



KTH Electrical Engineering

Quality aspects of Internet telephony

IAN MARSH

Doctoral Dissertation
Stockholm, Sweden 2009

TRITA-EE 2009:025
ISSN: 1653-5146
ISRN KTH/EE-09/025-SE
ISBN 978-91-7415-313-2

School of Electrical Engineering
KTH, Stockholm, Sweden

Akademisk avhandling som med tillstånd av Kungliga Tekniska Högskolan
framlägges till offentlig granskning för avläggande av teknologie doktorsex-
amen i telekommunikation fredagen den 5 juni 2009 vid KTH.

© Ian Marsh, april 2009

Tryck: Universitetsservice US AB

Swedish
Institute of
Computer
Science



Swedish Institute of Computer Science, SE-164 29 Kista, SWEDEN
SICS Dissertation Series 51
ISSN-1101-1335
ISRN SICS-D-51-SE

Abstract

Internet telephony has had a tremendous impact on how people communicate. Many now maintain contact using some form of Internet telephony. Therefore the motivation for this work has been to address the quality aspects of *real-world* Internet telephony for both fixed and wireless telecommunication. The focus has been on the quality aspects of voice communication, since poor quality leads often to user dissatisfaction. The scope of the work has been broad in order to address the main factors within IP-based voice communication.

The first four chapters of this dissertation constitute the background material. The first chapter outlines where Internet telephony is deployed today. It also motivates the topics and techniques used in this research. The second chapter provides the background on Internet telephony including signalling, speech coding and voice Internetworking. The third chapter focuses solely on quality measures for packetised voice systems and finally the fourth chapter is devoted to the history of voice research.

The appendix of this dissertation constitutes the research contributions. It includes an examination of the access network, focusing on how calls are multiplexed in wired and wireless systems. Subsequently in the wireless case, we consider how to handover calls from 802.11 networks to the cellular infrastructure. We then consider the Internet backbone where most of our work is devoted to measurements specifically for Internet telephony. The applications of these measurements have been estimating telephony arrival processes, measuring call quality, and quantifying the trend in Internet telephony quality over several years. We also consider the end systems, since they are responsible for reconstructing a voice stream given loss and delay constraints. Finally we estimate voice quality using the ITU proposal PESQ and the packet loss process.

The main contribution of this work is a systematic examination of Internet telephony. We describe several methods to enable adaptable solutions for maintaining consistent voice quality. We have also found that relatively small technical changes can lead to substantial user quality improvements. A second contribution of this work is a suite of software tools designed to ascertain voice quality in IP networks. Some of these tools are in use within commercial systems today.

Acknowledgments

Two pages of acknowledgments, “Oh please”.

The first line of the acknowledgment section in my 2003 licentiate thesis reads “Writing this part of the thesis is actually *enjoyable*.” Now, in April 2009, for my doctoral dissertation the best I can come up with is “Writing this part of the dissertation is actually *weird*.” I could *never* have imagined how much more effort there was still remaining, as well as the ups and downs that would accompany them.

Some people have been responsible for getting me close to the end and they are first and foremost Prof. Gunnar Karlsson who enrolled me, kept with me, and hopefully will see me graduate. Without him none of the eight years of PhD studies would have ever happened. Also thanks to Dr. Bengt Ahlgren, my boss and lab leader of the NETS group at SICS, again without whom I would not be at this point. Thanks to you both! Acknowledgments also to Prof. Gerald Q. ”Chip” Maguire Jr. whose input and influence is present within this dissertation.

I would also like to acknowledge Janusz Launberg the business manager and Dr. Staffan Truvé the CEO at SICS. Thanks for the support over the years. I would like to thank the many other people at SICS for the creative and relaxing environment. This includes all the support staff, which seem to be sadly overlooked in many acknowledgments. The group(s) within which one works are critical, therefore the folks of NETS (formerly CNA) and the chaps at LCN deserve a special mention, some of which have become good friends, which goes to show there is more to life than just research (but not much more). To the original LCN’ers we have (almost) made it.

Some of this work has been done in collaboration with people namely Olof Hagsand, Florian Hammer, Christian Hoene, Ingemar Kaj, Moo Young Kim, and Martín Verala it was a pleasure to work with you all. The students I was responsible for during the years (chronologically) are: Zheng Sun, Anders Gunnar, Fengyi Li, Juan Carlos Martín Severiano, Viktor Yuri Diogo Nunes and Daniel Lorenzo, all of whom have been successful in their post education lives. It was a great pleasure to be involved in your education and I hereby gratefully acknowledge your contribution in my PhD dissertation.

The Swedish PhD presents an opportunity to do research. It also presents an opportunity to develop highly needed technical skills in the form of

courses. Although I never quite got the right balance between coursework and SICS duties, the educational part of my PhD was the most enjoyable and character building. The skills and patience of the teachers need to be acknowledged by me here. I hope I remembered you all (alphabetically): Daniel Andersson, György Dán, Gunnar Englund, Viktoria Fodor, Anders Forsgren, Mikael Johansson, Ingemar Kaj, Supriya Krishnamurthy, Arne Leijon, Ali Ghodsi, Dan Mattsson, Lars Rasmussen, Mickael Skoglund, Lena Wosinska and Jens Zander.

Funding is critical for the continuity of a PhD, and I have been fortunate to receive financial support from SICS as well as from Vinnova, the EU, Telia AB, Nordunet, SSF and KK-Stiftelsen.

Special thanks are due to Prof. Henning Schulzrinne, the acknowledged expert within IP-based voice communications. It is an honour for me to have Prof. Schulzrinne as an opponent for this work. Also thanks to Doc. Christer Åhlund, Prof. Carsten Griwodz and Dr. Roar Hagen for agreeing to act as grading committee members.

As I have already found my post-doc life in Portugal and I would like to thank the following people for offering me positions, Prof. Manuel Ricardo at INESC Porto, Prof. Rui Aguiar in Aviero, Prof. Luis Correia in Lisbon, Prof. Edmundo Monteiro (plus crazy family of course!), Prof. Fernando Boavida in Coimbra and finally Prof. João Barros in Porto for agreeing to a post-doc position without formally a PhD (here is the dissertation though :-)).

Many people have helped me when things were not the easiest, and I am quite sure I would not be completing the thesis without their professional and unwavering support, in particular Drs. Lars Grahn and Nina Havervall.

To the many friends I met and enjoyed the company of during the years, to name just a few, Iyad, Ehsan, Ali, Jim, György, Ilias, Nacho, John, Evgueni, Henrik, Ibrahim, Petros, Katherine, Berit, Luiza, Kia, Katalin, Adrian, cheeky Ian (another one), Gary and the many others I have surely forgotten to mention.

To my family, especially my Mother, Ray and my fantastic grandmother for all the support and love over the long education. Years ago (I think 1988) I said I wanted to do a PhD and now its nearly done! Last and not least to my devoted and (very) long suffering girlfriend Margarida (alias 'baby Gui'), you deserve the biggest thanks **of all** for accompanying me along the ups and downs of the closing steps of a PhD education. Your crazy cat deserves the final mention in this all too long acknowledgment section for chewing just about every cable I ever owned:-)

I think I'll stop there.

Ian, April 2009.

Contents

1	Introduction	13
1.1	Internet telephony introduction	13
1.1.1	PC-based Internet telephony	13
1.1.2	Broadband Internet telephony	15
1.1.3	IP telephony and the Internet backbone	15
1.1.4	Wireless Internet telephony	17
1.1.5	Summary of the introductory sections	18
1.2	Dissertation outline	18
1.3	Dissertation motivation	19
1.4	The problem statement and its relation to the publications	21
1.5	Research methods used in this dissertation	23
1.6	Paper summaries and contributions	26
1.7	Conclusions	32
1.8	Future directions	33
2	Background	35
2.1	A voice journey across the Internet	35
2.2	IP telephony signalling	36
2.2.1	H.323	38
2.2.2	SIP	39
2.2.3	A comparison of H.323 and SIP	40
2.2.4	Non-standardised signalling	42
2.3	Firewall traversal	43
2.4	Speech encoding	44
2.4.1	Pulse Code Modulation (PCM)	44
2.4.2	Adaptive differential pulse-code modulation (ADPCM)	45
2.4.3	Low bit rate models	45
2.4.4	Modern codecs GSM, G.729 and iLBC	47
2.4.5	A (very) brief history of speech coding	48
2.5	Internetworking and voice	49
2.5.1	The Real-Time Protocol (RTP)	49
2.5.2	Addressing, routing, and timing constraints	52
2.5.3	Packet delay	53

2.5.4	Packet jitter	54
2.5.5	Packet loss and redundancy schemes	56
3	VoIP quality aspects	59
3.1	Quantifying quality	59
3.2	Measuring quality	59
3.3	Quality tolerances	60
3.4	Quality and noise	61
3.5	The ITU-T E-model	62
3.6	Perceptual Evaluation of Speech Quality (PESQ)	63
3.7	Other measures	65
4	Packet-switched voice research: A brief history	67
4.1	Pre-Internet days (1970-1980)	67
4.2	A decade of research (1980-1990)	69
4.3	Emergence of telephony applications (1990-1995)	69
4.4	Early deployment days (1995-2000)	71
4.5	Internet telephony comes of age (2000-present)	72
	Appendix: Included articles	89
	A: Dimensioning links for IP telephony	93
	B: Modelling the arrival process for packet audio	114
	C: Sicsophone: A low-delay Internet telephony tool	131
	D: Measuring Internet telephony quality:Where are we today?	145
	E: Wide area measurements of VoIP quality	156
	F: Self admission control for IP telephony using early estimation	168
	G: IEEE 802.11b voice quality assessment using cross-layer information	182
	H: The design and implementation of a quality-based handover trigger	199
	I: A Systematic Study of PESQ's Performance from a Networking Perspective	213
	List of SICS publications	228

Acronyms and terms used in this thesis

Acronyms and terms	Meaning
3GPP	3rd Generation Partnership Project
BGP	Border Gateway Protocol
E-model	ITU objective quality rating
E-UTRAN	Evolved UMTS Terrestrial Radio Access Network
EPC	Evolved Packet Core
FEC	Forward Error Correction
GAN	Generic Access Network
GPRS	General Packet Radio System
GSM	Global System
H.323	ITU Internet telephony signalling protocol
ICE	Interactive Connectivity Establishment
IEEE 801.11	Wireless unlicensed Local Area Network standard
IETF	Internet Engineering Task Force
IMS	IP Multimedia Subsystem
IPTV	Internet Protocol Television
ITU	International Telecommunications Union
LTE	Long Term Evolution
MBONE	Multicast Backbone
MDC	Multiple Description Coding
MOS	Mean Opinion Score
NAT	Network Address Translation
PCM	Pulse Coded Modulation
PESQ	Perceptual Evaluation of Speech Quality
PSTN	Public Switched Telephony Network
QoS	Quality of Service
ROHC	Robust Header Compression
RTCP	Real Time Control Protocol
RTP	Real Time Protocol
SEC	Selective Error Checking
SDP	Session Description Protocol
SIP	Session Intiation Protocol
STUN	Simple Traversal of User Datagram Protocol
TURN	Traversal Using Relay NAT
UMA	Unlicensed Mobile Access
VoIP	Voice over Internet Protocol
WiFi	Commercial synonym for IEEE 802.11 standard networks
WiMAX	Commercial synonym for IEEE 802.16 standard networks

Chapter 1

Introduction

1.1 Internet telephony introduction

Real-time voice communication using IP networks is the subject of this dissertation. The scope of this dissertation is broad and includes several different aspects of real-time voice communication. The effects of the public Internet on telephony sessions have been investigated. Also within our scope is the impact of the access network, and the influence of mobile users. This includes roaming users who can utilise both IEEE 802.11 wireless and cellular networks. The end systems have also been studied and include traditional computers as well as hand-held terminals. Finally, to explicitly include the user expectations in our investigation, we have devised a method to estimate speech quality from real-time network measurements and from off-line processing of sample blocks.

In order to give some background to this dissertation, the upcoming four sections (1.1.1 to 1.1.4) provide a brief description of IP-based voice services. They include four areas in which one encounters the technology - very much from a user perspective. Each section outlines the original impetus for the particular deployment, an introduction to its functionality as well as some possible future directions for each one.

1.1.1 PC-based Internet telephony

From a technological perspective, PC-based telephony came about due to improved CPU performance, permanent and high speed Internet connections, and notably better IP telephony software. Sufficient CPU performance is needed in order to encode the voice for transmission and to decode the received samples. Speech coding is discussed in section 2.4.

Permanent connections are needed to allow incoming calls. Current PC-based telephony software allows calls to be made independently of the local network configuration; this is important as firewalls and routers have caused

setup problems in the past. IP telephony software is now available for essentially all operating systems and hardware combinations including hand-held devices and mobile phones. With this new functionality the personal computer is transitioning from a computing device to a voice enabled communication device. Phone calls are not only limited to computer to computer with PC-based telephony, but using IP to phone gateways, regular phones can also be reached.

PC-based telephony was revolutionised by the popular SkypeTM application [30]. It is a cross-platform solution that became successful partly by embracing recent technological developments, and because it provided good, free and easy voice communication. The technological developments it embraced were: Internet-specific speech coding, a firewall bypass solution, a scalable call establishment system, and an intuitive graphical user interface. Skype has continued to add functionality such as inter-operability with the telephony system, a payment scheme, and conferencing capabilities. Recently, the developers have added video and SMS capabilities.

PC to PC communication has become a major success due to Skype and similar applications. The market looks likely to grow by considering the number of Skype online users, see Figure 1.1. As of 2006 VoIP accounted for approximately 20% of the world's telephony traffic of which 4.5% has been attributed to Skype¹. Therefore, with 80% of the world's telephony traffic still being carried by traditional telephony systems, the migration of voice traffic should further motivate VoIP research.

1.1.2 Broadband Internet telephony

Given the uptake of PC-based telephony, operators realised that similar techniques had a role in cost effective solutions for their voice customers. By leveraging the low cost of high capacity long distance IP links, operators could offer cost effective telephony solutions using the Internet. Different types of operators pursue different strategies: the larger incumbent operators seek to reduce costs, whilst new operators want to enter the voice market with relatively little capital. Both types of operator tend to bundle voice services with Internet access, as the return on providing voice services is falling.

The operator usually provides the customer with a modem into which the customer connects their existing phone and Internet connection. On powering up the modem it establishes the necessary connection, allowing users to make and receive calls using their regular phone. It needs to obtain a local IP address, discover if it is behind a NAT or firewall, and register itself with a server to permit bidirectional media flows. One important phase of this establishment is to locate the correct gateway (see section 2.2).

¹http://www.telegeography.com/cu/article.php?article_id=15656

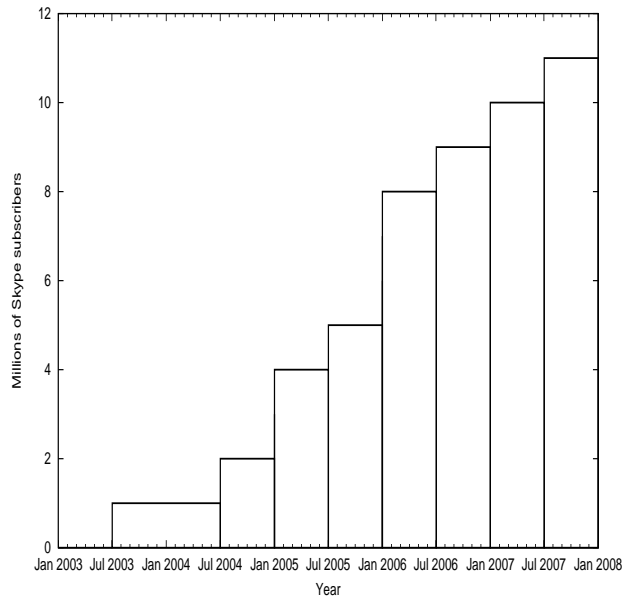


Figure 1.1: Skype usage from August 2003 to February 2008 (source www.wikipedia.org/Skype)

Call records are kept centrally and are used for billing as well as quality monitoring. Subscribers are largely unaware that their voice is partly being transported over the Internet.

Broadband telephony does not require a home computer, making it simpler, more accessible and cheaper than a PC-to-PC solution, and users do not need to be computer literate. Interoperability with the phone system is provided by the operator through a voice gateway. One problem with PC-to-PC solutions discussed in the last section, is that the caller cannot always be identified and located, which is a necessity for emergency calls. Broadband Internet telephony customers on the other hand are registered to an address and thus can make emergency calls.

Broadband telephony is growing, as customers seek to reduce their phone costs, both in terms of lower subscription charges and per minute tariffs. Additional impetus is created by the rising number of homes with broadband Internet subscriptions and (often) bundled voice subscriptions.

1.1.3 IP telephony and the Internet backbone

In the 1990s, research and small-scale tests showed that the Internet was capable of carrying real-time telephony traffic. This was demonstrated with the multicast MBONE transmissions that carried IETF meetings and space shuttle missions. Importantly, the sessions used intercontinental networks, which showed that a business case could be made for wide area real time

voice transport.

A service such as home calling was particularly popular amongst immigrant workers in the United States. Many of the schemes were (and still are) Internet based and prepaid. Traditional phones and local exchanges are used to relay the voice from the regular PSTN network to a gateway, from where the Internet carries the voice over long distance links; the phone network again provides the final leg. Thus the Internet serves as a voice bearer. Some companies saw opportunities in such services, Dialpad and Net2phone were two such examples. Importantly, they both had agreements with the long haul Internet operators. Many thousands of such companies now operate such voice services in most countries of the world.

