SIP or other protocol enabled soft phones.

Apart from trunk connections to other PBXs, the test platform can be directly connected to phones if relevant connection port is available. For example it can be connected to a POTS phone via a Foreign exchange office (FXO) card or an Ear and Mouth (E&M) card (not shown in the figure). It can also be connected to a soft IP phone via an Ethernet connection. Of course it is necessary to firewall the connection to prevent possible intrusions from the IP side because the monitoring remote connection is also connected via the remote IP connection.

Different channels can be joined inside the Asterisk platform for instance PSTN to SIP call or conference calls between them. The signalling translation and forwarding processes are handled in the Asterisk box as a part of the Private branch exchange (PBX) function. There are also transcoder models in Asterisk to handle codec translation and mixing in a call if the caller and callee are using different codec. Codec models are discussed in next subsection.

3.2.4 Speech Quality Measurement Feature

Speech quality measurement features in the voice quality test platform include the integrated PESQ measurement module, the result statistic and publication module and the archive index and search module.

The voice server platform allows us to perform speech connectivity test for mobile handsets and other products or services. It needs external modules to perform speech quality measurement task. The international standard for intrusive object speech quality measurement tool PESQ is integrated in the voice server to perform speech quality test for VoIP enabled mobile handsets.

The official PESQ license comes from Psytechnics Limited UK, the co-inventor of the

PESQ algorithm. For calibration purpose, another version of PESQ implementation from OPTICOM GmbH, Germany, who is the other co-inventor of the PESQ algorithm, is also tested. About 100 different speech samples from different sources are tested and the results are identical. This confirmed that the different implementation of PESQ will not introduce errors into the speech quality measurement results. The Psytechnics Limited version is selected as it is coming with a nice graphic interface for showing and analysing results.

PESQ is an intrusive objective user perceived speech quality measurement. It works by comparing the reference speech file with the degraded speech file and measuring the different and the possible impact to user perception to the distortion. When it is working in the voice server system, it needs to be given a reference speech file and the degraded speech file. The original speech file is pre-specified as Microsoft .wav format and pre-stored in both the voice server and the monitor PC, as shown in Figure 3.3. In this figure, the mobile is making a call to the voice server and the connected monitor PC is sending a speech sample to the UPLINK direction i.e. to the voice server.

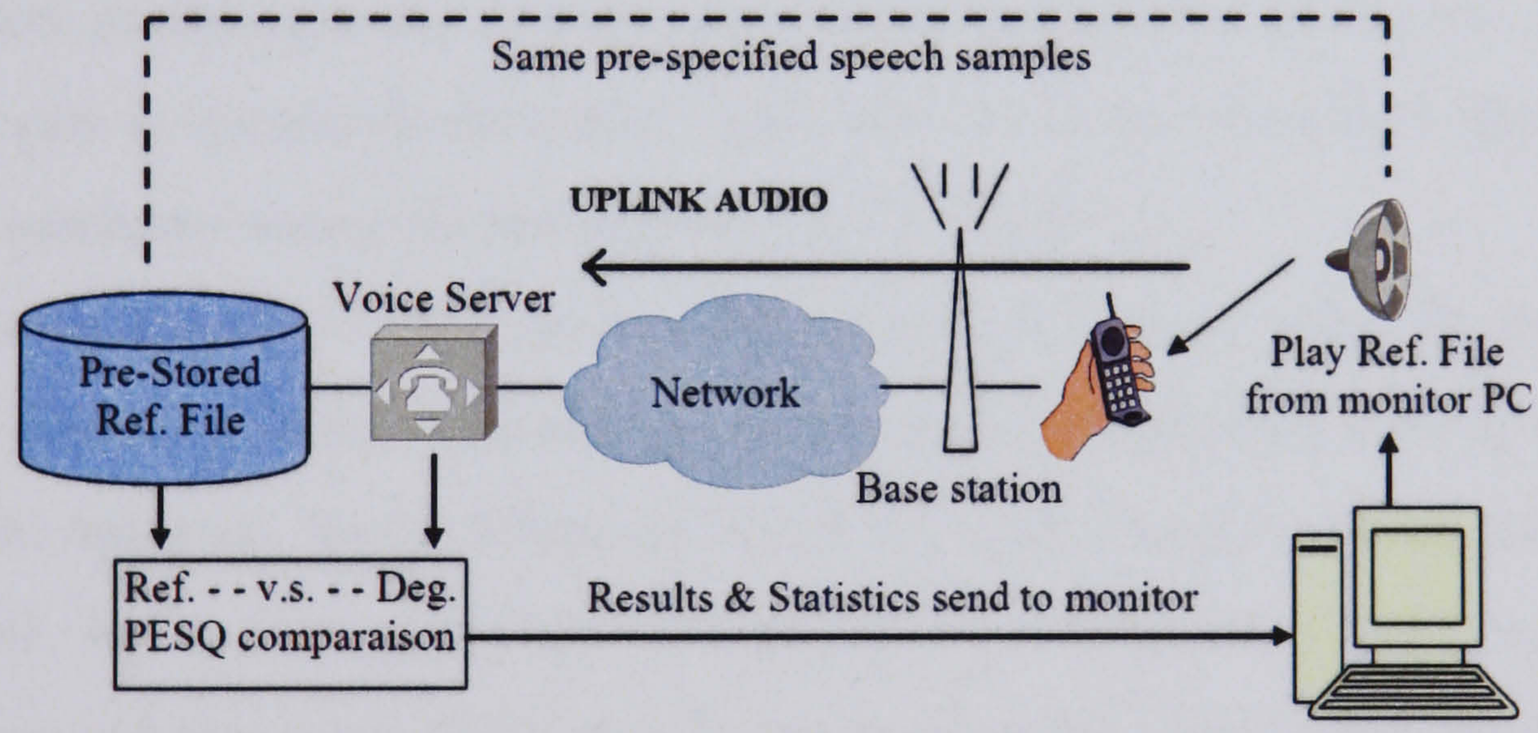


Figure 3.3: Uplink speech quality measurement

The speech file gets network distortion together with application level distortion, for instance codec introduce distortion and jitter buffer introduced impairments. Because the speech is traveling uplink, the jitter buffer distortion is introduced by the network carrier, before the VoIP traffic been converted to TDM traffic, where speech frames transmission are synchro-

nized. To test the performance of the jitter buffer in the mobile handset, a DOWNLINK direction test i.e. speech been played from the voice server to the mobile handset, is needed. Details of a DOWNLINK direction speech quality test is discussed in later sections.

The reference speech is stored in both the monitor PC and the voice server before the test starts. So in the uplink test, the receiver i.e. the voice server will have access to both the reference speech sample (from local storage) and the degraded speech sample (from the mobile to voice server call). This makes it possible to call PESQ function to calculate the objective user perceived speech quality and give a PESQ score.

In most of the cases, test speech samples will be repeated many times to reduce random effects and achieve an average of PESQ score. Those results can be summaries and send back to the monitor PC for reports and further studies. Individual recorded degraded speech sample files are stored in the voice server and be able to be accessed individually. There is a web interface to down load individual files. They are not sent back via the monitor link because the size of files are quite big and there is no need to look into specific files on most of the cases.

The web interface can also do index and condition searches for the recorded files. This makes it easy to find out abnormal files. Abnormal files include those have different size, different parameters and maybe extremely high or low PESQ scores.

In the tests process, it is find out that there are possible situations where PESQ gives out wrong or misleading results. Some of these situations are caused by misuse of PESQ or problems in the test setup. Some of these are caused by wrong settings or configurations of the test system. But there are some relatively small chances in which PESQ could get measurement wrong and gives wrong PESQ score on certain speech files. Those cases are discussed in Section 3.4 with detailed case study.

To use PESQ correctly and properly is very important if the researchers want to get robust and reliable measurement of mobile speech quality. A calibration process is needed to achieve the proper and correct measurement. In the next section, some calibration and tools preparation work is presented. Some potential problems are highlighted and solutions given as well.

3.3 Performance Calibration of the Speech Quality Test Platform

Field test is a critical test phase for mobile handsets. In the field test process, mobile handset's performance is tested on real live carrier network on the move. There are more test realistic issues need to be addressed in a field test. For instance a real electrical cable connection from a monitor PC to the mobile handset under test is need to be established because there is not enough processing power nor enough storage space in the VoIP enabled mobile handset to perform the play and record process. Even if it does have such capability, it is not fair for it to do so because that will bring the over all performance of the mobile operation system down thus possibly degrade the performance of the VoIP speech quality.

PESQ evaluation is an important function of the voice server. With the help of the voice server, the uplink and downlink speech quality can be measured in the field. It can also bring the features to measure end-to-end delays but can not perform that in real time manner.

Test platform setup and test case designs will be presented here in this section before the test cases because some of the test platform designs are critical for reliable quality measurement results. The analyze process of the measurement results will be discussed in reasonable details.

3.3.1 The Gappy Audio Problem and a Practical Solution

In current computer systems'architecture, hard disks are the main storage space for files. In most of the cases, the original speech files are located in a hard disk before they are played. And when degraded speech files are recorded, they need to be written to a hard disk for storage and for later access. At the beginning of the play process, there is a task to read the file out from the hard disk. Compare with other processes like play a wave files, this task is a high demanding task in terms of processing power and transmission speed from the hard disk, because the whole length of the wave file needs to be processed in the very short period of time. Similar heavy task is needed to write the recorded speech file back to the hard disk.

In the voice server implementation, speech files could be required by several players on the same time and most likely with different starting time. This is because different users may dial in to require the same original speech file to be played on different time and those starting time may be different but very close. Figure 3.4 demonstrates the concept of different users requiring the same speech file (speech file 1) on the same time. There could be overlap of play or record time. Therefore a read or write request could happen on the time when a speech file is played from the hard disk.

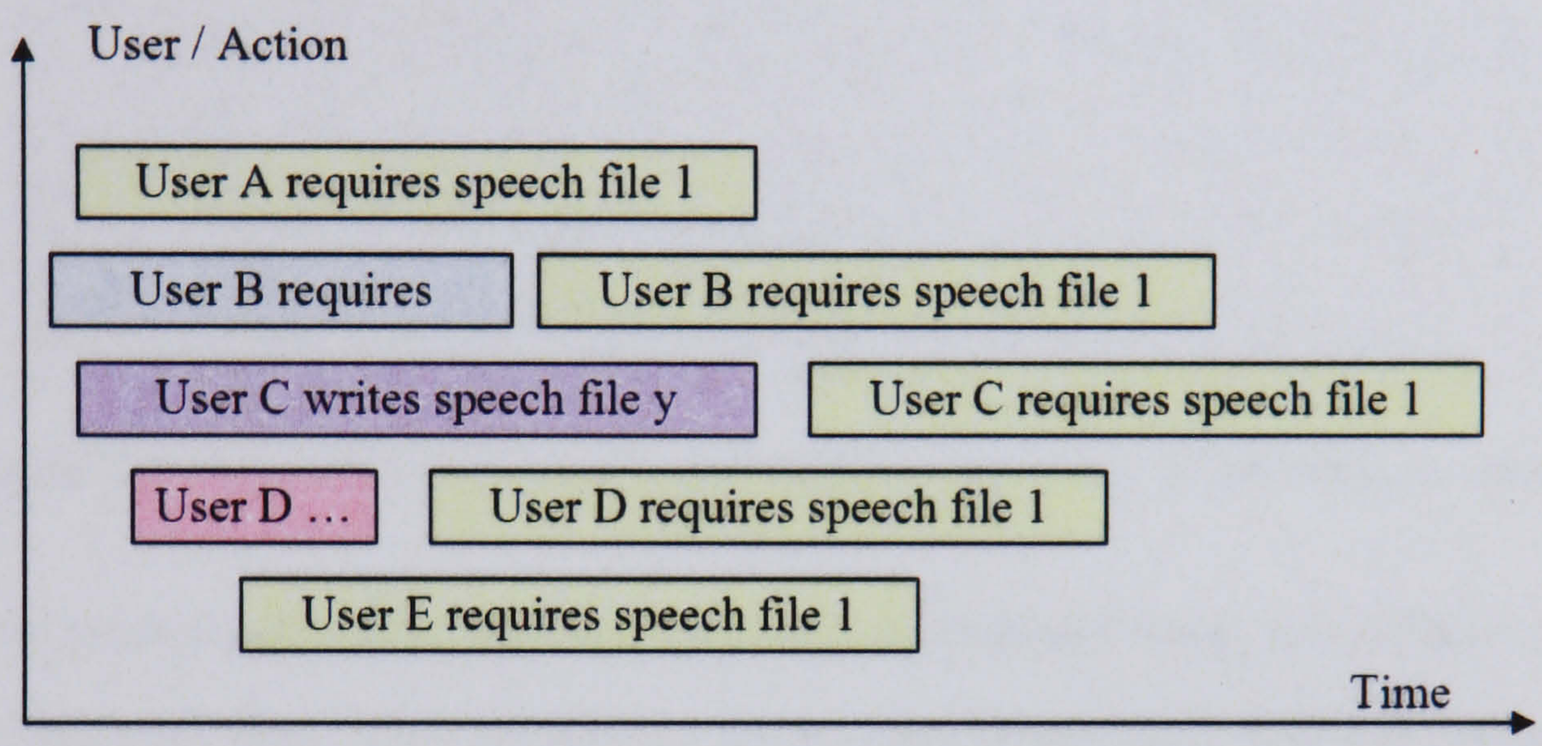


Figure 3.4: Different user may require the same speech file on different time

Although hard disks support random data access, the limitation of processing power of the hard disk and limitation of the file system bring some extra real time affects to the real time play and record process. The results of those extra real time affects are small silence gaps in the play out speech. This happens when a speech file been read out from a hard disk and play out on the same time. When the hard disk or file system can not deal with the speed of the play out, most likely when multiple read processes happening on the same time, the player will introduce extra silence gaps in the speech, hence introduce an unnecessary distortion to the transmitted speech.

Those small gaps can be picked out by PESQ from the frame by frame quality plot and can be confirmed by look at the speech wave form in details. Figure 3.5 shows a 2.6 ms silence gap

introduced in the transmitted speech when another copy of the same reference file is required by another user from another line. The gaps will also happen when the other user finishes a recording and writing back a degraded speech file to the disk.

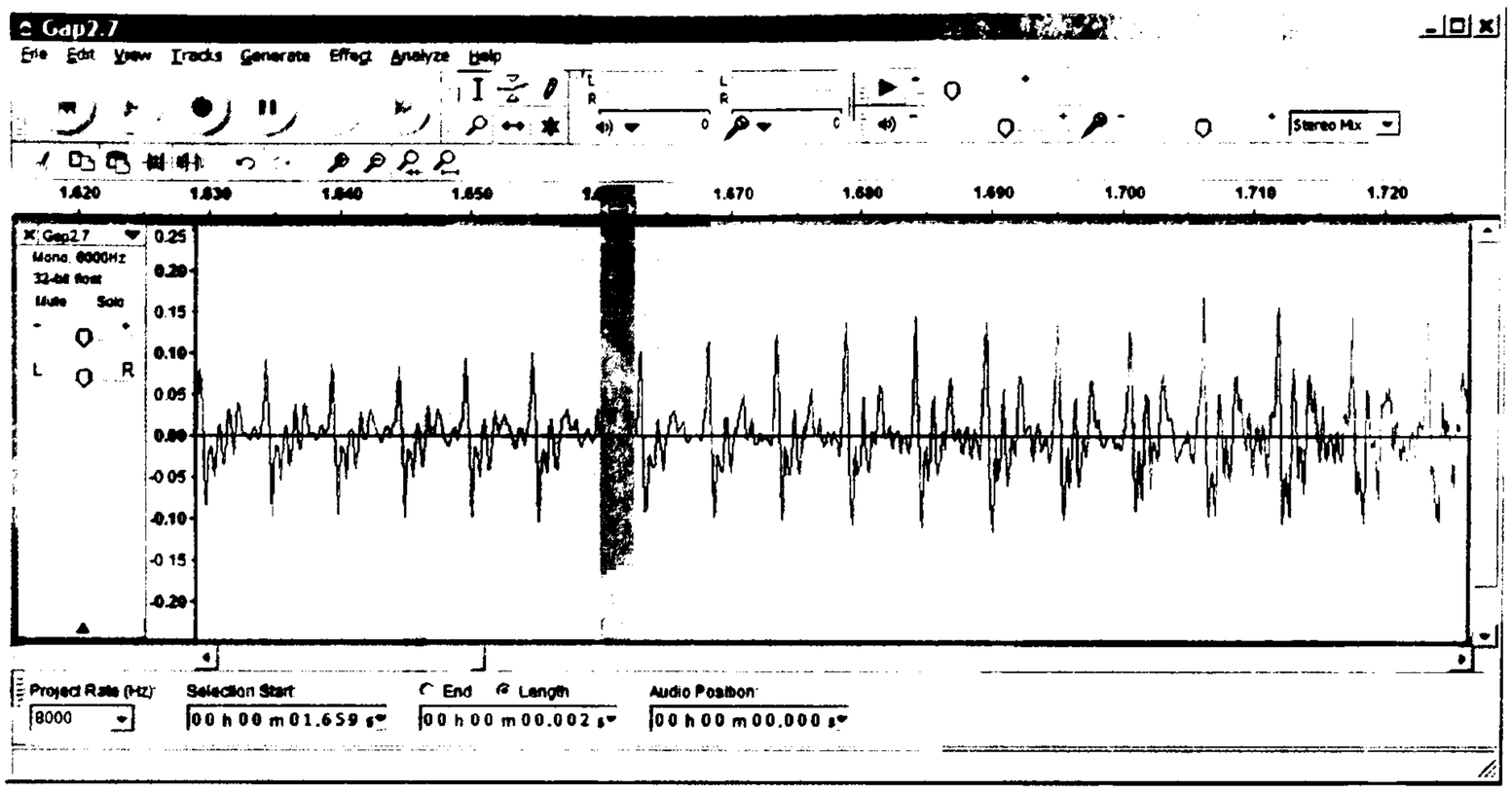


Figure 3.5: A 2.6ms silence gap in transmitted speech due to hard disk readwrite

Each of those small silence gaps is about a few milliseconds long. It is difficult to find out the exact reason for those small silence gap because they are not only relatively small, compare with a gap introduced by speech frame loss, which is strictly the same size of the speech frame, 20 ms for instance for the AMR codec, but also because they happen randomly due to the randomness of start/finish time of other line users.

To capture this gap problem, a small group of telephone lines are used and users are configured to request the same speech file to be played through the voice server. The user requests are to be fired strictly on designed sequence. Then the recorded speech are checked in fine details. Once the pattern of the gaps are discovered, it can be linked to the sequence of the read or write process. By analyse the pattern of the gaps in the recorded speech, the direct link between those small gaps and the read or write process to the hard disk are discovered. Therefore, the root cause of those small silence gaps is the hard disk read and write process.

The solution is relatively simple after the real root cause of the problem is discovered. All the speech files are pre-loaded to a virtual hard disk in the memory space of the computer

running Asterisk. To use a solid state disk, which provides a better speed, instead of a traditional disk is also a possible solution. By doing that, the read and write process take much less time and they are not affecting each other any more even they are performed on the same time. Of course its necessary to mention that not all of small gaps are related to hard disk interruptions. It is only safe to say read or write process to hard disk will cause small gaps but the reverse argument is not correct.

3.3.2 Mobile to Sound Card Connection Cable Calibration

When the PC plays a speech sample to the mobile, through the telecommunication network, eventually the speech will be recorded in the voice server, this form of test is called an uplink test. If a speech quality test is carried out in an uplink test, the codec performance and the network side jitter buffer performance is included. The reverse direction is called the downlink test, where the speech is played from the voice server to the mobile and eventually been recorded in the monitor PC.

To perform the uplink test, a electrical signal is needed to drive the microphone in the mobile handset instead of the vibration of the air. To perform the downlink test, the speaker output need to be converted into an electrical signal drive the recording mechanism in the monitor PC. The reason for this electrical cable interface to replace the speech wave transmission over the air is because that PESQ can only handle electronically transmitted speech samples but not suitable for air transmitted acoustic signals.

On the monitor PC, the equipment to handle the play and record process is a sound card. It will take downlink electrical signals for record and play electrical signal to drive the microphone in the mobile for record in the voice server. It is very important to use a high quality sound card and install the driver properly because some low end or on board sound card will introduce unwanted distortions like noise, gaps or even time drifts into the speech quality test and result to a misleading speech quality measurement result.

Some onboard sound card are tested in a laptop PC and the distortion it introduced to

the recorded speech are measured. The finding is that the distortion they introduced are not tolerable. They introduced millisecond level silence gaps to played out speeches randomly and the frequency of such distortion is much higher than expected, reaching a few times per second. And they also introduced some random noise clicks into the record speech, which might be caused by some internal electrical noise. This is only an extreme case but it highlighted the importance of the high quality sound card.

The sound card finally selected is a high end sound card with its customized driver for Windows platform. It is a broadcast level high end sound card and a calibration test is done on it to prevent instrument introduced system error. The software used to play and record speech samples should be robust and reliable as well. A Motorola made speech quality test tool using the customized driver from the sound card producer is used in this test platform. Implementation details of the test tool is transparent to this project but it is told that the driver has been optimized to fit the operating system and the application, otherwise it may not be able to deliver perfect real time performance, due to some operating system issues. The hardware interface of this sound card is a group of professional XLR connector. The connection cable is discussed in next subsection together with the solution of the voltage mismatch causing clipping issue.

The way used to calibrate the sound card is to connect the mic and speaker socket directly together, to make a local loop. Then a play and record function in an audio tool can be used to do a record of played speech sample. Its possible to check if the recorded speech is exactly the same as the played speech. When use PESQ to compare the recorded speech with the reference speech, the PESQ score keeps at 4.5 for our setup. Since 4.5 is the highest possible PESQ score, it is assured that the sound card and mobile connection is not going to introduce extra distortion to the speech quality test and the possible system error from the sound card instrument is eliminated. By using the "record while play" function of the audio tool, the sound card local loop back delay is measured at about 92 milliseconds.

3.3.3 Clipping Issue Caused by Voltage Mismatch

When connect the sound card directly to the mobile microphone and speaker, there is a voltage mismatch problem causing the uplink speech been clipped. When the echo in voice server test is done for the calibration of the sound card, it is found out that all recorded speeches are clipped due to a much higher volume than it should be. Since the echo function in the server is set to return every RTP packet to the sender side, the clipping problem should be coming from the sender side of the loop. A further record in the server confirmed the assumption.

Then a voltage mismatch problem between the sound card output and the microphone in the mobile is noticed. Because the sound card output positive and negative cable, which are suppose to drive a speaker, are connected directly to the microphone's positive and negative pins, the sound card output voltage is much higher than the input range of the microphone. So the microphone in the mobile are receiving much higher signal range and producing much higher volume with clipping as shown in Figure 3.6.

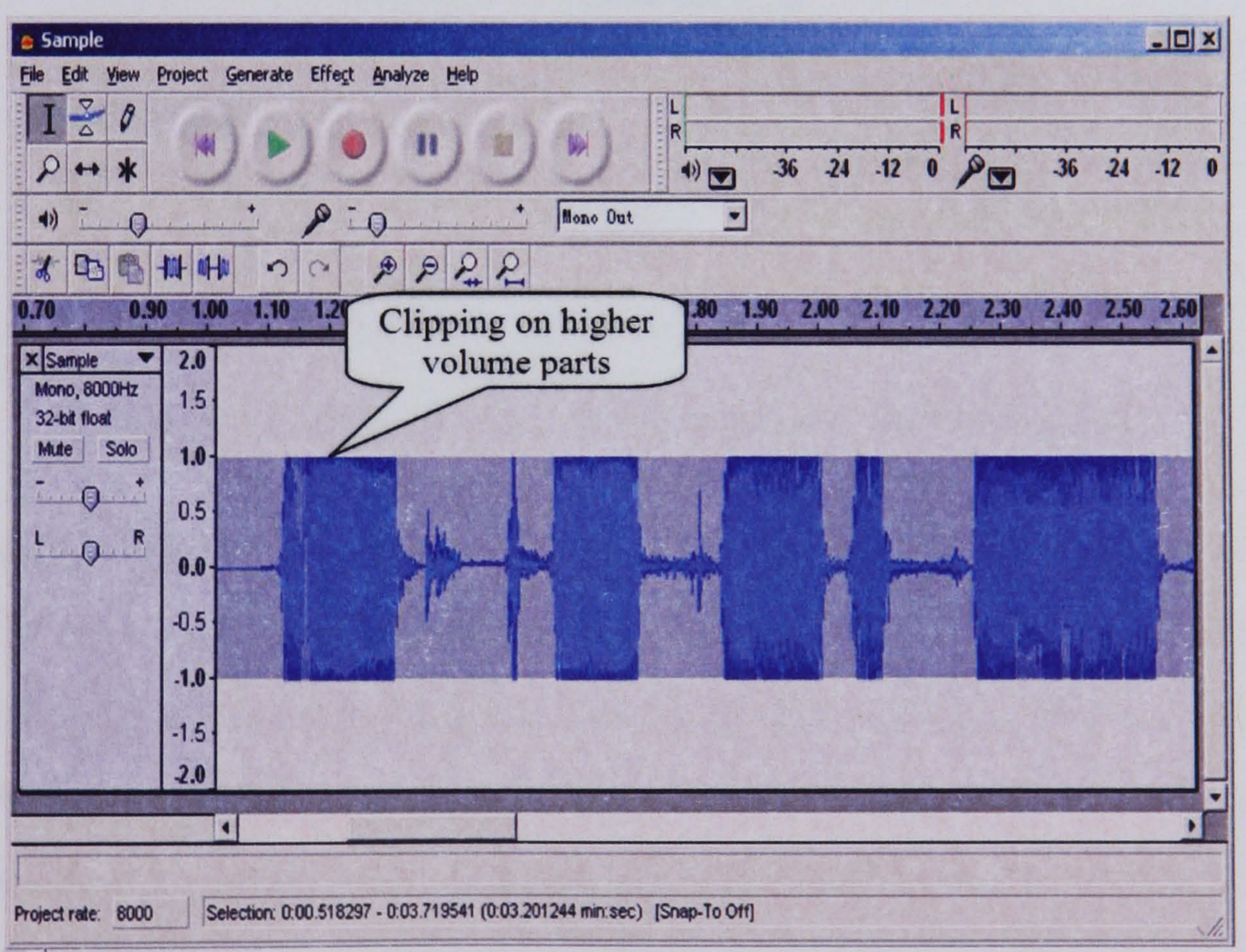


Figure 3.6: The clipping in uplink record caused by voltage mismatch

Since the voltage difference between the sound card out and the microphone range is too high, its not correct to adjust it in the player volume but needs to fix it with hardware. Volume adjustment is needed to get more accurate PESQ but the part will be discussed later. The sound card output cable (positive and negative) is connected to a resister network and the voltage reduced to about 1/50 of the original. The resister network connection is as shown in the following Figure 3.7.

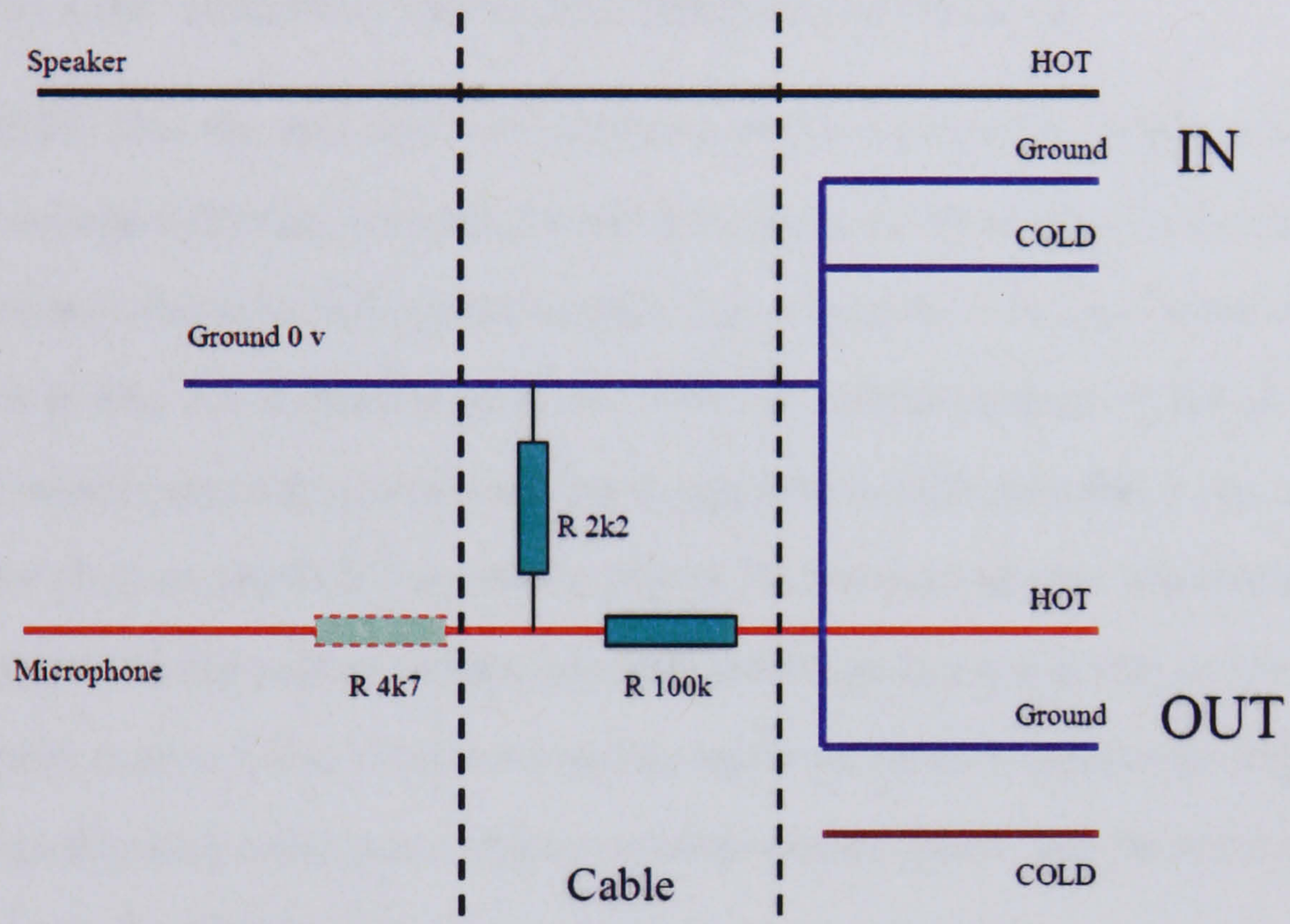


Figure 3.7: Resister network for reduce sound card output level

As shown in Figure 3.7, a voltage sharing resister network is added in the microphone side of the cable to cut the voltage of the sound card by about 1/50. There is a 4.7k Ohm resister in the mobile phone headset originally so only the 2k2 Ohm and the 100k Ohm resisters are added. There are microphone and speaker link because there are sharing the same ground. The resister networks are on the sound card output side and does not affect on the input side.

The connection from the EMU socket of the sound card to the mobile phone handset is modified from a headset cable, including the voltage mismatch solution above. By using the

modified headset cable, it is possible to connect the sound card electrically to the mobile handset and get round of the air acoustic interface. The only limitation is that the mobile's DSP in headset mode may not be behaving exactly the same as in the normal speaking mode or laud speaker mode. Since only the VoIP speech quality is interested in, the headset cable solution is good enough to fit the purpose.

3.3.4 Play Out Volume's Effect to PESQ Measurement

When PESQ does the user perceived speech quality measurement, it does not taken into account the volume difference between the reference speech and the degraded speech because it will preprocess volume on both speech samples before it starts to do the comparison between them. But it is find out in the test platform, different volume settings in the player on the monitor PC, which plays out reference speech through the sound card cable to the mobile, will have different effect on the PESQ test result. Figure 3.8 demonstrates the relationship between the player volume setting and the mobile to mobile speech quality test results in terms of PESQ score. Basically there is a best volume for certain hardware setup to achieve the highest PESQ score on given distortion conditions. Higher or lower volume away from the best volume level will both degrade the PESQ score.

In this best volume finding test for the mobile to mobile scenario, two mobile phones are made to make a mobile to mobile call, the reference speech sample are played to one mobile and then recorded from the other mobile. Then the PESQ score is calculated given the reference and the degraded (as shown in Figure 3.9). There is no test platform introduced extra distortions such as extra jitter in this volume setting test. By the mobile to mobile connection method, it is possible find the best mobile to mobile player volume setup for both directions. Those two mobiles connected in this mobile to mobile test is a GSM mobile and a VoIP UMA mobile.

As Figure 3.8 shows, the PESQ score increases to a peak and then decrease with the volume setting increases from -36dB to 39dB in both directions. The peak for VoIP mobile to GSM mobile direction is captured at 3 dB and the peak for the reverse direction is recorded at 9dB

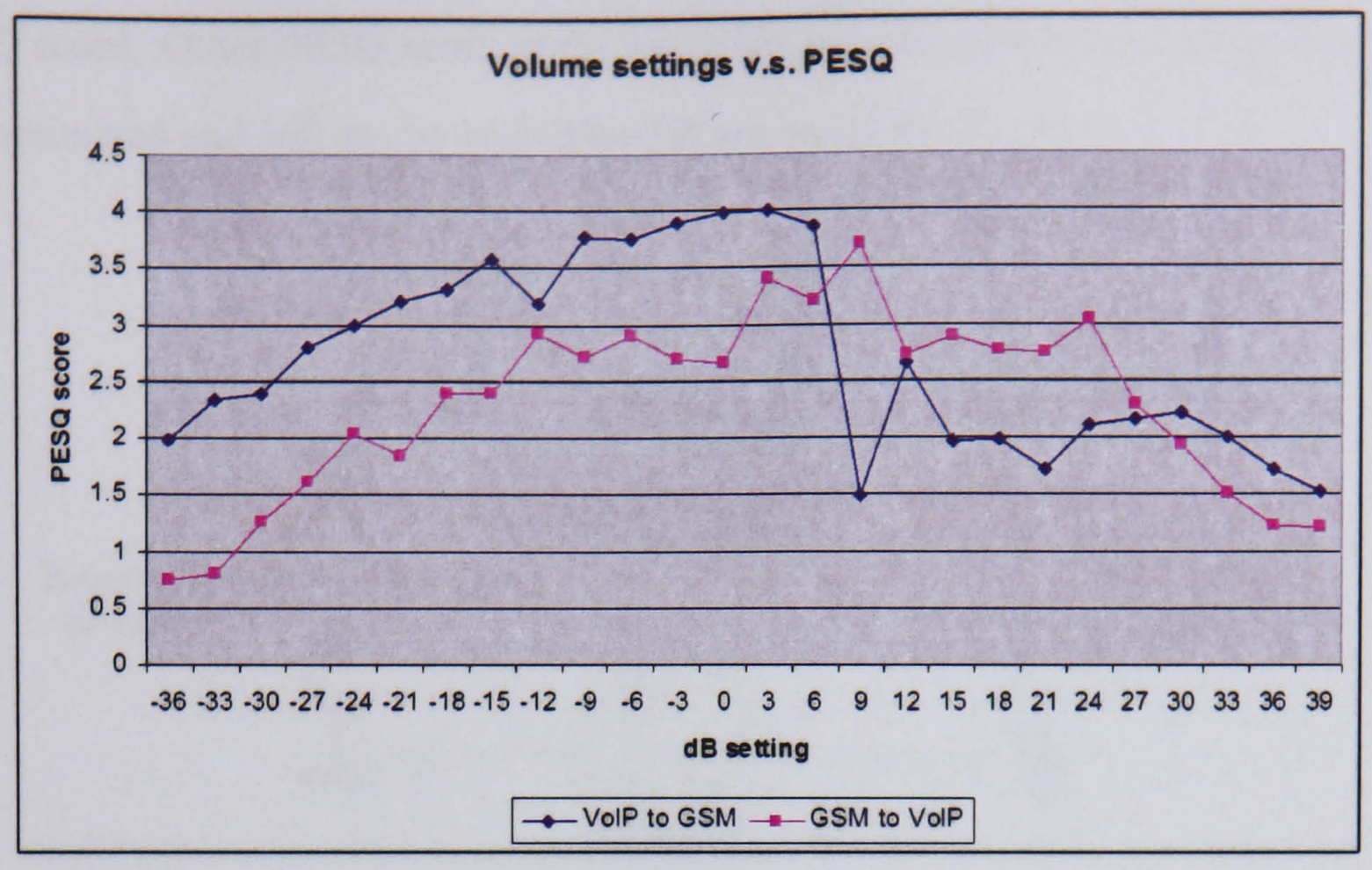


Figure 3.8: Volume setting affects PESQ score

in the player setting. The difference best volume setting between two directions is caused by hardware difference between mobile and cables, for example the resister could have 10% manufactory difference thus introduce different attenuation into the resister network described in Figure 3.7.

The reason for the PESQ score to increase in the first half of the test is that the signal noise ratio (SNR) increases with the signal level i.e. the player volume increases but white noise on the mobile and the cable stays the same. With a higher SNR, the noise distortion introduced to the recorded speech is decreased and PESQ score increases. But if the play out volume increases too much, the microphone and the encoder can not coop with the highest volume and the speech start to get clipped on the high volume part, PESQ will find out the clipping and thus give a lower score. The higher player volume setting goes, the more clipping happens and PESQ gives lower score.

To find the best suitable volume for each hardware set, which includes the mobile handset, the cable and if involved, the voice server, is justified because the user will adjust their volume in a conversation to the best comfort level and our volume setting is not artificially increasing

the PESQ score. Other PESQ result in the voice server test platform discussed this thesis are volume optimized and will not be highlighted if not specified otherwise.

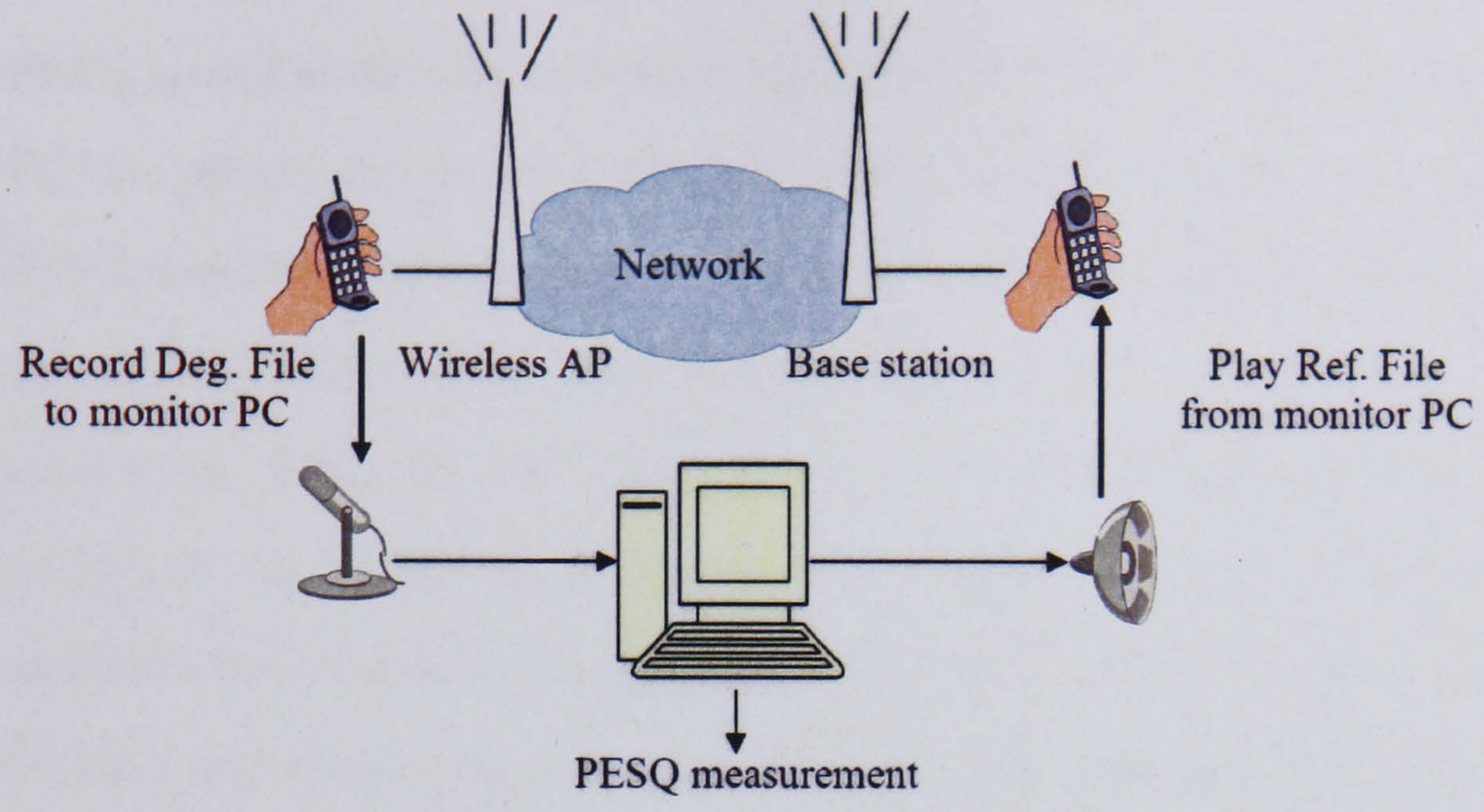


Figure 3.9: Mobile to mobile volume level test

By analyzing the result of this mobile to mobile speech quality test, it can be concluded that the VoIP mobile to the GSM mobile could achieve a better user perceive speech quality, at PESQ score 4.01 on the peak with volume set to 3 dB. As a comparison, the GSM mobile to VoIP mobile direction can only get PESQ score 3.71 on the peak with volume set to 9 dB. This is because the GSM mobile used is a stabilized mature product and the VoIP jitter buffer in the base station side seems to be working better. The VoIP mobile is still a prototype and the adaptive jitter buffer was under development when the test was done.

This result shows that play out volume does play an important role in PESQ measurement and it is necessary to find the best volume for each hardware set includes the mobile hardware, the server and the cable, otherwise the best accurate PESQ measurement may not be able to be achieved . The way to find the best play out volume is to run an increasing volume test on each hardware setup.

3.4 PESQ Error Cases in the Wireless VoIP Mobile environment

When PESQ is used as the only evaluation tool for user perceived speech quality test, calibration of PESQ's performance in new scenario is need to prevent misleading or error results. A couple of PESQ errors cases discovered in the new wireless mobile VoIP scenario are discussed and a few more real test cases are discussed here in details.

End-to-end voice quality for over 1800 mobile-to-mobile and mobile-to-PSTN calls are tested over a period of 3 months. Each of the speech samples are about 9 seconds long, with about 5 seconds of active speech. There are male and female speeches and some speech samples contains both male and female speech. The volume of the speech samples are either normalized or slightly lower than normalized volume. All tests are done in an office environment with interference from other mobiles around.

Its discovered that there exist some PESQ errors when its been used in wireless mobile VoIP environment. These problems are mainly caused by improper time alignment in the PESQ algorithm when there are silence gap and speech sample removal or insertion due to packet loss concealment and jitter buffer adjustment in mobile devices.

Subsection 3.4.1 presents the discovery and analysis of a PESQ error case caused by silence gap in the degraded speech. Subsection 3.4.2 presents the discovery and analysis of a PESQ error caused by delay shift in the degraded speech.

3.4.1 Discovery of PESQ Delay Measurement Error Caused by Silence Gap

A PESQ error case is presented in this subsection. The PESQ algorithm gives wrong relative delay measurement when a silence gap presents in the degraded speech.

As discussed in previous sections, relative delay i.e. delay difference between different parts

of a speech sample is a useful tool to measure end-to-end user perceived speech quality and it is very important for performance evaluation in new adaptive jitter buffer algorithm development. Relative delay is able to be measured by PESQ using the frame by frame delay function inside PESQ algorithm. The PESQ GUI program can provide a graphic interface of the relative delay reading.

In the mobile filed test, it is necessary to measure the relative delay in a very long period for example 2 hours. So a Motorola internal tool has been developed based on the voice server test platform and PESQ frame by frame delay function to fulfill the very long relative delay test. In the very long call test, a pre-defined long speech sample contains repeating 10 seconds unit of speech samples are played and recorded. The very long delay test can be done on uplink, downlink or mobile to mobile directions depends on the test requirements. In the adaptive jitter buffer algorithm's performance research, most of the very long call test is done for downlink direction because jitter buffers normally works in the downlink direction.

After careful cutting process, each individual speech sample of 6-9 seconds can be compared to their reference speech, and the relative delay can be calculated in PESQ. Then by putting the relative delay plot of each individual speech sample together followed by the played sequence, the relative delay in the very long call test is pieced up.

It is worth to note here that if the test is not a mobile to mobile test, i.e. the play and record speech are not absolute time synchronized, a time drifting issue may accrue in a very long call since the clock speed in the monitor PC and the voice server may have some slight difference causing the long speech samples to be cut into wrong pieces instead of correct individual degraded speech samples. The solution is to introduce some synchronized silence gap into the very long grouped speech sample and use it to indicate the right cutting place.

When measurement is carried out in a VoIP enabled UMA mobile handset, it is noticed that PESQ will sometime make mistake of the frame by frame delay time and gives out unreasonable error readings.

Relative Delay Spike

Figure 3.10 shows an example of incorrect relative delay readings given by PESQ in a very long call test.

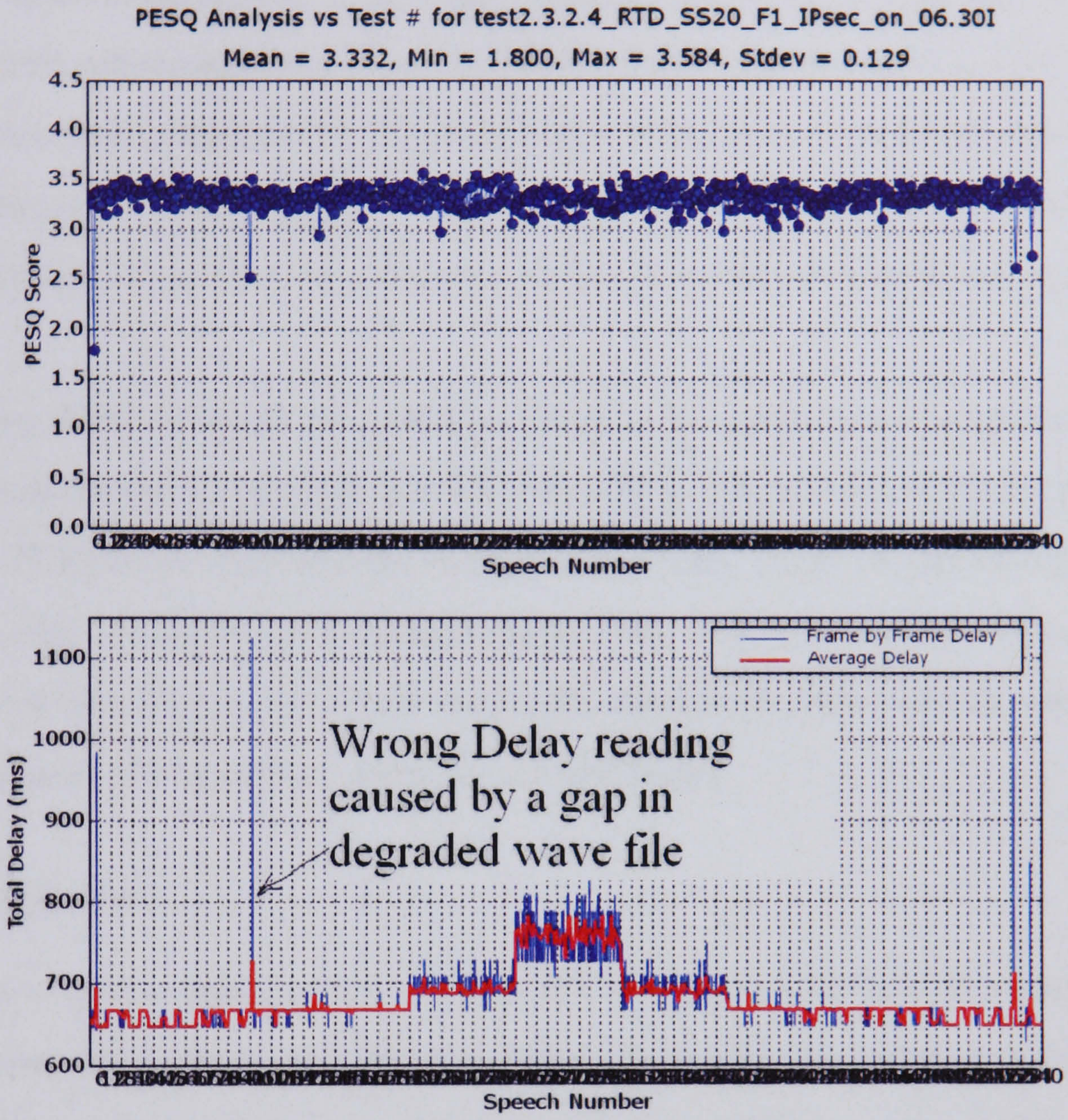


Figure 3.10: Example of wrong relative delay reading in PESQ

In the figure, each blue dot indicates an individual speech sample and its measurement. The top half of the figure shows PESQ score for each individual speech sample, using speech sample sequence for X axis. The bottom half of the figure has two indication lines, one is the blue one showing the combined frame by frame reading of each individual speech sample and the red line is the combined average delay reading for each speech sample.

The Y axis in the bottom half of the figure shows the total delay between the reference

speech and the degraded speech but because the record and play are not absolutely synchronized this ready is not reliable. But the blue and red lines still give indication of relative delay if the total delay readings are ignored.

In the figure, as can be seen on the marker pointed place in the delay plot, there is a spike of a few hundred milliseconds high. This spike indicated that the relative delay of a few frames in this speech sample is increased a few hundred milliseconds in a very short time and decreased to about the same as other part of the speech. This behavior is very abnormal and a closer look of it is necessary. There are a few other spikes in the plot showing similar big delay increase as well.

By using PESQ GUI to run the suspicious speech sample again, it can be seen that the delay measurement does have a big increase in this sample at the time point between 7 and 8 seconds, at least as PESQ shows. The following Figure 3.11 shows the incorrect PESQ reading of frame by frame delay. At the bottom of the figure is the frame delay plot showing a more than 400 milliseconds delay increase and it lasts about 1000 milliseconds. Please note the unit of the Y axes in the delay plot is 1/256ms, due to the software design.

Waveform Analysis

400ms or more in relative delay, is the result PESQ gives on this specific speech sample, but it is an incorrect reading. It is possible to check the real relative delay by comparing the reference speech sample waveform with the degraded speech waveform, as shown in Figure 3.12.

There are two waveforms of speech samples are shown in Figure 3.12. The top one is the reference speech sample and the bottom one is the degraded speech sample recorded in the downlink test, cut from a very long downlink speech quality test. As the highlight part of the figure shows, there is a piece of talk spurt missing from the recorded speech sample and the space is filled with silence. It is not exactly sure what happened in that about 1 second long period of time, but it is assumed that the silence is introduced by some network affect. Since the measuring of the performance of the VoIP enabled mobile handset's performance is done

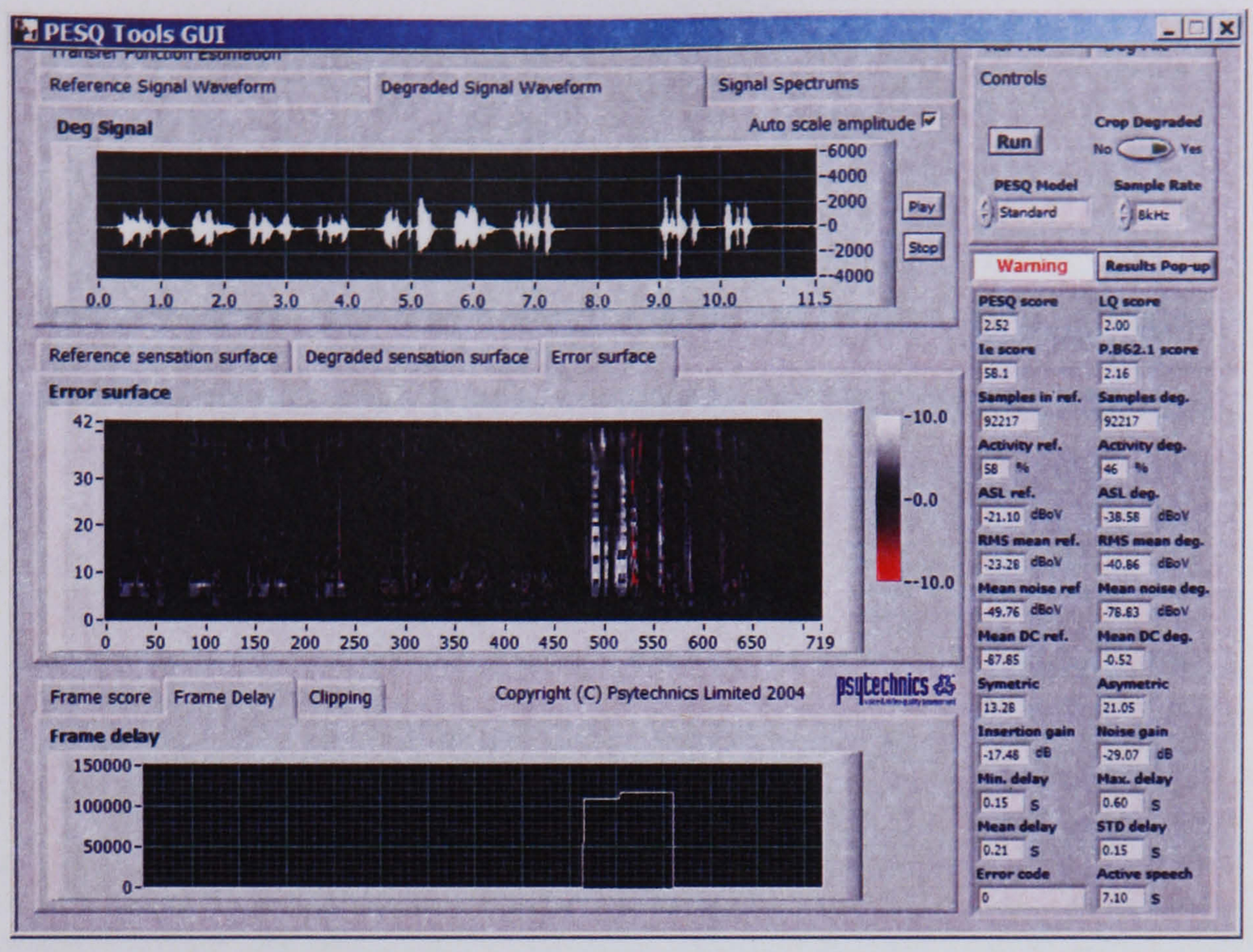


Figure 3.11: Incorrect PESQ reading of relative delay for a sample

on the live network, a short time disturbance in the network, such as a reroute of IP packets due to route change or a congestion caused by a sudden overload of traffic could make this kind of gap in the recorded speech.

By compare the degraded speech sample's waveform to the reference speech sample's waveform, it is clearly visible that there is no major time difference i.e. relative delay difference between those two waveforms. So the PESQ reading about relative frame by frame delay is not correct in this case.

In the scope of ITU-T description of PESQ standard [4], it is mentioned that PESQ is known to provide inaccurate predictions when used in conjunction with extremely temporal clipping, or is otherwise not intended to be used with extremely temporal clipping. The description of the extremely temporal clipping is "Replacement of continuous sections of speech making up more than 25% of active speech by silence". In our test, the speech clip is about 8 seconds long and the temporal clipping is obviously not exceeding the 25% limit. It is still a valid point to note that, in this kind of silence replacement of talk spurt situation, PESQ might give incorrect

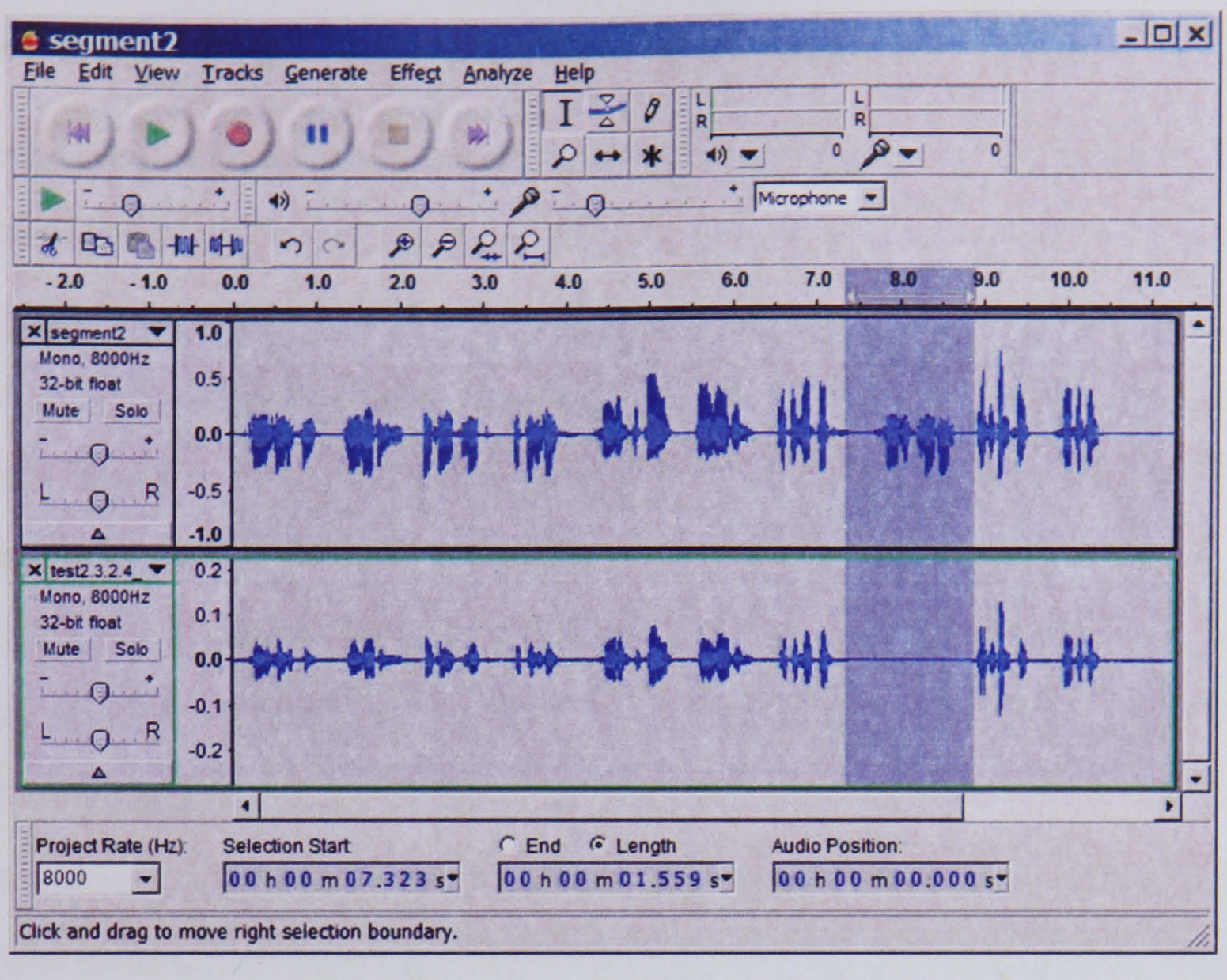


Figure 3.12: Waveform comparison for the same sample

relative delay readings and may also give incorrect PESQ score.

The incorrect result of relative delay as shown in Figure 3.10 is caused by the method PESQ algorithm used to calculated frame by frame delay. In the talk spurt replaced by silence case, the next talkspurt after the silence is incorrectly compared to the existing talk spurt in the reference, and the duration of the silence plus some silence gap between the talk spurts is been counted as the time difference i.e. relative delay.

A similar detailed investigation been carried out for the other delay spikes in Figure 3.10 and the result shows similar behavior. The silence in the wave form lead to PESQ synchronization problem. And thus same incorrect relative delay results are given. In other similar very long speech quality tests, this kind of incorrect reading is rarely seen because live network doe not give out same conditions every time. How ever, in lab environment, this behavior of PESQ is reproducible. To reproduce it, the tester can just edit a speech sample file, replace a talk spurt with silence and compare the edited speech sample with the original reference sample.

Solutions to Avoid PESQ Delay Error Caused by Silence Gap

Since there are 540 individual samples in the long duration call speech quality test and there are only 4 of this kind of incorrect delay readings, and all of the incorrect delay readings are due to missing long time of speech, which is not frequently happening in general VoIP environment, a conclusion can be made that PESQ is reasonably robust for relative delay measurement.

The mechanism to correlate reference and degraded speech sample is inside PESQ algorithm and it is not easy to modify it without change PESQ's over all performance. Because PESQ is the international standard for objectively measure user perceived speech quality, it is not easy to modify the algorithm. And even it is modified for research purpose, to verify the patch will need formal MOS test, which is not possible at this stage.

So two possible solutions to this PESQ incorrect relative delay problem is proposed. One solution is to manually check the delay plot, to find those spikes, investigate as described and remove those errors manually. This method is more reliable but very time consuming.

There is another approach to deal with this kind of rare relative delay errors. Because each sample will be repeated about 20 times in the whole test, it is possible to use standard deviation requirements in the test criteria to mask out those error spikes but still keep the valid over all results. Because this kind of delay spikes are not happening very frequently, masked result is hence reliable and robust.

Not only network events can cause this kind of silence gap but also bugs in the VoIP products can. It is necessary to investigate the VoIP path from sender to received and find out the root cause of the silence gap if it happens more frequently. PESQ is the main focus here so the root cause of the silence gap in this test is not investigated further.

3.4.2 Discovery of the PESQ Measurement Error Caused by Time Shift

Another PESQ error case is presented in this subsection. The PESQ algorithm gives wrong scores in some cases when time shift i.e. relative delay difference presents in the degraded

speech. The test scenario is first introduced. Then followed with the discovery of PESQ's error in the wireless mobile VoIP environment. Detailed analysis then proves and validates the reason of the PESQ error. A brief solution to avoid this error is discussed in the summary part of this subsection.

The voice server can be used for fine tuning of the performance of an adaptive jitter buffer method. It is important to investigate PESQ's performance in such scenario. Here the focus is on the incorrect PESQ score issue caused by PESQ's time alignment error.

As highlighted in the introduction section of this chapter, it is observed that in some wireless VoIP speech quality tests, PESQ scores were significantly lower than the user's opinion (obtained by a simplified MOS [5] test with 6-7 listeners). Table 3.1 summarizes the difference between MOS and PESQ for a few test cases. In the table, test sample number column represents the name of 4 different degraded samples, the names are taken automatically related with time of the test. PESQ column shows PESQ measurement result for these samples. And MOS column shows the mean opinion score of the small group of listeners. This group of samples are using the same reference file.

Test Sample Number	PESQ	MOS
23-38	3.09	3.6
24-37	3.26	3.7
25-46	3.71	3.7
26-09	3.09	3.7

Table 3.1: PESQ score different from informal MOS score

There are 4 samples in the table, PESQ gives 3 of them (1st, 2nd and 4th) abnormal scores. These samples are taken from a group of 60 speech samples in a test. The test is done in a live wireless IP mobile using AMR122 (full rate) and an adaptive jitter buffer. The following investigation shows that the abnormal scores are not correct.

Test Scenario for Adaptive Jitter Buffer Test

As been discussed in previous chapters, delay jitter is more seriously impacting end-to-end speech quality when compared with TDM or traditional circuit switched speech. Adaptive jitter buffer is an important element in the IP based UMA mobile VoIP solution for improving user perceived speech quality for the mobile handset. It can smooth out delay jitter introduced in the packet switched IP network and in the handset's internal process.

To balance better user perceived speech quality with conversational delay, it is important to adjust the buffer delay time accordingly. With the frame by frame delay measurement function, PESQ is used to measure speech quality and delay movement on the same time for the adaptive jitter buffer test.

To test the adaptive jitter buffer algorithms performance in a VoIP enabled mobile handset, a downlink speech quality test with the voice server test platform is established as shown in Figure 3.13. The VoIP mobile makes a call to the voice server via VoIP services, through the VoIP network, gateway to the TDM network and connect to the voice server. The server plays the pre-stored reference speech file down to the VoIP mobile, and the monitor connected to the mobile records the degraded speech. The reference and degraded speech samples are then compared by PESQ and the downlink speech quality score then given.

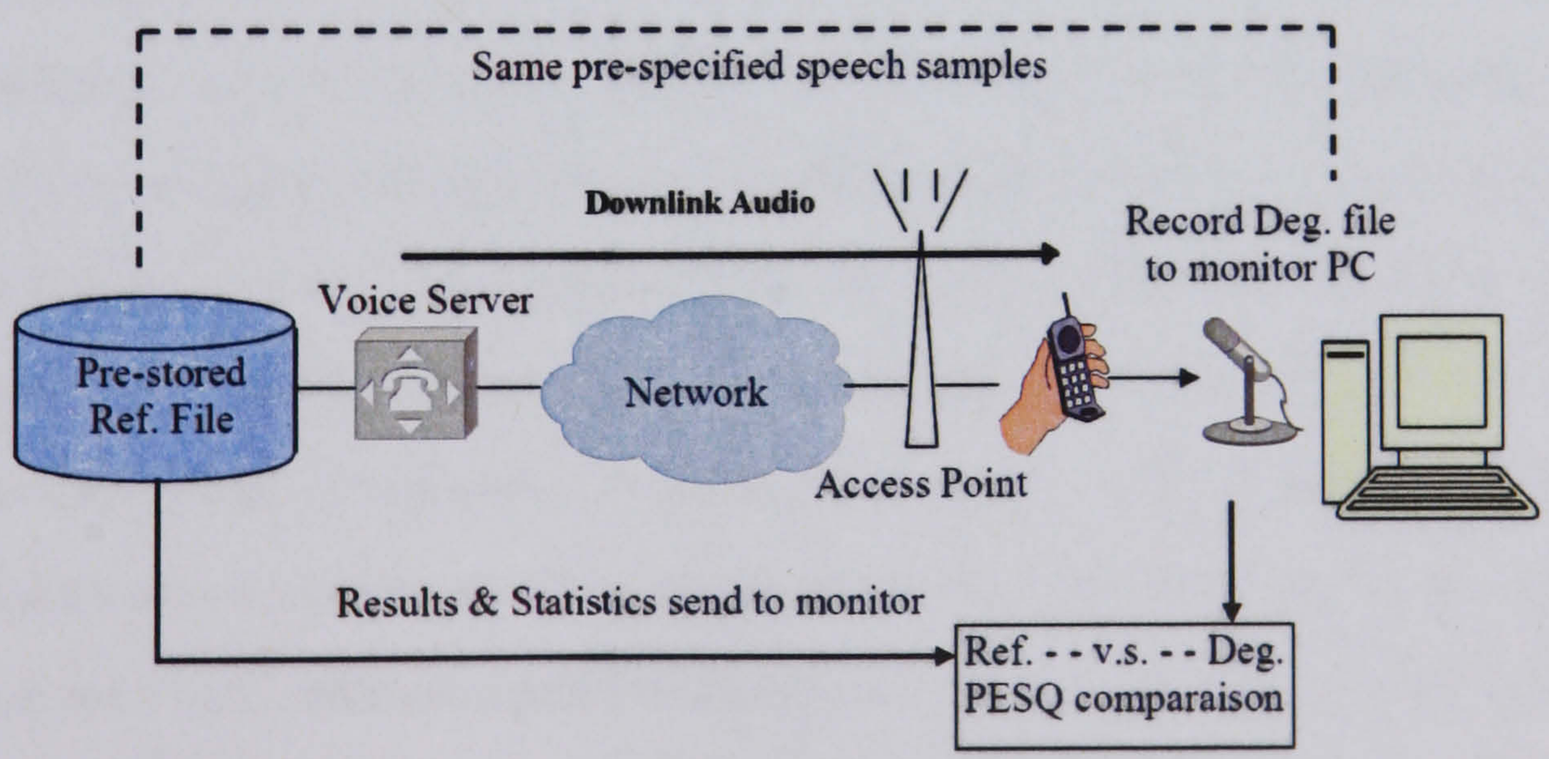


Figure 3.13: Downlink speech quality measurement for mobile handset

With the adaptive jitter buffer working, delay jitter introduced by the IP network and busy processing inside the mobile can be buffered and smoothed. An adaptive jitter buffer can move the buffer delay to a lower number in a low jitter condition to allow lower end-to-end delay and if the jitter goes higher, it is possible to adapt to a higher buffer size with longer delay to prevent more packet losses. By measuring the PESQ score and monitoring the delay, the performance of the adaptive jitter buffer algorithm in the VoIP mobile can be investigated.