From the user's perspective, there should be no major quality difference between telephony being carried by the Internet and a regular telephony network. From the operator's perspective on the one hand, it is important that the number of users on the IP network is controlled to avoid overload situations and hence disgruntled customers. On the other hand if a link is being leased for Internet telephony, then it makes financial sense to multiplex as many calls over that link as possible, subject to quality constraints of course. The telephone industry has a highly developed theory (and practice) to allocate calls onto high capacity trunks. This can largely be attributed to one man, A. K. Erlang who produced seminal research contributions from 1909 and onwards. The same theory can be applied to IP networks in order to deduce the allocation of calls per link.

One of the technology remnants from ATM is layer 2 switching: Multi Protocol Label Switching (MPLS) is a carrier technology for IP packets. Basically, MPLS switches labels that are added to IP packets at the ingress of a MPLS network. IP packets that belong to a call are all labelled identically and switched over a dedicated path. Therefore link dimensioning for IP telephony becomes much simpler using MPLS.

The Internet revolution initially bypassed the traditional telecommunications equipment manufacturers and operators. However, the 3rd Generation Partnership Project (3GPP), established in 1998, brought together a number of commercial, organisational and standardisation bodies to work on integrating IP into their solutions for mobile communication. 3GPP has already standardised the use of an IP based core network. Today telecommunication companies are deploying the 3GPP IP Multimedia Subsystem (IMS) to merge Internet technologies with mobile networks. So called 'Release 5' enables operators to upgrade their existing telecommunication equipment and allows a smooth transition to IP technology. IMS is based upon the Session Initialisation Protocol (SIP) which is described in section 2.2.2. The upcoming 3GPP Long Term Evolution (LTE) standard will use IP in both the access and core networks to carry data and voice traffic.

Currently local wireless IP voice services have not reached significant market penetration, as current handsets and infrastructure are dominated

by the telecommunication industry's 2nd and 3rd generation standard solutions. There can be voice quality issues with the current data-centric LAN technologies we have today. The problems are mainly due to coverage and heavy load situations. These are discussed in the next section.

1.1.4 Wireless Internet telephony

Ever more geographically local zones are being established. With the proliferation of dual-mode (local wireless and wide area cellular) telephones, local wireless based Internet telephony could represent an important opportunity for IP-based voice. The France Telecom UNIK service uses dual-mode telephones and a 802.11 gateway, France Telecom quote figures of 25,000 new subscribers per month. In the UK, British Telecom has a similar scheme and T-mobile will launch their own service in the US during 2009. The IEEE 802.11 standards are the current technology preference for local wireless access.

As far as voice traffic is concerned, there are two broad usage scenarios within local wireless networks. One is to only use the local wireless technology; voice calls are not continued should the user move from the coverage area. Therefore movement is restricted to within the coverage area. Note however that the coverage area may comprise several access points allowing some geographic area to be covered within one administrative domain. Further deployment and new technologies will allow for greater coverage in the future. Collectives are being formed based upon coverage and financial incentives to set up and share wireless networks, e.g. the Fon and Skype Zone initiatives.

Voice quality can suffer if there are radio coverage problems, interference from external sources, and excessive network load. The range for good quality varies from a few metres to a hundred meters depending on the equipment in use, obstacles, interference sources, and so on. Therefore the second scenario is to switch calls between the local wireless and cellular infrastructures in order to provide call continuity outside the coverage area of the wireless LAN. As mentioned, mobile phones and PDA's are now available with both cellular and 802.11 interfaces. This provides an option for switching to the cellular network when needed. Alternatively, if local wireless coverage is detected during a cellular call, a switch to the local network is possible, thus freeing cellular resources and potentially avoiding the cellular operator's tariffs. Entering a home or office area are typical scenarios in which a cellular call could be transferred to the local 802.11 network. The procedure of switching an ongoing call from one technology to another is known as a handover or handoff. Ideally the user should be unaware of the change, if this is the case it is known as a *seamless* handover. The current technological barriers for seamless handovers are the configuration and connection establishment mechanisms rather than the switching of the voice

stream. Switching a voice stream means receiving two parallel streams to the same terminal over different networks. Once running in parallel to the terminal, the initial stream can be stopped and the new voice stream played to the caller instead.

As we are interested in maintaining call quality, the timing of handovers from the WLAN to the cellular network is important. In the case of radio problems there might be insufficient time to initiate and start a call to the cellular network. In the case of handover due to the onset of congestion, the handover success depends on the rates of the other flows. This is due to the time needed to estimate the call quality and if need be, to initiate a cellular-based call. In the other case where a user would move out of the coverage area, there should be time to schedule the handover. The speed and path of the user movement can be tracked to estimate whether the user is moving out of coverage. In this case there is a design tradeoff: To maintain connectivity in the coverage area as long as possible to minimise the frequency of handovers on the one hand, or to reduce the probability of poor quality and switch early on the other. Therefore more conservative or aggressive switching algorithms can be envisaged.

Generic Access Network (GAN), formerly known as UMA (Unlicensed Mobile Access), is one possibility to provide seamless roaming between local and wide area networks [31]. GAN allows voice, data, and IMS/SIP applications to be accessed from a mobile phone. The operation of GAN is as follows: Once a local wireless network is detected (e.g. Bluetooth or 802.11) the handset initiates a secure IP connection through the local network to a gateway in the operator's network. A GAN server makes the handset appear as if it were connected to a new base station. Thus, when the handset moves from a cellular to a 802.11 network, it appears to the core network as if the handset is simply associated with a different base station. There is GAN support for 2nd and 3rd generation cellular technologies.

1.1.5 Summary of the introductory sections

Apart from the obvious human need to support real-time person-to-person communication over geographic distances, it is hopefully clear from the last four sections that Internet telephony has a permanent position in modern communication networks. As voice is a real-time conversational service, there are strict requirements on the end-to-end quality characteristics that the telephony operator must provide in order deliver a successful and robust service. We will now look at the task of fulfilling these requirements as research topics within this dissertation.

1.2 Dissertation outline

This section gives the motivation, problem statement, methods, conclusions and potential topics for future research. There are additionally short descriptions of the research contributions of each publication and the individual contributions of this dissertation's author.

The second chapter presents the major IP telephony building blocks. A short description of the path voice samples take from speaker to listener is given. This is to illustrate the typical processing that voice samples undergo. Subsequently, sections on signalling, speech coding, firewall traversal, voice Internetworking and human tolerances to digitised speech are elaborated upon further.

The third chapter of the dissertation concerns Internet telephony from a quality perspective. We go through some of the mechanisms used to assess speech quality, including measuring and estimating quality, plus an overview of two ITU proposals for objective speech quality assessment.

The fourth chapter is a research literature review from a historical perspective. It is divided chronologically, into episodes of the development of Internet telephony from early packet switched experiments to world-wide deployment.

The appendix of the dissertation is composed of the nine published papers. The structure of this dissertation is shown as an illustration in figure 1.2.

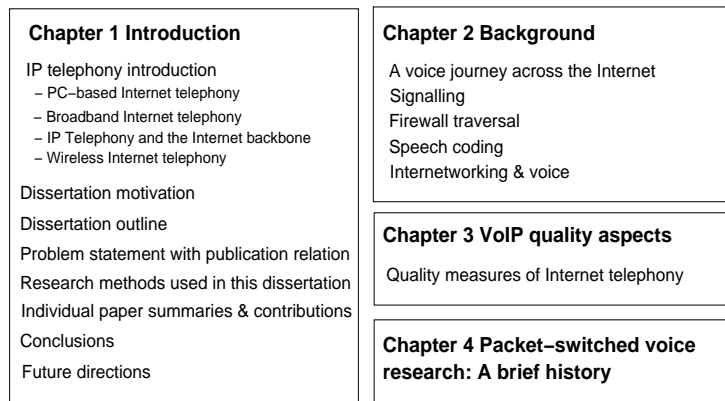
1.3 Dissertation motivation

In the previous sections we have seen various settings for IP-based voice in telecommunications systems. Although each has its own particular challenges when it comes to providing acceptable quality for its users, we can formulate a unifying motivational statement for this work: To carry real-time voice from speaker to listener with acceptable quality under a range of operating conditions.

This statement can be further subdivided into seven motivating reasons for this research.

Current relevance: Real-time voice is still the most efficient media to carry information quickly and unambiguously from person to person. Although email, instant messaging and SMS have become popular recently, the unequivocal importance of real-time voice communication remains.

Network challenges: Using IP networks to transport real-time voice can be challenging. The complex nature of bulk IP traffic makes a complete understanding of the aggregate behaviour difficult, especially when viewed



Included articles

<p>Paper A Dimensioning links for IP telephony</p> <p>Paper B Modelling the arrival process for packet audio</p> <p>Paper C Sicsophone: A Low-delay Internet Telephony Tool</p> <p>Paper D Measuring Internet Telephony Quality: Where are we today?</p> <p>Paper E Wide Area Measurements of VoIP Quality</p> <p>Paper F Self-admission control for IP telephony using early quality estimation</p> <p>Paper G IEEE 802.11b voice quality assessment using cross-layer information</p> <p>Paper H The design and implementation of a quality-based handover trigger</p> <p>Paper I A Systematic Study of PESQ's Performance from a Networking Perspective</p>
--

Figure 1.2: Dissertation structure

from different time scales. Where voice data is multiplexed with many data flows, the received speech sequence usually does not resemble the transmitted sequence. Traffic demands vary on the Internet to some degree according to popular applications and services, therefore there is no fixed target to design for. In addition to the traffic, there are differences in the operating environments, such as fixed and wireless access networks or transit and backbone networks. Despite the known user requirements for voice, the conditions under which it is delivered leads to a complex problem.

Implementation feasibility: New solutions can be introduced into IP networks. The relatively simple IP programming interface facilitates novel and innovative solutions. Whole or partial solutions are implementable using approximately 20 library functions. This is in stark contrast to the telephony system which requires detailed specialist knowledge for application development.

Subjective assessment: It is possible to assess perceptually the success or failure of IP-based voice research. In subjective assessments real people listen and indicate scores according to the quality of the speech. There are two forms of subjective tests, comparative or absolute. Comparative tests indicate the perceptual gain with and without improvement. Absolute tests simply ask whether the quality delivered is acceptable without a comparative signal. The major disadvantage with subjective tests is that real subjects are required, the trials should be conducted according to expensive standard procedures and eventually test subjects become tired. There are alternative objective measures, which we have used in our research, but are clearly less accurate.

Understanding broader traffic issues: There are two aspects to be considered in a broader sense. First, since the voice data is generated as a (nearly) periodic stream, it acts effectively as an active probe along the network path. The stream can reveal useful information of the path conditions by reporting properties such as the loss and delay distributions. Second, investigating the effect of large data volumes on “thin” voice streams may indicate what measures need to be taken to implement protection for delay sensitive traffic. In some respects understanding the behaviour of this mixed traffic is the key to better network planning. If network mechanisms are to be introduced to maintain balance, predictability and quality of service for voice and other time sensitive media, then the interplay of mixed traffic types should be investigated.

Terminal heterogeneity: Ultimately the voice must be replayed for a listener. Minimally, the timing information must be restored to produce the original speech pattern and (optionally) lost frames masked. The functionality of the receiver depends very much on the type of hardware, operating system, computational power, battery capacity, and the network to which the terminal is connected. The motivation of this work therefore, with respect to terminal heterogeneity, is that each solution needs careful tailoring for a particular hardware/software combination.

1.4 The problem statement and its relation to the publications

Let us begin with a non-problem. In *principle*, capturing, processing, transmitting and receiving real-time voice samples that use an IP infrastructure is non-problematic. Voice samples are captured, coded and sent at constant intervals. Samples are batched together as packets, addressed and sent across shared access, transit and backbone networks. The packets are received and are buffered in order to provide a continuous stream of samples

Paper	Title
A	Dimensioning links for IP telephony
B	Modelling the arrival process for packet audio
C	Sicsophone: A low-delay Internet telephony tool
D	Measuring Internet telephony quality: Where are we today?
E	Wide area measurements of VoIP quality
F	Self-admission control for IP telephony using early quality estimation
G	IEEE 802.11b voice quality assessment using cross-layer information
H	The design and implementation of a quality-based handover trigger
I	A systematic study of PESQ's performance from a networking perspective

Table 1.1: List of papers in the dissertation

for an application. The samples are removed from the packets, the timing restored and passed to the operating system for playout. The purpose of this brief explanation is to illustrate that no extraordinary processing needs to be performed in the absence of network, or end system abnormalities. In other terms, well dimensioned networks and capable end systems should be sufficient for ample quality voice communication.

The problem statement therefore is as follows: *Delivering a real-time good quality voice service over multiservice, multiplexed IP communication paths supporting stationary and mobile users using heterogeneous terminals.* Using the publications included in this dissertation (see Table 1.1), we will now discuss the problem statement and their relation.

In paper **A** we look at how to allocate resources for a single service voice network. The problem to solve is how to regulate the number of calls entering a system so that acceptable user quality can be delivered. The paper considers an IP network in which only voice is carried, somewhat similar to a telephone network. In relation to the problem statement we are looking at the *multiplexing effects* of IP-based voice streams.

The above scenario may be thought of as somewhat naïve in the IP case. In practice the networking (and computing) resources are often shared, thus disruptions in the voice stream are possible. Therefore paper **B** addresses the issue of modelling packet disturbances in order to reconstruct the *variance distribution* as observed by the receiver. Having a model of the variance helps the receiver in making more informed decisions on what actions to take as packets arrive. Modelling the variance distribution is complicated by the fact that packets can be lost and that silence periods are introduced into the stream when the speaker is quiet. For the model, it is assumed that the network delay distribution is estimated, measured, or indeed known.

Replaying voice streams on real end systems is the topic of paper **C**. This means buffering the arriving packets at the end system and reconstructing the original timing from the RTP packet header information. Not only should the process be accurate, but with the lowest possible delay (and loss). In this work, we provide a method that utilises the operating system and hardware efficiently. We have implemented, tested, and measured a new

approach to end system design for voice streams. In relation to the problem statement, we are addressing the problem of good quality communication.

To gain insight into the real-world aspects of Internet telephony we have undertaken two large wide-area measurement experiments. By large we mean using hundreds of generated calls in the first experiment and thousands in the second. Analysis of these experiments are described in papers **D** and **E**. The problem is to obtain representative measurements from the end systems we had access to. One issue with measurement tasks (generally) is to *completely* anticipate the needs before the upcoming analysis. As well as our own measurement experiments, we were aware the data would be used in related investigations, both by us (papers **B** and **F**) and by other researchers. Therefore acquiring all the necessary information for related studies requires a fair amount of foresight. As a simple example executing `traceroute` before and after each session might help backtrace why a session exhibited abnormal behaviour. As we have taken two partially intersecting sets of measurements taken four years apart, we would like to compare the results for any trends. In relation to the problem statement, we are studying the multi-service nature of Internet traffic.

Paper **F** explores the idea of terminating sessions early when poor quality can be predicted. This can be seen as a problem of self-admission, implying that a call should not continue if an estimate of the call quality is below a quality threshold. Using data from paper **E**, the problem becomes how to determine this threshold, as well as the time needed to reach a decision. In relation to the problem statement above, this paper addresses actions to be taken when conditions deviate from an acceptable operating range.

Wireless and mobile IP systems have their own set of associated challenges which can impact on the voice quality. In wireless systems, stochastic link conditions is one inherent factor. In addition, the radio frequency bands used by 802.11 interfaces are not licensed, and hence not regulated, so interference can occur from other devices. We have investigated VoIP quality over 802.11 networks using cross layer information in paper **G**. In relation to the problem statement we are considering the mobile, and hence wireless, users.

One solution is to use the 802.11 network where possible, but to handover a call to the cellular network when the link conditions are insufficient to support good quality as stipulated in the problem statement. How to schedule this handover has been addressed in paper **H**. Real-world voice handovers typically need time to initialise a parallel technology to switch to. As calls to the public phone network take in the order of five seconds to setup, estimation of deteriorating quality conditions in the 802.11 network must anticipate (at least) this interval ahead of the handover. The relation of this work to the problem statement is in the *heterogeneity of the systems* and providing *good speech quality* to the users.

Ultimately users must be satisfied with the quality of the voice recre-

Technique	Paper
Mathematical modelling	A, B
Discrete event simulation	A
Implementing proof-of-concept applications	A, C, H
Active measurements	E, D, G
Statistical analysis	B, F, I
Subjective user tests	I

Table 1.2: Summary of research methods used within this dissertation

ated from the incoming data stream. Missing parts of a sentence or lost keywords can easily lead to unintelligible phrases. The challenge of paper **I** is to understand how packet losses effect speech intelligibility. Our goal was to produce an estimator that can monitor packet losses and output a simple indicator of the speech quality. To be of any real practical use, our evaluation should correlate with that given by a person who listens to the same sequence. The advantage of having an objective measure is that the system can react to what it thinks is poor quality speech being delivered to the user (or ideally before). The relation of this work to the problem statement is *good quality* and *mobile users*.

1.5 Research methods used in this dissertation

We have used a number of different techniques to solve the problems discussed in the last section. The research in this dissertation focuses on real-world problems concerning quality aspects of real-time packetised voice. The techniques used and the paper letters are shown in Table 1.2. The upcoming paragraphs step through these methods one by one and state in which work, and to what degree, the methods were used.

Mathematical modelling: Within this dissertation, we model the statistical multiplexing of telephony calls in paper **A**. By modelling the multiplexing we can produce a tractable approximation of a telephony system consisting of packet streams from multiple callers arriving at a single queue. In this model the number of calls is governed by a Markov process and each packet stream as a Poisson process. The resulting flows at a multiplexer constitute a Markov Modulated Poisson Process (MMPP). The role of the model is to form a tractable approximation of the number of flows that can be allocated to a given link capacity, and the size of the buffer at the multiplexing point.