But even with an adaptive jitter buffer, there are still delay movements taking place in the degraded speech recorded in the mobile. Because of the jitter buffer size increase or positive adaptation introduced time gap is added into the played speech, the degraded speech could be longer than the reference speech. Alternatively, if the jitter buffer size decrease or negative adaptation introduced time gap is taken away from the played speech, the degraded speech could be shorter than the reference speech. In either case, the PESQ score will be affected. And PESQ is expected to capture the delay movement and on the same time provide valid PESQ measurement to represent the end-to-end user perceived speech quality.

Discovery of PESQ Measurement Error in AJB Tests

In the PESQ measurement process for the downlink speech quality test, it is find out that PESQ sometimes can not give the user perceived speech quality correctly, due to frame time alignment error. Once the frame time alignment is not done correctly, the PESQ measurement result could be obviously different from a informal human listening test result or small scale MOS test result. There are a few sample of this kind collected, for which human listeners give much higher result but PESQ measurement results indicate that their quality are not that high.

To compare PESQ score with user perceived quality, a small scale informal subjective speech quality test is carried out. The comparison results gathered from the small scale MOS test are not used as the definitive proof of PESQ's error but as a starting point leads to a more detailed investigation presented later in this section, involving delay analysis and partial comparison. It is the later investigation proves the PESQ error's existence so the limited scale of the

informal MOS test is not affecting the validation of the discovery and analysis. Due to limited time and funding, a full scale MOS test is not carried out.

This informal MOS test is taken place in our office with 6 ordinary engineers with no audio quality test experience. They are not audio or speech quality expert either. So the informal MOS test result can be used as an initial indication of the incorrect PESQ score problem. The comparison of the PESQ score to the informal MOS score is summarized in Table 3.1. The human opinions to those 4 samples are about 3.7 but PESQ measurements give them a wide variation from 3.09 to 3.71.

Same waveform comparison process carried for the investigation of the incorrect PESQ problem. More detailed analysis shows the reason of the PESQ reading gets wrong is that the time alignment mechanism is not handling some time movement well. So the degraded speech is compared to the wrong place of the reference, hence caused the wrong readings. Figure 3.14 shows the PESQ result plot of frame by frame delay and frame by frame score for sample 23-38 mentioned in the table above.

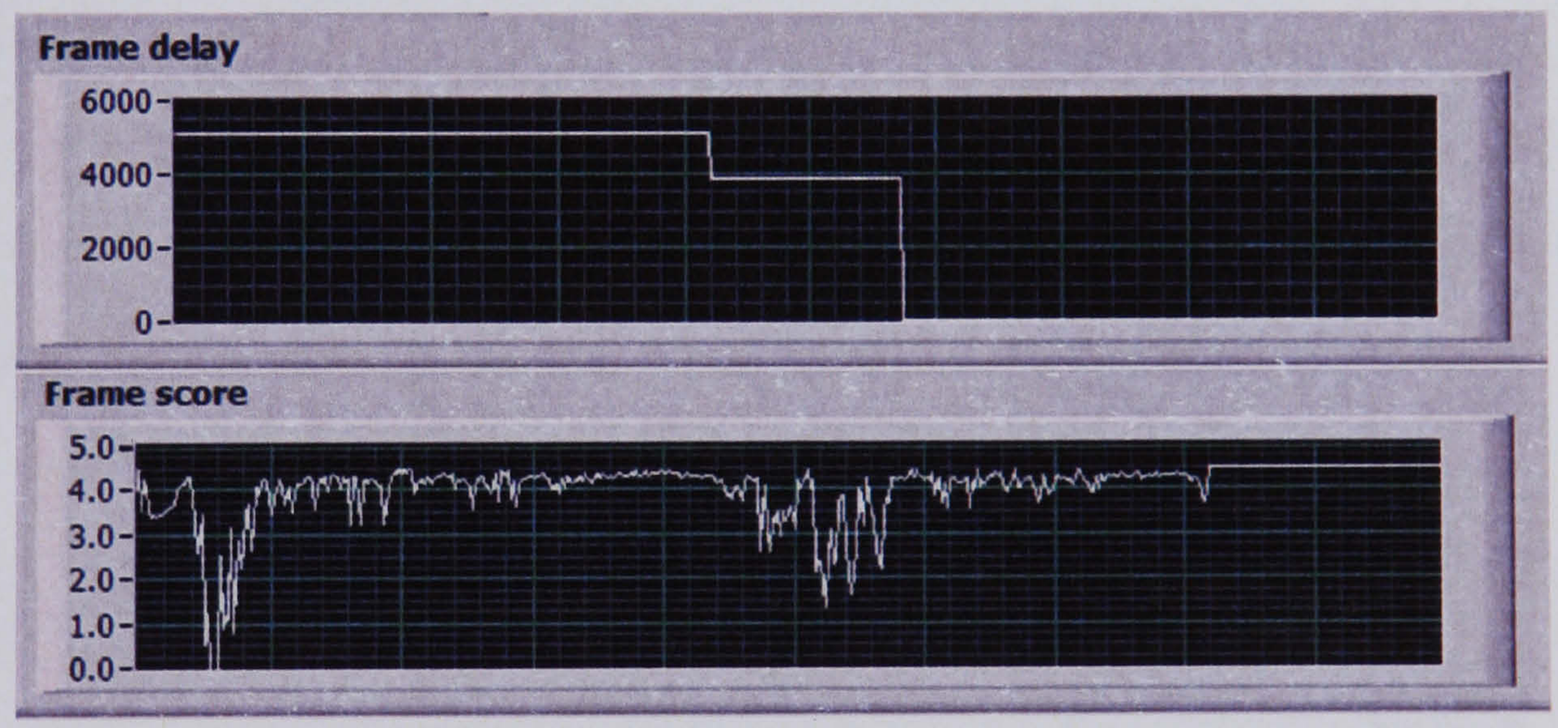


Figure 3.14: PESQ frame delay and frame score plot for sample 23-38

On the top is the frame by frame delay plot shows two negative shifts of about 4 milliseconds and 16 milliseconds. The bottom half of the figure shows the beginning part and the middle part of the speech sample is the worst degraded part of the speech. But after detailed analysis it can be proven that neither the PESQ frame by frame delay nor the frame by frame

score is correct in this case.

From the waveform relative delay analysis, it is clear that the degraded speech waveform has two major relative time movements compare with the reference speech but not one of them is located at the spot the PESQ frame by frame delay plot suggested in Figure 3.14. And the frame by frame PESQ score is not correct neither, caused by the wrong frame delay alignment. The next subsections give detailed analysis about the relative delay and PESQ analysis.

Degraded Waveform Relative Delay Analysis

By compare the degraded waveform with the reference waveform, the relative delay of the degraded speech sample can be identified in the waveform. In Figure 3.15, the first positive time shift is shown in waveform comparison form. The degraded sample, as shown in the bottom of the wave figure, is slightly longer than the reference, as shown on top of the wave figure, and it has to be cut to align with the beginning of the reference speech file. The 8 seconds reference speech file has 64000 samples and the sampling rate is 8 kHz, which means there are 8000 voice samples in a second. A positive time shift is shown on sample number 5656 of the reference. After that point, the recorded speech sample is delayed 20 ms i.e. 160 silence samples are added into the reference speech. That means every speech sample after that point is delayed 20 ms more then samples before that point, until next time shift happens.

Figure 3.15: 1st positive time movement in file 23-38 (at ref. sample 5656, 20ms)

The time shift represents only relative delay changes because the time in the player side and

the recorder side is not synchronized. Recorded speech samples are normally longer then the reference file because the recorder are started earlier then the player starts, to prevent speech been cut at the beginning part of the recording.

Please note the plot shows only a zoomed part of the speech sample and they are head aligned. In this comparison case, 4276 samples from the beginning of the degraded speech file are removed to establish the head alignment for analysis.

In Figure 3.16, a negative time movement is shown as marked. On the bottom of the waveform comparison figure, the degraded speech file has 160 samples relatively less when compared with the reference speech. That means every speech sample in the degraded speech is delayed 20 ms less than the previous part of the degraded speech from this point on, until next time shift.

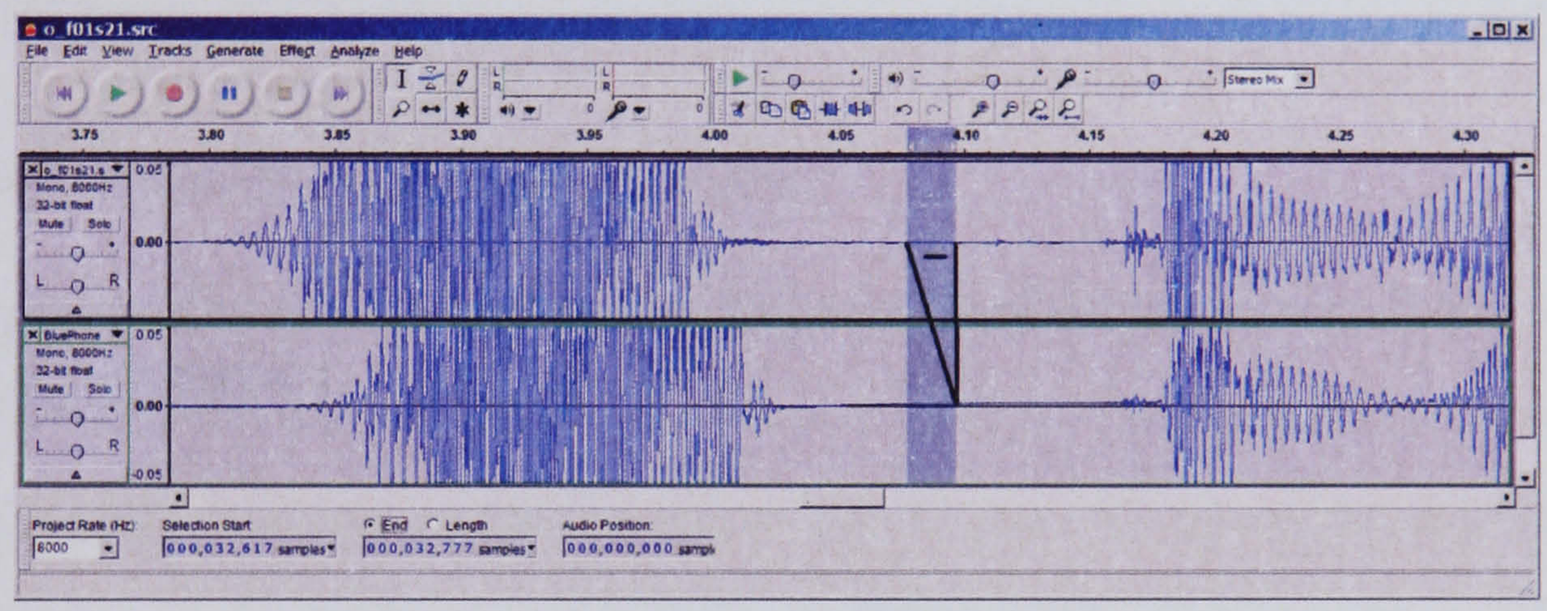


Figure 3.16: 2nd negative time shift in same file (at ref sample 32,617, 20ms)

The real waveform file is too long to be shown in a whole picture so the following Figure 3.17 shows the zoomed relation between the reference and degraded speech file. From the top to bottom are the reference speech file, the head aligned degraded speech file and the real speech file. Because the degraded speech file is longer, the head alignment process is needed by remove the blue part of the waveform. The correct frame delay plot is shown on the bottom of the figure. There is a raise of 20 ms when the gap introduced to the degraded speech (yellow to green) and it drops back by 20 ms when the 20 ms time been taken away from the speech file (green to pink).

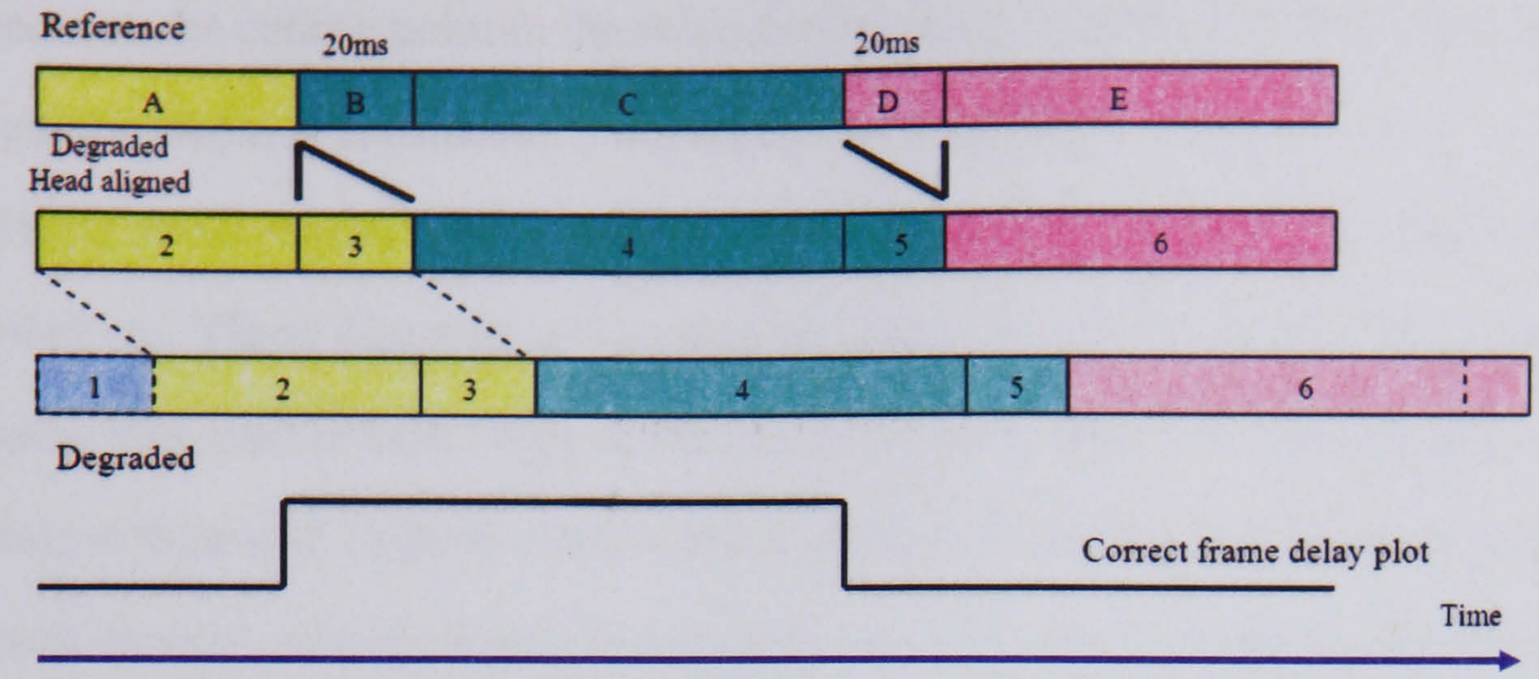


Figure 3.17: Time shift and the correct frame delay plot

Compare with the PESQ frame delay plot as shown in Figure 3.14, it is clear that PESQ's delay alignment process is not working correctly in this case. And this delay alignment error could be the reason lead to the incorrect PESQ score as the informal MOS test indicates.

Delay Shift Caused PESQ Difference Analysis

To prove the PESQ error is caused by delay shift presented above, the PESQ scores for speech with and without the relative delay shift need to be compared. The way to test the impact of the time shift in degraded speech is to cut the reference speech and degraded speech at the place the time movement happens and use PESQ to calculate those parts individually. By cutting the speech sample at the gap, the impact of time shift can be removed from the measurement thus we can compare the averaged PESQ score on cut speeches with the uncut speech's PESQ score. If the cut score shows significant difference when compared with the original uncut degraded speech, it can prove that the incorrect PESQ measurement is caused by the time alignment process.

Since there are two relative delay movements in the degraded speech sample investigated in Figure 3.17, it is necessary to do the cut and uncut PESQ comparison twice. In the positive time movement, i.e. the first time movement event in this case, there is a 20 ms gap added in to the degraded speech. It is necessary to consider the quality impact of this 20 ms in PESQ

measurement so the cutting point in the reference speech is overlapped 20 ms, to guarantee the 20 ms extra waveform is considered in the PESQ measurement.

Two PESQ measurements are carried out on Ref (A+B) v.s. Deg (1+2+3) and Ref (B+C+D+E) v.s. Deg (4+5+6). Those segments and cutting places are shown in Figure 3.17. Reference part B of 20 ms is measured in both cases so the gap of 20 ms's impact on PESQ score is measured but the positive time shift of 20 ms and its affect to the PESQ delay alignment is avoided by the two segment measurement method. As a comparison, the option of take the 20 ms of segment 3 away from both segments i.e. Ref (A) v.s. Deg (1+2) and Ref (B+C+D+E) v.s. Deg (4+5+6) is also tested. The results for PESQ measurement on the two segments cut from the 20 ms time shift place is compared in Table 3.2.

It is clear that both segments have higher PESQ score when compared as a whole speech sample. To include the 20 ms gap in the degraded speech for PESQ measurement or not does not have significant impact on the result. Because PESQ score is an average of frame by frame score, if two segments both have higher PESQ score than the whole speech test result, it is clear that the original whole speech quality test result given by PESQ is not correct.

Case	PESQ	Delay	Notes
Ref. v.s. Deg	3.09	Incorrect	
Ref (A+B) v.s. Deg (1+2+3)	3.35	Correct	Too short
Ref (B+C+D+E) v.s. Deg (4+5+6)	3.79	Correct	Even the 2nd shift problem is gone
Ref (A) v.s. Deg (1+2)	3.35	Correct	Too short
Ref (C+D+E) v.s. Deg (3+4+5+6)	3.79	Correct	Even the 2nd shift problem is gone

Table 3.2: Segment PESQ measurement comparison to avoid time shift impairment

The only difference between the whole speech test and the two segment test is that the time shift is removed from the segments. This is a clear proof that the incorrect low PESQ score, as the informal MOS test indicated, is caused by the time shift in the degraded speech file. The PESQ frame by frame delay alignment mechanism is not performing correctly in this case

caused the incorrect PESQ score.

The negative time shift in the degraded speech file i.e segment D in Figure 3.17 is tested under the same principle. The reference and degraded speeches are cut to two segments in where the 20 ms gap taking place. Because it is a negative time movement in second time shift case in the sample, the degraded speech is overlapped 20 ms in both segments. The first cut segment in reference speech includes part (A+B+C) and it is compared with the degraded segment includes part (1+2+3+4+5). And reference part (D+E) is compared to degraded speech part (5+6). The result is PESQ score 3.10 for the first segment and PESQ score 3.75 for the second segment.

The first segment still contains the time shift discussed above, therefore PESQ score 3.10 for the first segment is proved lower than it should be. Although the second segment has a PESQ warning message of sample too short, as the first segment in the previous test, it is clear in the analysis that the time shift is the main reason for the big decrease of frame PESQ score as shown in Figure 3.14.

To cut the whole speech sample into three segments and compare them separately gives similar result. Other speech samples analysis also confirms the finding that delay movement in degraded speech file can some time causes incorrect time alignment in PESQ frame by frame delay analysis and lead to an incorrectly low PESQ score.

There is no evidence has been found to support the idea that PESQ might get higher scores in some time shifting cases. If the incorrectly higher PESQ score case does exist, the impact to the test result is much lower because the false higher PESQ case population is much lower than the false lower PESQ case, and the possible higher range is much lower than the possible lower range.

Solutions to Avoid PESQ Error Caused by Time Shift

This research has proved that in adaptive jitter buffer test, PESQ can some times provide false measurement results due to false reaction to time shift in the degraded speech file.

The chance of false lower PESQ score is related to the behavior of time shift in the degraded samples and the adaptive jitter buffer algorithm. 1800 individual samples (with different AJB settings) are roughly scanned and the PESQ error rate are widely different between different jitter buffer settings. For one AJB setting, in a group of 60 speech samples investigated in the case study, there are 3 obvious false lower PESQ cases (as in Figure 3.1 shows) and no false higher PESQ case. Some less obvious cases are not counted in but may exists. For other jitter buffer settings, for instance a static jitter buffer, the PESQ error rarely happens because the time shift hardly exist in the degraded sample.

The result of the investigation leads to two improvement approaches. One is to manually investigate degraded speech samples and remove the incorrect cases accordingly. This approach is more accurate but very time consuming. The other approach is to use average and standard deviation to filter out odd cases and still get general idea about over all performance of a big group of tests. These two approaches can be selected to fit different test applications in different scenario.

Alternatively, it is possible to improve the time alignment mechanism in the PESQ algorithm and increase the accuracy in frame by frame delay measurement and solve the problem from the root. But that is out of the focus of this thesis and any findings will need to be evaluated by sophisticated MOS test before it can be contributed to the ITU-T for the new standard.

3.5 Summary

This chapter introduced the live VoIP mobile speech quality test platform i.e the voice server and based on it the user perceived speech quality test platform, especially the usage of voice server for objective evaluation of mobile speech quality. Some cases have been investigated in this chapter as well to highlight the technical details need to be focused when doing live speech quality measurement for handsets including some PESQ performance case study.

The contribution to knowledge includes the following:

(1) A couple of PESQ error cased are discovered and investigated in this chapter. A couple of case studies have been carried out to investigate some PESQ defects in certain test cases. The result shows PESQ will give incorrect frame by frame delay measure if a period of silence replacing an active speech part. And the time alignment error can cause PESQ error in some cases as well.

(2) Voice server speech quality test platform, especially the real mobile VoIP integration part and the live mobile VoIP connection part, and the detailed calibration process, includes the virtual disk solution to read/write caused gap in record problem, the resister network solution for the sound card cable to mobile microphone connection to solve the voltage mismatch problem and the best volume setting test solution to solve the play out volume affect PESQ score problem are presented in this chapter. It builds up the basis for further VoIP QoS research and solved some calibration problems.

This chapter shows the performance of the live VoIP mobile speech quality test platform and its application for PESQ performance evaluation. The investigation suggests that detailed calibration and performance investigation is needed when use PESQ for field VoIP mobile speech quality test.

Chapter 4

Perceived Quality Enhancement with a New Jitter Buffer Algorithm

As VoIP getting more and more popular in the telecommunication industry, mobile handset producers are more aware of the importance of integrating VoIP applications into mobile phone handsets. Industrial leaders and international standard bodies have developed an architecture frame for Unlicensed Media Access (UMA).

The UMA framework not only specified the architecture for integrating the network infrastructure to support VoIP services which include the core network IP support, gateway support and Access Point support, but also specified the architecture to integrate the wireless VoIP services into mobile handset. UMA is not the only solution to integrate VoIP into mobile handset and there are other methods vary from software SIP agent running on a handhold WiFi enabled pocket PC to voice over wireless broadband. This project is focusing on UMA implementations as it is widely supported by both mobile handset manufacturers and network operators and this project has access to a UMA platform on which solutions can be tested in a live network. The research results are not limited to UMA and they can provide useful improvements to other VoIP applications including SIP clients or VoIP gateways as well.

Jitter buffer plays an important role in VoIP applications because it provides a key mechanism for achieving good speech quality to meet technical and commercial requirements. The main objective of this chapter is to develop a new, simple-to-use jitter buffer algorithm as a

front-end to conventional static or adaptive jitter buffer algorithms to provide improved performance, in terms of enhanced user-perceived speech quality and reduced end-to-end delay. Supported by signal processing features, the new algorithm, the so-called Play Late Algorithm (PLA), alters the playout delay inside a speech talkspurt without introducing unnecessary extra end-to-end delay. The results show that the new algorithm achieves the best performance under different network conditions when compared to conventional static and adaptive jitter buffer algorithms. The results reported here are based on real mobile phone prototypes tested on live and emulated network conditions. The mobile phone uses AMR codec and supports full IP/UDP/RTP stack with IPSec function in some of the tests. The method for perceived speech quality measurement is based on the ITU-T standard speech quality evaluation tool (PESQ).

The remainder of the chapter is structured as follows. Section 4.1 introduces necessary background of this chapter. In Section 4.2, the performance of a number of leading VoIP services and WiFi/SIP VoIP hardware are characterized using an objective voice quality measure to establish an over view of quality problems inherent in existing VoIP solutions. In Section 4.3, the new proposed Play Late Algorithm to address some of the problems is presented. In Section 4.4, detailed tests taken on prototype mobiles and live networks are shown. Test results are presented in Section 4.5, together with some discussions. Section 4.6 concludes this chapter with discussions of some possible future developments.

4.1 Introduction

VoIP is a developing technology and it have certain limitations. Among those limitations, the speech quality issue is a very important one. Because the packet transmission nature of the IP network, speech quality is not guaranteed in the VoIP system. Currently there are lots of research activities working on optimization of speech quality in VoIP systems. The portable character of mobile handset restricted the VoIP applications' environment to wireless VoIP and the platform of running those VoIP applications is limited to a not so powerful handset

environment. The embedded mobile operation system and the supporting hardware can not achieve the almost unlimited processing power when compare with the stationary computer, which might have much higher IP network bandwidth as well. This makes the mobile VoIP environment a more challenging field for speech quality optimization.

To take the specific challenge introduced with the wireless VoIP environment, the wireless or mobile VoIP solution's performance need to be measured and optimized towards the overall user satisfaction. This chapter is focusing on the measurement and optimization of VoIP speech quality in UMA wireless mobile environment.

As discussed in Section 2.3, packet loss, delay and delay jitter are the main network impairments that affect user perceived speech quality in a VoIP system. Jitter buffer plays an important role in VoIP system as a means of compensating for the impairment introduced by jitter. A tradeoff between increased packet loss rate and jitter buffer delay is necessary for any playout buffer algorithm [90]. The longer the buffer delay the lower the packet loss rate (packets dropped by the jitter buffer are considered as contributing to the overall packet loss) and vice versa.

In practice, jitter buffer can locate before the depacketization process or after it but before the codec. Or there could be more than one jitter buffer in the receiver side. Our focus in this chapter is on the jitter buffer before codec. Any other jitter buffers before this point can be considered as an network element even sometimes they could be located in the mobile handset side of the connection.

Figure 4.1 shows the concept of jitter and jitter caused by packet loss or drop by the jitter buffer. In Figure 4.1, Tx line represents the time line of transmission time of packets 1,2,3,4,5 and 6. Packet 1 send time is denoted as t1 in the figure and similar for other packets. Rx line represents the time line of packets 1,2,3,4,5 and 6 received time by the jitter buffer. Packet 1 received time is denoted as r1 in the figure and similar for other packets. Playout line represents the actual play out time for each packets. The play out time for packet 1 is denoted as p1 and similar for subsequent packets. As shown in the figure, packets 3 and 4 arrive after their playout

time denoted as p3 and p4, and are dropped by the jitter buffer. This is a typical jitter buffer reaction for very late packets.

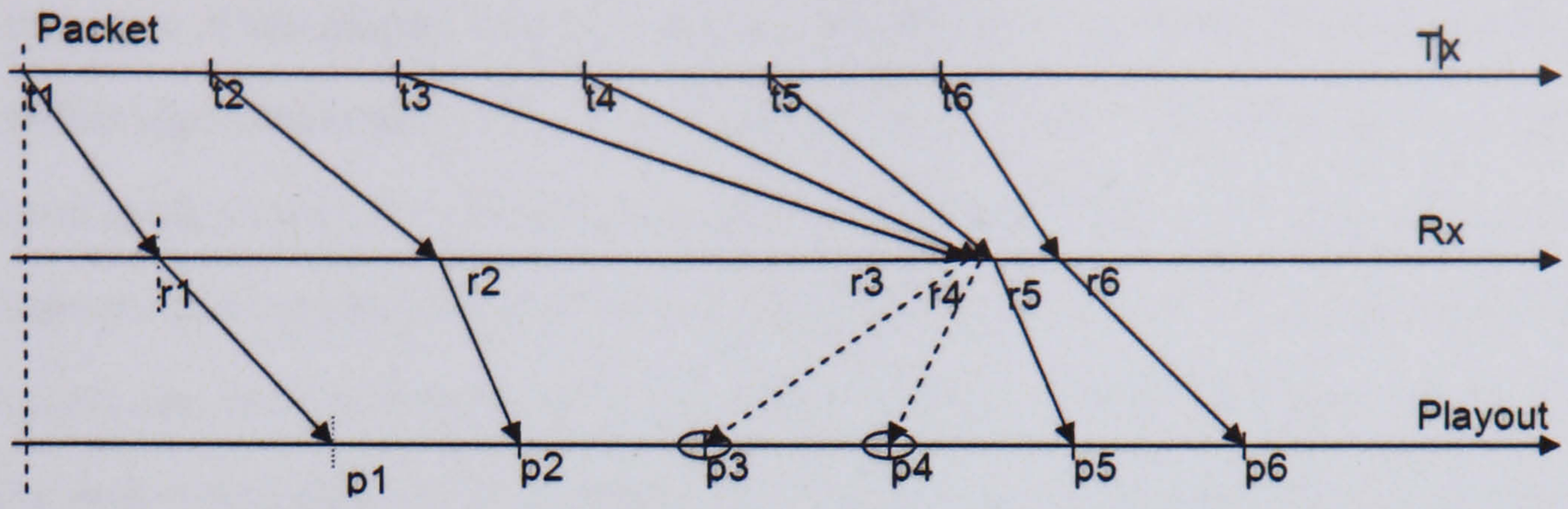


Figure 4.1: Tx, Rx and Playout time in a jitter buffer example

A characteristic feature of human speech, especially during speech communication through a mobile link, is that there are active talk spurts and silence time periods. When a person is not speaking but listening to the other party, the time is count as silence time. There are also gaps between sentences and even words can be counted as silence time [91]. Detailed active silence identification is not in focus of this research. Because of the active and non-active speech period interleaved behaviour of speech conversation, it should be possible to improve the performance of current jitter buffer algorithm, in terms of overall user perception, by making increases to the relative delay inside an active talkspurt or reduce it in silence period.

A number of jitter buffer algorithms have been developed to reduce the impact of delay jitter [35, 90, 92, 93]. Some of the algorithms focus on the adaptation of the buffer size based on the time of arrival statistics [92], but these do not consider the talkspurt character of human conversational speech. Some algorithms adjust the playout time at the beginning of the talkspurt (such as [90, 92, 93]), but these use fixed delay settings in a talk spurt and so they may not react well to delay spikes. Other approaches involve the modification of the wave form inside a speech frame, [35], which is computationally expensive. A common limitation of most existing algorithms is that they do not take into account user-perceived quality. In addition, they are based on simulation results and not on speech data from live mobile networks. The processing power requirement is relatively high in some of those methods.

To implement wireless VoIP in mobile handset has more challenge than ordinary VoIP on stationary IP networks because the resource restrains in the mobile handset environment. The main objective of this chapter is to present a new method to improve the performance of current jitter buffer algorithms and to provide an algorithm which can produce better end- to-end user perceived quality based on adapting the relative delay time inside a talkspurt efficiently. The new improvement method should consider not only improvement of user perceived speech quality, but also the implementation restrains in the specific mobile handset environment.

The main contribution of this chapter is to propose a new adaptive jitter buffer algorithm which can be used on top of many conventional static jitter buffer or adaptive jitter buffer algorithms to provide a better end-to-end user perceived speech quality measured by the ITU-T standard P.862 and P.862.1 i.e. PESQ and PESQ_LQ without adverse impact on the end-to-end delay and conversational speech quality and on the same time easy to implement and run in a existing mobile handset system.

Tests have been carried out using real prototype mobiles running over commercial and partly emulated IP networks. The results show that the new Play Late Algorithm (PLA) can provide better overall user-perceived speech quality without excessive impact to the end-to-end delay.

4.2 Jitter and Jitter Buffer's Impact on Existing VoIP Solutions

It is important to understand the issues before a solution can be find for it. In the VoIP market, there are a few VoIP services available, including some wireless VoIP solutions [94]. This section is developed to have an overview test on them and gain some understanding of their current performance in terms of end-to-end user perceived speech quality in reaction to network conditions such as delay jitter. As discussed in the introduction, jitter plays the most important role in network impairment and this research is focusing on improvement to jitter

buffer element leading to optimized speech quality so the speech quality is tested v.s. delay jitter. And the test results are discussed in Section 4.2.1. The following Section 4.2.2 discusses some current jitter buffer algorithms.

4.2.1 Characterizing Jitter's Impact in Current VoIP Products

A number of popular commercial VoIP equipments and services were selected and tested to characterize speech quality in current VoIP products. The VoIP services tested cover key VoIP products and include a popular VoIP service that provides free computer to computer calls using P2P technology and a popular VoIP chat software that provides free computer to computer calls. There are two hardware VoIP phones on the market tested as well, both of which use WiFi and SIP.

The objective in this section is to characterize the speech quality of the VoIP products and services in terms of user perceived quality. An objective measure of speech quality was obtained using PESQ. Different network conditions that range from clear live network to emulated high-jitter conditions were tested to provide a broader view of the overall performance of the services and products under different jitter conditions. The live network was used whenever possible to make the test as realistic as possible, this means for example the caller and the called party are not connected on the same switch but located in different labs and connected via a few IP hops.