In paper **B** we model the arrival process of a single IP telephony stream at a receiver. We consider two types of delays for a given packet: the delay caused by waiting behind previous telephony packets and the delay

introduced by cross traffic along the same path. The arrival process is modelled as a discrete time Markov chain. The function of the model is to reveal the delay distribution of the packets at the receiver.

Discrete event simulation: Discrete event simulation is used to model the propagation delay of the individual packets from multiplexed voice sources in paper **A**. Each packet is traced from source to destination. The simulator counts packet loss at the multiplexer. `ns-2` was used as the simulation framework and the goal of the simulation was to confirm or deny the accuracy of the MMPP model described above and an implementation described below.

Implementing proof-of-concepts: As well as the obvious working software, we have used a proof-of-concept in paper **A** to verify the accuracy of the model and simulation. The working implementation shows whether the theory and practice match, and whether the solution can be deployed into an operational network with some confidence. Proof-of-concept implementations also show which parts of the model are missing, either by design due to abstraction, or simply not accounted for in the problem formulation.

In paper **C** we have implemented a voice playout strategy to reduce the delay incurred by a VoIP receiver. The solution was implemented on a standard PC running different versions of the Windows operating system. The basic idea is to avoid copying the data from the operating system, to the application, then back to the operating system for playout. DirectX now provides similar functions to perform copying in this manner. The role of the implementation is clear, to test and measure the improved playout mechanisms.

In paper **H** we implemented an automated handover mechanism on a PDA running Windows CE. We estimate the call quality in the terminal based on network measurements and signal a third party application that the current call should be transferred from the 802.11 network to the cellular network. The handover was triggered when the quality fell below a quality threshold. Our implementation allowed automatic roaming from 802.11 to GSM networks. The goal of the implementation was to show proof of concept, as well as to judge differences in the speech quality at the time of handover.

Active measurements: Active in-band measurements have been used to sample the path properties during our standard call. The main goal of the measurement work was to report on the suitability of diverse paths with respect to real-time voice. Although limited to academic sites, we chose a wide range of path diversities in order to generalise the results as best we could. One additional reason for conducting the measurements was at

that time (1998 and 2002), no extensive public measurement data was freely available. The measurement work forms the core part of papers **D** and **E**. Some comparison between the two data sets was done to determine whether the quality improved or deteriorated between the measurement periods. We used a modified version of the tool described in paper **C** for the measurement work.

We made a comprehensive evaluation of 802.11 networks using active measurement techniques reported on in paper **G**. Since we had control over the network we were able to perform systematic tests starting from simple (line-of-sight ad-hoc) to complex (infrastructure with competing traffic) experimental setups. The main objective of the active measurements in this case was to capture and quantify the stochastic behaviour of the 802.11 network with respect to voice traffic. A secondary objective was to utilise cross-layer methods that are well suited to voice over wireless applications as demonstrated by the cellular solutions.

Off-line analysis: The active measurements have been used in our off-line analyses. Paper **B** modelled the arrival process of individual voice streams, where measurements from paper **E** were used to validate the inter-packet predictions of the model. Paper **F** used the measurement data from paper **E** in an attempt to estimate which calls would yield poor-quality conversations from the initial seconds of a call. The information from the rest of the call showed whether the decision was indeed correct or not. In paper **I** we used a tool standardised by the ITU (PESQ) to estimate the subjective effect of packet loss on standard eight second voice samples. Our results were used to map network losses to an approximation of the subjective quality. Due to the complexity of the PESQ algorithm in terms of the signal processing, such tests have to be done off-line.

Subjective user tests: In paper **I** we used test subjects to indicate a quality rating for pre-recorded speech samples. The subjects listened to several eight second degraded samples and rated their opinions on a nine point scale. We used 11 test subjects and set up the tests according to the P.862 ITU recommendation [135]. The objective of this recommendation is to ensure that tests are conducted systematically, with an appropriate test duration, warm up tests, deafness tests and so on. The goal of this work is to compare the subjective results with those given by PESQ. The role of such experiments within networking research is often underplayed where the results can be judged by real users.

We also used subjective user tests in paper **H**, where the quality of voice was rated before a handover from the 802.11 to the cellular network. Where the quality started good and ended up poor and a handover was suggested, we recorded this event as a positive result. Where the quality

started good and remained good, and a handover was not suggested we also considered as a positive result. In the two other situations the handover estimation was deemed a negative result. The total number of positive results, in comparison with the sum of positive and negative results gave the performance of our handover algorithm.

1.6 Paper summaries and contributions

Paper A

Bengt Ahlgren, Anders Gunnar (née Andersson), Olof Hagsand, and Ian Marsh. Dimensioning links for IP telephony. In *Proceedings of the 2nd IP-Telephony Workshop*, pages 14-24, New York, USA, April 2001.

Summary: The number of IP telephony calls that can be admitted to access networks is addressed in this paper. Link dimensioning based on packet loss is one method for dimensioning links for high utilisation of networking resources whilst providing acceptable user quality. Using this approach we also show how to select router buffer sizes. We validate and compare our approaches using a mathematical model, a discrete event simulation, and a laboratory-based implementation.

Contribution of this work: The contribution of this work is a planning tool for use in dimensioning networks for voice traffic. We have established a relationship between the important parameters of a packet voice network: namely the speech coding, the link capacities, the number of users, the buffer sizes, and the acceptable loss rates.

My contribution: The original idea to perform such a study was mine. I implemented most of the testbed environment and the traffic generator. Within the project I supervised a masters student, Anders Gunnar (née Andersson), who implemented the MMPP model in Matlab and corresponding simulation scripts in ns-2 [100]. Anders was co-supervised by Professor Ingemar Kaj at Uppsala university. We were assisted by Henrik Abrahamsson, Bengt Ahlgren, Olof Hagsand and Thiemo Voigt. I co-wrote the paper with Anders and presented it.

Paper B

Ingemar Kaj and Ian Marsh. Modelling the Arrival Process for Packet Audio. In *Quality of Service in Multiservice IP Networks*, pages 35-49, Milan, Italy, February 2003.

Summary: In this work, we model the arrival process of voice packets at a receiver. The assumption is that the original packet spacing has been disturbed by bulk data transfers and queuing behind packets of the same stream. The solution, based on a Markov model, models the delay variation of the speech packets. The packets are assumed to be subjected to network delays when travelling from source to destination. The waiting time in intermediary buffers is assumed to be exponentially distributed. The use of such a model allows silence suppression and packet losses to be incorporated; as they are independent of the network induced delay variation.

Contribution of this work: The contribution of this work is a model for the packet audio arrival process. A simple method to estimate packet loss based on observed interarrival times is also given, independent of whether silence suppression is used or not. The model was verified by measurement data.

My contribution: The idea was jointly conceived. My contribution was the measurement data and validation of the model data. I also wrote several tools to process the data. I co-wrote and presented the paper.

Paper C

Olof Hagsand, Ian Marsh, and Kjell Hanson. Sicsophone: A Low-delay Internet Telephony Tool. *IEEE 29th Euromicro Conference*, Belek, Turkey, September 2003.

Summary: All VoIP systems terminate with a receiver. It can be a PC, hand-held terminal, or phone. The terminal has an important role in the overall system performance. For the PC case, we look at how to reduce delay through a novel receiver buffering scheme. The solution uses the low-level features of audio hardware and a specialised jitter buffer playout algorithm. Using the sound card memory directly eliminates intermediate buffering. A statistical-based approach for inserting packets into the audio buffers is used in conjunction with a scheme for inhibiting unnecessary fluctuations in the system. For comparison we present the performance of the playout algorithm against idealised playout conditions. To obtain an idea of the system performance we give some mouth to ear delay measurements for selected VoIP applications. The proposed mechanism is shown to save 100's of milliseconds on the end to end path.

Contribution of this work: The contribution of this work is a sizable reduction in the delay incurred by the VoIP end system. Although many researchers have looked at optimising and reducing jitter buffer sizes, many

do not implement their ideas in a real system. An important byproduct of this work is Sicsophone, a fully functional VoIP application.

My contribution: I wrote the RTCP part of Sicsophone. I performed comparisons between the playout delay of Sicsophone and the optimal playout delay. I co-wrote and presented the paper.

Paper D

Olof Hagsand, Kjell Hanson, and Ian Marsh. Measuring Internet Telephony Quality: Where are we today? In *Proceedings of IEEE Globecom: Global Internet*, pages 1838-1842, Rio De Janeiro, Brazil, December 1999.

Summary: Users of Internet telephony applications demand good quality audio playback. This quality depends on the instantaneous network conditions and the time of day. In this paper, we describe a scheme for measuring network quality and motivate the development of a new metric for VoIP, *asymmetry*, to include into quality reports.

Contribution of this work: In 1999 we reported on the findings of our first VoIP measurement study. As far as we are aware of, the jitter and asymmetry results were new within the VoIP community. The number of downloads of the data from a COST Action web site exceeded 100.

My contribution: The idea, measurements, and paper were done by me. I wrote and presented the paper. The Sicsophone tool used to conduct the measurements was originally written by Olof Hagsand and Kjell Hanson with some modifications by me for the measurement work.

Paper E

Ian Marsh and Fengyi Li. Wide Area Measurements of VoIP Quality. *Quality of Future Internet Services*, October, 2003, Stockholm, Sweden.

Summary: We have investigated the network characteristics of loss, delay and jitter for VoIP streams that are transmitted over diverse Internet paths. Based on over 24,000 sessions, taken from nine sites connected in a full-mesh configuration, we reported on the average quality that can be expected by a user. The VoIP quality was acceptable for all but one of the nine sites we investigated. We also concluded that VoIP quality had improved marginally since the previous study in 1999 (paper **D**).

Contribution of this work: The contribution of this work is a comprehensive report on the quality of Voice over IP in 2002. We defined the quality in terms of the one-way delay, loss, and jitter. For three of the sites, we have been able to compare the quality from 1999 to find some trends in VoIP quality. More than 500 downloads of the data have taken place since they were made available. The data has been used papers **B** and **F** within this dissertation.

My contribution: The idea to improve on the measurements from 1999 (Paper **D**) was mine. I advised a masters student, Fengyi Li, to perform the measurement tasks. Further modifications of Sicsophone were done by me. I wrote a tool to process the measurement data. We jointly wrote the paper based on Fengyi Li's master thesis [87], I presented the paper.

Paper F

Olof Hagsand, Ignacio Más, Ian Marsh and Gunnar Karlsson. Self-admission control for IP telephony using early quality estimation. In *4th IFIP-TC6 Networking*, Athens, Greece, May 2004.

Summary: The idea is to use packet loss statistics from paper **E** to potentially identify poor quality calls given only the initial seconds of a call. The application is a self-admission control scheme, which will continue or terminate a call depending on a quality threshold. The threshold is determined by the acceptable loss rates of the speech coding used. If sessions themselves can determine whether entry into a system is worthwhile, given the early loss rates, then system resources and user frustration can be avoided.

Contribution of this work: The contribution of this work is a self admission control for IP telephony. The scheme does not require any network support or external monitoring schemes.

My contribution: My role in this work was in the initial discussions and providing the measurement data. Some filtering of the data was needed to begin the work, hence I wrote the initial version of the data parsing tool. We jointly authored the paper.

Paper G

Ian Marsh, Juan Carlos Martín Severiano, Victor Yuri Diogo Nunes, and Gerald Q. Maguire Jr. IEEE 802.11b voice quality assessment using cross-layer information. In *1st Workshop on Multimedia over Wireless*, Athens, Greece, April 2006.

Summary: The conditions that VoIP users can encounter in 802.11 networks is covered in this paper. It is measurement based and takes a methodological approach to understanding quality variations in 802.11b networks. We started with simple point-to-point VoIP experiments to determine the delays associated with the terminals and operating systems.

We progressed onto 802.11 infrastructure mode using line of sight and indoor measurements. Next non line of sight experiments were conducted and again re-conducted in the presence of competing TCP traffic. Some simple, but effective, mechanisms were proposed to maintain acceptable VoIP quality using 802.11 networks. We used the Sicsophone tool amended with modules for obtaining the MAC layer retransmissions and data rates.

Contribution of this work: The contribution of this work is a comprehensive study of 802.11b networks as far as voice is concerned. This includes the methodology we employed plus utilising cross layer techniques to obtain our desired results. Many of the lessons we learned were put to use in paper **H**.

My contribution: The ideas for the project were mine. Most of the work was carried out by two masters students, Severiano and Nunes, working on the MAC/IP layer interactions and on the IP/application layer interactions respectively. Gerald Q. Maguire Jr. co-supervised the students. We all authored the paper.

Paper H

Ian Marsh, Björn Grönvall and Florian Hammer. The design and implementation of a quality-based handover trigger. In *5th IFIP-TC6 Networking 2006*, Coimbra, Portugal, May 2006.

Summary: In this work we looked at the conditions under which an on-going call could be migrated from a 802.11 to a cellular network without perceivable loss in quality. We performed measurements on the 802.11 network in order to make workable predictions of the call quality. We implemented our solution on a hand-held terminal and performed 100 handover test trials of our handover mechanism.

Contribution of this work: The contribution of this work is one part of a fully working system that allows calls to be migrated from a 802.11 to a GSM network automatically.

My contribution: Björn Grönvall and I jointly conceived the initial idea and jointly performed the base experiments on which the automatic trigger

was designed. We co-implemented the solution. We also integrated our solution into software developed by Optimobile AB. Florian Hammer helped in the PESQ assessment of packet loss. Björn Grönvall and I wrote the paper and I presented it.

Paper I

Martín Varela, Ian Marsh, and Björn Grönvall. A Systematic Study of PESQ's Performance from a Networking Perspective. *Proceedings of Measurement of Speech and Audio Quality in Networks*, Prague, Czech Republic, May 2006.

Summary: The basic idea is to have a general function which maps losses into estimations of the quality due to packet loss. Using standardised samples distorted by network losses, we could utilise PESQ processing off-line to map packet losses into quality ratings over a range of operating conditions. We verified our results with real test subjects. We also compared the single sided measure (ITU P.563 [67]) to our own findings.

Contribution of this work: The contribution of this work is a real-time single-sided metric for estimating speech quality. A systematic study of the behavior of PESQ as a function of losses has also been performed. Also the variability of PESQ ratings under several different test conditions has been conducted. The PESQ ratings were compared to subjective scores for a range of bursty losses.

My contribution: I worked jointly with Martín Varela and Björn Grönvall on the idea. We conceived the idea together. Martín was responsible for most of the scripts, whilst we both analysed the data. The paper was jointly authored.

1.7 Conclusions

This dissertation addresses selected topics within real-time voice communication. Our focus is on the *quality aspects* of voice communication, since poor quality often leads to user dissatisfaction. The techniques presented in this dissertation attempt to solve the research problems independent of network QoS efforts.

Each of the publications draws similar conclusions, that is, reasonable quality Internet telephony can be offered, provided that the whole system is carefully engineered. This implies the introduction of mechanisms to preserve the subjective quality when impediments are, or are about to, occur. Some of the conclusions from our research are as follows: The network load

should be controlled for links that carry real-time voice. This means providing and dimensioning links with sufficient capacity, or alternatively, restricting the admission of voice calls to heavily loaded links. The monitoring of network conditions, in particular loss, should be used to signal potential quality problems on particular paths. We have presented a solution where the end system can do the monitoring where such network functionality is absent. Should we require earlier indications of impending problems, tracking the network delay or jitter at the end system can be investigated. This technique has been used in our handover studies, where several network parameters have been combined in order to schedule a handover. Continuing in the wireless case, we have proposed mechanisms for maintaining quality by switching to lower data rates, or even switching to an alternative technology where available.

Since the scope of this work is broad, we have taken different cuts through IP telephony research by looking at the access and backbone networks, using modelling, simulation and experimental techniques; we have considered both fixed and wireless networks using subjective and objective quality tests to obtain the most appropriate solution for a particular problem. We have also looked at systems with and without background traffic, used real-time and off-line techniques, and finally applied cross layer approaches that combine normally separated layers of the protocol stack.

The main contribution of this work is a near-complete system study concerning quality aspects of an Internet telephony system. We have looked at a number of different methods to enable adaptable solutions for maintaining acceptable quality. We have often found that relatively simple changes can lead to substantial user quality gains.

The tangible outcome of our research has been a number of software tools. These include an IP based voice measurement package, a handover algorithm for wireless terminals, a VoIP traffic generator and a PESQ processing package.

1.8 Future directions

Plenty of challenges remain within the area of IP-based voice quality. We will consider each one in the context of the research done within this dissertation, and later on discuss broader topics outside the scope of this work.

In-dissertation issues: In the area of network provisioning, a macro level investigation needs to be conducted on the suitability of the MMPP model for dimensioning tasks on an Internet scale. Our investigations were done and verified for links only up to 1.5Mbits/s. Therefore, one (ambitious) theoretical study could be to investigate migration of the world's telephony traffic onto the Internet. This would partly include capacity studies of the

existing voice traffic, separating voice traffic from TCP flows, and estimating the future demands of voice on the Internet, thus scaling up the dimensioning work to much larger network capacities.

On the individual flow level, research should be done on understanding the network delays for voice packets over different operating conditions and network types. Due to the increase of bandwidth-heavy applications such as P2P traffic and video streaming, the conditions for voice traffic needs to be reinvestigated. As far as the network is concerned, the arrival process for VoIP packets over wireless links should be reexamined. One reason is access to the medium is distributed, allowing multiple flows to become multiplexed at the first hop. Finally the concept of backoff timers in CSMA protocols has not been included in our model.

More work can be done on hand-held terminals to support voice applications. This is because smartphone type terminals currently offer insufficient voice quality on 802.11 networks. Essentially this is because terminals are computers and the 802.11 protocols have been designed for data transmissions. Furthermore, the networking interfaces are commodity items and do not provide sufficient handles for voice application designers. We have said earlier, voice applications on IP networks need careful engineering. The telephony side of some smartphones is separate and has a highly integrated system using techniques such as joint source and channel coding. Voice application writers do have the access to such technologies, they simply have a strict layered protocol stack to interface to. In the specific case of 802.11 networks, application writers would benefit (at least) from access to the MAC retransmission counters, precision RSSI signals, data rates and near instantaneous bit error rates at the link layer level.

As far as active measurements are concerned, additional investigations should target home users to include their usage patterns. This includes 802.11 based networks and telephony. Coordination and collaboration with ISPs would be beneficial in order to obtain a broader sample set of users, as well as important data on the network operational status. Some cities operate open 802.11 networks which could be instrumented to obtain better operational status. As the 4th generation networking technologies are almost upon us, investigations of voice over the newer radio access technologies (e.g. OFDMA) and the Evolved Packet Core (EPC) would surely be desirable for a new look at capacity planning on telecommunication networks.

We believe there is still much research to be done in the voice handover area, including monitoring the network conditions at the handset. As we have eluded to earlier, tight integration achieves the best results and in the case of dual-radio phones, prediction of impending problems is the key criterion. Not included in this research is the possibility to make use of tracking i.e. estimating the position or path of the user. This would greatly influence the decision of whether to switch a call to an alternate technology.

Further work needs to be done on objective quality assessment tools.

While PESQ and the single-sided measure methods exist, improvements can still be made. From our experience the performance of these methods deviates as the loss process becomes more correlated. Naturally, it is difficult to adjudge a series of samples with missing segments, with or without the reference signal, nevertheless, such loss processes are reality on many wireless networks today. Also one would like to include delay into the assessment, as current methods are loss-based only.

People adapt to delays, by less frequent interruptions in the conversation for example. Conversational quality models have been proposed, however their accuracy is still not clear.

Broader issues: Moving onto just one topic outside of this dissertation, we believe that higher fidelity telephony should be available in the near future. Although the technology for transporting bits has improved, the media stream itself has not changed since the introduction of 64 kb/s voice many decades ago. From the user's perspective the voice quality of a 13 kb/s stream is actually worse than that of traditional telephony. However, we are prepared to pay this cost in order to have mobile telephony, and, of course, the operator can squeeze more calls out of the system without substantial investment.

The drive to reduce bitrates for calls has been to multiplex more calls onto capacity constrained links. However, as ever more capacity is becoming available both on the cellular and Internet technologies, the time is right for a new type of voice experience. Therefore, one example would be to use *higher* fidelity than we are currently used to. This may be stereo voice, and would require headsets, but many mobile users already use such devices to listen to music.

Going one step further is 3D telephony. This will enhance the experience at the listener through capturing binaural signals at the speaker, optionally rendering them in 3D space, and replaying the enhanced signal at the listener. Capturing the signals at the speaker can be done by placing small microphones on the outside of the headsets, somewhat similar to what noise cancelling headsets do today.

Steps such as these would represent a new domain for telephony that has been thus far the preserve of specific environments such as audio conferencing. 3D telephony is very much under investigation, however significant challenges remain, particularly in the domain of noise cancellation, either at the sender or receiver, or both.

Chapter 2

Background

This chapter consists of two parts. The first is a short description of the path that voice samples take from a sender to a receiver as part of a VoIP system. The second part contains sections on some of the important building blocks of IP telephony: signalling, firewall traversal, speech coding and IP networking.

2.1 A voice journey across the Internet

Figure 2.1 shows the processing components (as blocks) typical for a stream of voice IP packets. The voice is captured by a microphone, sampled, digitised, and encoded into a format chosen by the application. Typically a voice frame is of 20 ms duration and contains 160 voice samples, where each sample is 8 bits of information sampled at 8000 Hz.

FEC/MDC (Forward Error Correction/Multiple Description Coding) can create redundant samples from the existing samples. The redundant samples are transmitted with a time shift from the original samples to reduce the probability for losing both the original and redundant data. The encoded voice is then packetised which means gathering the samples into one transmission block. Addressing information is pre-pended to the block which includes RTP, UDP, and IP headers. The packet is sent onto the local network via a network interface. A link local frame header is appended for each link traversed on the path.

The packet traverses one or more networks where multiplexing occurs. Once the packet reaches the receiver, the headers are removed and any FEC or MDC that was applied can be used to recreate lost packets. The packets need to be available for decoding in continuous blocks, therefore they are buffered and timing information in the RTP information used to generate the sequence. The application can also take action if the packet loss protection was not sufficient, voice frames can be created using a technique called packet loss concealment (PLC) where lost samples are masked by creating

approximations of the lost samples from those received. Finally the voice samples are transferred to the end terminal's audio output device (shown as speaker). In some cases the speech decoding and loss concealment may be combined in one algorithm.

Internet telephony can be broken into two distinct phases: signalling and voice transfer. The signalling phase is responsible for the initiation and control of the sessions, whilst the data transfer phase is concerned with the transfer of the speech content. The next section outlines the main features of Internet telephony signalling using two standard protocols, SIP and H.323, and one proprietary protocol used in Skype.

2.2 IP telephony signalling

Signalling is primarily responsible for enabling the communicating parties to 1) find each other, 2) establish a session, 3) agree upon session parameters, and 4) gracefully terminate the session.

In this section we will explain the basic operation of three protocols: H.323, SIP, and Skype's signalling protocol. H.323 and SIP are standardised signalling protocols, although by different bodies, and Skype's protocol is a non-standard proprietary protocol. There is a wealth of information available on H.323 and SIP in [53, 8, 101, 37]. For the Skype signalling protocol refer to [126, 131, 117, 45].

Much of the work of a signalling protocol is to maintain a consistent state within the communicating parties. Therefore a large part of a signalling protocol is devoted to ensuring correct operation of the system: which means operations that are initiated, terminate as expected. Indeed, during the protocol standardisation phase it is highly desirable to formally prove that inconsistent situations cannot arise during message exchanges. Signalling failures can manifest themselves as timeouts, looping behaviour, unacknowledged messages, and inconsistent states. Telecommunications standardisation bodies such as the ITU or ETSI have attempted to formalise the design/testing phase, whereas the IETF approach has traditionally been less formal, with validations done via early implementations and inter-operation tests.

Signalling utilises transport protocols such as TCP, UDP or SCTP. The selection may depend on the characteristics of the transport network. We will briefly discuss message loss in relation to these three transport protocols. Reliability at the transport layer is desirable, however TCP introduces signalling delay due to its three-way initial handshake. UDP on the other hand, requires only a one-way trip time to initialise an existing connection. Although this is attractive in delay terms, lost messages must be handled by the higher layers, increasing application complexity. The Stream Control Transport Protocol (SCTP) is a transport layer protocol similar to TCP,

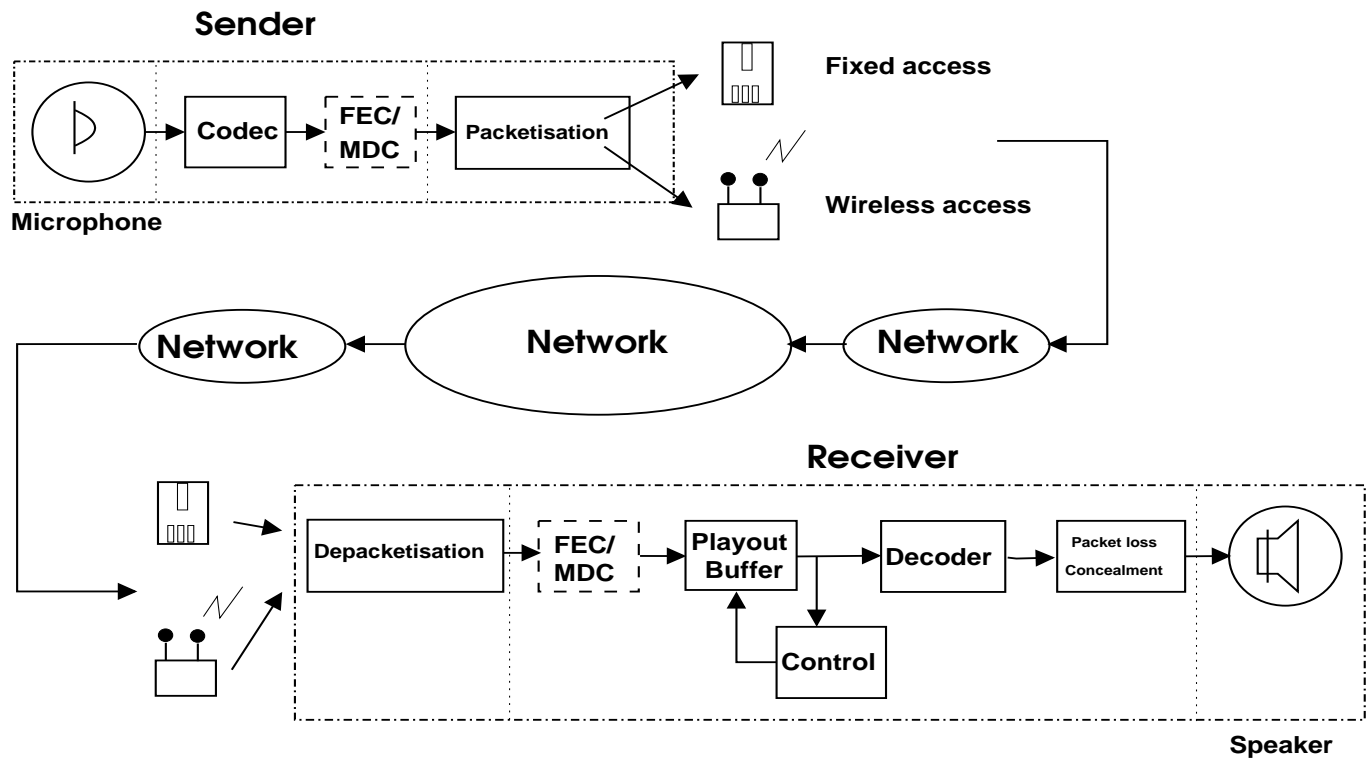


Figure 2.1: A voice journey's path

but supports complete and multiple message streams. It operates on whole messages rather than on single bytes such as TCP and UDP. SCTPs uses a 4-way handshake to initialise a session making use of a signed state cookie. This renders Denial of Service attacks more difficult to which TCP can be subject (i.e. a SYN flood attack). SCTP was originally designed to transport PSTN signalling messages over IP networks.

Two dominant signalling standards for Internet telephony have emerged during the past ten years, ITU-T's H.323 [69] and IETF's Session Initiation Protocol (SIP) [113]. The following sections will explain the basic setup operation of these two protocols plus give a short comparison of their major characteristics. Details of the particular protocol operations are however beyond the scope of this dissertation, and we refer the reader to the earlier references for further information.

2.2.1 H.323

H.323 is the result of standardisation by the International Telecommunications Union (ITU) standardisation body, the ITU-T. The ITU-T has been responsible for many standards that define operating practises within the global telecommunications industry. As the Internet has become more prevalent, H.323 has undergone a number of revisions. The current standard was approved in June 2006 (version 6). There are actually a number of separate components within H.323. It is in fact a complete protocol suite incorporating methods for both Internet and traditional telephony. More complex signalling has been necessary in H.323 to include legacy telephony.

Figure 2.2 shows an example of the signalling process between two nodes and a gatekeeper (server in the H.323 terminology). In the figure terminal A sets up a connection to terminal B. In phase I, terminal A initiates the communication to the gatekeeper, with registration, admission and signalling (RAS) messages. This part of the communication is indicated with dashed lines.

The gatekeeper provides information for A to contact B. In phase II terminal A sends a SETUP message to B on a well known signalling port. It negotiates which unit is the master and which is the slave in the pairing, establishes RTP port numbers, plus signals the logical channels. The logical channels are used for the media flows and are instantiated using a request and ACK exchange to the gatekeeper. In phase III terminal B responds with a CALL PROCEEDING message and also contacts the gatekeeper for permission to continue the call establishment. In phase IV an ALERTING message is sent from B to A via the gatekeeper once the phone is ringing at the callee. In phase V a CONNECT message is sent from B to A once the phone is answered. Both the ALERTING and CONNECT messages contain transport addresses (such as port numbers) to allow the terminals to open media channels. Phase VI uses the H.245 protocol to negotiate the codecs

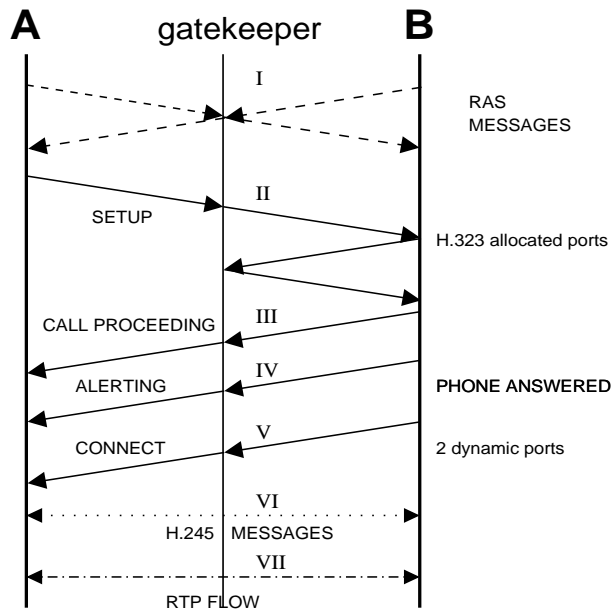


Figure 2.2: H323v3 call setup

for the session. Finally phase VII begins with the flow of the voice data. Even though this is a relatively simple H.323 setup operation, there can be a complex flow of messages. Much more material on the H.323 signalling protocol can be found in [85].

2.2.2 SIP

The Session Initiation Protocol (SIP) is a signalling protocol used in many different types of sessions. It is widely used in multimedia initialisation, but has also been adopted in presence, messaging and telecom applications. Developed within the Internet Engineering Task Force (IETF), SIP 2.0 was the first proposed standard version and is defined in RFC 2543 [51]. The protocol was further refined and published in RFC 3261 [113]. Unlike H.323 and its telephony origins, SIP is very much Internet based with its extensive use of existing IETF protocols plus an HTTP-like syntax. SIP inter-operates with external protocols such as the Session Description Protocol (SDP) for media description. SIP began life with a smaller feature set than H.323. However its adoption in other applications, notably instant messaging and 3GPP's IP multimedia system (IMS), has increased its size in recent years. For SIP material consult [18, 125]. Up to date tutorials can be found in [37, 55].

Figure 2.3 illustrates an example of a simple SIP session between two user agents, a user agent client (A) and a user agent server (B) plus a SIP proxy server. Clients are referred to as user agents in the SIP world.

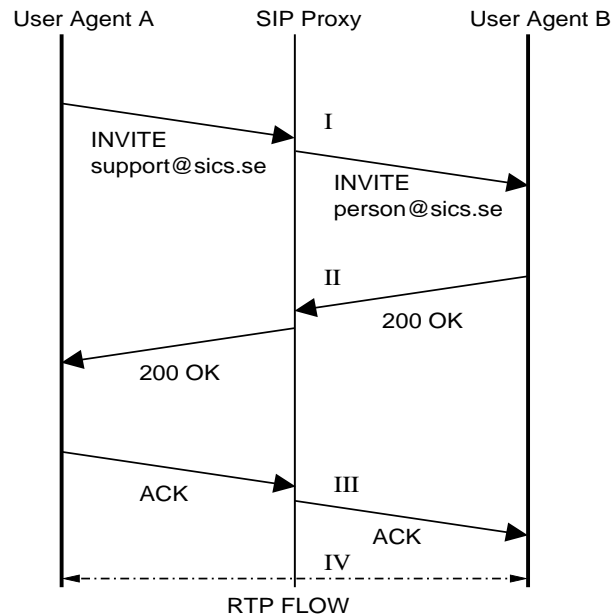


Figure 2.3: SIP setup

SIP servers play a central role as they provide inter-operation between SIP components and offer device, service and session mobility. User Agent A sends an INVITE message to a local SIP proxy. The SIP proxy then looks in a location database where “support” is registered, and an INVITE message is sent from the proxy to the user agent server B (Step I). User agent B responds with an OK message to the proxy which in turn sends back an OK to the initiator A (Step II). Within an INVITE message there are the details of the voice coding A is willing to accept, the path of the OK message must follow the same path of the original invites. The list of proxies are stored successively from sender to receiver in the original INVITE message. User agent A responds with an ACK, and if the capabilities are agreed upon (step III), the RTP media session can begin.

In both this and the H.323 case we have selected a simple scenario (A calls B). However it is clear to see the complexity difference between the two. Note, however in both cases users were assumed to be within reach of a single gatekeeper/server. In cases where redirects are needed, the number of messages needed in both protocols can increase significantly. There is a large difference between the simple call case and fully fledged telephony functionality, which has led to many add-ons and expansions to the original draft RFC.