Figure 4.2 depicts a conceptual diagram of the scheme for evaluating the speech quality of the VoIP products or services. The voice samples are transmitted through the selected VoIP softwares and the received speech compared with reference samples using PESQ to provide an objective intrusive measure of voice quality. A network emulator NISTnet is used to introduce different jitter conditions and emulate different network environment. It can add specified jitter and other network impairments to the IP traffic to emulate different jitter conditions. Considering the live network conditions are roughly the same for each product under test, the difference between each product under test should come from the extra jitter added. The

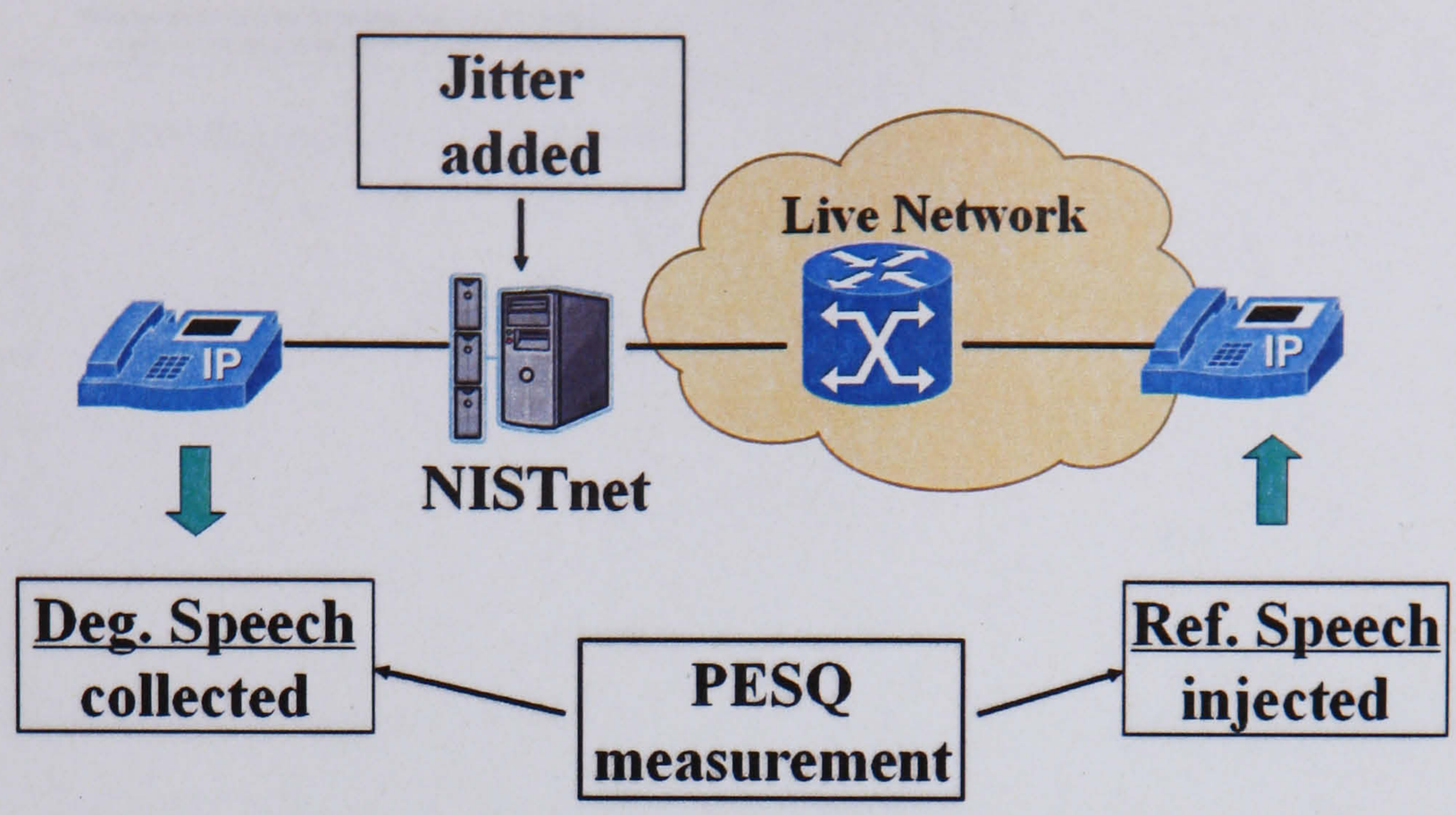


Figure 4.2: Evaluation of jitter's impact to current VoIP products

difference between live network condition and other systematic difference between products are not controllable in this setup so they are not considered in the test. This does not affect the aim of these tests, which is to prove the importance of jitter buffer in a VoIP system.

A similar method is used to test WiFi/SIP based VoIP handset hardware, the only difference being that an audio cable is used to connect the hardware to PC sound card so the speech samples can be played to and recorded from the VoIP phone.

To avoid commercial and legal issues, it is not appropriate to refer directly to the VoIP products or services analyzed here, but in stead they will be referred to as products A, B, C and D.

Figure 4.3, Figure 4.4, Figure 4.5 and Figure 4.6 give plots of PESQ scores on different jitter conditions for products A, B, C and D, respectively. As can be seen in the figures, some current products or services provide excellent quality under no jitter conditions, but can not cope with jitter; others provide acceptable quality over a wider range of jitter conditions.

Figure 4.3a shows the PESQ score for a 20 minute call for product A. In the figure, each point represents a speech sample of about 6-9 seconds long. In this case, no jitter is added in the IP path and so the jitter readings on the X axes are kept at 0 ms. To be consistent to

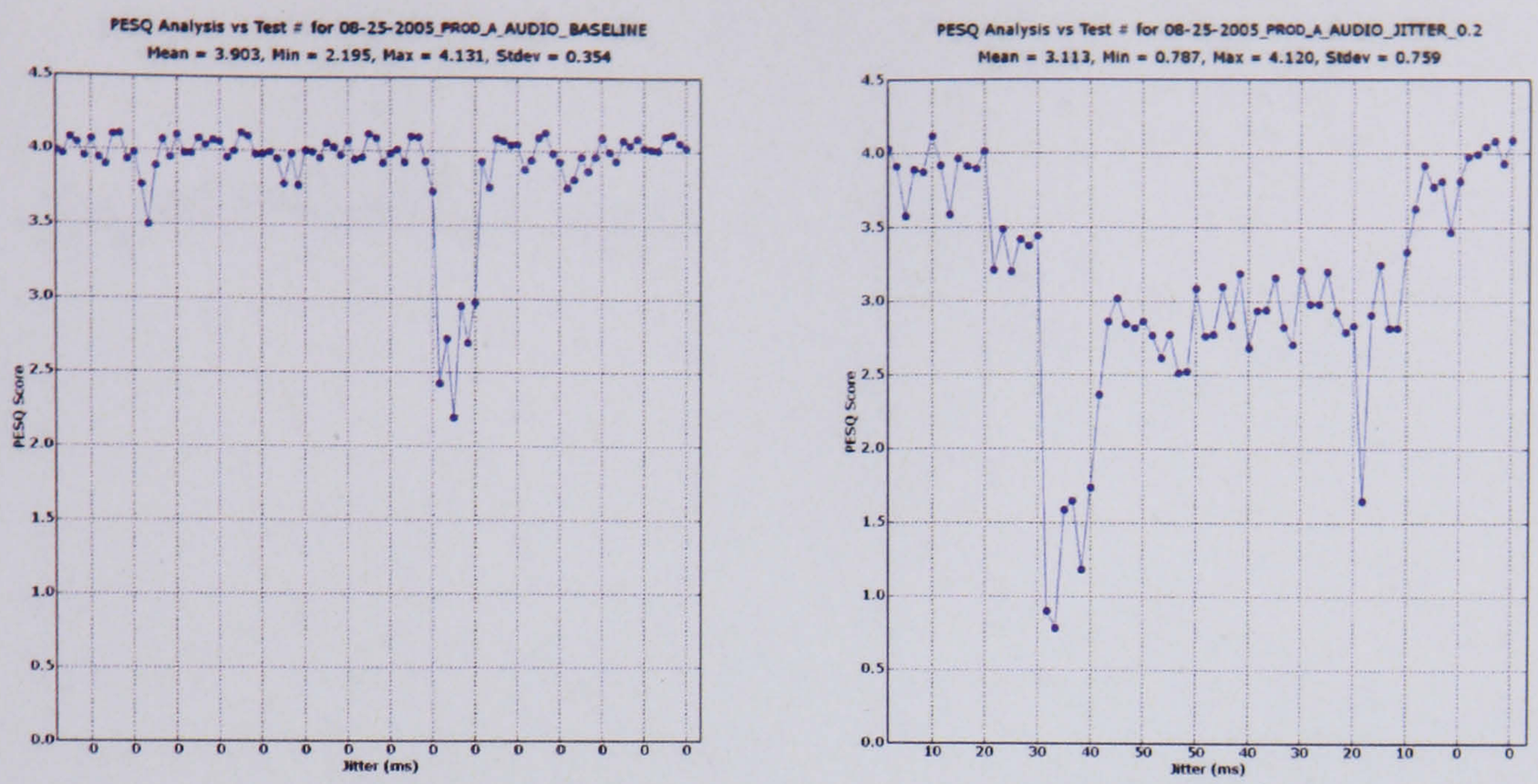


Figure 4.3: PESQ score for a) no jitter live and b) jitter conditions - product A

figures showing jittered cases, the X axes shows jitter conditions instead of case number. The PESQ values show that the average quality of product A is very good (PESQ score of about 4), although it has a period of bad quality in the middle of the call which lasts for about one and half minutes. The bad quality could be caused by some Internet disterbance or processing power over load in the operation system. But as the figure shows the product recovered from the bad situation and the quality of speech comes back from the low points.

Figure 4.3b shows the PESQ v.s. jitter plot for another 20 minutes call for product A, but this time some jitter is induced by NISTnet. As can be inferred from the X axis the jitter is increased from 0 ms to 50 ms in steps of 10 ms, and then reduced to 0 ms from 50 ms. The speech quality recorded shows a reasonable response and PESQ score of the jittery condition stays at about 3 apart from some extreme cases where the PESQ score is recorded as low as 1, which means in that 6 to 9 seconds sample, the user might think the conversation is meaningless. Over all, this product has a good average quality and responds reasonably well to jitter, but it does not appear stable enough.

Figure 4.4a and b show the performance of Product B under similar test conditions as Product A. The results indicate that on no injected jitter condition, the average speech quality

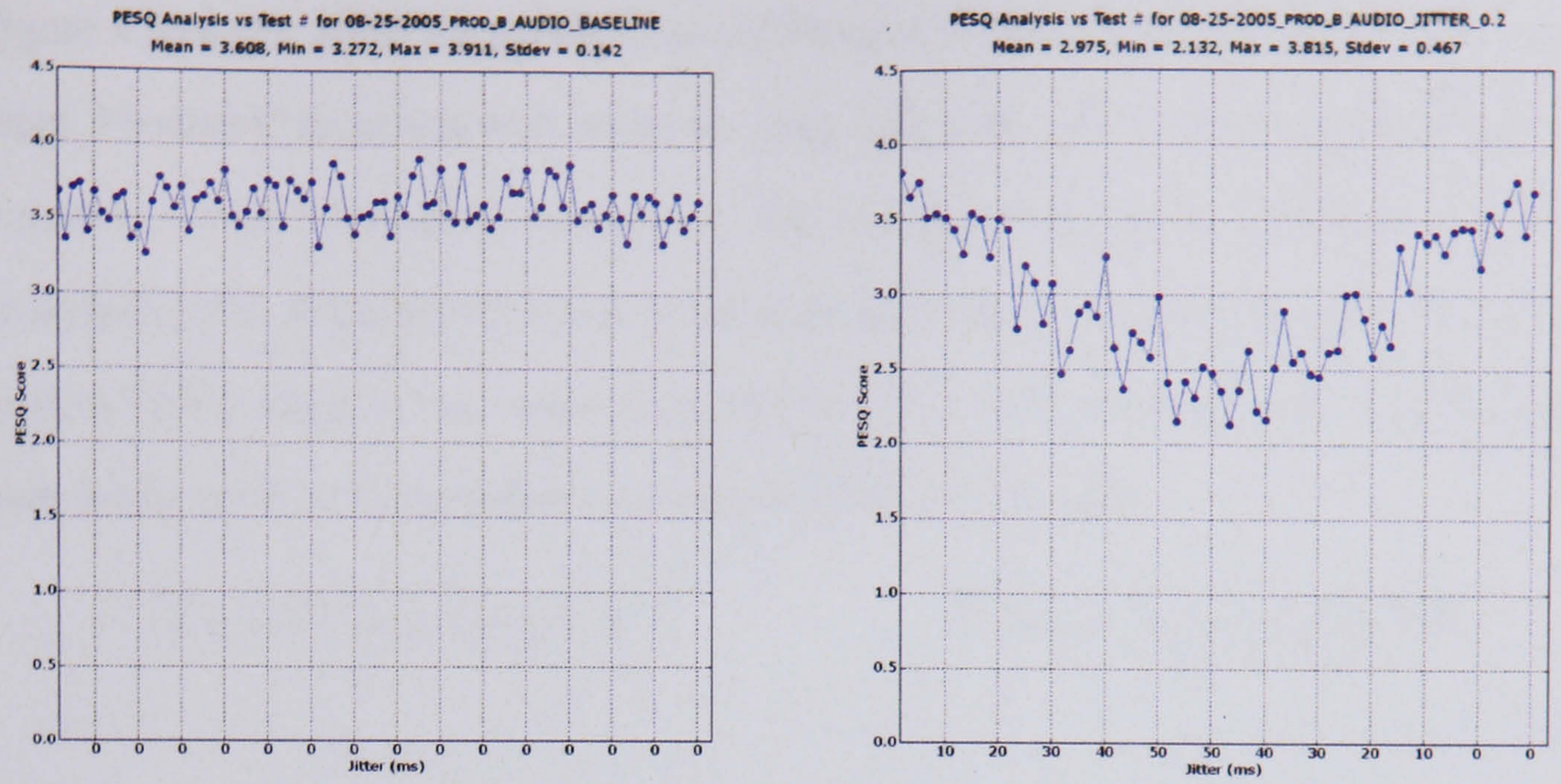


Figure 4.4: PESQ score for a) normal and b) jitter condition - product B

for Product B is not as good as Product A. The PESQ score for Product B is stable but averages at a lower point, about 3.5. The response of Product B to jitter, on average, is more stable than Product A, but the average speech quality is slightly lower and it seems more sensitive to jitter conditions as the curve changes more when different level of jitter is introduced. The lower average speech quality scores may be due to the use of a different codec or a different buffer algorithm.

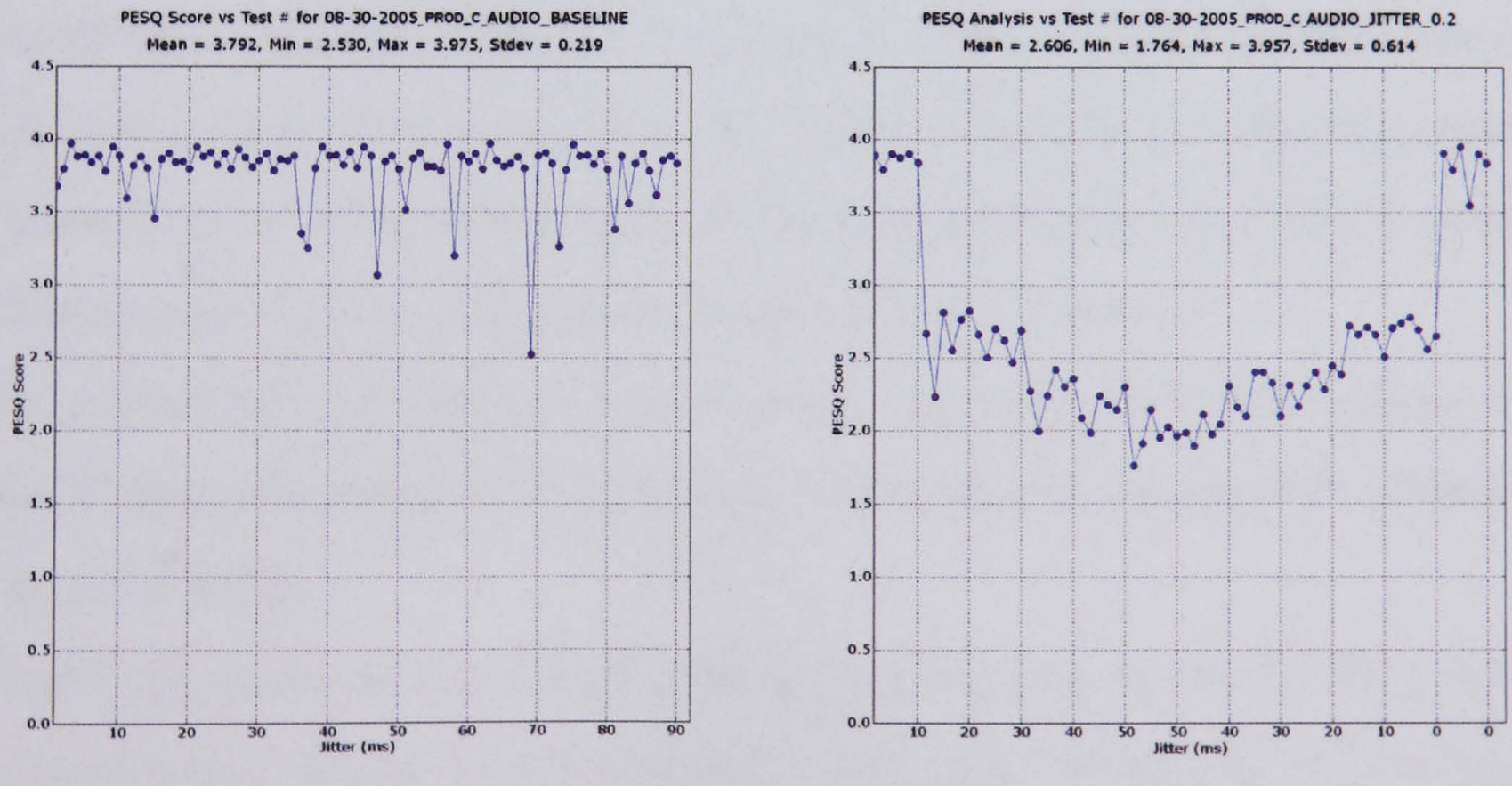


Figure 4.5: PESQ score for a) normal and b) jitter condition - product C

Figure 4.5a and b show the performance of Product C under similar test conditions to previous tests. Product C responds well under no jitter conditions, gives a very good average PESQ score near to 4 with a few quick lower spot. But as Figure 4.5b shows, Product C does not respond well to jitter. With even a low level of jitter introduced, the speech quality measurement dropped to PESQ score 2.5 or lower very quickly. This behaviour indicates that the length of the jitter buffer used in this product may not be sufficient enough.

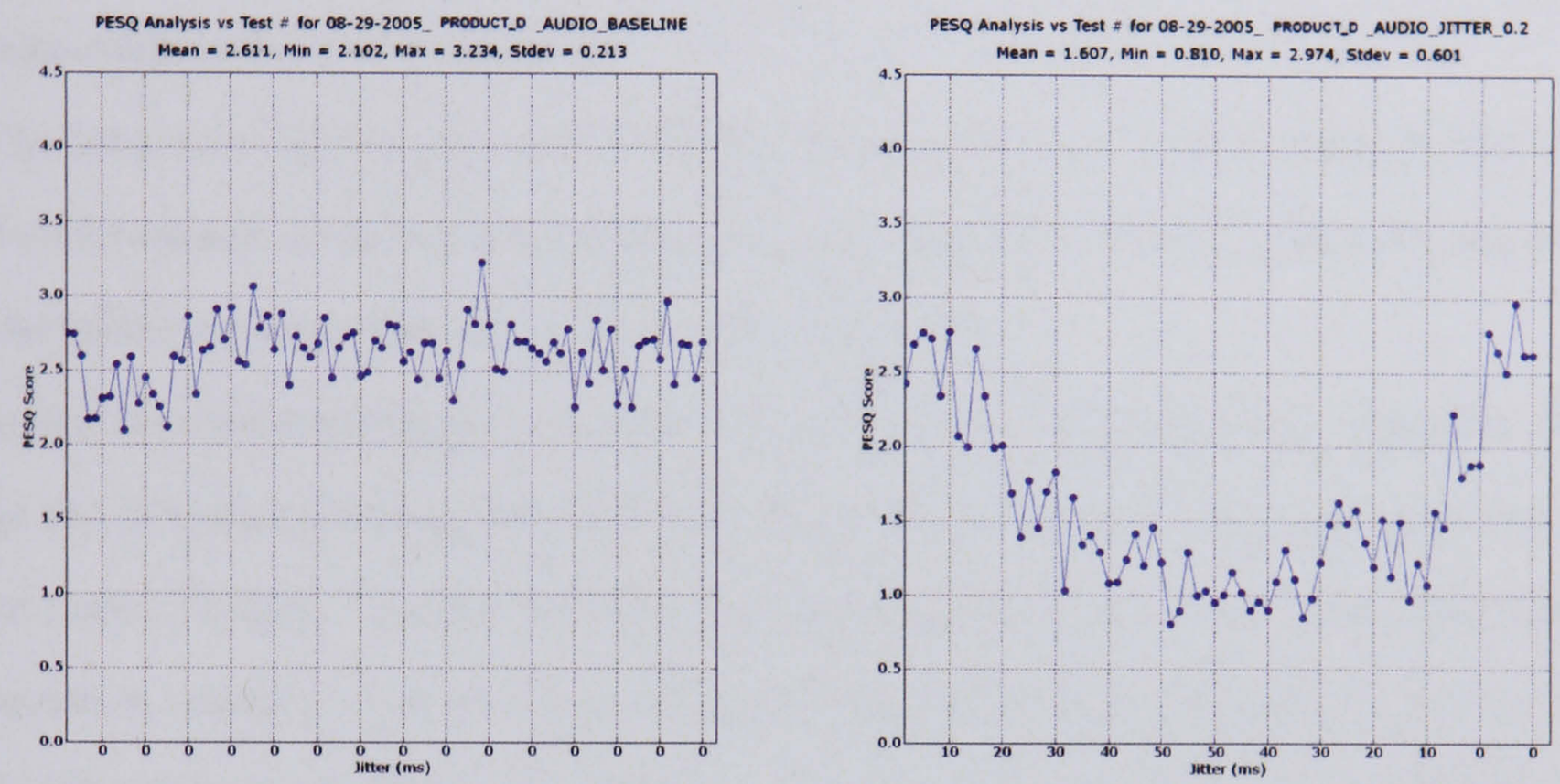


Figure 4.6: PESQ score for a) normal and b) jitter condition - product D

The PESQ v.s. jitter plot of Product D as shown in Figure 4.6a and b provides neither good average quality nor good response to jitter. The average of PESQ score on no jitter condition is only at about 2.5 and when some jitter introduced, the speech quality level fells to as low as 1, which means this service or product is not usable on jittery conditions.

The problem here could be due to a poor quality codec, poor jitter buffer designs or other reasons. Clearly, this product is not sufficiently well developed and users will not be satisfied with the voice quality.

Overall, the results indicate that jitter buffers play an important role in VoIP systems and more improvements needed in current products. Delay jitter behavior are analyzed in a typical mobile VoIP environment and some jitter buffer algorithms are discussed in the following

Subsection 4.2.2.

4.2.2 Analysis of Jitter and Jitter Buffer Algorithms

Jitter buffers are designed to trade off between end-to-end delay and packet loss or drop. Longer delays mean higher possibility of echo, longer response time, and lower conversational quality. Higher packet loss or drop rate means less information being transmitted and lower listening-only user perceived quality.

The behavior of jitter in the network differs in characteristics. In [95], traces in live network were collected and analyzed, and patterns of jitter classified. Normally, they are not random, and the spikes are more difficult to detect and to deal with.

In the telecommunication environment human speech has some special features (for instance the frequency band is from 300-3400 Hz). These features have been exploited to optimize codec designs. Certain features of human speech are not continuous especially in a conversation scenario. For example, the active rate is normally between 25-75% as ITU-T P.563 defined the limitations [96]. When not in active period, there is no need to transmit silence to the other end. Thus, dynamic transmission (DTX) method has been developed to avoid the transmission of silence. In the AMR codec [8], a function called SID (Silence Insertion Description) is used to encode only noise level in a smaller frame and to send the SID update less frequently than normal active speech frame. These functions can be utilized by jitter buffer algorithms to improve the overall user perceived quality. Jitter buffer is an important source of loss because it will drop packets if the packets are late.

To deal with packet loss from the network or packet loss due to drops by the jitter buffer, state of the art codecs and the use of DSP provide functions that make it possible to reproduce lost or dropped packets by replaying some of the previous packets or by using other more complicated methods. As the information contained in the lost or dropped packets is not available to the DSP functions, concealment methods introduces a distortion to the end-to-end listening-only quality the effect of which can be measured by PESQ.

Although jitter buffer algorithm affects the speech quality, it takes effect by changing the delay and loss rate of the stream. To evaluate the performance of a jitter buffer, the delay and loss rate needs to be considered. A test has been carried out to show the results of PESQ values v.s. packet loss rate for the AMR codec (AMR122 with simulated random NISTnet packet loss rate). Figure 4.7 shows different AMR codec rates' response to different packet loss rate.

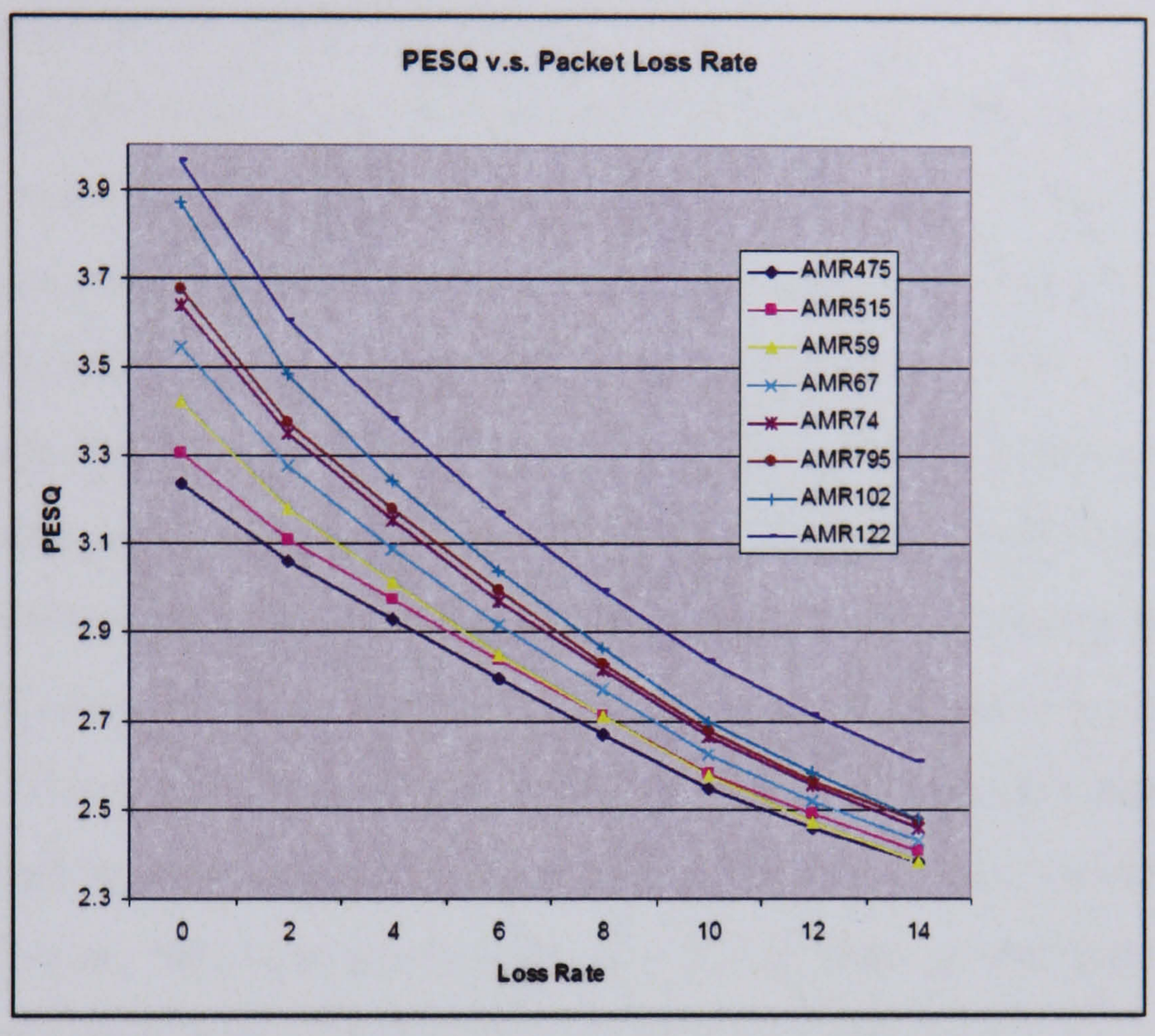


Figure 4.7: PESQ drops when Packet Loss Rate increases

Each of the point in the figure represents the average of 20 test measurements. Each measurement is one PESQ reading on one speech sample going through the coder, network loss simulation and decoder process. There are some gender difference between male speech samples and female speech samples. Normally male speech samples have higher PESQ score on the same packet loss rate. But considering the statistic of caller should have no major gender difference, 10 male and 10 female speech test are averaged together. The reason to test 20 times is to reduce the random effect of the random packet loss because the location of the loss may cause some difference of PESQ result. Those 20 tests will have different packet losses

on different place of the speech stream thus average the loss locate spread. As can be seen in Figure 4.7, packet loss rate or the amount of information lost is closely related to the user perceived quality.

Researches on jitter buffer, including [95], show that the behavior of jitter in the network varies significantly. Thus, the jitter buffer needs to cope with these variations in order to deliver optimized end-to-end user perceived quality.

Jitter spikes [35] are difficult to detect or treat because the jitter buffer normally uses statistical methods to determine buffer delay. Thus, if a short or single spike occurs, the jitter buffer normally cannot make a clear decision whether to discard the single late packet as an exception or to extend the buffer delay to try and capture more of this kind of packets.

An example of such a jitter behavior is shown in Figure 4.8. As the figure shows, there is normally a packet in the stream that is delayed for a long period which prevents a group of later packets from being delivered on time. In this figure, audio frames are transmitted twice as a redundant FEC method each 20ms so the delay between each point should be 0-20-0-20-0-20-0 and so on. This pattern means there is no delay between the 1st and 2nd packets (they are duplicates), then the 3rd and 4th packets transmitted at 20ms later and then 5th and 5th at 20ms after that and so on. But as pointed out above, once a spike occurs, it delays later packets for 168ms leading to more than 10 packets arriving almost at the same time.

This behavior is caused by the operating system of the device which can not schedule jitter buffer related thread when it is needed. It is mostly due to lack of processing power in the device and thus less obvious from traces taken by powerful computers. Please note a temporally over loaded high powered device could cause the same affect to jitter. Our new algorithm can deal with this type of jitter pattern better and more efficiently than traditional jitter buffer algorithms. This will be demonstrated in section 4.5 in details including a wave form analysis from a live mobile call.

A brief comparison of the concept of a number of state of the art jitter buffer algorithms and our new algorithm is presented here.

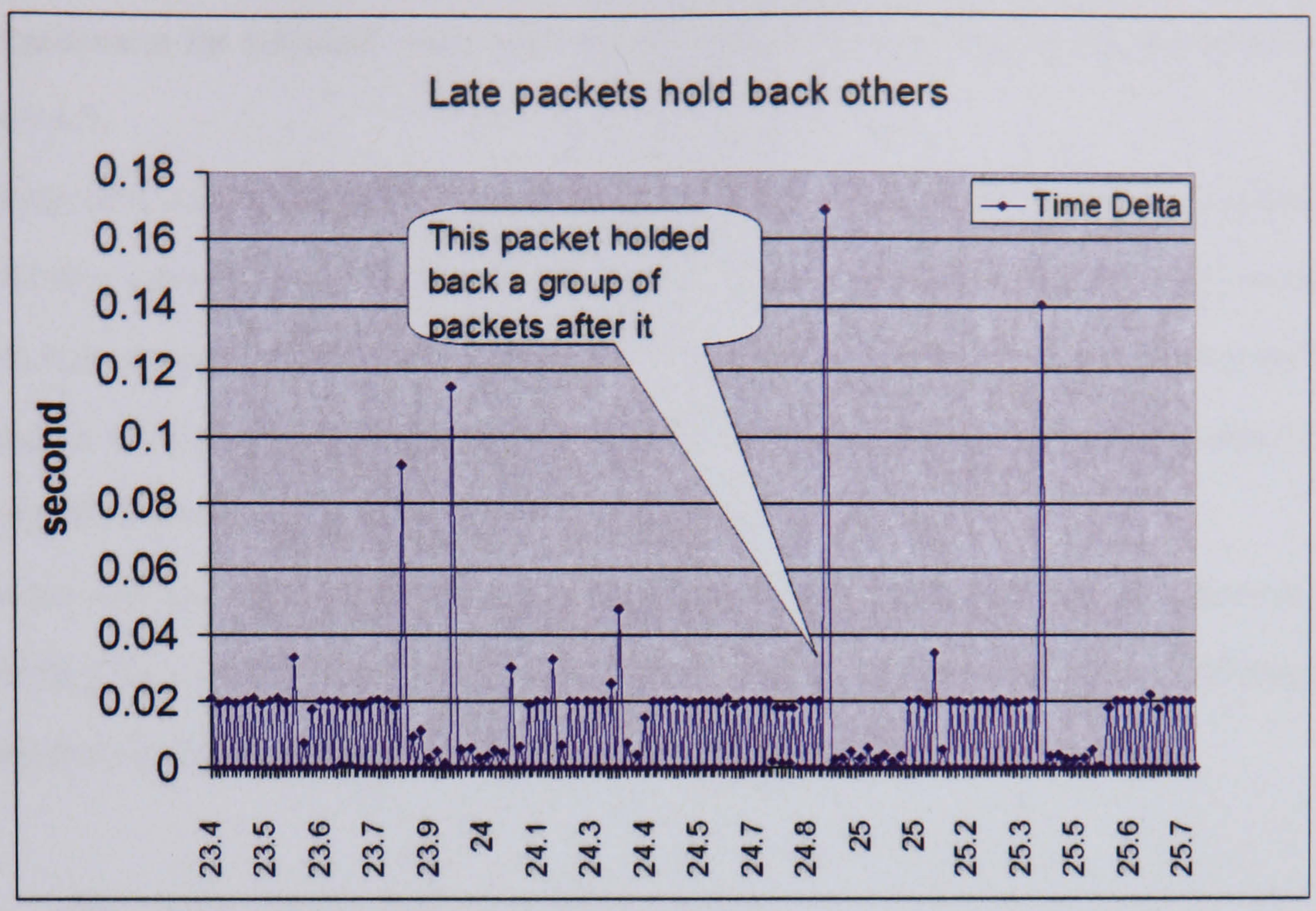


Figure 4.8: Relative delay plot shows late packets hold back later ones

Reference [97] describes a jitter buffer algorithm which can increase or decrease the jitter buffer size based on the played status, but the algorithm does not consider SID and no-SID adaptation. When packets are discarded, some jitter buffers do not do so during silence period but in stead drop the packets at any point and this may cause unnecessary additional distortion to speech quality.

Reference [98] describes another buffer algorithm with buffer size adaptation. Compared to our algorithm, when packets are discarded, this algorithm does not consider discarding during silence periods and so packets could be dropped during talkspurt and would cause higher distortion.

The adaptive jitter buffer algorithm described in [93] inserts frame only during silence period but our method inserts duplicated frames during speech if it is necessary.