2.2.3 A comparison of H.323 and SIP

We will now give a brief comparison of H.323 and SIP functionality. First their similarities, H.323 and SIP offer essentially the same set of services. They both provide call setup, control, and tear down. Both have basic call features such as call waiting, transfer, identification and so on. Both protocols typically rely on well known servers (or gatekeepers) for registration. Both can operate in either stateless or statefull mode and can use TCP, UDP, or SCTP for their message exchanges. A SIP user agent registers with a proxy server and H.323 terminals register with a gatekeeper; both can use IPsec or TSL. At a higher level they focus on different domains, but increasingly SIP is addressing telephony-like functionality and connectivity to the PSTN, whilst H.323 is becoming more IP compatible. This trend will probably continue from their once clear domains until the point where they cannot be easily distinguishable anymore.

We now move onto their major differences. The following discussion refers to H.323 version 6 and SIP version 2.0. SIP is under the auspicious control of the IETF whilst H.323 protocol is defined by the ITU-T. This is reflected in H.323 still being telephony based with its ISDN influence and ASN.1 coding, whilst SIP is TCP/IP based with its HTTP-like syntax. The capabilities exchange is more complex with H.323 than with SIP. The latter relies on the session description protocol SDP. SIP+SDP can issue a single request that contains most of the necessary information to initiate a session. H.323 defines its own mechanisms for such functions. SIP provides better personal mobility e.g. redirection of a callee to different locations and better support for caller preferences. H.323 has better internal developed multimedia session capabilities such as whiteboards, video, and data collaboration facilities based on the T.120 specification.

SIP is somewhat better at adding new features with its call processing language (CPL) and SIP-Common Gateway Interface (SIP-CGI). SIP also allows a third party to control a session, which is not presently possible with H.323. Due to SIP's modularity, it can more easily support a wider range of applications as we have already mentioned. H.323 has to use the ITU's H.450.1 supplementary service creation.

For Quality of Service (QoS), SIP relies more heavily on external functionality and can use any reservation protocol (COPS, OSP, RSVP) whilst H.323 recommends RSVP for bandwidth reservation, however admission control is still controlled by the gatekeeper. In terms of security SIP supports MIKEY and SRTP while H.323 relies on H.235.

Generally H.323 is more complex relying on hundreds of components such as those mentioned above. SIP initially only defined a small set of primitives (32 headers in the base specification), however in recent years, it has become rather large and the existing base standard document now extends to over 2000 pages.

As of late 2008 it is difficult to say if one protocol will become dominant, however the adoption of SIP into IMS may have an influence, depending of course on the success of IMS. Predicting the dominance of one standard over another becomes less necessary with the presence of protocol translators from Asterisk, VOCAL, and Yate which can translate H.323 and SIP messages. Additional discussions of the two protocols can be found in [28, 77] as well as those given in [68, 103, 26].

2.2.4 Non-standardised signalling

Signalling does not necessarily need to be standardised. Commercial developers find it advantageous to keep their systems proprietary for monopolistic reasons, and often cite issues such as security, complexity and performance as reasons to develop closed solutions.

We will now discuss one proprietary protocol for IP telephony, specifically the application layer Skype protocol. Unlike SIP and H.323 there is no centralised server/gatekeeper, there is however a central login server. Within the Skype network there are two classes of nodes: normal nodes and super nodes. We will first discuss normal nodes. Normal nodes are typically a home owner's PC and are usually behind a home firewall and/or an ISP's NAT. These nodes typically have a private IP address allocated to them. A private address is not globally routable and is defined by specific ranges which routers do forward data from. The CPU processing of normal nodes is also assumed to be somewhat limited.

Super nodes on the other hand are well connected machines and must possess a public IP address. A typical example might be a UNIX computer on a university network. Due to their connectivity and processing capabilities, super nodes perform routing and forwarding of Skype signalling messages. The load on the super node is carefully monitored so Skype message processing does not interfere with the normal operation of its host. Usually users are not unaware that the computer has been elected to super node status. The software distribution for normal and super nodes is actually identical, with different routines being invoked after initialisation. The super nodes also forward login requests on the behalf of the normal nodes, if the normal nodes cannot reach the login server.

On the first invocation of Skype, a normal node uses a pre-configured list of permanent super nodes, it then receives an update of more recent super nodes. The directory of Skype users is decentralised. Skype uses its Global Index technology to find a user with encrypted (256 bit AES) messages. In order to locate a user the procedure is as follows. A normal node sends a request to one super node, if it doesn't know itself the location of the callee. That super node then responds with four additional nodes to be queried if the person was still not found. The normal node then queries these four nodes. If the user is not found, an exchange occurs again with the same

super node. The super node then responds with eight new (and different) nodes. This is repeated several times until the user is found. Here we have assumed that the normal nodes has a public address for simplicity, in the case where it has a private address this negotiation is done by a super node on the normal node's behalf. Search results are also cached at intermediate nodes for subsequent searches.

Non-standardised solutions need to use protocol translation services if they are to inter-operate with existing solutions. Protocol translation involves taking a message from one protocol and generating a (near) equivalent message in the second protocol. We briefly mentioned some names of known translators for H.323 and SIP in the previous section. For a closed protocol the developer themselves must create a translator for the desired interoperability.

There have been many publications and presentations on the Skype protocol. Prestige in being the first to reverse engineer a closed (and widely used) protocol often acts as an incentive for such efforts. Some of these can be found in [131, 117]. A more basic introduction to the operation of Skype at a somewhat higher level can be found in [126].

2.3 Firewall traversal

We will briefly look at two protocols for traversing NATs and firewalls. These are STUN (Simple Traversal of User Datagram Protocol) [114] and TURN (Traversal Using Relay NAT) [111]. Some assumptions to begin with. An external STUN server needs to have a public IP address and STUN assumes that UDP datagrams are not blocked by NATs.

The basic principle of STUN is to ascertain if a client is behind a NAT or a firewall, and what type of NAT or firewall it is behind. By requesting a reply from different servers and by requesting different ports the client can learn the bindings applied by its NAT. The Skype application has a built in STUN client, which sends a number of requests to the external STUN servers to find these bindings. STUN, however has been criticised for being unreliable and opening up security problems, a draft RFC [78] suggests that STUN is not sufficient as a complete NAT traversal mechanism.

TURN is used for a client to traverse symmetric NATs by contacting a relay NAT. A symmetric NAT works by each request from the same internal IP address/port pair to a specific destination IP address/port pair is mapped to a unique external source IP address and port. The client can use either TCP or UDP connections. TURN is normally used by clients behind a symmetric NAT that want to receive a single connection (only). It is designed so that the internal client can be on the receiving end of a connection also requested from behind a NAT. TURN is more reliable than STUN, but is more costly in terms of traffic to and from the TURN server.

This is because the server must receive and forward all the media traffic in both directions. In the case of symmetric NATs, STUN is often tried first and then TURN by clients.

2.4 Speech encoding

Human speech occupies a fundamental frequency in the range of 85-155 Hz for men and 165-255 Hz for women. Higher tones or harmonics can be heard up to 10 KHz. Encoding this full frequency range would require a sampling rate of at least twice this frequency to faithfully reproduce the speech. In a voice transmission system, the speech is sampled and then digitised according to the quality required (or restrictions) of the transmission system. In a system such as the traditional telephony system, this capacity is not sufficient to faithfully accommodate human speech's full frequency range.

2.4.1 Pulse Code Modulation (PCM)

In narrowband telephony, the frequency bandwidth is restricted to 3100 Hz, ranging from 300 to 3400 Hz. Voice in the fixed telephony system has therefore to be reduced from its original range to this 3100 Hz range (a reduction of about one third). The lower frequency of the human range is lower than that of the telephony system. This is not as problematic as it may seem, due to the perceptual system's ability to reconstruct the lower tones from the overtones. Traditional telephony does not use the low frequencies as they are very hard to reproduce with inexpensive loudspeakers.

Quantising the sampled waveform can either be done using constant steps between the sample levels or using non-constant steps, such systems are known as linear and non-linear quantisers respectively. From a 12 bit linear input signal, an 8 bit companded signal can be produced which has a similar signal to noise ratio as the original. Non-linear quantisation has the advantage that the quantisation performance is independent of the signal loudness. Its disadvantage is lower accuracy for larger amplitude signals. Two (similar) examples of non-linear quantising encodings are known as the A and μ -law companders. There are three main methods of implementing the μ -law algorithm:

- One is using an amplifier with non-linear gain to achieve companding entirely in the analogue domain.
- The second is to use an analogue to digital converter with quantisation levels that match the μ -law algorithm.
- The third is to convert the 12 bit linearly quantised representation to μ -law coding entirely in the digital domain.

In Europe A-law coding is used. The A-law algorithm provides a slightly larger dynamic range than the μ -law version at the cost of worse proportional distortion for small signals. By convention, A-law is used on an international connection if at least one country does. The G.711 standard encapsulates the A-law and the μ -law formats into a single standard [65]. G.711's simplicity (and the low SNR) makes it the default choice in the non-wireless telecommunications infrastructure.

2.4.2 Adaptive differential pulse-code modulation (ADPCM)

Differential (or delta) pulse-code modulation (DPCM) encodes the PCM values as differences between the current and a predicted value. An algorithm predicts the next sample based on previous samples, and the encoder transmits only the difference between this prediction and the actual value. If the prediction is reasonable, fewer bits can be used to represent the same information. For speech, this type of encoding reduces the number of bits required per sample by about 25% compared to PCM. Adaptive DPCM (ADPCM) is a variant of DPCM that varies the size of the quantization step to allow further reduction of the required bandwidth for a given signal-to-noise ratio. The rate of ADPCM is 32 kb/s.

2.4.3 Low bit rate models

Speech that is sampled and encoded using A or μ -law at 8000 samples per second with 8 bit resolution for each sample produces a data rate of 64 kb/s. Current speech coding techniques can produce encoded voice with rates as low as 16 kb/s which are indistinguishable in quality from 64 kb/s codec. We will discuss some of these schemes soon, however it is first necessary to explain how humans produce speech, in order to understand the technique known source filter modeling.

Human production of sounds: The lungs produce a stream of air that enters the vocal tract. The vocal tract is the pharynx, mouth, and nasal cavities. There are essentially two types of sounds: voiced and unvoiced sounds. Voiced sounds such as /a/ or /e/ are produced by the vocal chords. Unvoiced sounds have two types, the first type is fricatives such as /s/, /sh/, or /f/ which are produced when the vocal tract is constricted. The second type of unvoiced sounds are known as plosives, and include sounds such as /p/, /k/ or /t/. They are produced when the end of the vocal tract is closed, pressure is built up, and the pressure is released suddenly. There are actually additional types of sounds such as the nasal /n/ sound, but we will omit these from the following discussion.

Voiced and unvoiced segments: In order to encode and transmit speech at low bit rates, it is necessary to differentiate between the voiced and unvoiced sounds. As we will see, these sounds constitute different parts of a source filter model, and are actually transmitted separately. In order to separate them different techniques are available:

- Spectral flatness: calculated by the geometric mean of the power spectrum divided by the arithmetic mean. Unvoiced frames (typically 20 ms long) are flatter than voiced frames. The spectral flatness can also be measured within a specified sub-band of frequencies as well as across the whole frequency band.
- Energy: the square of the spectrum values of the sampled frame. Voiced frames have greater energies than unvoiced frames.
- Zero crossing points: counting the sign changes in the signal, voiced frames exhibit fewer crossing points than unvoiced frames.

Source-filter models: The most popular technique within source filter models is based on linear predictive coding (LPC). The basic idea is to model the speech generator as produced by the human vocal system, described in the previous section. The generator is a simple buzzer at the end of a tube. The space between the vocal chords (called the glottis) produces the buzz. It is characterized by its intensity and frequency (pitch). The vocal tract (the throat and the mouth) forms the tube, which is characterized by its resonances, these are known as formants.

The parametric coding process: Low bit rate coders estimate the formants, remove their effects from the speech signal, and then estimate the intensity and frequency of the remaining buzz. The process of removing the formants is called inverse filtering, and the remaining signal after the subtraction of the filtered signal is called the residue. The formants and the residue can then be transmitted to recreate the voice at the receiver. Another term for this process is vocoding, a contraction of the words voice and coding.

Decoding or synthesising the speech signal is done by reversing the process. The buzz parameters are used together with the residue to create a source signal. The formants are used to create a filter (which is the tube), and the source is run through the filter reproducing the original speech. The spectral information is well suited for vector quantisation. Compression algorithms often differ in how the residuals are treated. Typically 30 bits are used to code the 10 coefficients for basic LPC quality, and up to 18 coefficients can be used for improved fidelity.

Code excited linear prediction (CELP): In an attempt to improve on the robotic sound of early LPC schemes, a number of improvements were made that have led to methods used in modern codecs (see section 2.4.4). Multi-excitation linear predictive coding (MELPC) is based on LPC but instead of using a periodic pulse train for the voiced segments and white noise to represent the unvoiced segments, it uses mixed periodic and aperiodic pulses, a pulse dispersion filter, and spectral enhancement. The multi-pulse linear predictive coder (MPLPC) is an analysis by synthesis approach where each excitation vector consists of a number of pulses where their amplitudes that have been derived from closed loop optimisation. CELP uses an codebook (sequence) of excitation pulses as the excitation rather than the multi-pulses of MPLPC. The optimum sequence is chosen to minimise the distortion between the derived signal and the original one. At the decoder the sequence of excitation signals is passed through a long term filter and a LPC vocal tract filter to produce a block of reconstructed samples. The bitrate of CELP coders is usually in the range of 5 to 15 kb/s.

Transform coders: Transform coding tries to draw the best from waveform tracking techniques used in the PCM encoders, but also include models of the human production of speech as the source filter models do. Knowledge of the speech signal is used to select which information to discard in order to lower the bandwidth of the signal. Transform coding derives its name from frequency based techniques to code the transform coefficients in a manner suitable for voice.

Different transforms have been suggested for speech compression, we will briefly consider just two: the Karhunen-Loève transform (KLT) and the Discrete Cosine Transform (DCT). The Karhunen-Loève transform offers optimal coding performance (in terms of minimum square error) if the input samples are Gaussian distributed and the coefficients are scalar quantised. However the Karhunen-Loève transform is difficult to implement and its performance is signal dependent. The DCT is signal independent, but is sub-optimal (compared to the KLT) in that it cannot completely decorrelate the transform coefficients. The DCT is attractive since there are computationally efficient algorithms to compute it, and it retains the formant structure of the speech. The bitrate of transform coders is in the range of 10-20 kb/s, but can produce better fidelity speech.

2.4.4 Modern codecs GSM, G.729 and iLBC

GSM networks employ a LPC-based speech encoding technique called Code-Excited Linear Predictive (CELP) coding. The significant difference between CELP and LPC is that the excitation signals are not simply generated based upon a voice or unvoiced sound, but taken from stored codebooks. There are two types of codebooks, fixed and adaptive which are used in

conjunction to code the signal. ETSI's GSM has defined different rate voice codecs ranging from 6 kb/s (half-rate) to 13 kb/s (full-rate). GSM was further enhanced in the mid-1990s by the GSM-EFR codec (effective full-rate) which is a 12.2 kb/s codec that uses a full-rate GSM channel. GSM is one of the preferred speech coding schemes for wide area radio links. EFR is a fixed rate codec, however some GSM networks now use Adaptive Multi-Rate (AMR) coding [7]. AMR uses link adaptation to select from one of eight different bit rates depending on the instantaneous link conditions.

G.729 is another example of a LPC-based encoder, again a CELP codec. The coded stream consists of linear prediction coefficients, the excitation codebook indices, and gain parameters. Technically it is known as variable bit rate conjugate structure algebraic code excited linear-prediction scheme (CS-ACELP). The standard rate of G.729 is 8 kb/s. It requires 10 ms input frames and produces an 80 bit output frame. It also includes a 5 ms lookahead, producing a 15 ms algorithmic delay. Annex B of the recommendation (G.729B [61]) also describes a silence compression scheme and a voice activation scheme. It also has a discontinuous transmission module, which estimates the background noise at the sender and can use a comfort noise generator at the receiver. G.729 is popular within VoIP applications, due to its low data rate and the features just mentioned. A Skype call initiated from the Internet and terminating at a PSTN connection uses G.729 for the Internet part of the path. It was developed by the University of Sherbrooke (Canada), the Nippon Telegraph and Telephone Corporation of Japan and France Telecom in 1995.

The iLBC encoder from Global IP Solutions is a block-independent orientated LPC coder [4]. Whereas LPC schemes have a memory that lead to error propagation in the case of lost packets, iLBC encodes each frame as a separate block. It therefore has a controlled response to packet loss and exhibits a robustness similar to PCM with respect to packet loss concealment [66]. The CPU resources when using iLBC are comparable to that of G.729A, but it yields higher basic quality. Although a narrow-band speech coder, iLBC uses the full 4 KHz spectrum unlike most 300-3400 Hz codecs, thus producing better fidelity. iLBC is popular in PC to PC communication and is found in tools such as Skype and GoogleTalk.

2.4.5 A (very) brief history of speech coding

The vocoder was invented in the late 1930's and is an implementation of the model of the human sound production system. Vocoders are often known as analysis-synthesis systems, where the input speech is passed through a multiband filter and each filter is passed through an envelope follower. The signals from the envelope followers are transmitted, and the decoder applies the amplitude controlled signals to corresponding filters in the synthesizer. The main motivation for this type of system was to cryptographically encode

the signals during transmission. Delta modulation appeared in 1952, it is the simplest form of differential pulse-code modulation (DPCM) where the difference between successive samples is encoded into a one bit stream. Also in the 1950s the Lincoln Laboratory at MIT conducted a study of pitch in speech detection, which led to vocoders designed to reduce the speech bandwidth. The first LPC ideas came about in 1966 from work done at NTT in Japan. In the late 1960's early real-time versions of LPC coders were implemented. The first workable LPC encoder was the US government's LPC-10 coder developed in the early 1980's [133]. The ten in LPC-10 signifies the number of coefficients it used. 1964 saw the standardisation of PCM waveform coding for fixed telecommunication networks. The implications of this choice is still with us today.

Moving forward a number of years, warped LPC was first proposed in 1980 which is a variant of LPC where the spectral representation of the system is modified. This reduces the bitrate required for a given level of perceived audio quality/intelligibility. In 1985 the Code-Excited Linear Predictive (CELP) codec was introduced [89]. The ITU's G.729 was standardised in 1996 [62]. In 1997 the Enhanced Full Rate (EFR) codec was standardised. More recently intelligent multimode terminals have appeared that can adapt their configuration to different rates, quality and robustness. These are known as adaptive multirate AMR codecs which was standardised in 1998. For an account of the early vocoder history research consult [38].

2.5 Internetworking and voice

This section deals with the networking aspects of real-time voice communication. It explains how media synchronisation is achieved at a receiver, describes formats for transporting voice, how addressing and routing effect voice streams, as well as outlining the main quality detractors in IP voice communication.

2.5.1 The Real-Time Protocol (RTP)

The RTP protocol has been developed for end to end transport of real-time media, including unicast and multicast network services. RTP can also synchronise multiple streams arriving at a single receiver. Often the RTP protocol is used with the UDP datagram service and is used in conjunction with signalling protocols we have discussed, H.323 and or. RTP was first published as a standards track document by the IETF in 1996, more recent developments have been made up to July 2003 when it became a standard [121].

The primary role of the RTP protocol with regard to voice streams is to ensure intelligible playout of the speaker's words for the listener. Without RTP, disturbances in the stream may result in incorrect playout, for

example the voice might be reproduced too fast or slow or even with parts of the sentence clipped. Recall that VoIP systems do not (normally) use synchronised clocks, therefore the timing information needs to accompany the data so that the original voice stream can be recreated at the destination. Thus far we are only discussing the operation of the synchronisation protocol, network losses only serve to compound the problem.

To recreate the spoken pattern of words and silence periods, the sending application notes where there is speech activity and when there is none. The periods of speech activity are known as talkspurts. The start and stop times of these talkspurts are recorded using media dependent timestamps into the RTP packet. The RTP timestamp is based upon the sampling instant of the first sample to be put into a data packet. The clock frequency used to derive the sampling instants is dependent on the payload media (see table 2.1). This means the system clock is not directly used, rather some function of the media rate. In the case of 8000 Hz fixed-rate PCM sampling, the clock is updated 8000 times per second (once per sampling instance). If an audio application reads blocks of 160 sampling periods (i.e. every 20 ms), then the timestamp would be increased by 160 for each packet.

In addition to the synchronisation functionality, RTP is responsible for a number of other functions such as source identification, packet sequencing, stream profiling, payload identification, and multiple source multiplexing. Out of order delivery is permitted by RTP, if the application reassembles the stream from the sequence numbers. The RTP header is shown in figure 2.4. Here PT stands for payload type and is filled in by the application.

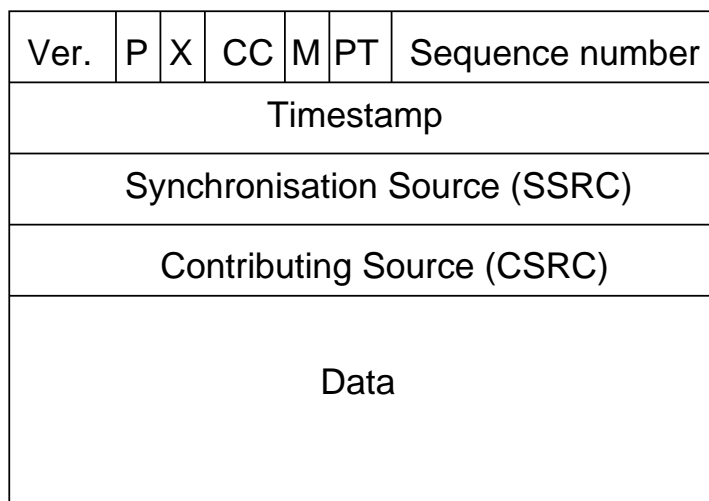


Figure 2.4: RTP header structure

A number of payload types have been specified by the IETF as shown in Table 2.1. The sequence number field is used to store the current packet

Payload type	Name	Type	Clock rate (Hz)	Audio channels	References
0	PCM-Ulaw	Audio	8000	1	RFC 3551
7	LPC	Audio	8000	1	RFC 3551
2	G.721	Audio	8000	1	RFC 3551
3	GSM	Audio	8000	1	RFC 3551
31	H.261	Video	9000	-	RFC 2032

Table 2.1: Some examples of RTP payload types

number in the stream. It is incremented by one by the sender for each data packet transmitted. The receiver uses it to calculate packet loss as well as to restore packet sequence if packets arrive out of order. A random value for the sequence number is selected at the start of a session in order to make DoS attacks on the session more difficult. The synchronization source (SSRC) identifies the synchronization source. This value should be unique and is also chosen randomly, with the intent that no two pair of synchronization sources within the same RTP session will have the same SSRC. The contributing source (CSRC) identifies the contributing sources for the payload contained in the packet. The CC field indicates the number of CSRC identifiers.

Refinements to the RTP protocol have primarily focused on header compression. The purpose is to reduce the combined size of the IP, UDP and RTP headers. In 802.11 networks, headers can be significantly larger than the payload itself. This is partly due to the large 802.11 frame headers. One proposal has been the Robust Header Compression (ROHC) scheme specified in RFC 3095 [132, 80]. The scheme is called robust as it can deal with relatively high error rates. Note that ROHC and similar schemes do not compress the payload, only the headers. The typical compression rate is from 40 bytes down to 4 bytes. Also note that shorter headers reduce the possibilities of bit errors in the frame, since they constitute fewer bits in the air.

In header compression schemes, a compressor and decompressor exist before and after the link where compression is needed. The basic idea is to send a complete header at the start of a session, and from then on only updates to this complete frame, called delta frames. Often more than one identical delta frame is sent to allow for low numbers of losses. There are different states within the compressor, such as full state, first order state where the static fields have been detected, second order state where dynamic fields are suppressed and replaced by logical sequences, partial checksums so the receiver can predict and generate the next sequence number and so on.

The Real Time Control Protocol (RTCP) is a companion protocol to RTP. RTCP is used in one-to-one or multi-party sessions by receivers to

inform the sender of the stream quality they are receiving. Observed packet loss, delay, and jitter are fed back to the sender. RTCP can generate data based on the start of the session or from the last report arrived. To calculate the round trip delay, the sender transmits a report containing the time the report was sent. On reception of this report the receiver records its current time. Therefore two times are now recorded within the report. When transmitting the report back to the sender, the receiver subtracts the time it held the report from the time it initially put in the report, therefore accounting for the time it held the report. Using this information, the sender can calculate the round-trip delay and discount the time spent processing the reports at the end points. This can be done in both directions if asymmetry problems are suspected.

Further extensions have also been proposed to RTCP [22, 102, 35]. These basically extended the information put into the reports to include end systems artifacts such as buffer levels and estimations of the quality received.

2.5.2 Addressing, routing, and timing constraints

The end-to-end delivery of voice packets is the joint responsibility of the networking and terminal equipment. This includes the end systems, access points, layer 2 switches, firewalls, NATS, IP routers and interchange points (iX's). In an MPLS network one has the label switch routers at the center of the network and the label edge routers at the extreme points of the network. To some degree, all network elements affect the end-to-end transmission of voice data, in particular delay. It has been argued for some time within the Internet community that several classes of traffic (including telephony) deserve higher priority than other data in order to reduce delay within equipment that queues or stores packets [3, 15, 9].

Simplistically the destination IP address is used to route each packet toward the target terminal. The UDP port field is used to demultiplex the data at the receiver to the correct application. The source identification field (SSRC) is used to locate the correct RTP flow within a session. For the actual routing of IP voice data, the normal IP routing mechanisms apply. Path information from a company, home, or university network is provided to the backbone using interior link-state routing protocols such as IS-IS or OSPF. In the backbone network, routes are determined by peer agreements and the inter-domain routing protocol BGP. In a MPLS network voice traffic may be given its own path through the label switched path if the particular operator has sufficient traffic for it to be worthwhile.

It is possible that a router may send some packets of a single stream via one route and other packets via another route. A more likely event however is that the route between two parties may differ over longer time intervals. That is to say, there is less path stability over longer durations. Route changes are an inherent fact of the IP infrastructure. However the issue

for voice traffic is that the delay requirements are not exceeded. A typical route from Europe to the US consists of approximately 20 intermediate routers. Routes are not necessarily symmetric, which means that the number of traversed routers is not the same in each direction. The end to end implications for voice depends on the traffic on each link and router, not purely on the number of hops.

Also it is possible that some of the voice packets will be lost, due to congestion in the routers, discarding algorithms such as RED, or link problems. Again, loss is unwelcome in telephony-like applications. Correlated losses are more likely to cause problems for the voice receiver which are more prevalent in wireless networks. Additionally non-licensed spectrum technologies are more prone to disturbances and losses of frames [16]. However, 802.11 provides a link layer retransmission protection that can alleviate frame loss on wireless access links to some degree at the expense of a little delay. Other sources of problems for IP-based voice are heavy traffic loads on shared links, poorly dimensioned links, long-delay link technologies (e.g. satellite links) and misconfigured equipment.

Over-provisioning and priority schemes can make acceptable quality IP telephony sessions possible. Lost packets cannot be retransmitted due to the overall delay budget for conversations, and therefore protection in the form of redundancy can be introduced at the sender, and concealment at the receiver can lessen the audibility of losses by interpolating small gaps in the sample sequence.

2.5.3 Packet delay

The *network delay* is the time taken for a packet from the operating system boundary at the sender to the operating system boundary at the receiver. The operating system boundary is usually thought of as the interface between the user/supervisor modes. In UNIX this would be user/kernel boundary. The *end system* delay varies widely from operating system to operating system and between VoIP applications. The delay incurred by an end system can vary from 20 ms up to 1000 ms, irrespective of the stream characteristics [47].

Since real-time voice has constraints on the end to end delay for the samples to reach the listener, we will now consider the constituents of the delay. From a routing perspective the path with the lowest delay is desired. This implies a propagation delay based upon distance. In reality finding the length of a link is not trivial, as the links can traverse non-obvious paths, be split into different paths and so on. This delay constitutes the *deterministic delay*, even if it is non-trivial to obtain. There are processing and queuing delays along the path too. Each packet needs to be processed by several routers. In most cases this means looking at the IP address within the header and finding the correct interface to forward the whole packet to.

Deciding upon which interface to select depends on matching the IP address in the header with a routing table. The path with the longest matching prefix is chosen. Whilst forwarding or processing, the packets behind it must wait, causing random delays. The instantaneous queuing delay at a router depends on: the traffic arriving at that instant, the processing rate of the router, the length of the packet and the number and lengths of packets waiting ahead of it. Due to processing and queuing delays, the original voice packet stream becomes distorted requiring resynchronisation at the receiver. The processing time for a voice packet is generally constant, but the queuing delay is variable, as it depends on the factors just mentioned.

Measuring one-way delays is not trivial without synchronised clocks [93]. One-way delays may be important from an operators perspective, but cannot be heard or distinguished by the users. Therefore it is easier to consider the round trip times. This is because most spoken words are responded to, creating feedback in the speech pattern between the two (or more) people. Only when the response is heard can a speaker have some idea of the delay. The tolerable round trip delay is typically in the order of 400 ms. Therefore the processing and queuing time per router should not exceed 10 ms if there are 40 router hops in the end to end path (20 in each direction). The interactivity of the conversation is affected by the round trip time, however defining an interactivity metric is not that simple, due to the human ability to adapt to varying delays. Conducting tests with pairs of people is more demanding than with individuals.

We have measured the network delay using the RTCP protocol, which is part of the RTP standard [121]. Because the sender and receiver exchange time reports it is possible to calculate the networking delay, by subtracting the time reports were held at the end host. Since these reports are exchanged every few seconds, the delay variations can also be found. This can be done in both directions to see if any significant asymmetries exist.

2.5.4 Packet jitter

Packet jitter is simply the variation in the delay. If isochronously sent packets arrive at the receiver with differing delays, the end to end transfer has introduced jitter into the voice stream. Jitter can have undesirable effects in a system. In voice systems it can lead to lengthened delays, due to the need to capture late packets. Loss can be incurred if the packet jitter is greater than the receiver buffer at that instant. One positive aspect of having a buffer in a voice system, is that it allows for a tradeoff of loss against delay. This means that the system is tunable to some degree, by using a buffer length that induces loss and reduces delay or increases delay and decreases loss. In a voice system the loss/delay balance should be based upon the acceptable round trip delay and the acceptable loss rate of the coding scheme.

Voice jitter is compensated for by re-aligning the timing of the packets to their recorded times. The jitter definition by the IETF is stated to be the mean deviation (smoothed absolute value) of the difference in packet spacing at the receiver compared to the sender for a pair of packets [121]. This is shown in equation 2.1.

$$J_i = J_{i-1} + \frac{(|D_{i-1} - D_i| - J_{i-1})}{16} \quad (2.1)$$

J_i is the current jitter value

J_{i-1} the previous jitter value

D_i is the current delay between two successive packets

D_{i-1} is the previous delay between two successive packets

16 is the smoothing constant

The jitter units are the timestamps used in the RTP packets, which is typically the packetisation interval multiplied by the sampling rate. If S_i is the RTP timestamp from packet i , and R_i is the time of arrival in RTP timestamp units for packet i , then for two packets i and j , $D_{(i,j)}$ (where j is sent after i) may be expressed as:

$$D_{(i,j)} = (R_j - R_i) - (S_j - S_i) = (R_j - S_j) - (R_i - S_i) \quad (2.2)$$

$D_{(i,j)}$ Delay for packet pair (i,j)

R_i Reception time for packet i

R_j Reception time for packet j

S_i Send time for packet i

S_j Send time for packet j

In practice one can calculate the jitter as the difference in the relative transit time for two packets. This is because the S_i and S_j are sent at (roughly) constant intervals, i.e. the time difference between two successive packets in RTP timestamp units. The jitter value is sampled and sent in RTCP reports, so that the sender has a quantitative notion of the packet delay variability in successive reporting intervals.

One other measure closely associated with jitter is the difference in the inter-arrival times. This is simply the difference between the arrival times of two consecutive packets. Given the packetisation time it is simple to calculate by how much the packet separation has been distorted. One other method for measuring the separation is to consider the difference in the time between when the packet arrived and when it *should* have arrived. Note that this measure can be either positive or negative.

2.5.5 Packet loss and redundancy schemes

Packet loss is the major quality detractor in Internet telephony as far as the network is concerned. Packet loss implies lost speech frames. Packets from a stream can be discarded from router queues, either due to buffer space restrictions or by explicit congestion alleviation algorithms. Some algorithms implement a random mechanism for discarding packets in order to ensure fairness between flows. Under-dimensioned as well as poorly administered networks often yield higher loss characteristics.

Transmission errors on fixed modern networks are rare, while frame losses are still prevalent for wireless networks. In wireless networks the interference from competing transmissions and weak signal conditions are the main causes of frame loss. Switching between base stations or access points also leads to bursts of lost packets. Wireless networks usually implement mechanisms for link layer retransmissions; nevertheless conditions may arise that lead to IP packet loss once a number of retransmissions of a frame has been unsuccessfully attempted.

Depending on where in the phrase the losses occur, relatively large differences in the intelligibility can be perceived [56]. The speech may be encoded to make it more resilient to loss. Redundancy optionally adds delay to the system as additional packets need to be received if the receiver is to recreate lost packets. The delay incurred depends on the lost packet's position in the redundancy block. The size of the block should be chosen so as to optimise a delay-quality tradeoff function as in [112].

Sample data from lost packets can to some degree be masked by speech codec-specific algorithms. That is, lost speech frames can be masked by the speech decoder where gaps are detected. Frames are created from those frames that are present, usually the ones just before and after the missing frames. Recreating lost frames is desirable as replying any sound has been shown to be perceptually more tolerable than just silence. Understanding the impact of losses in perceptual terms is not a trivial task. PESQ is one solution to assessing the influence of loss on a phrase, another is subjective user tests. A controlled response to packet loss is desirable from a speech coding point of view, thus this has been one goal of iLBC, G.729 and GSM. Studies by researchers in the 1990's advocated the use of forward error correction since losses were correlated but often only by a small amount [10]. Forward error correction (FEC) and multiple description coding (MDC) are techniques to reduce the probability of gaps in the decoder input. In IP networks FEC and MDC redundancy packets are sent time shifted from their originals. In voice communication the receiver buffer algorithm needs to make a delay calculation how long to wait for the redundant copies, assuming the originals were lost. More sophisticated scheme can feed back this information to the sender to regulate the amount of redundancy. This can be done in Reed-Solomon FEC coding for example. A comparison of the

rate distortion for these techniques can be found in [82]. Other techniques for loss analysis are [74, 73, 75]

The widespread deployment of local wireless access has changed this view somewhat. This is because the loss pattern in this setting is quite different than the fixed wired Internet (where losses were not correlated to the same degree). Unfortunately, in wireless systems correlated losses are more common, primarily due to poor signal reception at the receiver.

In practice, packet losses can be detected using the sequence numbers in the RTP header and loss ratios over time can be reported using the RTCP protocol. Further quality extensions have been proposed, such as sending back the loss distribution or finite-state model parameters (such as the Gilbert 2-state model) of the observed loss pattern. More meaningful information by the receiver (or indeed an access point) can lead to better solutions, whether it be over the last link or end to end.