Paper [35] introduced an adaptive jitter buffer scheme. In this scheme the network delay is estimated from past statistics but our algorithm does not need to keep past states for arriving time. The processing power demand for this scheme is much higher as well because

the requirement for calculate wave form inside speech frames. This will be discussed later in Section 4.3.

Paper [99] stated a complicate algorithm of adaptive jitter buffer aiming to optimize the playout time according to estimated MOS. This is a user perceived speech quality oriented approach, but compares it with our algorithm, the processing power requirement is much higher because the calculation of playout delay involves a few performance demanding processes including MOS estimation and an optimization process.

In the next Section 4.3, our Play Late Algorithm is presented in details. This new algorithm avoids the shortcomings of existing algorithms and can be used to optimize the performance of current jitter buffer algorithms to achieve a better performance.

4.3 The New Play Late Algorithm

The design aim of this new jitter buffer algorithm is to improve the performance of current jitter buffer algorithms and to provide an algorithm which can produce better end-to-end user perceived quality based on adapting the relative delay time inside a talkspurt efficiently. The new improvement method considers not only improvement of user perceived speech quality, but also the implementation restrains in the specific mobile handset environment.

4.3.1 Algorithm Description

The new playout algorithm, Play Late Algorithm, exploits a feature of human speech and is designed to enable existing jitter buffer algorithms (static and adaptive) to achieve better speech quality in VoIP networks. The pseudo code for the new algorithm is given in Table 1. Please note because coded speech frame number in each packet is fixed and in most of the cases equal to one, the word packet in the following text is considered the same as speech frame.

Pseudo code of the new algorithm is as follows:

Algorithm 1 Pseudo code of the new algorithm

1. Receive Packet N
2. Check SID status. If in SID session
3. Check if jitter buffer size need to be reduced. If no, play SID packet on time
4. If yes, adjust playout time
4. If not SID, Set playout time for Packet N (Tn)
5. Check if Tn is past or not. If no, insert Packet N into the jitter buffer; if yes
6. Check if there is any packet serial number higher than N have been played out.
7. If yes, discard Packet N and check if up to i continuous packets been dumped
8. If continuous number reached i, reset the queue; else go to (11)
9. If no later then N been played, play it out now
10. At Tn output Packet N to playout
11. Wait for next Packet N+1

Note: continuous packet dropping counter i should be set based on link conditions. i was set to 10 in the mobile implementation, which means after 200ms worth of packets been dropped from the buffer, it needs to be restarted.

There are four key points in the new Play Late Algorithm:

1: Play Late if There Are No Later Packets Being Played (Increase Length of Talkspurt if Necessary)

When a packet arrives later than its planed playout time, it is normally discarded or dropped in conventional jitter buffer algorithms. As shown in Figure 4.1, packets 3 and 4 are discarded because the scheduled playout times are past when they arrive at the jitter buffer.

In the "Play Late" algorithm, the late arriving packets are not dropped if there are no later packets have been played out. Instead, a simple logic decision is used to control if it can be played out. The playout decision is based on whether there are earlier packets i.e. packets which contain speech frame with a higher sequence number that have been played out. If there are no such packets have been play out, jitter buffer can play out the late arriving packets. Otherwise, the jitter buffer will drop the packet and does not send it for play out.

In Figure4.9, packets 3 and 4 will not be dropped because when they arrive, the scheduled

playout time past but there are no later packets have been played out. So they can be played out at the next time. This is equivalent to increasing the jitter buffer size on demand.

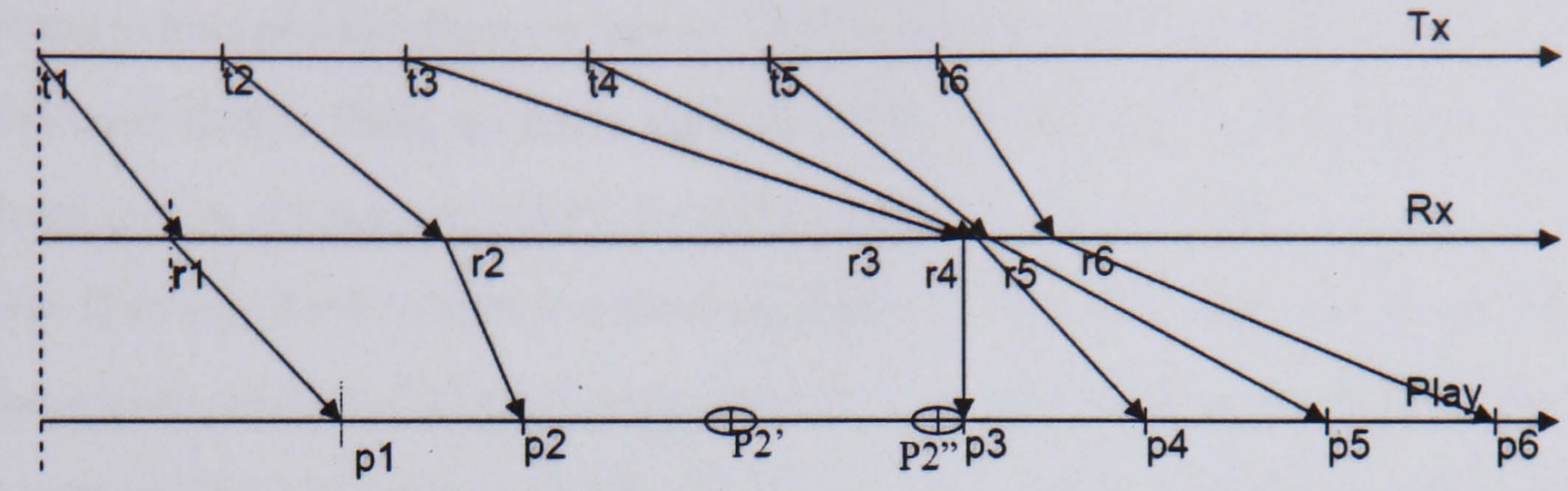


Figure 4.9: "Play late" algorithm will shift play time later on certain conditions

2: Recalculate Talkspurt Delay Time After a Silence Period (Increase or Decrease Buffer Size)

As a basic function of an adaptive jitter buffer, the Play Late Algorithm adapts the jitter buffer size at the beginning of each talkspurt. This process is done after a SID session. When the SID session occurs, the frames are sent not every 20ms but every 160ms. Thus, if the buffer size needs to be adjusted for the next talkspurt, insertion or removal of time slots from the stream will not cause significant changes to user's perceived quality or to the MOS scores from PESQ measurement. The method to adjust jitter buffer size for each talkspurt can be statistic based as in our implementation, or they can be based on other algorithms. The advantage of adjusting the jitter buffer size during SID sessions is that it will not damage talkspurt thus reduces the effect on user perceived quality.

3: DSP Mechanism to Cooperate with the Late Packets Caused Silence Gap (LPC)

After packets have been unpacked and queued through the jitter buffer, they will be sent to the digital signal processor (DSP) as audio frames. The DSP then takes the encoded audio frames, decodes, performs necessary audio processing operations and then plays them out as

sound samples. The DSP has a function to compensate for missing frames from the stream if it knows there are missing frames. It tries to generate low volume frames based on the information from previous frame or frames. This function is normally recognized as loss packet concealment (LPC). There are many algorithms that can be used to implement the LPC for different codecs. An example of LPC for AMR codec is given in [100].

The DSP can obtain information about the state of a SID session from the frame but it will not have know about the RTP time stamp because this information is removed from the header before the audio frames are sent to the DSP. Because the DSP does not have access to the RTP timestamp information or sequence number, the only way it can detect a missing packet is when it can not pick up a frame from the output of the jitter buffer. It will consider this as a packet missing and it will then start the concealment process.

4: Buffer "Error Exit" to Deal with Spike Jitter in Case Single Late Event Blocks Jitter Buffer

Once the Play Late Algorithm is added on top of a common jitter buffer, there is an uncertain risk of jitter buffer error, due to bad software implementation. This key point 4 is designed as an error catch to reduce impact of such error.

The Play Late Algorithm will allow the DSP to repeat the previous frame until the next available frame arrives. This will introduce an equivalent amount of delay into the jitter buffer. Once the next available packet arrives with a group of blocked packets, they could fill the jitter buffer up to the maximum size. After that, the jitter buffer is forced to drop later packets. Later packets could be continuously dropped until a SID session is reached even if they arrive within the required 20ms delay or no jitter. To deal with the jam condition, an "error exit" method is introduced in our algorithm. This lets the jitter buffer to check continuously cases of packet drop and to flush the whole jitter buffer if necessary. By discarding all packets in the jitter buffer and restarting the queue, the jam will be removed and the performance will return to normal. By a careful selection of parameters, this kind of "error" event will not happen frequently. And

in our tests, introduction of this "error exit" method will not affect the over all user perceived quality in terms of PESQ measurement.

4.3.2 Processing Power Estimation

The algorithm can be added to any traditional jitter buffer method as a simple plug-in and its complexity is very low because there is not much calculation involved as in other statistical based adaptive jitter buffer methods. It can also be added on top of a static jitter as well. As described in the next section, this algorithm has been tested with AMR codec with DSP concealment and it should work for most of the popular codecs including G.711 and so on.

To estimate the computational cost of the add-on algorithm, both CPU side processing and DSP side processing need to be considered. The computational cost i.e. the complicity of the proposed can be estimated by calculate the procedures it need to take to process each incoming packet.

On the CPU side, the design of the algorithm decided that the processing cost of the "Play Late Algorithm" is linear to the number of packets. If we denote number of incoming packets as N, the whole processing power needed in the CPU side will be k*N if we denote k as the average processing power needed for each incoming packet. In the DSP side, the processing power needed for each speech frame is the same. Once the play out time of a packet is set, the DSP will play the whole frame without extra processing. So the complicity of the process in DSP will be j*N if we denote j as the DSP process needed for each packet and N as total number of incoming packets. The overall processing power needed for speech segment including N packets will be (k+j)*N.

Compare with the algorithm in Liang's paper [35], the modification of samples i.e. time scaling inside a single speech frame involves much higher processing power in either DSP or the CPU. The maximum complicity of scaling a 20ms frame sampling at 8kHzm is 24,000 multiplications plus 24,000 additions according to Liang's paper. If the processing power needed for a multiplication process is denoted as x and the processing power needed for an addition

action is denoted as y, for a speech segment including N packets, the overall processing power will be (k+j+24,000*x+24,000*y)*N. It is still highly complicated for an embedded mobile phone hardware.

4.4 Test System Setup

A set of tests have been carried out to prove the concept of "lay late" algorithm. To achieve the most reality and reliability, tests cases were carried out in a test mobile and on live IP network. The following section describes the test set up and detailed test procedures.

4.4.1 Test System Structure

To make it easier to follow, the structure of the whole system is first introduced, followed by a more detailed structure of the mobile and then the integration of jitter buffer plus playout method.

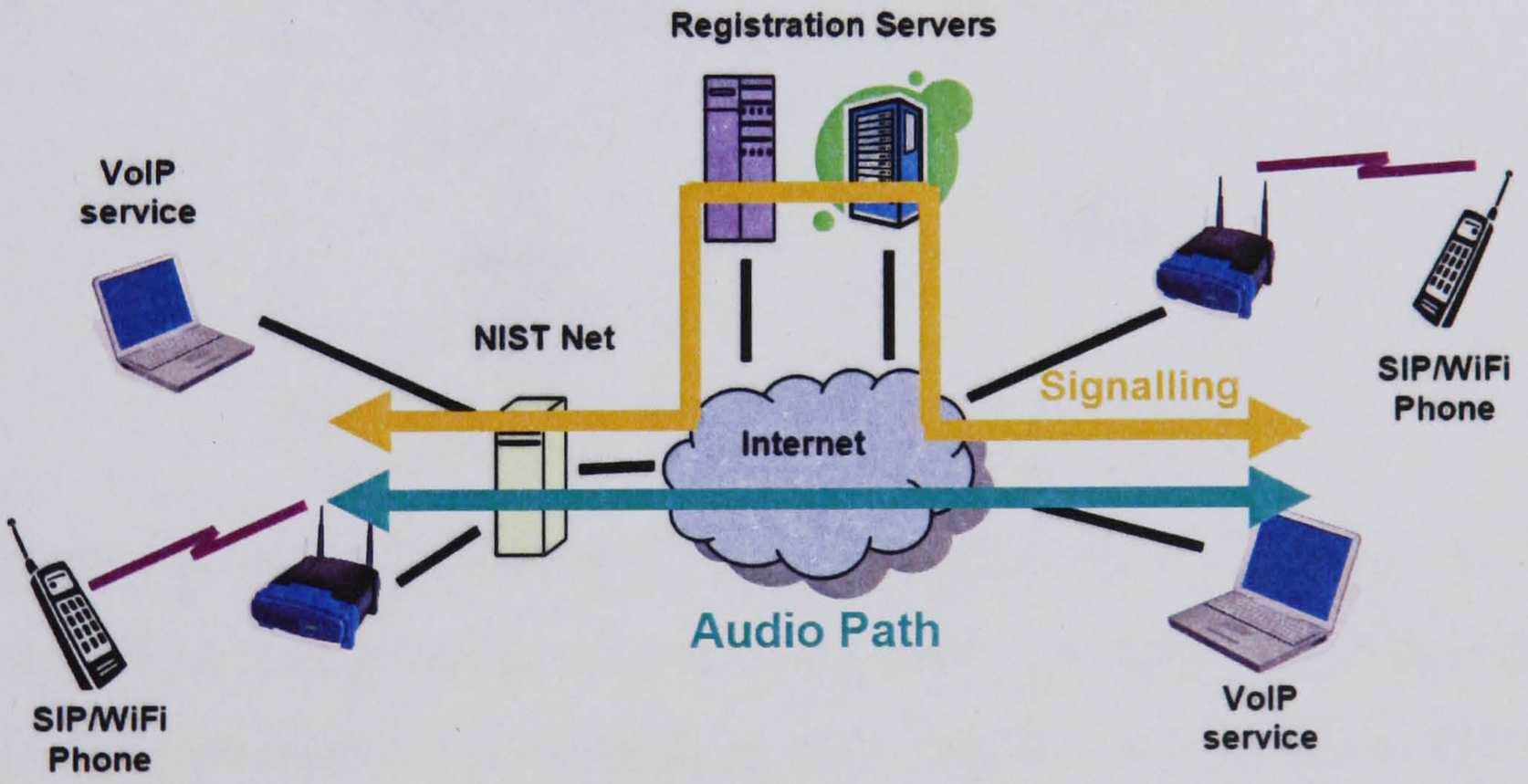


Figure 4.10: VoIP system structure

The basic structure of a VoIP communication system is shown in Figure 4.10. VoIP calls are made on live networks during the test. Controlled jitter can be added for test via the NISTnet

box. Reference speech is send through the live VoIP call and degraded speech is collected so the speech quality of the call can be measured by PESQ.

To implement VoIP in a GSM/IP dual mode mobile is more complicated than an ordinary VoIP phone because the reuse of elements in the mobile. To focus on the topic of this chapter, only the relevant parts will be described. And to focus on the jitter buffer algorithm, the VoIP path is described briefly for the downlink direction only.

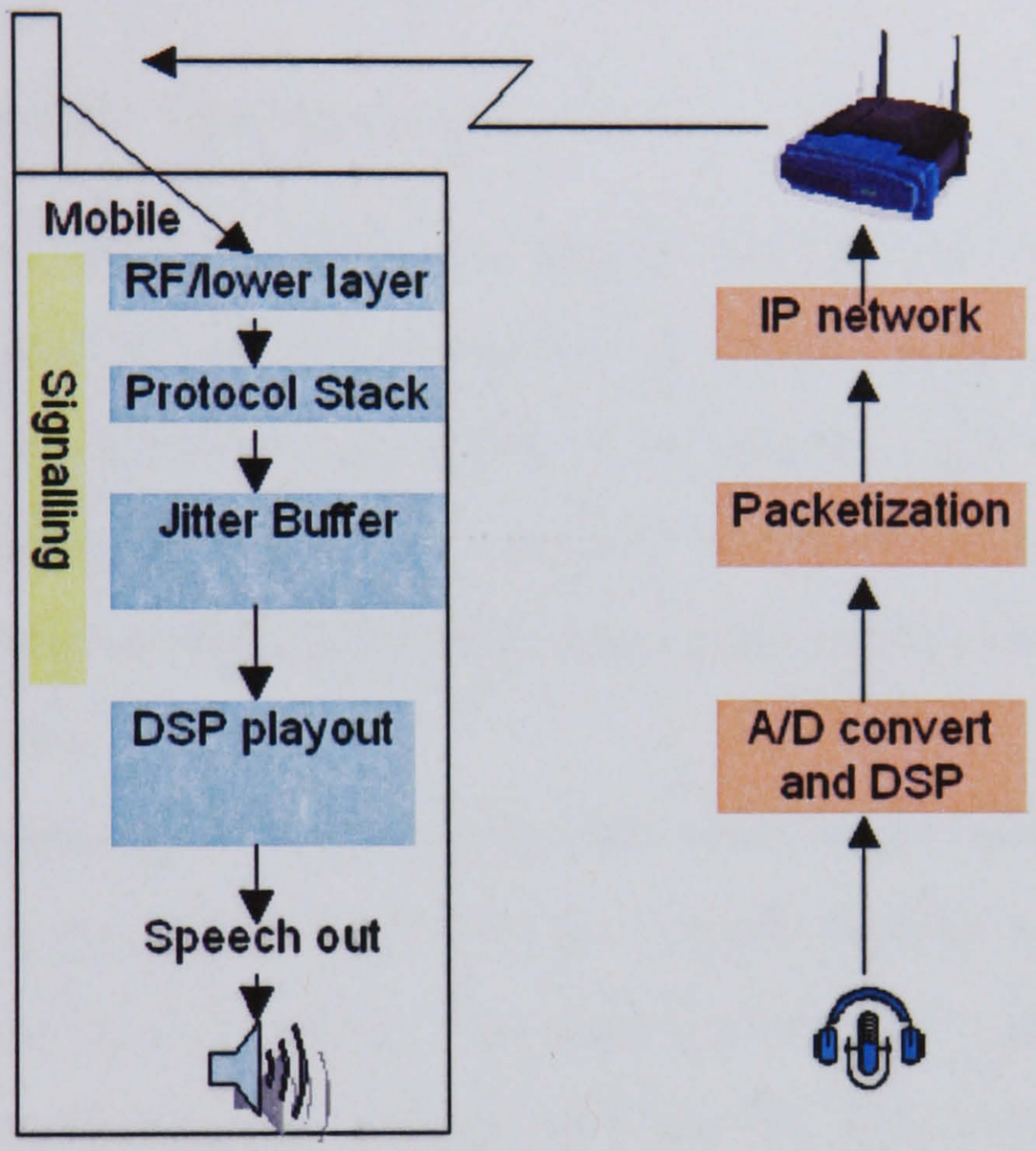


Figure 4.11: Basic audio path of mobile VoIP system

In Figure 4.11, the left hand side shows the basic structure of the VoIP parts of a test mobile. When a VoIP call is setup, speech frames are packed and sent through the wireless channel. At the receive side these are picked up by the radio frequency (RF) module of the mobile, processed and passed to the network module (protocol stack) for IP level processing. They will then be a hand over to the multimedia module to process the RTP and multimedia level information (including RTP header, sequence number, time stamp and so on). In the downlink direction i.e. the receive direction, the jitter buffer will work on the multimedia module to

provide a smooth playout time for each audio frame. The output of the multimedia processes are system messages containing audio frames. These audio frames have been removed from RTP packets so that there are no timestamps or sequence numbers attached on to them. These audio frames are sent to the DSP for playout. As described above, the DSP checks every 20ms to get a new speech frame and if the frame is available, DSP plays it out. Otherwise, jitter buffer takes care of the gap filling process.

4.4.2 Audio Quality Test Structure

Figure 4.12 shows the audio quality test setup for the evaluation of the performance of the Play Late Algorithm. To setup a downlink test for the jitter buffer performance, a speech source is needed to provide a reference speech. A few speech samples are stored in a voice server connected to the public network and are played out through an ISDN line. There is no analog line in the whole system so impairments introduced by analog links (such as noise and 4/2 echo) are eliminated.

There are a few building blocks in the path to send speech samples down to a VoIP- enabled mobile. The speech is played by the voice server via a digital telephone line (one voice channel of ISDN or E1/T1) to the VoIP gateway. The gateway then converts the PCM samples into AMR frames, packs and sends them through the IP network from a backbone to a last mile ADSL connection. On the receiving side of the link, there is an IP based ADSL modem with wireless access point which can forward the AMR/RTP/UDP/IP packets to the mobile via a wireless IP link. In the test environment, the wireless link is a Bluetooth link. In this path from the VoIP gateway via IP network, ADSL link and wireless IP connection to the mobile, there are existing different network impairments such as delay, jitter and packets loss which will affect the speech quality in the mobile. The jitter buffer built in the mobile is designed to minimize the impact of these effects and to deliver the best possible user-perceived speech quality.

To record sound from a mobile, a sound card is connected to the headphone socket of the

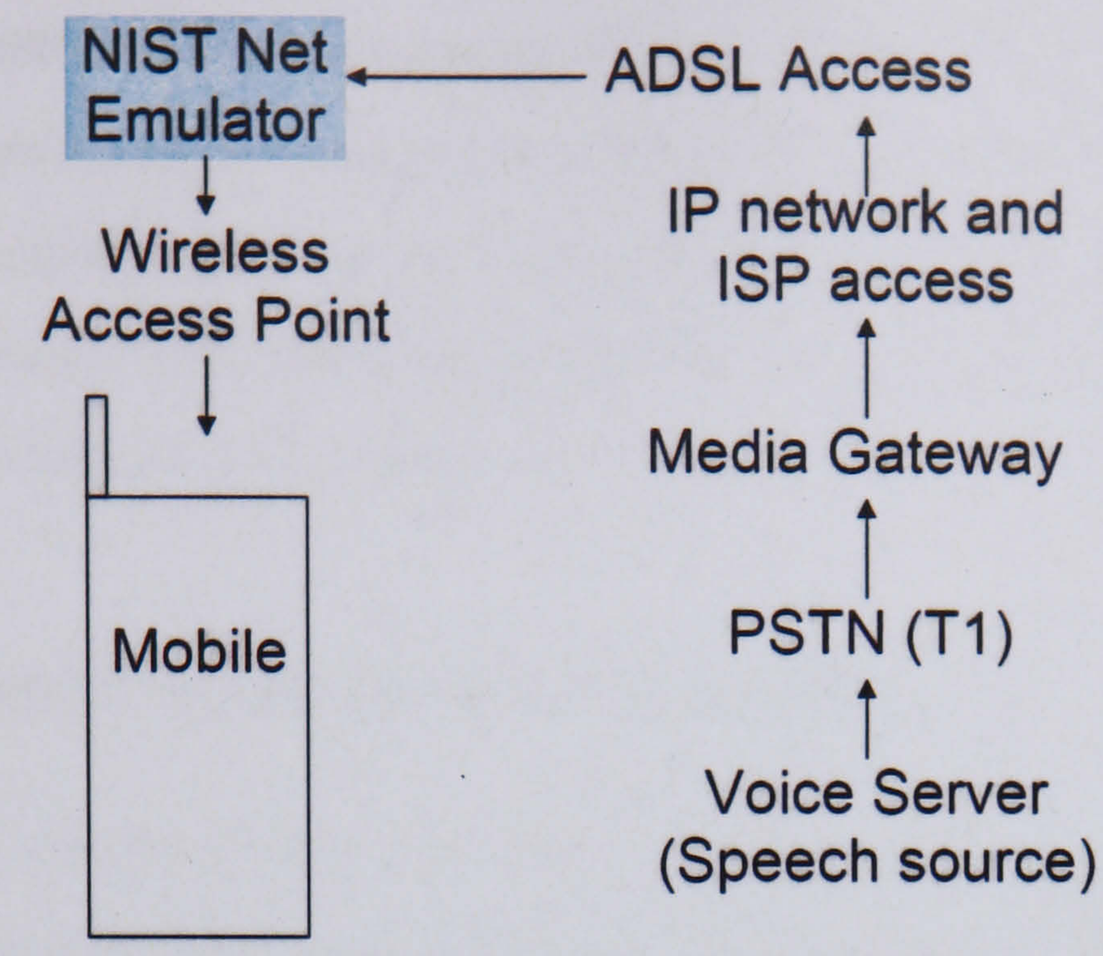


Figure 4.12: The structure of the mobile speech quality measurement system

mobile. Most internal sound cards or cheap external sound cards have problems such as time shifts when recording or there are small gaps during play or record. A broadcast class sound card is used to avoid possible gaps or clips introduced by the sound card or play/record device. For the downlink test, it is used to record the speech which the mobile received from the remote voice server.

In a controlled good coverage area, VoIP traffic may not be impaired by the network apart from a reasonable and stable delay. But in low coverage area and busy home Internet condition, the jitter of the network will introduce more degradation to the voice quality by introducing more jitter to the IP transmission. To reproduce this jitter condition, an emulation tool, NIST-net [101], is used in the test bed. Installed in a Linux PC with two network cards, NISTnet is a software that provides emulated network impairments (such as delay, jitter and packet loss) to the IP traffic going through it. By using the NISTnet emulator and other logging tools, the test is repeatable and easier to analyse.

There are other tools that are available in the mobile networks which can provide a log of different parameters in real time or non-real-time mode (such as current average jitter, relative time of each arrived packets and so on). Some tools can capture and reproduce audio frames be-

fore and after the DSP decoding process so the effect of the play-late algorithm can be analysed in detail. These tools include mobile phone logs which show current arrival time difference between packets and calculated average jitter, DSP logs which show DSP data starvation and thus the concealment process, and network tool such as Ethereal which shows traffic on the network and hence provides data about the statistics of jitter.

4.5 Test Result Analysis and Discussion

To demonstrate the improvement of the Play Late Algorithm when compared with current static jitter buffer algorithms, a group of tests have been performed. The aim of the tests is to prove that the Play Late Algorithm can provide better perceptual voice quality compared to classic static jitter and adaptive jitter buffer algorithms.

As described in the introduction, we can use PESQ score and the end-to-end delay to provide a measure of perceptual end-to-end speech quality.

The jitter buffer algorithm is implemented in a test mobile and then evaluated in a live network. Jitter is introduced via a NISTnet emulator inserted between the mobile and the live network. Due to the constraints of the setup, some parts of the setup including the live network elements were not under our control and this caused the overall PESQ score to be lower than the ideal condition. These problems are awarded and been reduce to a minimum by averaging repeated tests. Because all these tests for the new Play Late Algorithm and the comparison group tests are done on the same setup, the uncontrollable affects are applied to them at the same time and same level. Thus the performance comparison are based on the same ground and valid for the relative comparison purpose.

The following sections present the performance results of the new Play Late Algorithm with other jitter buffer algorithms on different network jitter conditions.

Figure 4.13 shows the performance of an adaptive jitter buffer without the Play Late Algorithm.

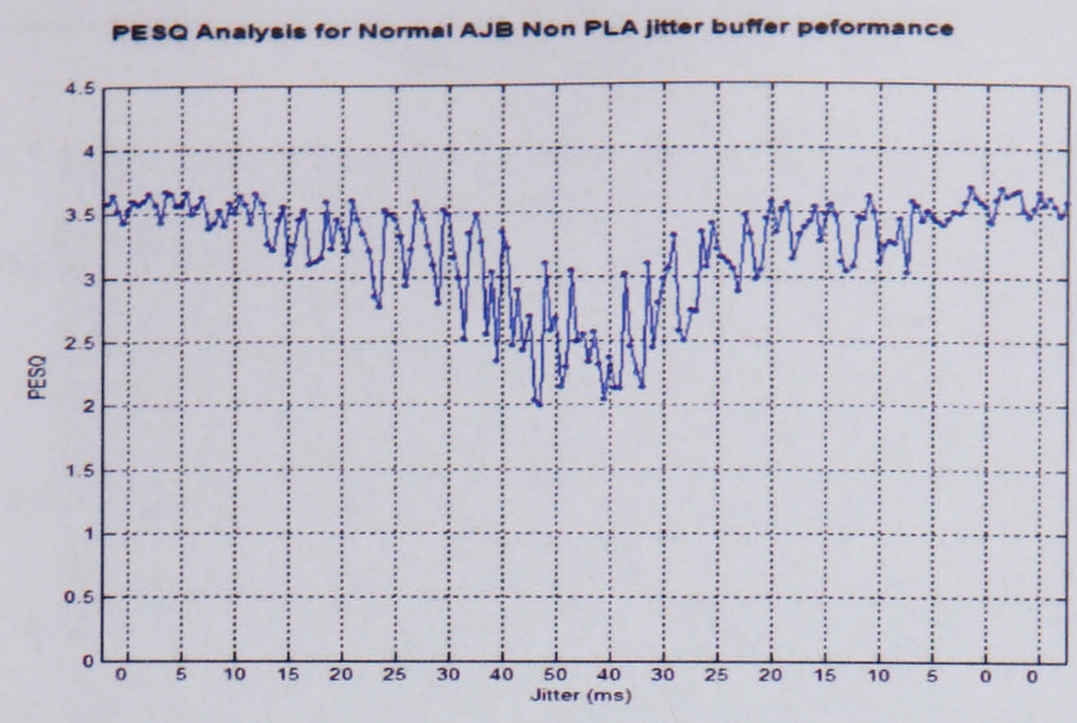


Figure 4.13: PESQ score for adaptive jitter buffer without PLA (Jitter 0-50ms)

In Figure 4.13, Y axis shows the PESQ score for each 10 seconds speech sample and X axis shows the jitter conditions of the samples, from 0ms to max of 50ms then back to 0ms gradually. For each jitter condition (0ms, 5ms,10ms, 15ms, 20ms, 25ms, 30ms, 40ms and 50ms), 10 samples are transmitted during the jitter increasing period and 10 samples are transmitted during the jitter decreasing period. Each dot in the figure represents one 10 seconds speech sample. Average PESQ score of the 20 samples (10+10) for each jitter condition is a better indication of the jitter buffer's performance. Test results shown in Figure 4.13 have been converted into Figure 4.14 and Figure 4.15 for better comparison.

4.5.1 Compare Static Buffer with Static Buffer + Play Late Algorithm

Figure 4.14 shows the average PESQ score results for the case of a static jitter buffer and the case of the static buffer with the Play Late Algorithm as an add-on .

The static jitter buffer length is 40 ms in both static and static plus PLA algorithm. In the static jitter buffer case, the length of the buffer is static and not adapting. In the static plus Play Late Algorithm case, the jitter buffer starting size is the same as the comparison case, which is 40ms. How ever, with the add-on of Play Late Algorithm, the buffer delay can be extended and reduced to get a better perceived quality.

Each data point in Figure 4.14 and Figure 4.15 is obtained as an average of 20 test samples

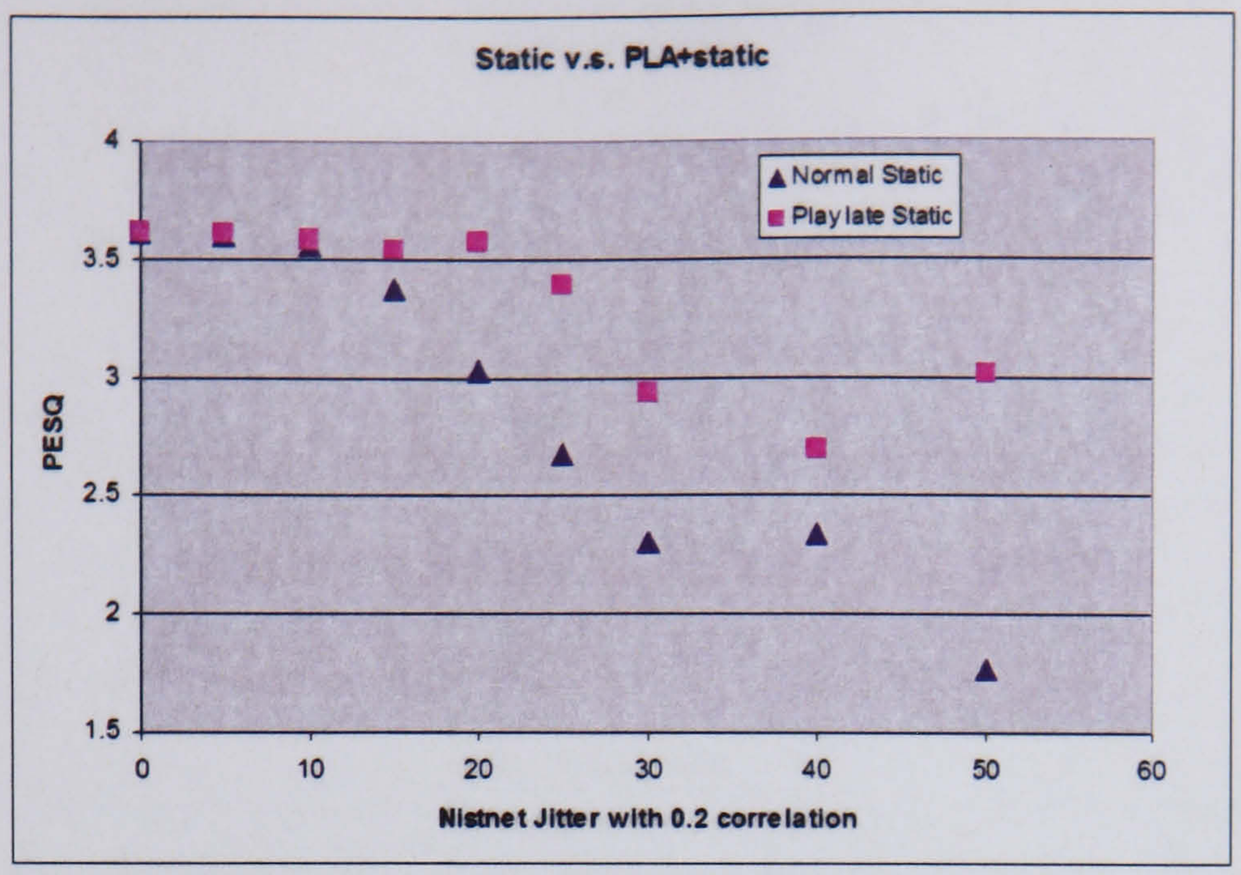


Figure 4.14: PESQ score of Static v.s. Static + Play Late

so it should be accurate enough to represent the trend. In the test, each sample is about 10 seconds long, levels and activity rate are within the limitations of PESQ measurement. The jitter introduced by NISTnet is from 0 to 50ms and the correlation parameter is set to 0.2.