Chapter 3

VoIP quality aspects

This chapter is divided into two parts, quantifying quality and some standardised approaches for calculating it. As tools and methods have been developed for the telephony industry, it seems natural to re-use them for Internet telephony where appropriate. We will introduce two standardised methods for estimating VoIP quality as they are used within this dissertation. For a more in-depth treatment of objective and subjective methods consult [106].

3.1 Quantifying quality

Although most people have a good feeling of what good quality (or more accurately fidelity) means during electronic communication, it is not straightforward to translate this into measurable parameters of a system. First the system we are dealing with is a distributed system and each component has its own individual attributes. Second people are involved in the assessments, and add inevitable human variations. Third, people are adaptable, therefore ratings tend to change over time and finally the situations differ from environment to environment.

The simplest form of quality rating for speech would be something descriptive, for example 'EXCELLENT' for a speech sequence that was almost glitch-free down to 'POOR' which was barely understandable. Different words could be used, or any number of intervals between the extremes choices, however studies have shown, in a descriptive setting, three intermediary steps are reasonable. Numerically, it is somewhat easier to get a finer scale, however more than ten intervals often leads to fuzziness between the intervals.

3.2 Measuring quality

Determining an accurate quantitative measure for human speech fidelity is desirable, but impossible. The best one can achieve is a qualitative rating that has been established in a rigorous and controlled manner. Typically test listeners and controlled auditory conditions are used for people to rate speech coder performance for example. It can be expensive and time-consuming. There are tools and methods that map qualitative assessments to quantitative values, however they will always be, to some degree, approximate. If one can show however, that there is reasonable correlation between the qualitative and quantitative results, and under what conditions the correlation holds, then this solution may be acceptable to some users. Some objective tools, such as those which use signal processing techniques, have shown this correlation and hence have found acceptance within the community. Therefore with some degree of confidence, the software developers can justify their techniques have proven success and give results as real people would.

3.3 Quality tolerances

When human speech is uttered, the time taken from when the pressure waves leave the mouth to the sensation of hearing is a fraction of a second for a nearby speaker. We have evolved to expect, and actually need, to hear our own voice. This is in order to be sure that we are saying what we really want to. The development of the human speech and hearing recognition has however taken place via face to face meetings. Thus, extra visual or body cues are available when uncertainty is present. An example of such ‘understanding’ is when a language is being spoken that we do not understand. We can sometimes guess the meaning from gestures, facial expressions and intonation.

On the other hand, impaired speech requires extra concentration from the listener, that is we are not used to processing distorted or missing segments, visual and auditory clues are more difficult to interpret. Somewhat similarly is communicating with people from afar, we don’t receive the original speech samples and visual cues are harder to see.

In IP voice communication systems the visual cues are not existent, thus making intelligibility more important. In order to hear one’s own voice a very short delay is introduced between capturing the recorded voice and replying it for the speaker. This is particularly applicable when using headsets. The introduced delay is in the order of 5 ms.

As far as the delay in the system is concerned, it is obviously desirable to keep it below some maximum. This is in the order of half a second. Delay is discussed from a networking perspective below. Recent results have shown that delay is not as significant as once postulated, at least in VoIP systems.

Traditional telephony standards have been much stricter with respect to delay budgets [134]. If one is not in a highly interactive conversation, then higher delays can be tolerated than those suggested by telecommunication standards. This is particularly true in situations where people use computers, delays are expected by users (operating system hiccups) and therefore their delay expectations also become relaxed from the communication system.

If users are engaged in quick voice exchanges, delays will frustrate their conversational style. Therefore, introducing the factor of interactivity into an objective quality measure is still under research. The following studies have looked at conversational interactivity [136, 46, 48, 107, 49]. The last reference in this list proposes the potential impact of interactivity on the perceived quality for Internet telephony services.

Where delays and losses are experienced at the same time, it has been shown that the influence of losses is much more significant with respect to the perception of quality degradation than the influence of delay. This implies that people are able to make a transition from highly interactive scenarios to a more measured communication style. In fact this transition appears to be somewhat bilinear, that is, the quality degradation from an interactive mode to a simplex conversation mode occurs in two linear steps, with the break at about 400 ms. Varying delays can be disturbing, due to the listener not being allowed to settle into a single mode of operation. For more information on the influence of delay on Internet telephony see [14].

3.4 Quality and noise

The quality of voice communication actually depends on many (independent) factors. The effect of noise, be it in the electrical circuitry, or in the surrounding environment can be a determining factor in the perceived quality.

The quality of the components is a key issue in voice systems. Lower quality components can leave voice sounding thin i.e. a lack of bass in the speech. Background noise, caused by poor grounding or shielding of the analogue components is frequently experienced as low frequency humming in the system. Internet telephony systems that use on-board sound cards can introduce noise of this nature into the signal. USB headsets are helpful, and they also alleviate the need for echo suppression.

The environment is another factor, whether a noise source is remote (distant from) or local (close by) to the speaker. In the remote case, the non-speech parts of the voice should be suppressed so as not to interfere with the spectral analysis of the voice processing. Undesirable noises from similar frequencies and volumes will be encoded into the signal, sent, and reproduced for the listener. Often listening to a remote speaker in the pres-

ence of background noise is more difficult than when background noise is present locally.

Research in the signal processing field has studied the issue of noise in systems [115]. Important speech parameters such as the intelligibility, clearness, or naturalness of speech can be improved by signal processing using digital, analog, or hybrid solutions. A robust, low complexity, speech enhancement algorithm has been proposed to show the advantages of a purely digital, purely analog, and a hybrid digital-analog implementation in [116].

In terms of testing systems with controlled noise, the ITU conducts tests with standardised background noises. These are known as mean noise reference units (MNRU) [63]. Typically well defined noise patterns of fixed modulated noise are presented at the beginning of each test. Each sample represents an example distortion corresponding to a five grade impairment scale (excellent to poor). The MNRU has been used extensively in subjective performance evaluations of conventional telephone and wide-band voice systems.

3.5 The ITU-T E-model

The E-model is intended as an off-line planning tool. Due to its simple form it has found applications into on-line assessments as well. Network planners can input parameters from a system and obtain a numerical value (between 1 and 100) representing an estimate of the perceived quality. One important point of the E-model is that loss, delay, jitter, speech coding and echo parameters are combined *linearly* to calculate the so called impairments that result in the score. The E-model assumes the parameters are *independent*. Another important (selling) point of the E-model is that the numerical scores correlate well with subjective tests, indicating that this estimation is indeed possible. Since the linear combination is simple, and most of the parameters are easily measurable, the E-model has been popular for a number of years.

The E-model also indicates how network impairments and speech coding can be combined to give an approximate estimate of voice quality. It is important to state that there are many tunable parameters included in the model, 19 in fact, not including the different speech encodings and loss concealment methods. Interestingly, jitter is not explicitly included as an input parameter. As jitter can affect whether packets arrive in time for playout or not, late packets for a real-time audio application are akin to network loss or delay, which are included in the model.

Table 3.1 shows scalar values known as the R-value derived from the computational model. They are relatively consistent with subjective scores, i.e. real user estimations of the speech quality, shown by their respective mean opinion scores (MOS). Mean opinion scores are derived by replaying samples to a naïve set of listeners who rank the quality on a scale from 5

User satisfaction	R-value	MOS score
Very satisfied	90	4.3
Satisfied	80	4.0
Some users dissatisfied	70	3.6
Many users dissatisfied	60	3.1
Nearly all users dissatisfied	50	2.6

Table 3.1: The ITU's E-model and MOS scores

(best) to 1 (worst). The R-value is defined as shown in equation 3.1.

$$R = R_o - I_s - I_d - I_{e-eff} + A \quad (3.1)$$

R = rating value

R_o = signal to noise ratio (noise sources)

I_s = voice impairments to the signal (side-tones and quantisation distortion)

I_d = delay and equipment impairments

I_{e-eff} = packet loss impairment (including random packet losses)

A = advantage factor (compensation of 'other' factors)

Each of the factors is calculated and subtracted from the maximum of 100 to obtain the R-value. The impairment due to the delay is denoted by I_d . Two different values are defined, $I_d = 0$ if the absolute delay (T_a) is less than 100 ms, i.e. no impairment or an increasing I_d if the delay is over 100 ms. A number of amendments have been to incorporate non-random losses into the model [1, 27]. The effect of packet loss on the R-value is given by the I_{e-eff} term. The I_{e-eff} is defined in the E-model as:

$$I_{e-eff} = I_e + (95 - I_e) \cdot \frac{P_{pl}}{P_{pl} + B_{pl}} \quad (3.2)$$

P_{pl} = packet loss probability

B_{pl} = packet loss robustness

For G.711, $I_e = 0$. This means for situations without loss, G.711 provides the best speech quality. The advantage factor A , is a value that indicates how tolerant users can be when using telecommunication equipment. It can be seen as a willingness to trade quality for operational convenience. One example is with mobile telephony, where users accept lower quality since they have the luxury of being mobile. One other example could be an advantage factor, as mentioned, where higher delays are tolerated when using a computer as a communicating device rather than a telephone.

3.6 Perceptual Evaluation of Speech Quality (PESQ)

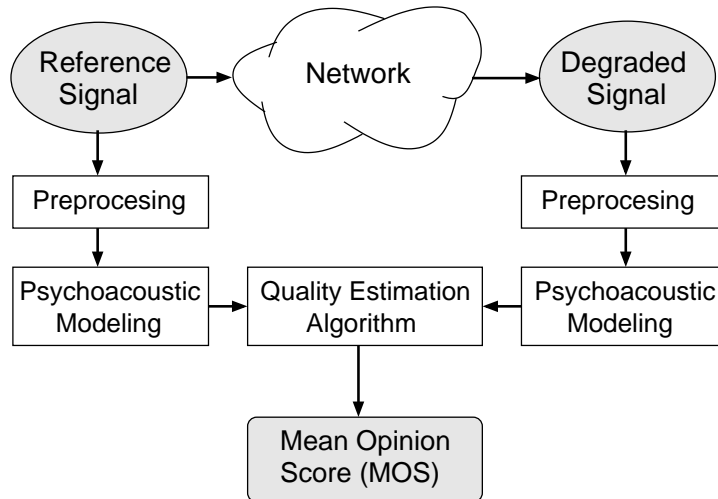


Figure 3.1: The PESQ processing structure

PESQ MOS	Linguistic equivalent	Quality degradation
4.5	Excellent	None
4	Good	
3.5	Good/Fair	Moderate
3	Fair	
2.5	Fair/Poor	Severe
2	Poor	
1	Bad	

Figure 3.2: A quality degradation scale

Although the E-model is popular for estimating quality using network parameters, it has shortcomings. As we have seen, the bursty effects of packet loss on speech quality are not well addressed in the E-model. A later development by the ITU was to develop a scheme that could improve on the E-model by estimating the impact of speech coding and losses on the original speech signal itself. The solution, the “Perceptual Evaluation of Speech Quality” or more commonly known PESQ, addresses these issues [135].

The idea is to estimate the *degradation* of the coding and loss on a speech sample using a model of the human auditory system. Figure 3.1 shows the functional units of PESQ. A reference speech signal is transmitted through a network that results in a quality degradation corresponding to the coding

used and the network losses. PESQ analyses *both* the reference and degraded signal and calculates their representation in the perceptual domain based on a psychoacoustic model. The disturbance between the original and the degraded speech signals is calculated by a quality estimation algorithm and a corresponding subjective mean opinion score (MOS) is derived. The evaluation of speech quality using PESQ is performed off-line due to its computational complexity. If one assumes a 20 ms packetisation and an eight second sample, the sequence would then be 400 packets. As an indication of the time needed to compute a PESQ score, a sequence with ten losses requires approximately two seconds of processing time for G.711 coded speech on a Pentium III computer. G.711 yields the maximum PESQ score (4.5) in the absence of loss, however it is particularly sensitive to packet loss even when concealment is used.

PESQ's validity has been shown by its ratings being sufficiently correlated to subjective ratings as we discussed in the introduction of this chapter. More recent research that correlates PESQ with subjective scores, shows that some small transformations are needed to better align PESQ to MOS [108].

3.7 Other measures

Recent work by Hoene et al. proposes a real-time implementation of PESQ called PESQlite [54]. The idea is that using PESQ in real-time is too slow for real-time use. Hence PESQlite reduces the complexity by making simplifications to the PESQ algorithm, e.g. by using constant length test samples and non-time alignment of the degraded samples. PESQlite is currently only available for G.711 coding.

One other alternative for an objective measure is to use machine speech recognition as a MOS predictor [76]. The technique uses a word recognition ratio metric to reliably predict perceived quality. This ratio is speaker-independent, whereas the absolute word recognition ratio of a speech recogniser is speaker dependent. The relative word recognition ratio is obtained by dividing the absolute word recognition ratio with the value at 0% loss. The results show that human and machine based recognition techniques are correlated, although not linearly. It is also been found that human-based word recognition ratio does not degrade linearly once packet loss exceeds 10%, due to performance limits of the codec.

Chapter 4

Packet-switched voice research: A brief history

In this chapter we provide a brief history on the development of Internet telephony. The focus is primarily on research related activities, rather than commercial ones. We will however mention standardisation efforts, as they are significant in the history of Internet telephony.

4.1 Pre-Internet days (1970-1980)

Well before the modern Internet was devised, people were investigating alternatives to the traditional telephony system for carrying voice. The earliest accounts of packet switched networks can be found in the signal processing community. Researchers and engineers were looking for computationally efficient methods of compressing voice for transmission over low bandwidth links. In fact, advances in low data rate coders and the deployment of a distributed packet switched network led to some of the earliest findings [91]. The details of the networking are often omitted, but the idea was to block-code voice for transmission. Much of the focus was on LPC and entropy methods. Blankenship et. al described the Lincoln Laboratory digital voice terminal system in a technical note published in 1975. Accounts of the early days of vocoder work can be found in [42] and the small amount of networking in [25].

In 1973, the Network Voice Protocol (NVP) was developed by Danny Cohen, then at the University of Southern California. NVP was used to send speech between distributed sites on the ARPANET using LPC coding. The protocol was implemented in two parts: a control protocol and a data transport protocol. The control protocol is the equivalent of today's signalling solutions (H.323/SIP), and the data transport akin to packet transport using pure UDP without RTP.

William Naylor in 1974 published "A status report on the real-time

speech transmission work at UCLA” [97]. In this work he expresses the need to smooth out the variable delays of a stream of packets (speech, for example) on packet switched networks. “This is to preserve the continuity of the stream”.

In 1977, Cohen wrote that the packetisation algorithm and data rate should be varied according to the network load [24]. Interestingly, this adaptive approach of reacting to the network load become popular many years later. Cohen also states that the time spent at the receiver (called the ‘waiting period’ in his paper) should be a function of the network performance. Indeed, Cohen states that the parameters in a real-time voice communication system are heavily dependent on the network performance, and a systematic method of predicting it must be developed. In the same year (1977) Naylor published his doctoral dissertation “Stream traffic communication in packet switched networks” [98].

James Forgie published “Speech communications in packet-switched networks” in [34]. In the first half of [38], Gold gives a background of speech coding techniques available in 1977. In the second half of the paper he gives an explanation of packet speech experiments performed across the ARPANET. He considered the delays both in coding and the transmission of different sized packets, as well as the variation in the delay. His conclusions were that reassembly needs to be done at the receiver via buffering, however vocoder techniques could be used without significant loss in the speech quality.

Among the other early works in this period was John Gruber’s “Variable delays in a shared network environment handling voice traffic” in 1979 [43]. His vision was a hybrid packet and circuit switched network called ‘Transparent message switching’ for handling both voice and data traffic. The ideas were novel, preliminary, and pre-cursory to both ATM and today’s multiservice Internet. The basic entities processed are messages rather than calls. The messages belong to an established call, however they may be completed or blocked at the network periphery. Voice messages are given priority when delays are excessive, however when loss is being experienced, voice messages could be discarded. This observation is interesting in that, delay was seen as a more critical issue than loss in those days.

In December 1984, Warren Montgomery published “Techniques for Packet Voice Synchronization” in [94]. He considers the local and wide area network scenarios separately to synchronise VoIP receivers. The paper discusses four types of delay calculations: blind delay as the worst-case assumption, round-trip measurement as estimated by the sender, absolute timing using a master clock, and accumulated variable delay using a time stamp as synchronization methods for packet playout at a receiver. Round trip estimates are sufficient for the local area case, whilst more sophisticated methods are needed for the wide area case. He suggests that the addition of timing information and incorporating extra delay at the receiver should be sufficient to

yield satisfactory voice quality in the wide area case. This is the approach taken by most modern real-time packet voice applications, as it is effective, relatively simple, and cheap to implement.

4.2 A decade of research (1980-1990)

The early eighties produced a flurry of packet-based voice research. Probably fuelled by other developments in IP research. This period is sometimes referred to as the golden age of IP networking research.

Much of the voice focus was on solutions, mostly theoretical, for buffer design and sizing [6, 5]. Work by Naylor and Kleinrock described general design methodologies for the design of jitter absorbing buffers [99]. Some early performance evaluation papers were also published, being both theoretical [128] and simulation studies [130]. Mackie et. al even considered a complete system [90]. Weinstein [140] and Adam [2] gave accounts of experiments using the ARPANET and the Cambridge ring LAN respectively. These relatively early works gave some valuable insights into the issues we face today.