The comparison result shows the static jitter buffer and the static plus Play Late Algorithm case has similar average PESQ score when the jitter is lower than 10ms. But the static plus Play Late Algorithm jitter buffer can achieve a significantly better average PESQ score in higher jitter cases. The reason for the higher PESQ score at the 50ms test point remains unclear and is a subject for future work.

4.5.2 Compare Adaptive Buffer with Adaptive Buffer + Play Late Algorithm

Similar as Figure 4.14, Figure 4.15 shows the difference in the average PESQ score between a normal adaptive jitter buffer and the same adaptive jitter buffer plus the Play Late Algorithm in a test mobile running in live and emulated network.

As in the previous figure, each of the points represents an average of at least 20 individual tests. The individual tests are about 10 seconds long and are individually scored by PESQ.

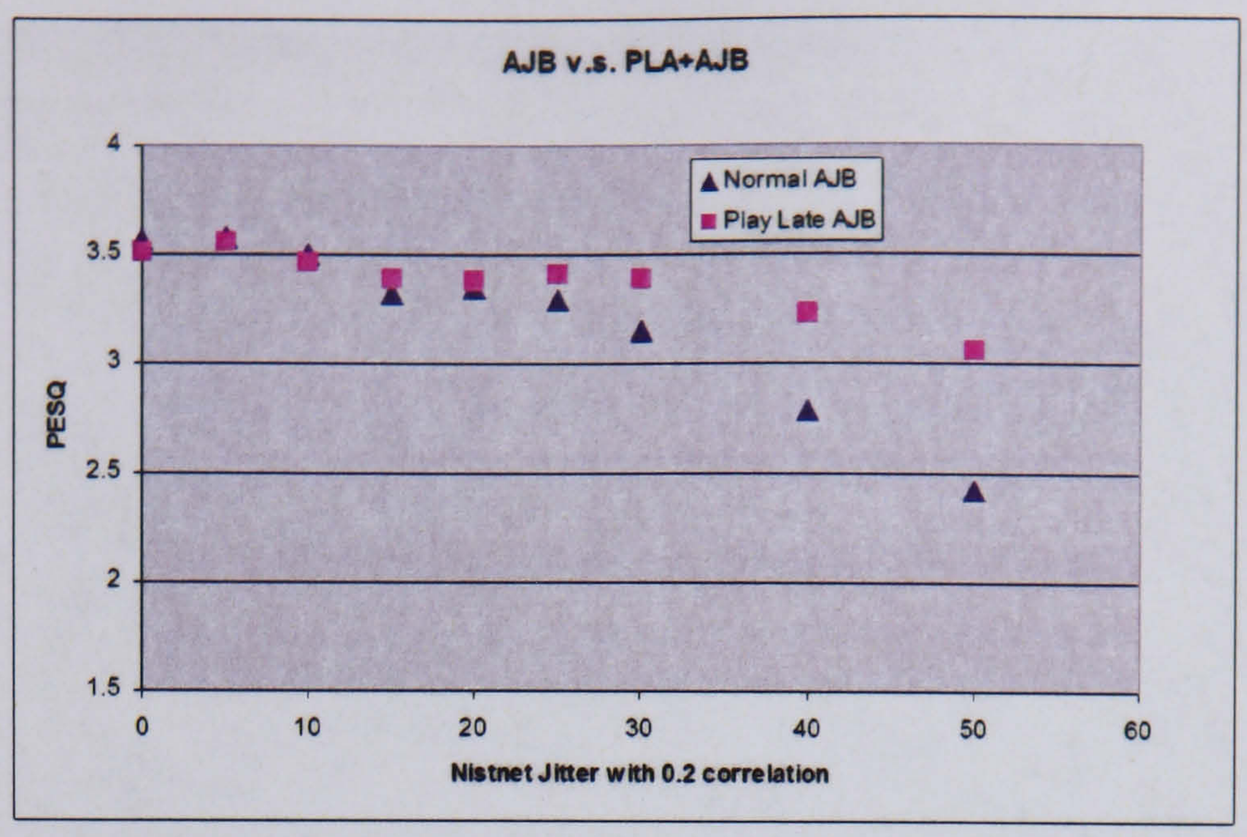


Figure 4.15: PESQ score of a non-PLA AJB v.s. AJB + Play Late Algorithm

As the result shows, when the Play Late Algorithm is active in a normal AJB, the over all PESQ score improves, especially when jitter increases to 20ms or more. The difference between adaptive jitter buffer and adaptive jitter buffer plus the Play Late Algorithm is not as significant as the difference between the static jitter buffer and the static jitter buffer plus the Play Late Algorithm is because the adaptive jitter buffer provides a better performance when compared with static jitter buffer.

4.5.3 Delay Analysis

It is also necessary to verified that the Play Late Algorithm does not introduce excessive delay time. Delay analysis is done by looking at the recorded loopback wave file under the same condition. As discussed above, because the extra delay added in a talk spurt can be removed once a SID session starts, the overall delay the end user experiences will not be adversely affected. Figure 4.16 shows what happens when extra delay is added in a talk spurt and then removed when the talk spurt is finished.

Figure 4.16 is a screen capture of a PESQ analysis tool that compares the reference speech file with a degraded speech file. The degraded speech is been transmitted through the new Play Late Algorithm jitter buffer. In this figure, the bottom part is interested in, where the frame by

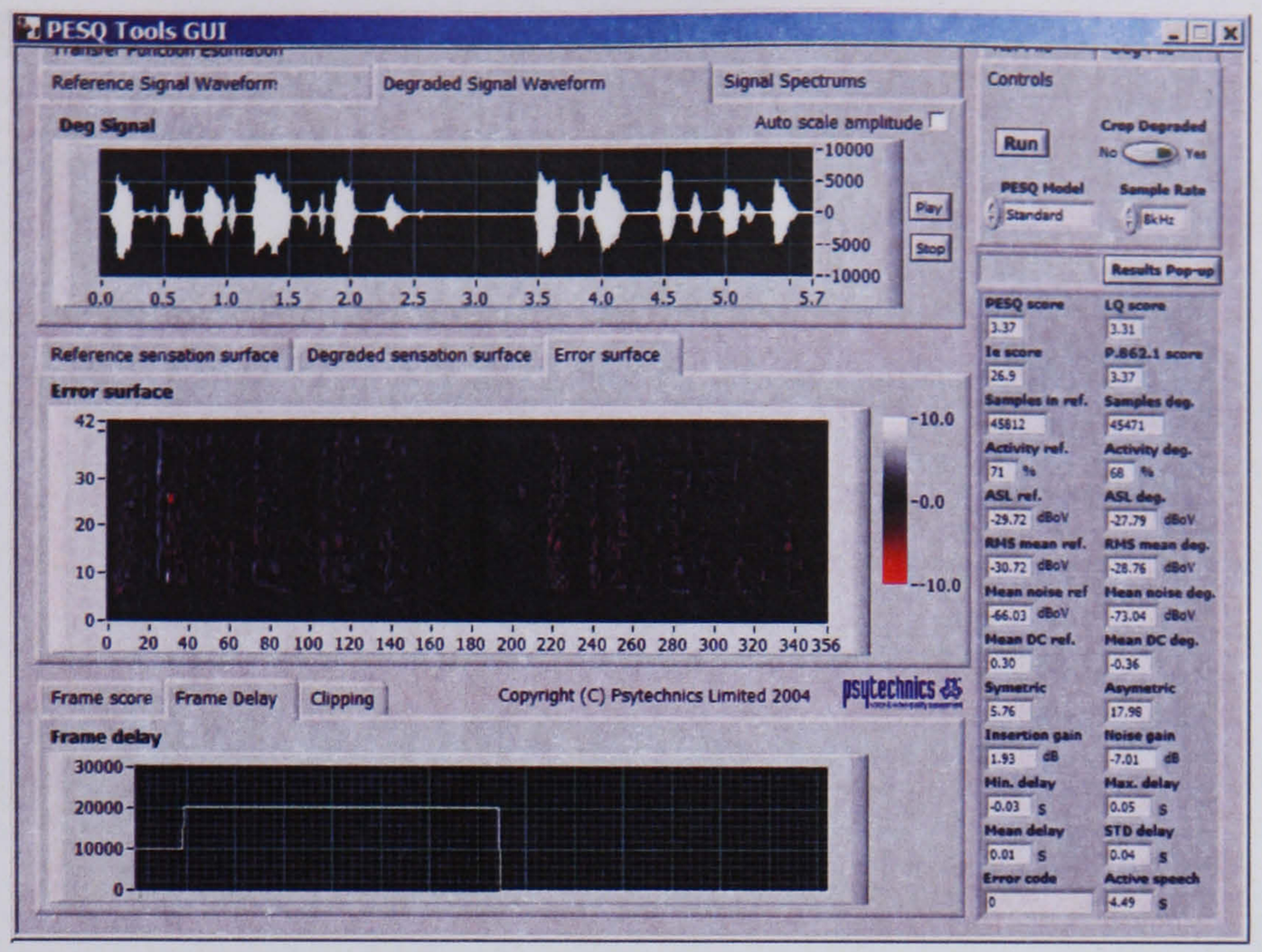


Figure 4.16: Frame by frame delay alignment plotted by PESQ tool

frame delay is shown. The PESQ tool performs time alignment for each PESQ frames which is 16ms for 8K Hz sample rate. The frame by frame delay plotted at the bottom of the figure shows that the delay is increased by about 40ms in the middle of the first talk spurt and then reduced by about 80ms in the next silence period. The unit of the Y axes in the plot is 1/256ms.

To show the relative delay increase and decrease in the waveform, the wave record is compared in Figure 4.17. The upper part of the figure is the original wave and the lower part the degraded speech which has gone through the transmission path and the AJB. The starts of the two waves are aligned so that the delay changes can be shown. The shadowed part is where the Play Late Algorithm enabled jitter buffer was waiting for the late packets (two of them in this case, 20ms each) and after the silence period, the delay retrieved about 80ms. The extra 40ms decrease should be associated with a decision of reducing jitter buffer size for the next talk spurt.

Compared to an adaptive jitter buffer which does not change the relative delay time in a talkspurt, the Play Late Algorithm can improve the user perceived quality because it reduces information loss due to reduced packet drop rate.

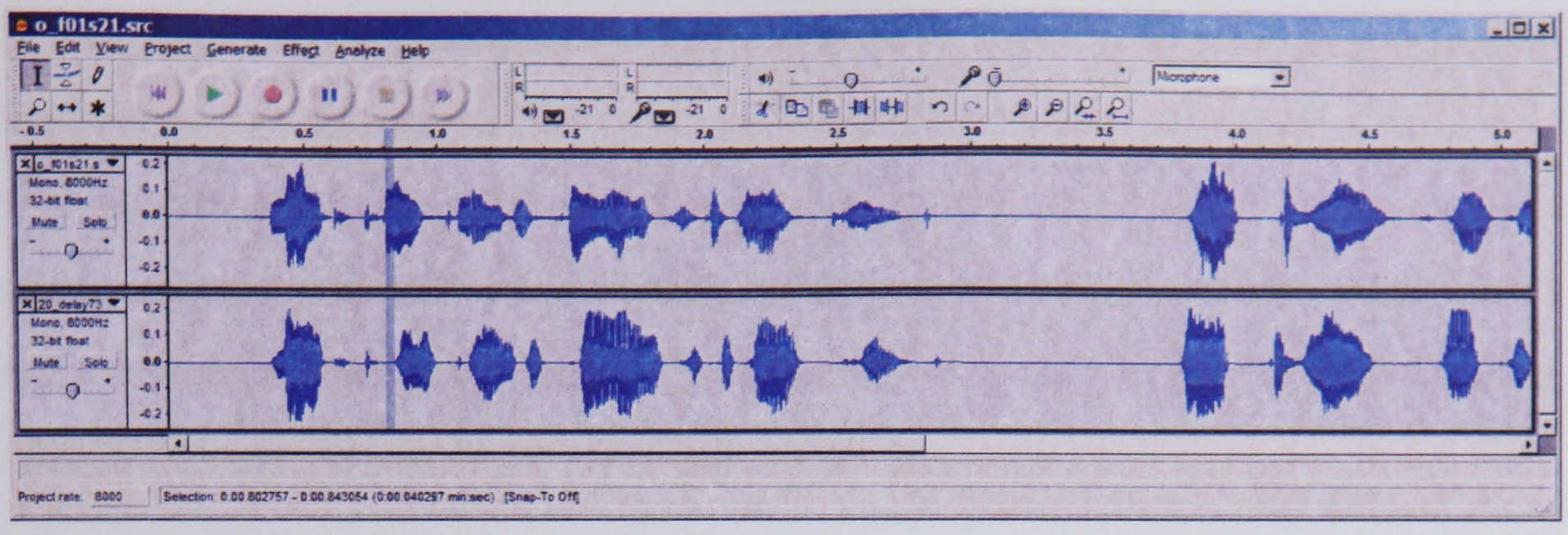


Figure 4.17: Delay time shifts shown in waveform

4.6 Summary

The Play Late Algorithm presented in this chapter is a simple but efficient plug-in for most speech jitter buffer algorithms to improve user perceived quality. There are four key elements in the algorithm. The first is to allow late packets to be played if there are no packets with an earlier sequence number being played; the second is to allow delay insertion during silence period; the third is to replace concealment frames while waiting for late arriving packets to reduce the effect of inserting delay; the fourth is to reset the whole queue in a buffer jammed condition.

Our results show that once the new Play Late Algorithm is added on to both static and adaptive jitter buffer, they can achieve better performance under different network conditions when compared to classic static jitter and the original adaptive buffer algorithms without introducing excessive delay. The algorithm is easy to implement and easy to run without demanding extra processing power.

Although the algorithm is only tested on the AMR codec at the highest rate of AMR122, it can be used on other speech codecs as the idea to wait for late packets is still valid in any speech conversation environment. Unfortunately, this algorithm can not be used for other media transmissions such as continuous audio or video streaming because the idea is based on the characteristics of human speech conversation.

It is worth to point out that the voiced part of speech may have different quality sensitivity to packet loss when compared with unvoiced part. And the beginning or end parts of a talkspurt have higher sensitivity to the impact of packet loss in terms of user-perceived speech quality when compared with middle parts [102]. It will be interesting to study adjustments in the delay insertion in the lower sensitivity parts. However, this may involve increased processing operations in the DSP and processer element of the hosting system.

Chapter 5

A Perceived Quality Enhancement Method Based on Combined Features from Different Layers

As discussed in previous chapters, Quality of Service (QoS) optimization is an important issue in Voice over IP (VoIP) systems because of the need to meet technical and commercial requirements. Current implementations of QoS enhancement mechanisms need to be further optimized and improved. This leads to the aim of this chapter: to enhance current Quality of Service optimization mechanism for VoIP services.

The main objective of this chapter is to propose a new QoS enhancement scheme that combines the strengths of rate-adaptive control method and speech priority marking technique to provide a superior QoS enhancement performance, in terms of perceived speech quality.

A second objective is to propose the use of an objective measure of perceived speech quality (i.e. objective MOS score) for adaptive control of sender behavior as this provides a direct link to user-perceived speech quality, unlike individual network impairment parameters (such as packet loss and/or delay).

The results show that the new combined QoS enhancement method achieved the best performance under different network congestion conditions compared to separate adaptive sender rate or packet priority marking method. The results also show that the use of an objective MOS as the control parameter for the sender rate adaptation improves the overall perceived speech

quality.

The results reported here are based on a simulation platform that integrates DiffServ enabled NS-2 network simulator, a real speech codec (AMR codec) and the ITU-T standard speech quality evaluation tool (PESQ).

5.1 Introduction

QoS enhancement mechanisms for VoIP should aim to make optimum use of available network/terminal resources and to minimize the effects of network impairments on voice quality. Several approaches exist to realize 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 [103–105] 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 from the network is not 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 [47, 106, 107]. In this approach, voice packets that are perceptually more important are marked, i.e. given priority, and thus less likely to be dropped than packets that are of less perceptual importance, in case of 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 research 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 enhancement scheme which delivers optimum QoS in terms of voice qual-

ity. 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. DiffServ 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 about to occur. With a proper feedback mechanism, this information can be used to control the send bit rate.

The main contributions of this research are twofold. First, a new QoS enhancement scheme is proposed, which 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, it is proposed that the use of an objective measure of perceived speech quality (i.e. objective MOS score, calculated by PESQ as in [108]) instead of individual network impairments (for instance packet loss and/or delay) to control sender behavior 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 [109], adaptive speech codec (AMR) and an objective perceived speech quality measurement system, which is based on the ITU-T speech quality evaluation standard PESQ.

5.2 QoS Enhancement Schemes

There are several QoS control schemes available for VoIP speech quality optimization. The new combined method is based on two existing schemes i.e the rate-adaptive QoS control

scheme and the priority marking QoS optimization scheme. They are introduced here before the new proposed combined scheme.

5.2.1 Rate-Adaptive QoS Enhancement Scheme

User perceived speech quality is related with several balancing complications such as packet loss rate and location in the speech, delay and delay jitter, codec type or codec rate etc. These complications are interlinked and sometimes conflicting. The aim of the adaptive control mechanism is to improve over all user perceived speech quality on a given network condition.

The Relationship Between Codec Rate and Packet Loss Rate in a Bottleneck Link

One common conflict between complications is that in a limited bandwidth environment, to increase codec bandwidth will increase packet loss rate and hence reduce the over all speech quality. And to decrease codec bandwidth i.e. lower bandwidth rate will decrease user perceived speech quality due to the nature of lower bandwidth codec.

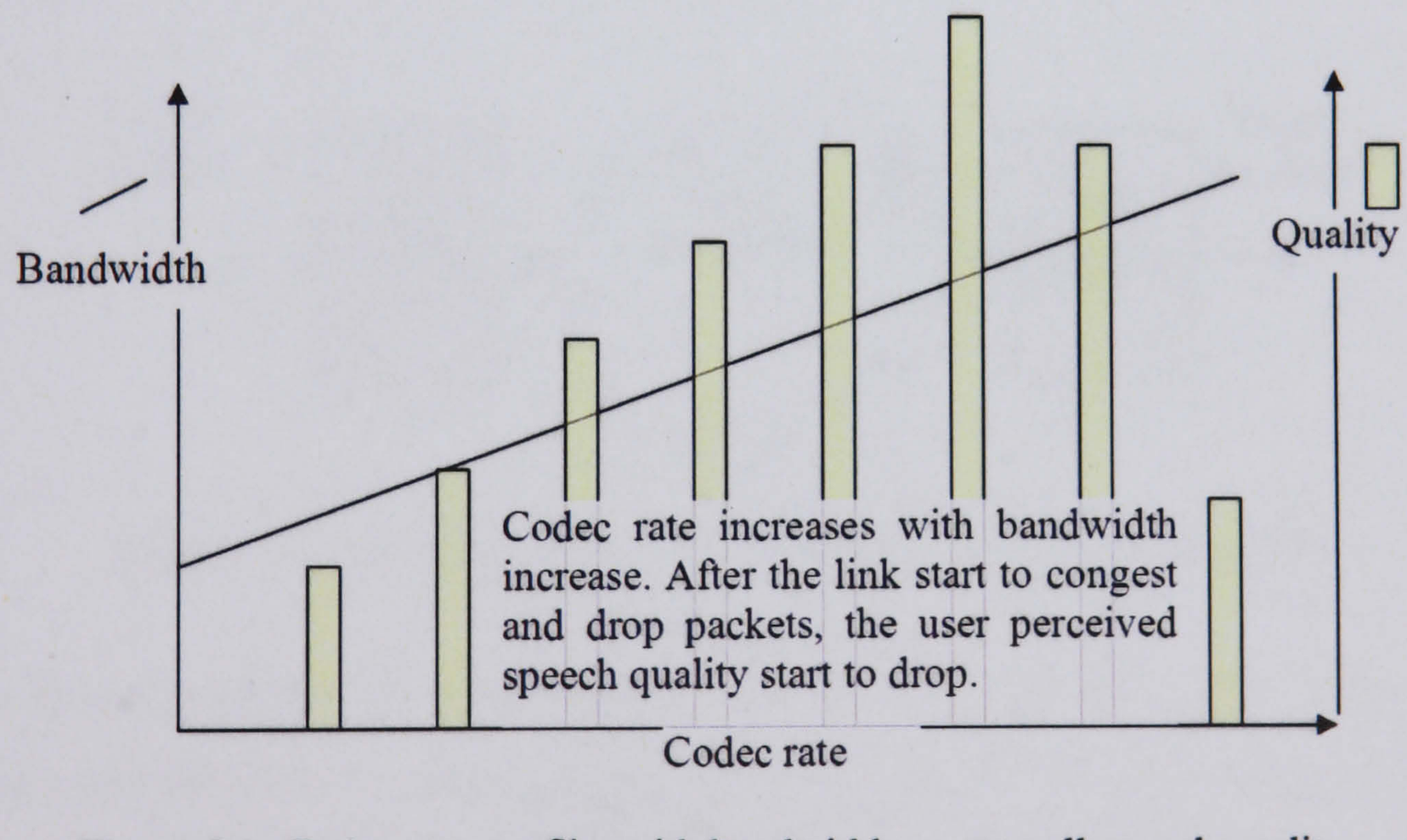


Figure 5.1: Codec rate conflict with bandwidth on over all speech quality

As shown in Figure 5.1, ideally there should be a balanced codec rate on which the codec

rate reached its maximum and have not cause significant packet loss due to the network congestion. Of course this balance is a dynamic balance because the network condition and user behaviour are changing dynamically. So the control mechanism should be able to optimize the codec rate dynamically on a real time basis.

PESQ v.s. Different AMR Rate and Packet Loss

The adaptive rate QoS control scheme is based on the AMR speech codec. The AMR codec was developed by ETSI and has been standardized for GSM. It has been chosen by 3GPP as the mandatory codec. It is a multi-mode codec with eight modes (MR475 to MR122) with bit rates between 4.75 and 12.2 Kb/s. Mode switching can occur at any time (frame-based). Thus, the AMR codec is well suited to rate control.

Packet loss in the coded speech stream will lead to degrading of the user perceived quality of the decoded speech. On the other hand, codec rate have effect on user perceived quality as well. So it is worth to study the speech quality reaction to different packet loss rate on different codec rate.

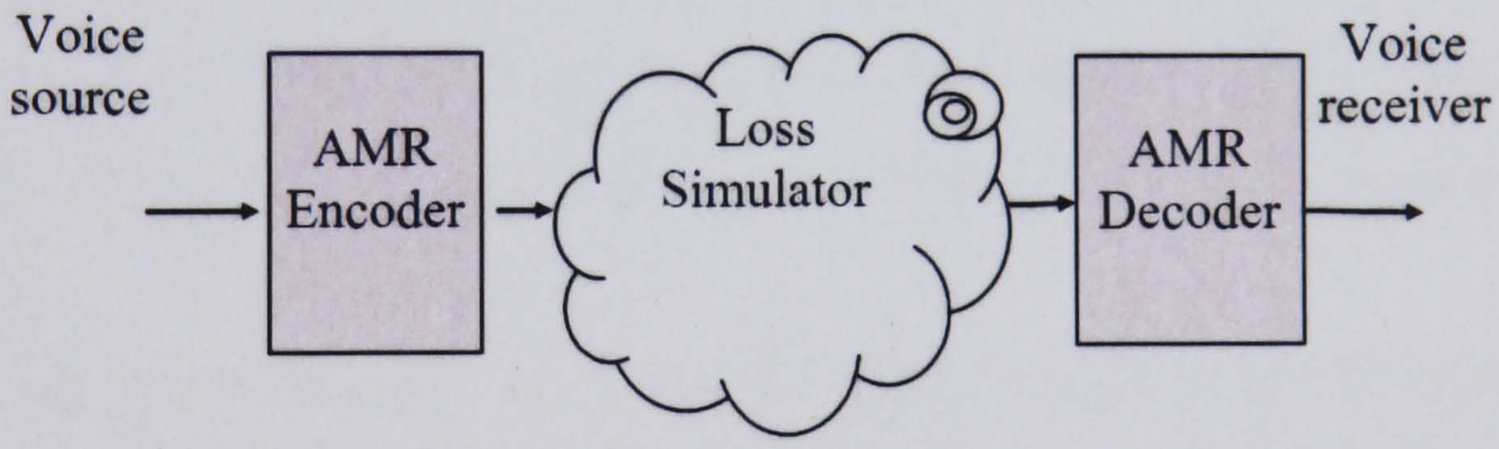


Figure 5.2: Test setup - AMR codec performance v.s. packet loss rate

As shown in Figure 5.2, speech samples are coded by an encoder, send through a loss simulator and decoded to degraded speech. The loss simulator handles simulated packet loss effect to the coded speech frame i.e. remove frames from the stream following packet loss rules. In this simple simulation, packets are just randomly removed to simulate a random loss with different loss rates.

The results are summarized in Figure 4.7, which shows the PESQ measurement result to speech samples are decreasing when the packet loss rate increases. And lower AMR codec rate gives lower PESQ score on the same speech sample when the loss rate condition is the same. This plot give us a good view of the balance between packet loss rate and codec rate and can help us to design the control mechanism for over all performance optimization.

It is worth to note that in some environment, the packet loss rate is not related to bandwidth utilization directly as a bottleneck link does. The performance of the adaptive scheme needs to be further investigated in such environment.

Feedback Facility in RTP

There is a feedback facility called Real Time Transport Control Protocol (RTCP) specified in the RTP protocol, designed to carry receiver side feedback information to the sender as introduced in Section 2.1.2.

The specification of RTP stipulates that the RTCP traffic does not exceed 5% of the whole traffic and that the time between the reports is at least 5 seconds. So the bandwidth over head for transmitting RTCP packets should not introduce too much affect to the over all performance of the system. And the feedback time is frequently enough compare with the average life time of a normal real time stream, which is about a few minutes, the same as a normal telephone conversation.

With the help of RTCP reports and AMR, the concept of adaptive rate QoS control scheme is introduce and as shown in Figure 5.3. In the scheme, the send rate of the AMR codec is adjusted in accordance with the network condition feedback to achieve the best possible QoS.

Bit Rate Control Mechanism

The bit rate control mechanism is based on individual network parameters (such as packet loss rate and delay) or on the predicted user perceived speech quality (such as the predicted MOS score as PESQ given). In VoIP applications, the feedback information can be sent via

RTCP reports. Of course the RTCP report packets need to be customized to fit those information picked.

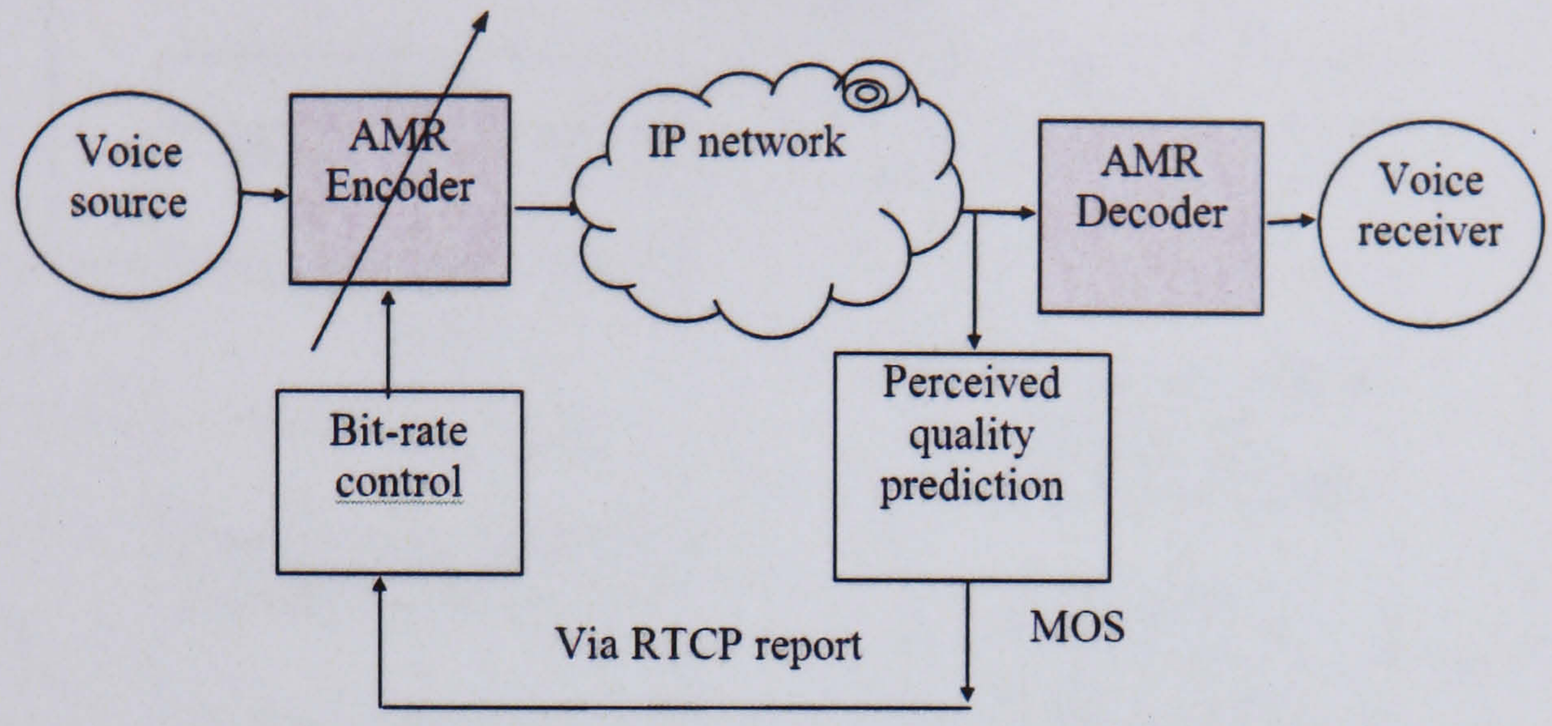


Figure 5.3: Rate-adaptive QoS control scheme

The two important QoS optimization modules in Figure 5.3 are the bit-rate control module at the send side and the perceived quality prediction module at the receive side. Approximately every 5s (the time interval between RTCP reports), a measure of the perceived conversational speech quality (i.e. PESQ score) is predicted from network parameters (such as packet loss and delay) using a PESQ based method, same as the mechanism proposed in [108].

The bit rate control module is used to adapt the send bit rate in accordance to the feed-back RTCP information send by the perceived speech quality prediction module. The adaptive algorithm used in the module follows the 'additive increase/ multiplicative decrease' concept that has been successfully employed in other congestion control algorithms, such as TCP and ABR [105].

The basic idea is that the AMR codec can reduce its bit-rate (if possible) when there is network congestion and increase its bit-rate when no congestion is detected. The AMR rate is then used to predict the packet loss rate and the MOS. The predicted MOS score is compared with the existing MOS and the controller will choose the best step to adapt to if necessary or keep existing AMR rate.

The detailed control mechanism used is presented in Figure 5.4. As the figure shows, in

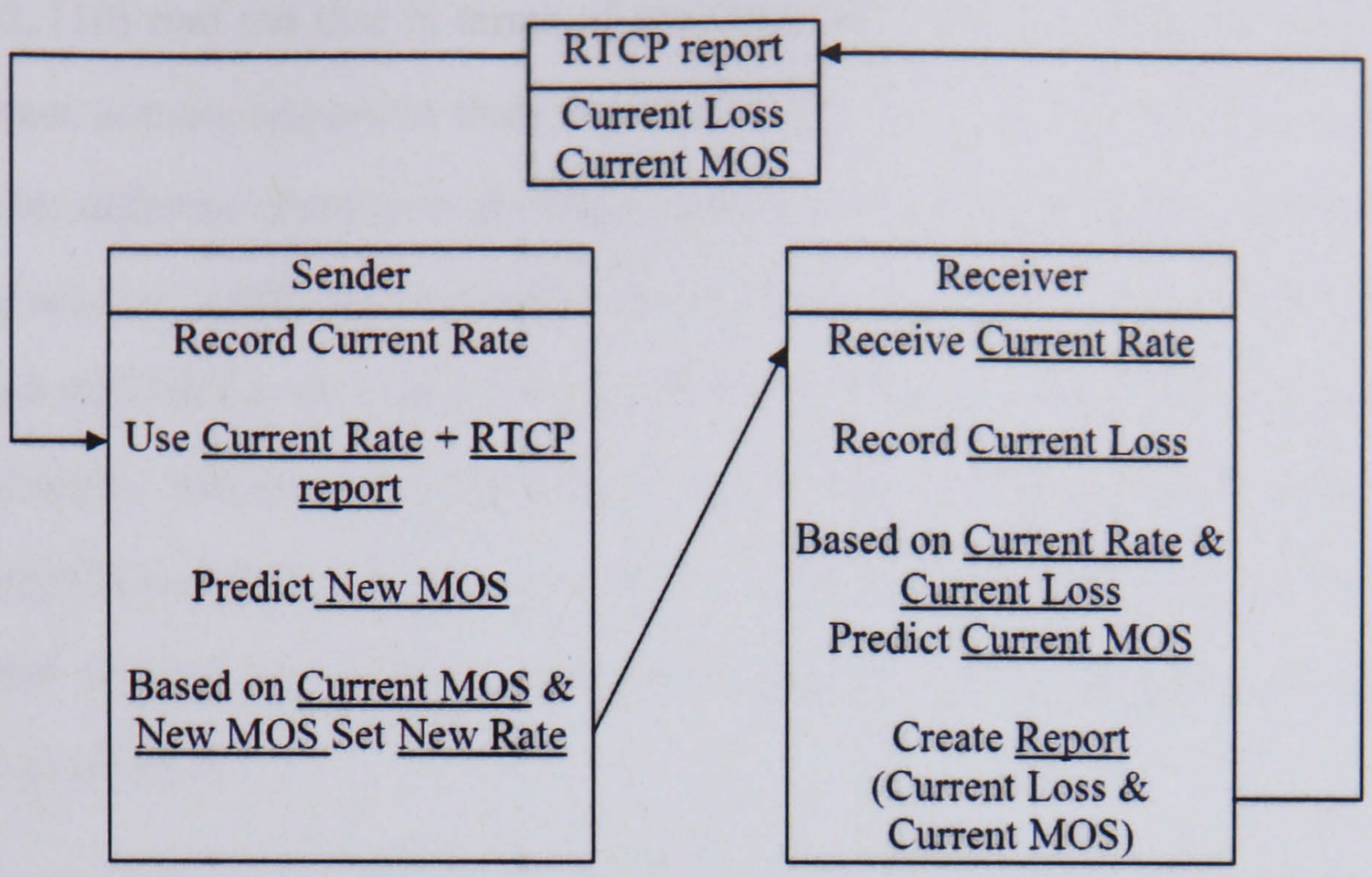


Figure 5.4: Feedback loop of the MOS driven rate adaptive control method

each feedback circle, on the sender side, the AMR rate is set by the control mechanism based on predicted user perceived speech quality, derived from current AMR rate and the feedback information transmitted via RTCP report. On the receiver side the AMR bit stream updates the current AMR rate to the receiver because each AMR packet is contains the rate information. And then the receiver side can predict the user perceived speech quality using the current AMR rate and recorded current loss rate in this report period. The current loss and current user perceived quality information can then be transmitted back for feedback and finishes the control loop. Each of the control loop take about 5 seconds as the interval of RTCP report is 5 seconds. If any RTCP report packet is lost in the network, the sender will keep the rate unchanged and adapt it when next RTCP feedback arrives.

5.2.2 Priority Marking QoS Enhancement Scheme

An other speech quality improvement method is the priority marking QoS enhancement scheme. In a telephone conversation, speech part of the voice stream is obviously more important than the silence part, when the peer of the conversation is listening. Previous re-

search [102, 110] find out that in terms of user perceived speech quality, some parts of the speech stream is more important than the other parts. The fact that human conversational speech carries different importance in different parts of the speech forms the basis of the priority marking scheme. AMR can detect active speech in the encoding process to help the marking process. And the DiffServ network QoS control mechanism provides the cooperation in the network to protect the important speech frames from loss. The aim of the priority marking scheme is to optimize the user perceived speech quality by protecting the important parts of the speech with the help of AMR and DiffServ. Detailed mechanism of the method is presented in the following subsections.

Perceptual Importance Difference in Human Conversational Speech

In rate-adaptive QoS control scheme, it is assumed that all the packets within a flow are equally important perceptually. But in the world of VoIP applications, the stream is transmitting conversational human speech. The nature of human conversation decided that not every part of the speech stream has the same perceptual importance to the user. Previous research has shown that some speech segments are more important than others. For example in a conversation, active speech segment is more important than the silence part, and in an active speech segment, the beginning part of the talk spurt is more important than other parts of the active speech segment. The importance is measured by user perceived quality, which means user is more sensitive to distortions happened in the important part of the speech stream and more likely to give lower opinion score compared with the same kind of distortion happened in less important part of the speech stream.

Marking Important Frames by Utilizing Codec Output Information

In the process of AMR coding, the codec will mark the speech as active speech or silence speech based on energy level difference. And these information is embedded in the coded speech frame and available to the packetisation process. It is possible to consider beginning

frames of the active speech segments as important frames and mark them as high priority, and mark the other parts as lower priority. Of course if there are more detailed importance levels, the marking can be more specific and lead to a more sophisticated control method. For example it is possible to consider the beginning of each talkspurt as the most important parts, give it the highest protection level, i.e. highest DiffServ service class, and the rest part of the talk spurt has the second highest importance level. The silence part of the voice stream then has the lowest priority and has less protection by marked as the lowest service level in DiffServ.

Protecting Important Packets by Using DiffServ

Priority marking QoS control scheme can be implemented in networks that support Differentiated Services (DiffServ) architecture [47]. DiffServ allows the user to set priority to packets and those marked packets can be treated differently in the network. The advantage of DiffServ is that the marking is done on the entering point of the DiffServ zone and there is no time consuming end-to-end negotiation process at the beginning of the call. And because there is no virtual circuit to maintain in the network, it is relatively easy to implement and deploy. The disadvantage of DiffServ include that some user may improperly mark their packets to higher level and lack of end-to-end guarantee if the packets are transmitted over more than one DiffServ cloud. The user over priority problem can be solved by pricing policy and it should not happen in a telecommunication context where the carrier have dominating control of priority marking at the entering point to the DiffServ cloud. And even the packets need to travel through several DiffServ clouds, for instance in an international call, the SLA between carriers should solve the priority difference problem respectively.

There are several detailed implementations of DiffServ protocol, for instance a simplified 2-bit marking DiffServ implementation [111]. There are different queue management method can be used in DiffServ as well. They include the Fair Queue, Weighted Fair Queue, Random Early Detection Queue, Weighted Random Early Detection and so on. Each carrier will choose from those methods accordingly and the detailed implementation is not in focus of this research.

The Priority Marking Scheme

The difference of importance in the same speech stream forms the basis for the priority marking control scheme, which is depicted in Figure 5.5. In the transmission side, after codec encoding, each speech frame in the coded speech stream is marked differently depending on its perceptual importance. For example, the priority-marking module marks the beginning of a voiced segment (for instance the first 5 or 10 frames of a voiced segment for the AMR codec) as high priority (for instance marked as a 'premium' class), while others are marked as perceptually unimportant (for instance marked as a 'best-effort' class).

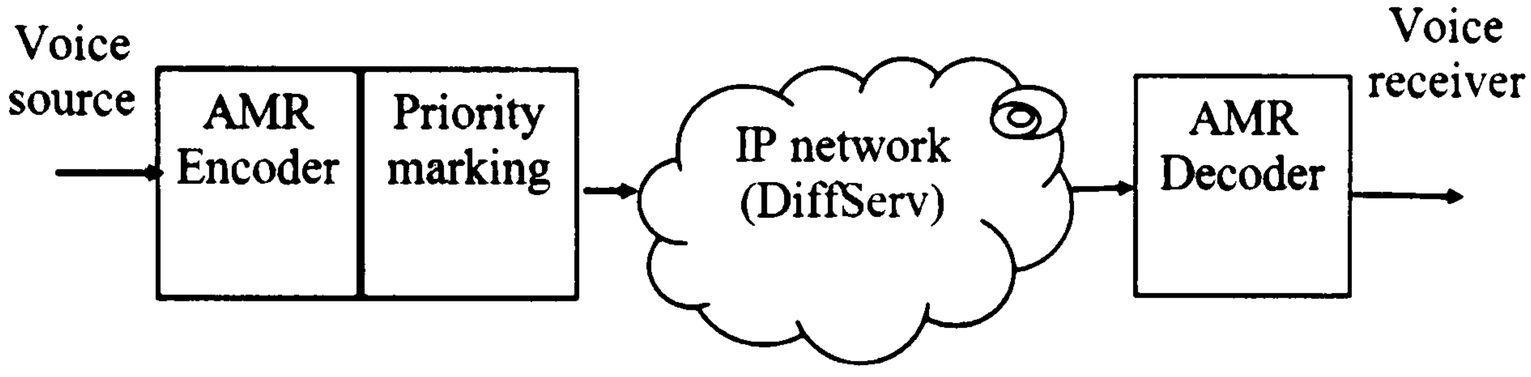


Figure 5.5: Priority marking QoS control scheme

When there is network congestion, the perceptually unimportant frames have a higher drop probability. 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.

5.2.3 The New Combined Rate-Adaptive and Priority Marking Scheme

As discussed in above subsections, the rate-adaptive QoS control scheme is based mainly on an objective perceived speech quality prediction at the receiver derived from current loss rate and current codec rate. In the PESQ evaluation process, packet loss rate contributed significantly to the measured or predicted objective MOS score. And in a coded speech stream, different frames could have different importance to the over all user perceived speech quality. Those important frames can be marked by codec and made available to packetization process and Quality of Service control mechanism. So it is worth to try a method to reduce packet loss

rate, especially to reduce packet loss rate for those important packets.

There are always bottleneck links exist in the network where packets need to be queued and might be dropped if the queue is full. The most basic queue is the simple single drop tail queue in which all the packets enter a single first in first out queue on the arriving sequence and leave the queue on the same sequence if they are not dropped. If the queue is full, later arrived packets will be dropped.

The limitation of this simple drop tail queuing mechanism is that there is no priority difference in the queue for time critical streams such as VoIP real time speech stream. Because every packets follow the same first in first out queue procedure, packets in real time streams not only need to wait the whole queue time same as other packets which time is not so critical for them, but also have the same possibility to be dropped if the queue is full.

To solve the Quality of Service issues above caused by simple over buffered drop tail queue, the network operator commonly use some congestion notification and control mechanism in the network. A queue management system can provide such Quality of Service feature.

Random Early Detection/Drop (RED) queue is an Active Queue Management (AQM) method proposed in [112]. It is originally designed to cooperate with TCP-friendly congestion control but now widely used to provide basic congestion notifications to the sender. In this case, this function can be used to adapt the sender rate to suit the network condition. There are other more sophisticated queue management methods developed to provide even better congestion notification and control but they are not going to be covered here.

Priority marking should reduce the loss or delay of important packets when compared with lower priority packets; at the same time, the packet loss counter in the received side has no knowledge about priority of the packets and it counts lost packets all the same.

The use of a queue management method in the communication system makes it easy to link sender rate adaptive control with packet priority marking by setting different queues or virtual queues in a queue management system to provide different treatment for different priority packets.

An important objective of this project 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 5.6.

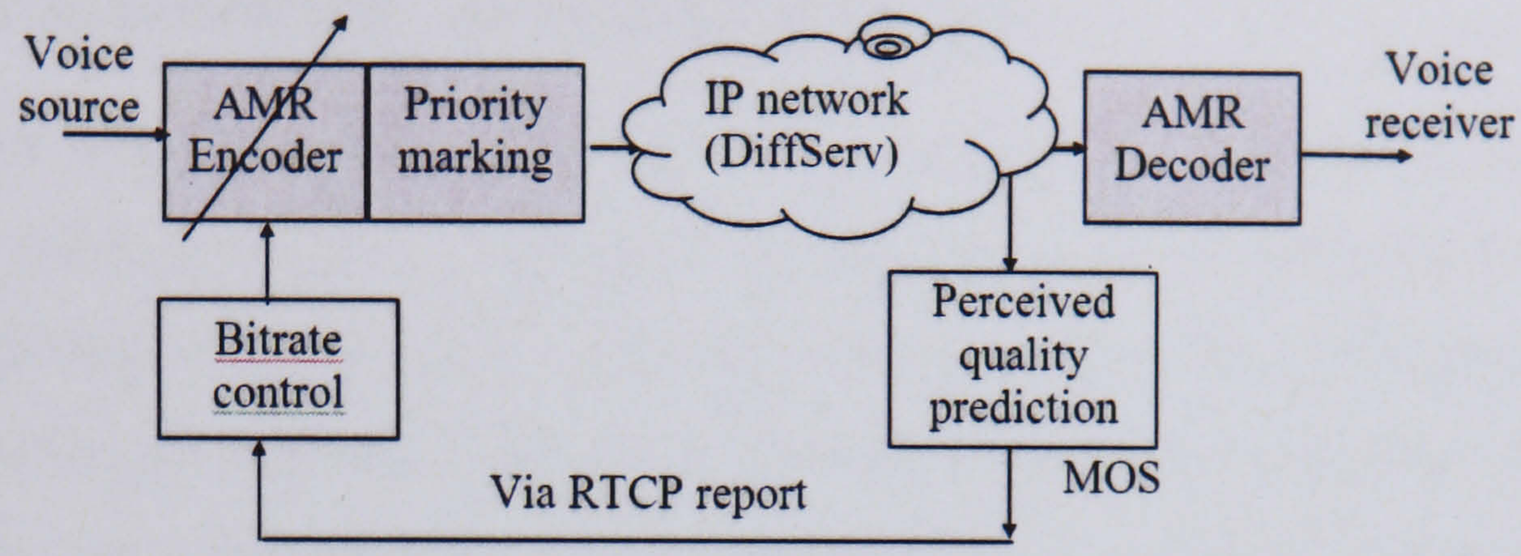


Figure 5.6: Combined control mechanism

As shown in Figure 5.6, the bit rate of the AMR codec is adjusted in accordance with the objectively 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.

A simulation test is carried out to evaluate the performance of the proposed combined control mechanism compared with those two individual control methods. Speech samples are injected in to the simulated network and the performance of the control mechanism is measured by the objective user perceived speech quality measurement method PESQ. The following section describes the detailed simulation setup and the experiments.

5.3 Simulation Systems and Experiments for the Combined Scheme

To evaluate the performance of the new proposed combined speech quality control method, some experiments need to be carried out. This section introduces the simulation tool model,

the quality prediction and performance evaluation model, the algorithm integration into the test system procedures and the test experiments running with real speech and simulated network conditions.

5.3.1 Simulation Tool - the Network Simulator

Network Simulator (NS-2) [109] used to run the simulation of the VoIP system and the proposed control mechanism. It is an open source project started in 1995 by the Information Sciences Institute of the University of Southern California. In the years of development and with help of the open source community, different projects using this system have added function blocks such as new types of protocol, new types of links or traffics in to the system.

Although in most of the cases, there are sufficient amount of defined objects including links and nodes can be used to build up a simulation network, users can adapt the software to their needs by declaring new objects such as new types of links or nodes.

Network Simulator allows the simulation of complex networks with various nodes, various types of traffics and various types of links on a single computer. The complexity of the simulated network is mainly limited by the processing power of the simulation computer. But due to the power limitation of the PC used, simulation time of a complicated network may take a significant amount of time to finish. And it increases dramatically when the network get slightly complicated.

The simulator is written in C++. And the simulations are controlled by a script written in OTcl, an object oriented extension of the Tcl script language. User with a basic knowledge of programming can use it to create a simple simulation example to start with. The central feature of the simulation system is the scheduler. It allows the user to define various events such as the start of traffics, or the change of parameters, simply by arranging them into a time table. After simulation starts, those events will kick off follow the sequence in the time table and the simulation goes on.

There is a graphical interface called Network Animator (NAM) developed for the simu-

lation visualization and control. It can demonstrate movement of packets in the simulated network with visualized effects like queuing and packets loss for example.

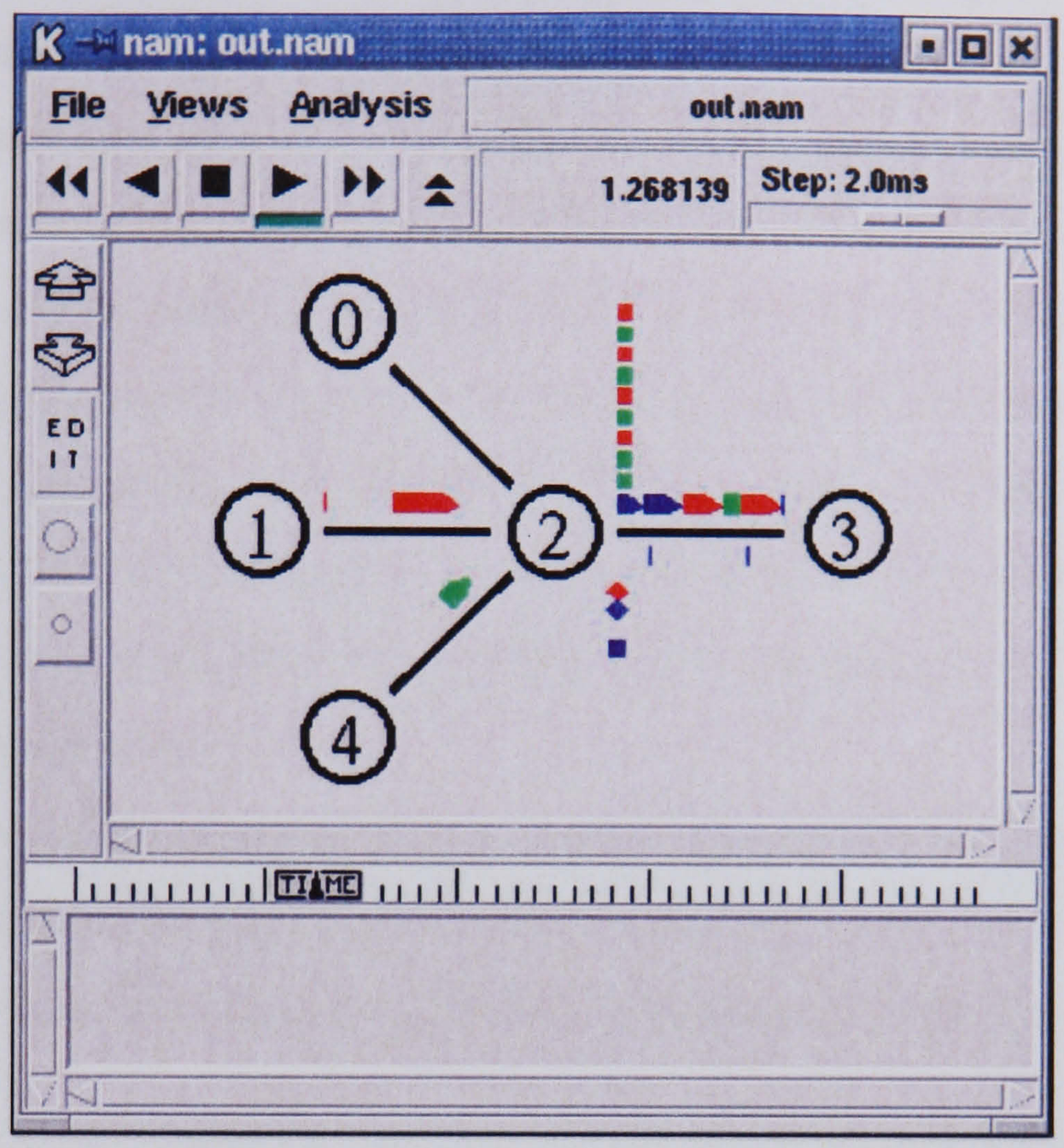


Figure 5.7: A typical simulation topology in NS-2

Figure 5.7 shows a typical simulation topology in Network Simulator. One node (node 3) receives packets from 3 different traffic sources. The node 1 and node 4 are sending a constant bit rate packet stream using UDP, denoted by the red and green traffic in the figure. The traffics congested at node 2 with a file transfer between node 0 and node 3, which is a bulky traffic using FTP over TCP, denoted by the blue traffic. Due to insufficient bandwidth of the link connecting node 2 and node 3, packets have to be queued in node 2 and if the queue is overloaded packets can be dropped.

5.3.2 Packet Loss Simulator

Network simulator is used to simulate network behavior with out sending real data through the simulated links but it is possible to inject real data in to the simulated network or extract traces from the simulated network and study the effect to real data. This is a good way to link simulated system to the real world. A packet loss simulator is designed to do the job.

Speech files can be coded using the real codec, and thus create a file represents the bit stream being send through the simulated network. By modifying the bit stream file correspond to simulated network impairment such as packet loss or wrong sequence, the simulated network effect can be linked to real world speech. After decode the modified bit stream file, the degraded speech can be reconstructed. Then the impairment of the simulated network can be measured by PESQ or a human listener.

This process is based on assumption that a jitter buffer is used to convert delay jitter into packet loss. So in the simulation, a certain format of jitter buffer need to be set in each receiving node. If packets arrive later than the buffer limit, it will be dropped thus cause packet loss in the bit stream.

The basic model for packet loss is Bernoulli loss model, which drops the packets randomly up to a given loss percentage. There are other more sophisticated models such as the 2-state Gillbert model or the n-state Markov model based on the calculation of the loss probability of packets depending on the loss probability of the last one or n packets. Details of those models can be find in [113] and [114].

Figure 5.8 shows the concept of loss simulator. On the left incomes the original speech file. Once it been encoded to encoded bit stream, it can be processed by the loss simulator following certain simulation rules. The resulted packet loss impaired bit stream can then be send to decoder and the degraded speech file can be generated. In the simulation tests, gender difference in speech samples is considered. The result are either averaged for male and female speech, or include balanced male and female speech parts in one speech sample.

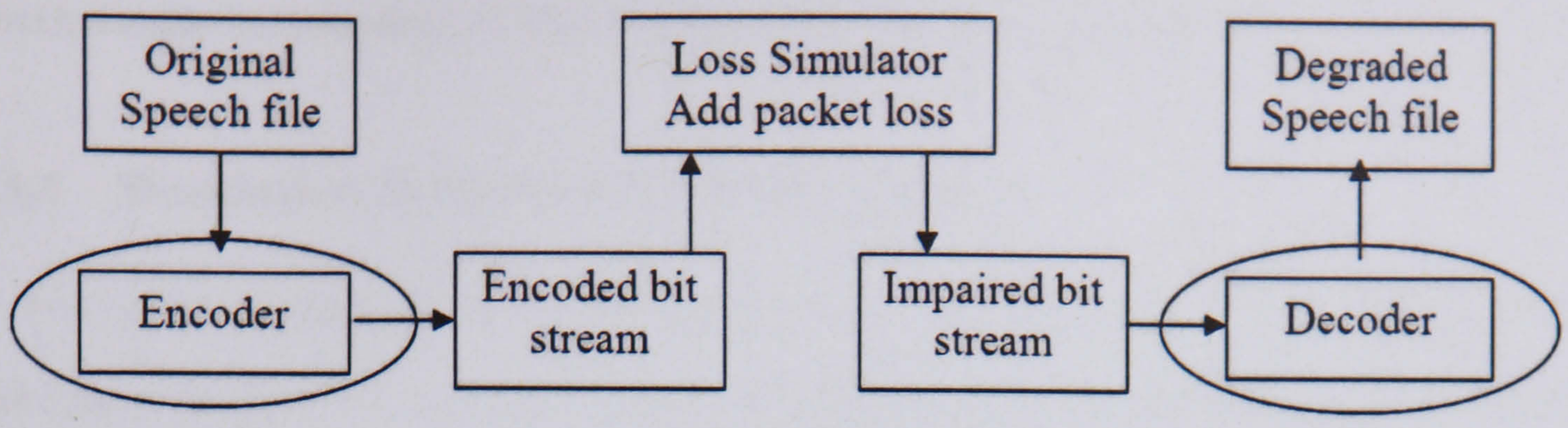


Figure 5.8: Connect loss simulator to real speech samples

5.3.3 The PESQ Based User Perceived Speech Quality Measurement System

As discussed in previous chapters, PESQ is an intrusive objective user perceived speech quality measurement method. It can objectively compare the difference of reference speech and the degraded speech. The distortion to speech files is measured by PESQ at speech level, which means the network impairments are transparent to the PESQ. This make it easier for us to setup a simulated network to perform designed control mechanisms but still measure the network impairment in real speech level.

So if the AMR rate in a certain period of speech stream is known, together with the loss rate of that period, it is possible to use the loss simulation system to generate degraded speech. If the process is repeated enough times to average out the impact of randomness in loss location, it is possible to build the link between user perceived speech quality prediction i.e. PESQ score with AMR codec rate and loss rate. This link is denoted as:

$$MOS = PESQ\{AMRrate, lossrate\} \tag{5.1}$$

It is worth to note that although the network simulator can simulate real time situation in a simulated network, PESQ measurement can not measure absolute delay between the sender and the receiver. The delay impact to end-to-end user perceived speech quality need to be considered separately. How ever, PESQ can be used to study the relative delay difference in a

speech sample i.e. introduce of silence gaps or omit of speech samples.

5.3.4 Simulation Setup and the Experiment

The combined QoS control scheme was set up as shown in Figure 5.9. It consists of three main parts: (I) An NS-2 network simulator to simulate multiple VoIP flows and IP networks with congestion; (II) a VoIP simulation system to simulate VoIP flow, which includes an AMR encoder/marker, loss simulator, decoder, and a user perceived speech quality driven control module, which can switch between the adaptive rate control, priority marking method and the combination of the adaptive rate control method and the priority marking method, and (III) a perceived quality evaluation system to provide a measure of the overall speech quality and quantify the performance of each control method.

Network Bottleneck Simulation

A one hop bottleneck network topology is simulated using NS-2, as shown in Figure 5.9(I). The reason to simulate the one hop bottleneck is that in most telecommunication system, congestion happens in a single bottleneck for instance the wireless hot-spot access point or an international gateway.

A total of N adaptive AMR sources were simulated for VoIP traffic. This assumed that the available bandwidth was shared among these UDP sources. In the real world, the Internet is shared with all kinds of applications but a managed telecommunication carrier IP network is simulated, where the traffic types are mostly VoIP applications because the entering point to this VoIP network is controlled by the telecommunication company. The telecommunication company will be interested to see how to improve user perceived quality without increase resource limitation.

All the sender sources were set as constant bit rate (CBR) UDP source in order to match with the simulation of VoIP flow in part (II) of Figure 5.9. Voice activity detection function VAD for AMR codec was not activated there. If the VAD function is on, the AMR could

change the sending rate and make the bit stream not constant bit rate anymore. The sender bit rate (plus header) was set according to the required bit rate for adaptive AMR codec i.e. when the control mechanism request the sender rate adjust to the highest rate AMR122, the CBR rate will change to the related rate (plus header over head) accordingly. In the NS-2 simulator, CBR source can change the send rate by request command but still using the name CBR. All flows sent by traffic source was traced by the network simulator. The loss location information was collected and sent back to the loss simulator in part (II).

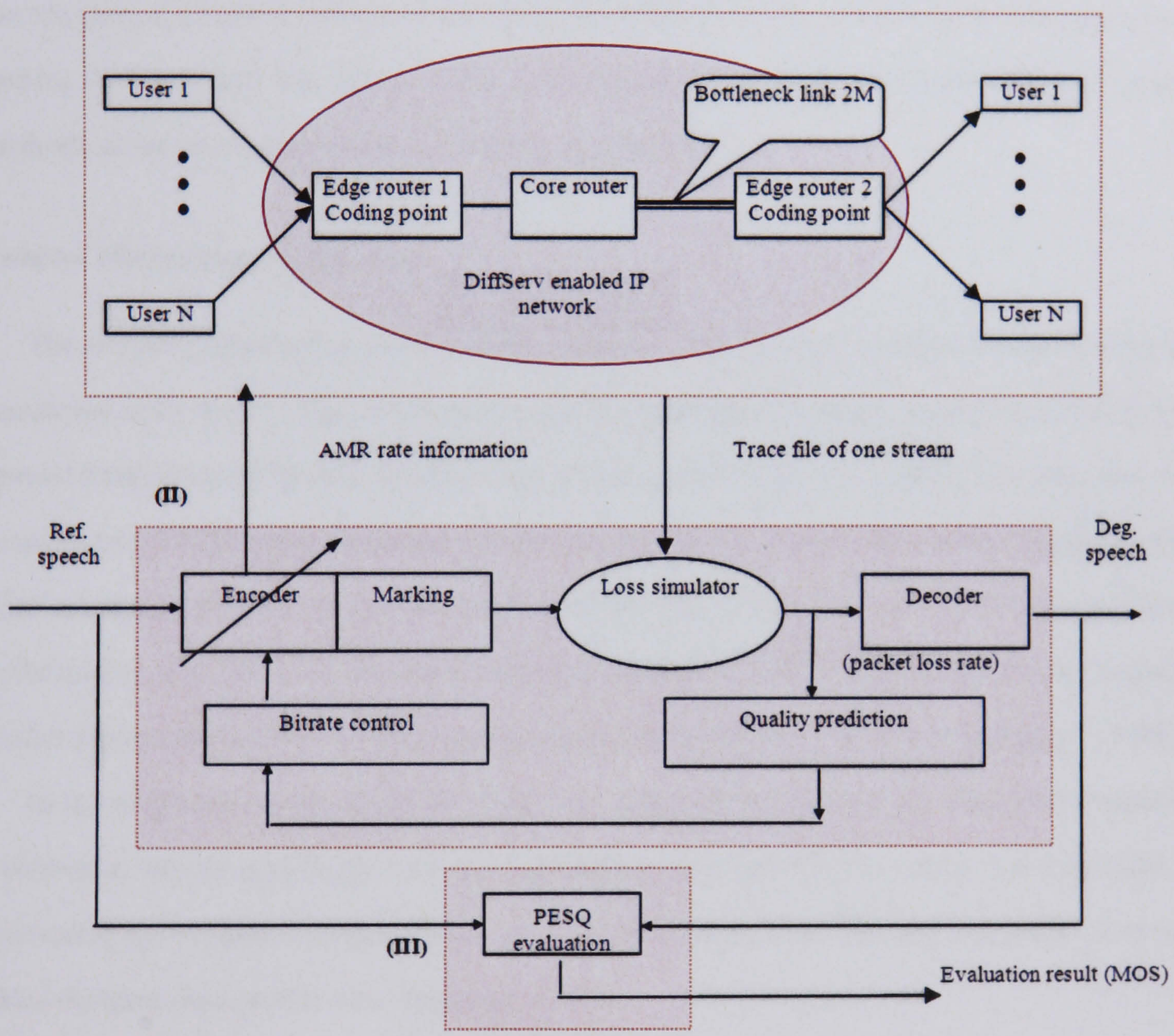


Figure 5.9: Simulation system for combined QoS control scheme

As described in Subsection 5.3.2, a VoIP flow was simulated via encoder with marking and rate control scheme, loss simulator and decoder. The loss information was also sent to

the quality prediction module to obtain a user perceived speech quality score, in this case the predicted PESQ measurement score. The MOS score was then fed back to the sender side for bit rate control.

A single hop of 2 Mbit/s bandwidth, representing a bottleneck link, was set in a DiffServ enabled IP network. 2 Mbit/s is a typical E1 connection bandwidth so it was selected to represent the bottleneck link. Different number of users joined the bottleneck link in different stage of the simulation, which represents the bottleneck link getting more and more busy and the congestion situation getting worse. With the increase in the number of simultaneous users sharing the bottleneck link, it is possible to investigate the performance of different QoS control methods under increasing network congestion situations.

Control Mechanism Integration

The overall performance of each of the different QoS control methods is evaluated by the evaluation system (III). Perceived speech quality prediction is based on the ITU PESQ (Perceptual Evaluation of Speech Quality). For every control loop (i.e. every 5 seconds, the time between two RTCP report message), speech quality (MOS) is calculated in the 'quality prediction' module, depending on the network packet loss and AMR mode. In practice, this predicted voice quality could be used non-intrusively as described in [108]. The predicted speech quality is also used to evaluate the overall quality of the control schemes as shown in Figure 5.9(III).

In the simulation study, the MOS value was computed with the ITU-T PESQ by comparing a reference sample speech file with the degraded sample speech file, which was generated by processing the reference sample speech in accordance with the AMR rate and packet loss rate. More detailed explanation was described in [76].

Simulation of the Priority Marking Method

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 param-

eter that indicates whether it is voiced/unvoiced for each frame was extracted directly from the decoder's voiced_hangover flag for simplicity. Reference [106] and [107] give more detailed explanations 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. For simplicity, the DiffServ policyer used for our simulation is a Time Sliding Window with 2 colour marking policer (built-in function supported by NS-2). It uses TSW2CM policer in the network simulator. In the TSW2CM policer, Committed Information Rate (CIR) and a drop precedence of two levels are used. The basic idea of the TSW2CM policer is that a lower precedence is used when the CIR is exceeded. The default scheduling mode is Round Robin in the simulation.

Simulation of the Perceived Quality Driven Rate-Adaptive Control Method

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. Perceived speech quality MOS score is used as a control metric to drive the control mechanism. MOS is predicted at the receive side for each RTCP interval (for instance 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 Algorithm 2.

MOS prediction is based on the ITU PESQ measurements for a given AMR rate and packet loss rate (As described in Subsection 5.3.2. The predicted packet loss in Equation 5.1 is based on the Equation 5.2:

$$lossrate = (MR * N - BW)/(MR * N) * 100\% \tag{5.2}$$

Algorithm 2 Rate-adaptive control loop pseudo code

```
1: For each RTCP report
2: { bitrate_old = bitrate_new;
3:   MOS_old = MOS_new; // get the new status
4:   get MOS_new from RTCP report;
5:   // compare MOS scores
6:     if (MOS_new > MOS_max) goto NOCHANGE
7:     else if ((MOS_new – MOS_old > threshold1) && (bitrate_old != bitrate_max))
8:        bitrate_new=next_higher_bitrate ; //increase the bit rate
9:     else if ((MOS_old – MOS_new) > threshold2)
10:       bitrate_new=next_half_lower_bitrate; //halve the bit rate
11:    else if (( threshold1 < MOS_old – MOS_new < threshold2 ) && (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:    get MOS_predicted from PESQ{AMR_rate, lossrate_predicted}
18:    if (MOS_new>MOS_predicted)  bitrate_new=bitrate_old; // no change of bit rate
19:    eles send bitrate_new to sender;  //control the sender
20: }
```

Where loss rate is the predicted packet loss rate for next control step, MR is the next step AMR rate, N is the number of users and BW is the bandwidth they have 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 predict MOS driven control idea.

In the simulation, the threshold1 in the Algorithm 2 was set to 0.2 in order to avoid unnecessary fluctuation, and the threshold2 was set to 0.5 to indicate an obvious decrease in perceived speech quality. The maximum MOS (MOS_max) achievable for the 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 send bit rate was sent back to NS-2 simulator to adjust the source bit rate. For simplicity, it is assumed that all N sources in the NS-2 use the same AMR sender bit rate at the beginning and are adjusted to the same bit rate when adaptation occurs.

Simulation of the 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 evaluation results of the simulation are discussed in next Section 5.4.

5.4 Simulation Result and Discussion

In order to investigate how the QoS control schemes (predicted user perceived speech quality driven rate adaptive control method, priority marking QoS enhancement method and the new proposed combined control method) affect perceived speech quality under different network conditions, different network congestion scenarios are simulated using the network simulator NS-2. The bandwidth of the bottleneck link was set to a fixed value (2 Mbps) with a delay of 1ms. The number of the streams sharing the link was increased from a small number of 70 to a large number of 140 to simulate different congestion scenarios.

5.4.1 Performance Comparison of Four Control Methods

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 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. In this case, packet loss rate for non-control scheme was measured more than 40% and the speech quality was unrecognizable. The latest investigation about PESQ method's performance in high packet loss situation [61] suggested that MOS score is much lower from PESQ result so the increase of the user number after 140 is stopped as the result

will be meaningless and useless if continued.

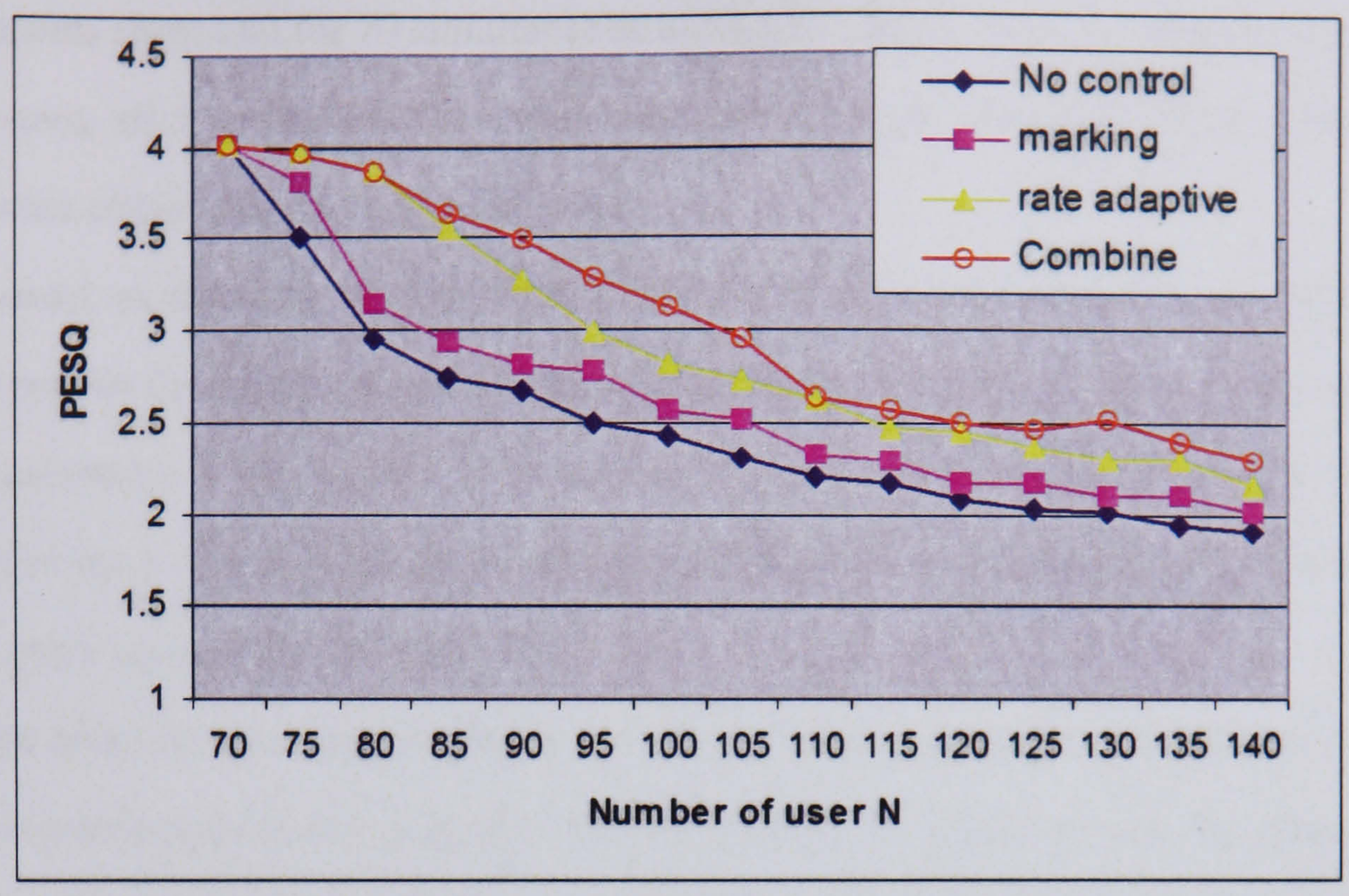


Figure 5.10: MOS v.s. number of users, with different control and non-control methods

In order to compare the performance between the different QoS control schemes and a "no control scheme", the priority marking only scheme is also implemented, together with the rate-adaptive only scheme and no control schemes. They are just subset of the new combined method so it is possible to remove relative parts of the simulation setup to represent the individual methods and it is not necessary to discuss simulation system details here again.

For the priority marking and no control schemes, the send bit rate of the AMR codec was set to a fixed rate of 12.2 Kb/s, which is a common practice in the telecom industry when AMR is first introduced. For instance Motorola was using the AMR code rate AMR122 for some mobiles and set it to constant 12.2Kb/s in some early stage mobiles to reduce complexity. Of course AMR has been fully implemented now.

For the 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 described. Figure 5.10 compares the results for all four

methods.

The results show that for 70 simultaneous users, i.e. when there is no congestion happening in the system, 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 5.10, the drop of the user perceived speech quality follows a similar pattern for all four schemes when the bottleneck link is shared by increasing number of coinstantaneous VoIP streams, because they all suffering from the increasing packet loss happened in the bottleneck link (see Figure 5.9) although the loss rate maybe different due to different optimization mechanisms used.

For the adaptive rate control scheme, the drop of speech quality is less steep compare with the "non-optimization-at-all" scheme. This is because the MOS driven rate adaptation can choose the best-optimized AMR rate in an almost real time fashion for each control loop to minimize the affect of codec rate decrease and packet loss increase and provide better service to the end users.

For the priority-marking scheme, the improvement over the non-control scheme is stable although not very significant. This is most likely because although the DiffServ method can be used to treat different packets with different priority (i.e. differentiate higher priority packets with lower packet loss rate), there is no absolute guarantee that every higher priority packets will arrive in certain time. And higher priority packets still have chance to be dropped, especially when the congestion is higher than the specified Committed Information Rate (CIR) in the queuing system .

In this simulation, the whole active speech part are marked as important and high priority but there is possibly space of improvement by using more sophisticated method to mark the importance in better detail (related to the speech character of individual speech) and map them to a finer classified DiffServ system.

As shown in the simulation result summarizing Figure 5.10, the performance of the new proposed combined scheme is always better than those two individual control schemes and the

non-control scheme. This suggest the combined user perceived speech quality driven codec rate adaptive control method and the speech priority marking scheme method does give higher over all performance in terms of user perceived speech quality i.e. PESQ score.

5.4.2 Performance of Loss Rate Driven v.s. MOS Driven Method

A packet loss driven rate adaptive control mechanism is also simulated to compare with the predicted MOS driven rate adaptive control method. The single packet loss rate control uses the same idea as shown in Algorithm 2 but use packet loss rate instead of predicted PESQ score as the parameter to control the sender rate. Figure 5.11 gives the results.

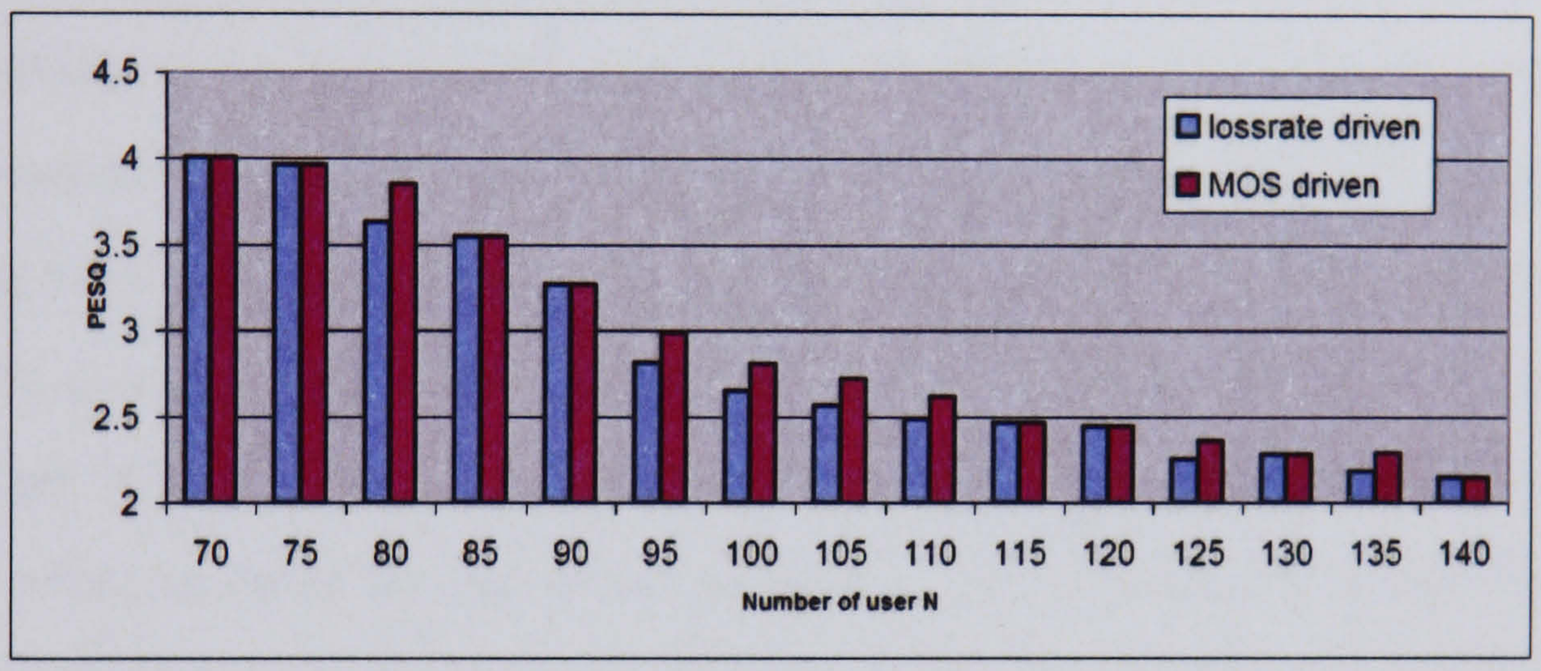


Figure 5.11: MOS v.s. Number of users, with two different rate adaptive control

The results show the over all performance of the predicted MOS driven method is slightly better than the single parameter driven rate adaptive control approach. Same as in the Figure 5.10, the result curve is not smooth because the control steps and AMR rates are not smooth. It is worth to point out that on some situation point, the predicted MOS driven method gives the same next rate decision as the single packet loss rate driven method, thus the result of the measured speech quality are the same for both methods.

In the MOS driven rate adaptive QoS control scheme, only the AMR rate and packet loss rate are considered in the MOS prediction, this is why the improvement from loss rate driven to MOS driven method is not very significant. When more parameters are considered (such

as network condition including other traffic sharing information, delay condition, delay jitter condition and multiple bottleneck information etc.) in the MOS predicted process in a more sophisticated rate adaptive control module, a more accurate user perceived speech quality can be predicted, thus the predict MOS driven rate adaptive control performance should provide a better performance.

5.5 Summary

In this research, a new QoS enhancement scheme is proposed, which combines the strengths of codec rate adaptive control method and priority marking QoS enhancement method, and uses a predicted objective measurement of user perceived speech quality as a control parameter.

This research 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-T international standard PESQ. The predicted perceived speech quality metric (measured by PESQ), instead of individual network parameters such as packet loss rate, are used to control the AMR codec's 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. Our results also show that the use MOS as a parameter to control the send rate adaptation can improve the overall perceived speech quality.

Currently only RTP type of stream is considered in the test setup. It could have more traffics in a real network other than RTP. In the future, the research could be extended to investigate the new combined method in a TCP/UDP mixed environment. The effects of delay in the DiffServ model and the use of conversational speech quality as metric to control rate adaptation could also lead to a further investigation.

Chapter 6

Discussion, Conclusions and Future Work

6.1 Introduction

VoIP is a rapidly developing technology and it is now widely used in telecommunication systems including mobiles. Due to the packet switching nature of the IP network and the complexity of the VoIP system, speech quality is adversely to be affected by different factors in the VoIP system.

Thus, the need to accurately measure and enhance voice quality in VoIP applications is an important requirement for all parties in the telecommunication market, including equipment providers, network operators, service operators and end customers.

The Mean Opinion Score (MOS) is the international standard for subjective speech quality measurement method. But subjective measurement is expensive, time consuming and difficult to repeat. Thus, objective speech quality measurement method is more attractive to meet the increasing demand for accurate yet efficient speech quality measurement.

Compared with non-intrusive speech quality measurement methods (such as E-model), intrusive speech quality measurement methods, such as PESQ is more accurate and efficient. To use PESQ correctly and accurately in a real live network, a calibration process is important for test results to be meaningful.

With the help of an accurate speech quality measurement test platform, a speech quality improvement method can be developed and the performance enhancement measured and com-

pared.

The main aims of this research are: (1) to build a live speech quality measurement platform including live VoIP network and real wireless VoIP mobile, and to undertake a fundamental investigation of PESQ's performance in the live platform and improve the measurement accuracy, (2) to exploit the nature of human speech and features of codecs in the design of novel jitter buffer algorithms for VoIP applications, the wireless mobile VoIP application in particular, and (3) to apply the knowledge of the network and application layer QoS mechanism, feature of codec and human speech into development of novel and efficient Quality of Service control method for VoIP system.

This chapter concludes the main contribution of this work and highlights the novelty, together with some limitations discussion and future work suggestions.

6.2 Contribution to Knowledge

In summary, there are 3 points of main contributions to knowledge from this project:

(1). The discovery of PESQ errors in wireless mobile VoIP environment.

PESQ is the basic of the international objective speech quality measurement standard and widely used in PSTN land line speech quality measurement and other areas. However, its performance needs to be investigated when been used in new VoIP environment such as the wireless mobile VoIP environment.

This research discovered that PESQ may give incorrect measurement result under specific conditions and found those error scenarios during the test, which investigated more than 1800 mobile-to-mobile and mobile-to-PSTN calls over 3 month, and pointed ways to avoid the wrong PESQ scores.

The error scenario includes the long silence in speech sample caused wrong relative delay reading and wrong PESQ score cases, the relative delay misalignment caused wrong PESQ

score case and other issues found in the test platform calibration part, which including the volume mismatch issue and the audio feature impairment issue.

This detailed investigation can help further research to avoid misleading PESQ scores in certain scenarios and improve the speech quality measurement systems accuracy and reliability.

This work is presented in Chapter 3.

(2). The development of a novel application layer VoIP speech quality enhancement mechanism i.e. a new jitter buffer algorithm, which can work as an add-on to static and adaptive jitter buffers

To enhance user perceived speech quality is an important aim in VoIP system development. Jitter buffer plays an important role in the speech quality improvement process. In some VoIP environment such as the wireless mobile VoIP application, processing power is another important issue need to be addressed for speech quality enhancement.

The new proposed adaptive jitter buffer algorithm exploits the talk silence interleaving nature of human speech, the silence detection feature of codec and the repeat frame feature of the DSP module to provide a better yet processing power efficient solution to the speech quality enhancement needs. This new method can work alone or be added on to existing static or adaptive jitter buffer and improve their performance.

The results are tested on live and emulated VoIP networks with real wireless mobile VoIP applications. The result shows that the new adaptive jitter buffer algorithm can improve perceived speech quality with little processing power consumption and does not introduce excessive end-to-end delay. The algorithm is simple enough to be ported into other existing static or adaptive jitter buffers as well. This part work has already attracted some business interest.

This work is presented in Chapter 4.

(3). The development of a combined rate adaptive and priority marking control scheme to enhance user perceived speech quality

The VoIP speech quality may be affected by different impairments, which are from different layer of the VoIP system. The new combined VoIP speech quality enhancement mechanism can make use of the application layer information, such as codec rate and speech content importance, and the network layer information including DiffServ settings, IP packet loss rate and other RTCP statistics. The new combined control mechanism can also utilize the PESQ algorithm's speech quality measurement function to predict the user perceived speech quality and make control decisions. The experiments are carried out with real speech samples on simulated network conditions. The result confirms the new proposed combined speech quality driven control mechanism can provide better overall speech quality compared with individual control methods i.e. the priority marking method and the rate adaptive control method. The user perceived speech quality driven control mechanism is proved to perform better than a network parameter (i.e. packet loss rate) driven control method.

This work is presented in Chapter 5.

6.3 Limitation of Current Work

There are number of limitations of the current work that need to be highlighted.

(1). Limitation of Various Application Environment

The work of the new adaptive jitter buffer algorithm is carried out with a focus on wireless mobile VoIP environment. Although the detailed analysis shows the method is generic and can be used in other VoIP applications, the validation is necessary before it can be widely deployed into live products. The parameters could be fine tuned to react to the local environment and provide best performance. To automatic adjust the parameters in an adaptive jitter buffer according to local network environment is however a future work focus.

The work of the new combined rate adaptive and priority marking control scheme is carried out with an assumption of a limited bottleneck environment. The environment scenario could

be different if this mechanism is deployed in different network conditions so the algorithm's performance needs to be validated before wide deployment.

Although some of the work is developed on real VoIP mobiles and evaluated on live Internet conditions, the performance is valid only if the Internet behaves the same as the test time because the live internet performance is out of control of the test platform.

With more infrastructure been developed and build, it is most likely that the Internet traffic condition is improving from time to time. Thus the parameters of control mechanisms may need to be adjust to the improved condition.

(2). Limitation of Combined Control Mechanism

In the new combined control mechanism, an assumption is made that an modified RTCP protocol can be developed to carry the control information across the network and layers. This work is not done in this study and it is not a straight forward task.

To develop a protocol maybe not difficult but to deploy it in the Internet and get the cooperation of other parties in the telecommunication system is not only a technical challenge but also a commercial challenge because other parties including network operators and telecommunication carriers may not want to modify their existing control protocol to suit a specific algorithm.

(3). Limited Validation of Results

This thesis has presented a few new speech quality enhancement methods including a new jitter buffer algorithm and a combined speech quality driven control mechanism. The performance of the new methods are evaluated mainly by PESQ, which is limited as an objective speech quality measurement method.

By compare the new methods' PESQ result with existing methods' PESQ score, advantages of the new methods are evaluated. Although PESQ result may not be exactly consistent with mean opinion score, the evaluation of the quality improvement is still valid.

For all speech quality enhancement methods, the ultimate approval should come from the human user. The ultimate evaluation should be based on group of human subjective tests, as the ITU-T.800 standard recommended. This is an expensive and time consuming task and not possible to be solved in this thesis.

6.4 Conclusions

An important conclusion of this project is that user perceived speech quality in VoIP systems can be enhanced by various methods such as exploiting feature of human speech and the features of IP telecommunication system. The objective speech quality measurement method PESQ is an efficient tool for telecommunication systems' speech quality enhancement provided its use has been accurately calibrated. The calibrated objective speech quality measurement test platform can be used to measure and enhance user perceived speech quality, not only for VoIP systems but also for other speech communication systems.

A detailed investigation about PESQ algorithm's performance in VoIP environment, especially wireless mobile VoIP environment, is carried out in this research and two PESQ error cases are discovered in the wireless mobile VoIP environment.

An user perceived speech quality measurement test platform is developed and specifically calibrated for the wireless mobile VoIP scenario. A significant part of the work and effort of this study has gone into understanding the feature of human speech and related codec features, the speech quality measurement method, the test platform and calibration, the VoIP network especially the wireless mobile VoIP system. Based on the understanding of these, two speech quality enhancement mechanisms are proposed. The Play Late jitter buffer algorithm and combined control algorithm are tested on specific real or simulated system environment. The performance of the proposed methods need to be validated before they can be migrated to other environments. The user perceived speech quality test platform can be used for other speech based communication system without significant modifications.

6.5 Future Work

There are 4 main areas of future work.

(1). To Improve PESQ Algorithm's Accuracy

As described in the thesis, PESQ algorithm's performance is not accurate enough in some real world test environment, especially in the wireless VoIP environment. The performance issue is mainly due to the delay estimation algorithm. To improve the accuracy the delay estimation for each frame is a challenge research topic and its worth to pursue. Some basic ideas may include to introduce more sophisticated frame by frame delay alignment method. The statistical method introduced in the thesis could provide the basis for the more accurate new method.

To develop an intelligent method which can adapt to different scenario of speech impairment is interesting as well but that might be another speech quality measurement method.

(2). To Extend the Adaptive Jitter Buffer to Consider Speech Content Importance in an Efficient Manner

Another further research area is to extend the adaptive jitter buffer algorithms to take into account the speech content importance.

It is clear that different part of the speech content has different importance to user's perception of speech quality. It is worth to extend the jitter buffer algorithm to adapt taking into account the importance of the inserted or removed speech frames or samples.

However, this introduction of speech content importance evaluation on real time jitter buffer size adjustment is resource-wise expensive considering the high processing demand and memory requirement.

Therefore, how to improve the adaptive jitter buffer algorithm with consideration of speech content importance in an efficient manner becomes an interesting research topic to push ahead.

(3). To Extend the Combined Control Mechanism over Various Network Conditions

The current combined control mechanism can be extended into more complicated real network environment such as wireless internet access point or base station with multiple network hops.

The performance of the algorithm need to be calibrated and more improvements is possible because more network information could be available with variety of network condition report mechanisms including simple network management protocol and other private feedback mechanism. All of these should require further field experiments and evaluation and investigations of the implementation of the algorithm.

(4). To Explore the Perceived Quality Driven Mechanisms in Video and Multimedia Services

There is an increasing demand for objective quality measurement method of multimedia services including speech, audio, video, gaming and so on. The measurement requirement is not only for individual media form but also including their inter-operation such as synchronization and interaction. Multimedia service quality improvement and enhancement can be carried out efficiently when the quality measurement framework is developed.

There are video quality measurement algorithm such as PEVQ available and in process to be integrated into the test system. However there are lots of research and development work can be done, for instance to study the impact of video and speech synchronization issue (lip synch issue) to improve the overall multimedia service quality measurement work.

The main objective user perceived quality measurement area will keep develop as new services emerge and the research in this area becomes very interesting and worthwhile.

References

[1] R. J. B. Reynolds and A. W. Rix, "Quality VoIP - An Engineering Challenge," *BT Technology Journal*, vol. 19, pp. 23–32, Apr. 2001.

[2] 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.

[3] 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.

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

[6] Z. Qiao, L. Sun, and E. Ifeachor, "Case study of PESQ performance in live wireless mobile VoIP environment," in *Proceedings of the PIMRC 2008 - VoIP Technologies Workshop*, (Cannes, France), Sep. 2008.

[7] Z. Qiao, R. Venkatasubramanian, L. Sun, and E. Ifeachor, "A new buffer algorithm for speech quality improvement in voip systems," *Springer Wireless Personal Communications*, 2007.

[8] 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.

[9] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson., "RTP: a Transport Protocol for Real-time Applications," *RFC 3550, IETF*, July 2003. `ftp://ftp.ietf.org/rfc/rfc3550.txt`.

[10] 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), pp. 1473–1477, June 2004.

[11] B. Douskalis, *IP Telephony: The Integration of Robust VoIP Services*. Upper Saddle River, NJ07458, USA: Prentice Hall PTR, 2000. ISBN 0130141186.

[12] J. Postel, "User Datagram Protocol," *RFC 768, IETF*, Aug. 1980. `ftp://ftp.ietf.org/rfc/rfc768.txt`.

[13] 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`.

[14] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, and E. Schooler., "SIP: Session Initiation Protocol," *RFC 3261, IETF*, Jun. 2002. `ftp://ftp.ietf.org/rfc/rfc3261.txt`.

[15] 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.

[16] Information Sciences Institute, University of Southern California, "Internet Protocol," *RFC 791, IETF*, Sep. 1981. `ftp://ftp.ietf.org/rfc/rfc791.txt`.

[17] M. Yajnik, J. Kurose, and D. Towsley, "Packet loss correlation in the MBone multicast network," *Proc. IEEE Global Internet Conf.*, Nov. 1996.

[18] S.Kent and R. Atkinson, "Security Architecture for the Internet Protocol," *RFC 2401, IETF*, Nov. 1998. `ftp://ftp.ietf.org/rfc/rfc2401.txt`.

[19] M. Spencer, B. Capouch, E. Guy, F. Miller, and K. Shumard, "IAX: Inter-Asterisk eXchange Version 2," *draft-guy-iax-04, IETF*, March 2008. `tools.ietf.org/id/draft-guy-iax-04.txt`.

[20] M. Arango, A. Dugan, I. Elliott, C. Huitema, and S. Pickett, "Media Gateway Control Protocol (MGCP) Version 1.0," *RFC2705, IETF*, Oct. 1999. `tools.ietf.org/rfc/rfc2705.txt`.

[21] Xiao Lei Wang and Leung, V.C.M., " Applying PR-SCTP to transport SIP traffic," *Proc. Global Telecommunications Conference, 2005. GLOBECOM '05. IEEE*, Dec. 2005.

[22] R. Braden, L. Zhang, S. Berson, S. Herzog, and S. Jamin, "Resource ReSerVation Protocol (RSVP)," *RFC2205 IETF*, Sep. 1997.

[23] C. Perkins, O. Hodson, and V. Hardman, "A survey of packet loss recovery techniques for streaming audio," *IEEE Network*, vol. 12, pp. 40–48, Sep. 1998.

[24] A. Tanenbaum, *Computer Networks*. Pearson Education, Aug. 2002. ISBN0130384887.

[25] Microsoft, "Netmeeting," http://windowshelp.microsoft.com/Windows/en-US/Help/54a96def-4ac6-42f3-bd15-574fdf21200f1033.mspx.

[26] Twinkle Project, "Twinkle," http://www.twinklephone.com/.

[27] CounterPath Corporation, "X-Lite Softphone," http://www.counterpath.com/x-lite.html&active=4.

[28] Microsoft Corporation, "MSN Messenger," http://download.live.com/?sku=messenger.

[29] Google Corporation, "Google Talk," http://www.google.com/talk/about.html.

[30] Zyxel Communications, Inc., "V630 VoIP Wi-Fi Phone," http://www.zyxel.com/web/product_family_detail.php?PC1indexflag=20040520161246&display=7963&CategoryGroupNo=D3A209A7-3CD9-46F9-B733-1073B9548461.

[31] M. Emmelmann, S. Wiethoelter, A. Koepsel, C. Kappler, and A. Wolisz, "Moving toward seamless mobility: state of the art and emerging aspects in standardization bodies," *Wireless Personal Communications*, vol. 43, pp. 803–816, Nov. 2007.

[32] European Telecommunications Standards Institute, "Generic Access Network (GAN)," *3GPP TS 43.318 version 8.3.0 Release 8*, Jan. 2009.

[33] 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.

[34] 3G TS 26.091, "AMR Speech Codec; Error Concealment of Lost Frames," *3GPP*, June 2002. V5.0.0.

[35] Y. J. Liang, N. Farber, 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.

[36] 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.

[37] 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.

[38] H. Nyquist, "Certain topics in telegraph transmission theory," *Trans. Am. Inst. Elec. Eng*, vol. 47, pp. 617–644, 1928.

[39] International Telecommunicaion Union, "40, 32, 24, 16 kbit/s Adaptive Differential Pulse Code Modulation (ADPCM)," *ITU-T Recommendation G.726*, December 1990.

[40] 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.

[41] 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.

[42] International Telecommunicaion Union, "Coding of speech at 16 kbit/s using low-delay code excited linear prediction," *ITU-T Recommendation G.728*, September 1992.

[43] J. Bolot, S. Fosse-Parisis and D. Towsley, "Adaptive FEC-based Error Control for Interactive Audio in the Internet," in *Proceedings of IEEE INFOCOM 1999*, 1999.

[44] W. Jiang and H. Schulzrinne, "Perceived Quality of packet Audion under Bursty Losses," in *Proceedings of IEEE INFOCOM 2002*, 2002.

[45] R. Rajavelsamy, V. Jeedigunta, B. Holur, M. Choudhary, and O. Song, " Performance evaluation of VoIP over 3G-WLAN interworking system," in *Proceedings of Wireless Communications and Networking Conference, IEEE WCNC 2005*, vol. 4, pp. 2312–2317, Mar. 2005.

[46] J. Wroclawski, "The Use of RSVP with IETF Integrated Services," *RFC2210 IETF*, Sep. 1997.

[47] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An Architecture for Differentiated Services," *RFC 2475, IETF*, 1998.

[48] W. C. Hardy, *QoS Measurement and Evaluation of Telecommunications Quality of Service*. John Wiley & Sons, 2001. ISBN 0-471-49957-9.

[49] 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.

[50] 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`.

[51] 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.

[52] L. Sun, *Speech Quality Prediction for Voice over Internet Protocol Networks*. PhD thesis, University of Plymouth, January 2004. UK.

[53] International Telecommunicaion Union, "Objective Quality Measurement of Telephone-band (300-3400 Hz) Speech Codecs," *ITU-T Recommendation P.861*, Feb. 1998.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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 Tele-

phone Networks and Codecs," in *Proceedings of 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2001.

[61] 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.

[62] International Telecommunicaion Union, "In-service, Non-intrusive Measurement Device - Voice Service Measurements," *ITU-T Recommendation P.561*, Feb. 1996.

[63] International Telecommunicaion Union, "Analysis and Interpretation of INMD Voice-service Measurements," *ITU-T Recommendation P.562*, May 2000.

[64] 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.

[65] European Telecommunications Standards Institute, "Speech Communication Quality from Mouth to Ear of 3.1 kHz Handset Telephony across Networks," *Tech. Report. ETR 250*, 1996.

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

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

[68] 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.

[69] R. G. Cole and J. Rosenbluth, "Voice over IP Performance Monitoring," *ACM Computer Communication Review*, vol. 31, pp. 9–24, April 2001.

[70] 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.

[71] L. Atzori, M. Lobina, and M. Corona, "Playout buffering of speech packets based on a quality maximization approach," *IEEE Transactions on Multimedia*, vol. 8, Apr. 2006.

[72] J. Allnatt, "Subjective Rating and Apparent Magnitude," *International Journal Man - Machine Studies*, vol. 7, pp. 801–816, 1975.

[73] 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.

[74] International Telecommunicaion Union, "Definition of Categories of Speech Transmission Quality," *ITU-T Recommendation G.109*, Sep. 1998.

[75] International Telecommunicaion Union, "Method for Objective Measurement of Perceived Audio Quality," *ITU-R Recommendation BS.1387*, Nov. 2001.

[76] A. E. Conway, "A Passive Method for Monitoring Voice-over-IP Call Quality with ITU-T Objective Speech Quality Measurement Methods," in *Proc. of IEEE ICC*, vol. 4, (New York, 2002), pp. 2583–2586, April 2002.

[77] 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.

[78] 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.

[79] Renaud Cuny and Ari Lakaniemi, "VoIP in 3G networks: an end-to-end quality of service analysis," in *Proceedings of IEEE Vehicular Technology Conference, 2003*, vol. 2, pp. 930–934, 2003.

[80] A. Barbaresi and A. Mantovani, "Performance Evaluation of Quality of VoIP Service Over UMTS-UTRAN R99," in *Proceedings of IEEE International Conference on Communications, IEEE ICC 2007*, pp. 634–639, Jun. 2007.

[81] Batu Sat and Benjamin W. Wah, "Playout scheduling and loss-concealments in voip for optimizing conversational voice communication quality," in *Proceedings of the 15th international conference on Multimedia*, pp. 137–146, 2007.

[82] 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.

[83] L. Ding, A. Radwan, M. S. El-Hennawey, and R. A. Goubran, "Performance Study of Objective Voice Quality Measurement in VoIP," in *Proceeding of The 12th IEEE Symposium on Computers and Communications, 2007 (ISCC2007)*, (Aveiro), July 2007.

[84] M. Varela, I. Marsh, and B. Gronvall, "A systematic study of PESQ's behavior(from a networking perspective)," in *Proceeding of MESAQIN 2006*, (Czech Republic), June 2006.

[85] C. Hoene and E. DulamsurenLalla, "Predicting Performance of PESQ in Case of Single Frame Losses," in *Proceeding of MESAQIN 2004*, (Czech Republic), June 2004.

[86] ITU-T Study Group 12, "Report and analysis of PESQ under-prediction of EVRC family of speech codecs and proposal for PESQ enhancement for CDMA speech codecs," *Qualcomm Inc., Sprint Nextel, Verizon Wireless, Motorola Inc.*, 2007. Contribution 121.

[87] Ditech Networks, "Limitations of PESQ for Measuring Voice Quality in Mobile and VoIP Networks," 2007. http://www.ditechnetworks.com/.

[88] International Telecommunication Union , "Mapping function for transforming P.862 raw result scores to MOS-LQO," *ITU-T Recommendation P.862.1*, Nov. 2003.

[89] M. Spencer and et.al, "Asterisk:The Open Source PBX and Telephony Platform." `http://www.asterisk.org`.

[90] 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.

[91] P. Brady, "A Model for On-Off Speech Patterns in Two-way Conversation," *BT Tech. Journal*, vol. 48, pp. 2445–2472, Sep. 1969.

[92] 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.

[93] K. Kramer and C. Forrester, "Jitter buffer management," *US Patent*, 2003. US6,658,027B1.

[94] B. Sat and B. W. Wah, " Evaluation of Conversational Voice Communication Quality of the Skype, Google-Talk, Windows Live, and Yahoo Messenger VoIP Systems," in *Proceedings of IEEE Workshop on Multimedia Signal Processing 2007*, Oct. 2007.

[95] 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), pp. 1478–1483, June 2004.

[96] International Telecommunication Union, "Single Ended Method for Objective Speech Quality Assessment in Narrow-Band Telephony Applications," *ITU-T Recommendation P.563*, May 2004.

[97] E. Liu and G. Shen, "Self-adaptive jitter buffer adjustment method for packet-switched network," *US Patent application publication*, 2005. US 2005/0058146A1.

[98] A. E. Eckberg, "Techniques for jitter buffer delay management," *US Patent application publication*, 2003. US 2003/0202528A1.

[99] K. Fujimoto, S. Ata, and M. Murata, "Adaptive Playout Buffer Algorithm for Enhancing Perceived Quality of Streaming Applications," *Spinger: Telecommunication Systems*, vol. 25, pp. 2337–2342, March 2004.

[100] F. ZAFIROPOULOS and C. S. XYDEAS, "Model based packet loss concealment for AMR coders," *Proceedings of the IEEE international conference on acoustics speech and signal processing*, vol. I, pp. 112–115, 2003.

[101] NIST Internetworking Technology Group (ITG), "NistNet homepage," 2000. `http://snad.ncsl.nist.gov/itg/nistnet/`.

[102] 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.

[103] 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.

[104] 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.

[105] 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.

[106] 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.

[107] J. C. De 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.

[108] 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.

[109] Information Sciences Institute of the University of Southern California, "The Network Simulator - NS-2," `http://www.isi.edu/nsnam/ns`.

[110] Henning Sanneck and N. Tuong Long 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.

[111] K. Nichols, V. Jacobson, and L. Zhang, "A Two-bit Differentiated Services Architecture for the Internet," *RFC 2638, IETF*, July 1999.

[112] S. Floyd and V. Jacobson, "Random early detection gateways for congestion avoidance," *IEEE/ACM Transactions on Networking*, vol. I, pp. 397–413, Aug. 1993.

[113] H. Schulzrinne, J.Kurose, and D. Towsley, "Loss correlation for queues with bursty input streams," in *Proceedings of IEEE ICC*, (Chicago, USA), pp. 219–224, 1992.

[114] 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.