A loss concealment scheme was published in Jayant and Christensen's 1981 article "Effects of Packet Losses in Waveform Coded Speech and Improvements Due to an Odd-Even Sample-Interpolation Procedure" [72], which was a form of Multiple Description Coding (MDC) using separate descriptions of the speech signal. Psychological and quality aspects were also beginning to emerge, Goodwin authored a book about the interaction of 'speakers and hearers' in the early 1980s [39].

Some researchers looked at quality aspects particularly for packet-voiced systems [44, 83]. Holtzman looked at the interaction between queuing and voice quality in variable bitrate packet voice systems [57]. Network delays [40] and statistical multiplexing of voice [86, 127] also appeared for packet voice, along with some early priority schemes for voice traffic [96]. The ITU released numbers of important specifications [59, 65, 58, 60]. Importantly in this decade, IP and ATM were competing technologies, with ATM keeping voice foremost in its multiservice solution. Basically the ATM Forum proposed five different circuit emulation services, depending on the capacities required. Although both IP and ATM were technically viable for both voice and data, the flexible data transport structure of IP, plus the development of the HTTP protocol, and lower hardware costs effectively sealed the fate of IP over ATM.

4.3 Emergence of telephony applications (1990-1995)

In the early nineties, Domenico Ferrari's group at the University of California at Berkeley published a number of significant papers about the effect of jitter

and delay on real-time communication applications [138, 32]. Their work proposed a distributed mechanism for controlling the delay jitter in a packet-switching network. They argued that if the advantages were sufficiently high, then the implementation was worthwhile. Although no such scheme was deployed, their work is still widely referenced as seminal.

Events such as IETF meetings and the space shuttle missions helped popularise conferencing over the Internet [21]. The space shuttle sessions were reception only, whereas people actively participated in the IETF meetings. Both showed that the Internet could support sessions of thousands of people, both passively and actively using IP multicast. The impact factor was significant, however it was only really realised by the Internet community at that time. It was the first time voice and video could be seen by normal users, via mechanisms other than radio or television. Additionally the MBONE sessions permitted group participation. Research continued on IP multicast, although it never really caught on for large scale deployments. A suite of real-time applications were produced, notably VIC, VAT, and wb (whiteboard) from the Network Research Group at LBL, USA [71] and at GMD, in Germany with Nevot [119].

Between 1993-1996 Jean Bolot wrote a series of papers that reported on, and characterised the loss and delay of audio packets on the Internet [10, 11, 12]. They were largely theoretical studies supported by experimental evidence that advocated the use of techniques such as redundancy protection against packet loss. In the late 1990s, a tool called Freephone was developed by the Rodeo group at INRIA in Sophia Antipolis, France which implemented FEC mechanisms [109]. At that time all the applications were UNIX based, as this was the only (open) operating system for Internet applications.

These earlier works led to a standardised transmission protocol, RTP for use with real-time media flows. One of the authors was Van Jacobson, who gave a Sigcomm tutorial in London 1994 entitled “Multimedia conferencing on the Internet” [70]. In this presentation he suggested using a simple synchronisation protocol to restore the original timing information at the receiver and a small adaptable buffer to absorb delay variations. Although the idea had been suggested by others previously, the presentation was influential and moulded the approach taken by researchers for many years. It also promoted the development of the RTP standard.

Henning Schulzrinne’s 1993 PhD dissertation “Reducing and characterizing packet loss for high-speed computer networks with real-time services” studied congestion control, scheduling, and loss correlation for real-time traffic [120]. Schulzrinne highlighted the practical importance of scheduling packet audio within the operating system. He was also one of the main contributors to the RTP protocol and has produced the most seminal research (over 50 publications) and prototypes within voice research, including SIP and RTSP.

Tools such as the Robust Audio Tool (RAT) came from UCL in London in 1995 [52]. RAT, with its simple redundancy scheme, sending one compressed version of the packet in the following one, was an simple example of utilising redundancy. RAT was intended for both group, and one-to-one conferencing. Somewhat surprisingly, RAT and VIC were still being maintained today as part of the AVATS project (formerly SUMOVER) at UCL, London.

4.4 Early deployment days (1995-2000)

Internet telephony seemed to succeed as a business, therefore many researchers took to looking at the core issues again. Some of the important VoIP papers appeared in 1995. The problems of packet loss was addressed in [12, 52, 118]. Packet jitter and playout were readdressed in [129, 95, 124]. Some fundamental design issues for the Internet were proposed in [123], and a book on speech coding and synthesis was published by Kleijn [84] (one of the creators of the iLBC codec). Also one of the first papers on IEEE 802.11 and VoIP was published in the same year [139].

In 1994, an IETF developed QoS mechanism arrived, called Integrated Services [15], it influenced many researchers and their real-time media agendas. Arguably, the proposal of new QoS mechanisms stifled pure packet-based research. That is, research on understanding core VoIP issues and receiver-based mechanisms for optimising the perceivable quality. This seemed to be the case both during the frantic Integrated Services and Differentiated Service periods. The first RFC for RTP appeared as late as 1996 [121] even though RTP was being used in the VIC and VAT tools since 1992.

As for speech coding, the first G.729 standard was released in 1996 [62]. As noted earlier, G.729 is an 8 kb/sec LPC-based coder still used in many VoIP applications today. This includes the Skype application when using IP to telephony services e.g. in the SkypeIn and SkypeOut services. As the load on the Internet grew, studies of error recovery were being published [13, 110]

Two forwarding-looking articles were also published in 1997, [79] and [23]. The first suggested that the research community should concentrate on: quality issues for voice, in particular the effect of consecutive losses on speech quality, RTP multiplexing, and multicast. The second details how the Internet needs to be modified to host IP telephony applications.

The ITU-T E-model was first proposed in 1998 [64]. Some important IETF standards were also first published in the same year namely RTSP [122] and SDP [50]. Some methods for recovering lost VoIP packets are summarised in [104], with FEC techniques summarised in [105].

4.5 Internet telephony comes of age (2000-present)

As IP voice entered the mainstream, Internet telephony research became more focused as well as standardised. We will begin with the standardisation of Internet telephony, then move onto the research efforts.

During the past nine years, the signalling protocols have become established, SIP and H.323, continue to be developed. Today, SIP also plays a central role in IMS [19]. In the non-standard protocol realm some researchers have reverse engineered the Skype signalling protocol and published their findings [131, 45].

One of the more active research areas has been in wireless voice services. Focus has mainly been in the areas of throughput and capacity issues of IEEE 802.11 networks. Casetti et al. present a framework that assumes variable rate speech coders at rates of 64 kb/s, 13 kb/s, and 8 kb/s [20]. Their rates are determined by an end to end control mechanism, based on measurements of packet delay and loss rates. Another approach is to look at the MAC protocol directly. Dong et al. propose and examine selective error checking (SEC) at the MAC layer of 802.11 [29]. They make use of the fact that speech bits can tolerate errors, but should be protected for optimal quality reproduction. Simulation results showed that the speech quality can be substantially improved by modifying the MAC layer with SEC to suit the Narrow-Band Adaptive MultiRate (NB-AMR).

Filali looks at a MAC tuning approach [33]. He exploits the properties of multimedia applications in IEEE 802.11-based wireless networks by limiting the number of retransmissions of a data frame by a source until the reception of a link-level acknowledgement from the destination.

In 2005, the IEEE approved QoS service enhancements for local area network applications called IEEE 802.11e. Garg et al. examines using the IEEE 802.11e protocol for voice applications [36]. The Enhanced Distributed Coordination Function (EDCF) has been proposed as a MAC protocol. EDCF assigns four different priority classes for incoming packets at each node which are called access categories (AC). Each AC has its own channel access function. This is in contrast to the standard Distributed Coordination Function (DCF) where packets all use the same access function to the channel. Access functions for different categories means assigning delay times, minimum contention windows, and the number of back-off stages for each type of service.

Garg et al. looked at 802.11e's ability to fulfil the goals of improved QoS and higher channel efficiency. They investigated the response of the protocol to options in the protocol parameters and showed that the Hybrid Coordination Function (HCF) reduces channel contention and provides improved channel utilisation. Both MAC coordination functions, EDCF and HCF, are sensitive to protocol parameters which are dependent on the scheduling algorithms. They conclude that further investigations need to be conducted.

Kawata et al. propose a dynamic Point Coordinator Function (PCF) for improved capacity [81]. They suggest two new media access schemes, dynamic point coordination function (DPCF) and modified DPCF (DPCF2). The claim is that the capacity of VoIP traffic can be increased by up to 20% in 802.11b networks. They show how a significant improvement in the end-to-end delay with mixed VoIP and data traffic can be achieved. Delay is maintained at approximately 100 ms in heavily loaded traffic conditions, whilst at 60 ms in normal traffic conditions.

Lindgren et al. [88] evaluate four mechanisms for providing service differentiation in IEEE 802.11 networks. The evaluated schemes are the PCF of IEEE 802.11, EDCF of IEEE 802.11e extension, Distributed Fair Scheduling (DFS), and Blackburst. Using simulation they looked at throughput, medium utilisation, collision rate, average access delay, and delay distribution for a variable load of real time and background traffic. The simulations showed that the best performance is achieved by Blackburst. PCF and EDCF are also able to provide good service differentiation. DFS can provide relative differentiation and consequently avoids starvation of low priority traffic.

Currently voice occupies relatively little of the IP wireless access capacity and the majority of voice traffic is carried by the cellular networks. Research in combining these two has been published within the context of voice roaming [17, 92]. Exploring voice quality in IP networks continues to be an active research area [41, 137].

Bibliography

- [1] ITU-T Study Group 12 Delayed Contribution 27. General Prediction of the Impairment due to Dependent (Non-Random) Packet Loss for Inclusion in the E-model, January 2005. ITU-T SG 12.
- [2] Chris Adams and Stephen Ades. Voice experiments in the UNIVERSE project. In *IEEE Record of the International Conference on Communications (ICC)*, pages 927–935, Chicago, Illinois, June 1985.
- [3] Paul Almquist. Type of service in the Internet protocol suite. RFC 1349, Internet Engineering Task Force, July 1992.
- [4] Soren Vang Andersen. iLBc - a linear predictive coder with robustness to packet losses. In *IEEE Workshop on speech coding*, pages 23–25, October 2002.
- [5] Giulio Barberis. Buffer Sizing of a Packet-Voice Receiver. *IEEE Transactions on Communications*, 29(2):152–156, February 1981.
- [6] Giulio Barberis and Daniele Pazzaglia. Analysis and Optimal Design of a Packet-Voice Receiver. *IEEE Transactions on Communications*, 28(2):217–227, February 1980.
- [7] B. Bessette, R. Salami, R. Lefebvre, M. Jelinek, J. Rotola-Pukkila, J. Vainio, H. Mikkola, and K. Jarvinen. The adaptive multirate wide-band speech codec (AMR-WB). *IEEE Trans. on Speech and Audio Processing*, 10(8):620–636, November 2002.
- [8] Uyles Black. *Internet Telephony: Call Processing Protocols*. Prentice-Hall, November 2000.
- [9] Steven Blake et al. An architecture for differentiated services. *Request for Comments (Informational) RFC 2475*, Internet Engineering Task Force, December 1998.
- [10] Jean Bolot. Characterizing end-to-end packet delay and loss in the Internet. *Journal of High Speed Networks*, 2(3):305–323, 1993.

- [11] Jean Bolot. End-to-end packet delay and loss behavior in the Internet. In Deepinder Sidhu, editor, *ACM Symposium on Communications Architectures and Protocols*, pages 289–298, San Francisco, California, September 1993.
- [12] Jean Bolot, Hugues Crepin, and Andrés Vega-Garcia. Analysis of audio packet loss in the internet. In *Proceedings International Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV)*, Lecture Notes in Computer Science, pages 163–174, Durham, New Hampshire, April 1995. Springer.
- [13] Jean Bolot and Andrés Vega-Garcia. The case for FEC-based error control for packet audio in the Internet. In *ACM Multimedia Systems*, 1997.
- [14] Catherine Boutremans. *Delay Aspects in Internet Telephony*. PhD thesis, EPFL, December 2003.
- [15] Robert Braden, David Clark, and Scott Shenker. Integrated services in the Internet architecture: An overview. *Request for Comments (Informational) RFC 1633*, Internet Engineering Task Force, June 1994.
- [16] Kim Byoung-Jo, R. Shankaranarayanan, N.K. Henry, P.S. Schlosser, and K. Fong. The AT & T Labs broadband fixed wireless field experiment. *IEEE Communications Magazine*, 37(10):56–62, October 1999.
- [17] Andrea Calvagna, Giacomo Morabito, and A. Pappalardo. WiFi mobility framework supporting GPRS roaming: Design and Implementation. In *IEEE International Conference on Communications*, pages 116–120, 2003.
- [18] Gonzalo Camarillo. *SIP Demystified*. McGraw-Hill Professional, August 2001.
- [19] Gonzalo Camarillo and Miguel-Angel Garca-Martn. *The 3G IP Multimedia Subsystem (IMS): Merging the Internet and the Cellular Worlds*. Wiley, February 2006.
- [20] Claudio Casetti and Carla-Fabiana Chiasserini. Improving fairness and throughput for voice traffic in 802.11e EDCA. In *IEEE PIRMC'04*, Barcelona, Spain, September 2004.
- [21] Stephen L. Casner and S. E. Deering. First IETF Internet Audiocast. *ACM Computer Communication Review*, 22(3):92–97, July 1992.
- [22] Alan Clark, Robert Cole, and Kaynam Hedayat. RTCP extensions for voice over IP metric reporting. *Internet Draft, draft-clark-avt-rtcpvoip-01.txt*, July 2002.

- [23] David Clark. A Taxonomy of Internet Telephony Applications. In *25th Telecommunications Policy Research Conference*, Washington, DC, September 1997.
- [24] Danny Cohen. Issues in Transnet Packetized Voice Communications. In *Proceedings of the Fifth Data Communications Symposium*, pages 6–10–6–13, Snowbird, Utah, September 1977.
- [25] Dany Cohen. Realtime networking and packet voice. In *ACM Sigcomm Tutorial*, August 1999.
- [26] SIP / H.323 Comparison. Numera. <http://www.nuera.com/applications/sipH323pfv.cfm>.
- [27] ITU-T Delayed Contribution D.221. E-Model: Additivity of Burst Packet Loss Impairment with other Impairment Types, March 2004.
- [28] Ismail Dalgic and Hanlin Fang. Comparison of H.323 and SIP for IP telephony signaling. In *Photonics East*, Boston, Massachusetts, September 1999. SPIE.
- [29] Hui Dong, I.D. Chakares, A Gersho, E. Belding-Royer, and J.D. Gibson. Selective bit-error checking at the MAC layer for voice over mobile ad hoc networks with IEEE 802.11. In *WCNC*, March 2004.
- [30] The Economist. The end of the line, October 2006.
- [31] European Telecommunications Standards Institute. Generic Access Network (GAN). *3GPP TS 43.318*, 2005.
- [32] Domenico Ferrari and Dinesh C. Verma. A Scheme For Real-Time Channel Establishment In Wide-Area Networks. *IEEE Journal On Selected Areas In Communications*, 8(3):368–379, 1990.
- [33] Fethi Filali. Dynamic and efficient tuning of IEEE 802.11 for multimedia applications. In *IEEE PIMRC 04*, pages 910–914, Barcelona, Spain, September 2004.
- [34] James W. Forgie. Speech communications in packet-switched networks. *91st Journal of the Acoustic Society of America*, 59(1), April 1976.
- [35] Timur Friedman, Ramon Caceres, and Alan Clark. RTP extended reports (rtp xr). *IETF Internet Draft (work in progress)*, *draft-ietf-avt-rtcp-report-extns-05.txt*, April 2003.
- [36] P. Garg, R. Doshi, R. Greene, M. Baker, M. Malek, and X. Cheng. Using IEEE 802.11e MAC for QoS over Wireless. In *Proceedings of*

the 22nd IEEE International Performance Computing and Communications Conference (IPCCC 2003), Phoenix, Arizona, April 2003.

- [37] Gerald Q. Maguire Jr. Practical Voice Over IP (VoIP): SIP and related protocols. In <http://www.it.kth.se/courses/IK2554/VoIP-Coursepage-Fall-2009.html>, 2009.
- [38] Bernard Gold. Digital speech networks. *Proceedings of the IEEE*, 65(12):1636–1658, December 1977.
- [39] Charles Goodwin. *Conversational organization: Interaction between speakers and hearers*. Academic Press, New York, 1981.
- [40] Prabandham M. Gopal and Barath Kadaba. A simulation study of network delay for packetized voice. In *Proc. IEEE Global Telecommunications Conf. (GLOBECOM)*, December 1986.
- [41] Volodya Grancharov. *Human Perception in Speech Processing*. PhD thesis, Royal Institute of Technology (KTH), June 2006. TRITA-EE 2006:016.
- [42] Robert M. Gray. The 1974 origins of VoIP. *IEEE Signal Processing Magazine*, 22(4):87–90, July 2005.
- [43] John G. Gruber. Delay Related Issues in Integrated Voice and Data Networks - A Review and Some Experimental Work. In *6th Data Communications Symposium*, pages 166–180, Pacific Grove, California, November 1979.
- [44] John G. Gruber and Leo Strawczynski. Subjective Effects Of Variable Delay and Speech Clipping In Dynamically Managed Voice Systems. *IEEE Transactions on Communications*, COM-33(8):801–808, 1985.
- [45] Saikat Guha, Neil Daswani, and Ravi Jain. An Experimental Study of the Skype Peer-to-Peer VoIP System. In *IPTPS*, 2006.
- [46] Marie Guguin, Valrie Gautier-Turbin, Latitia Gros, Vincent Barriac, Rgine Le Bouquin-Jeann, and Gard Faucon. Study of the relationship between subjective conversational quality, and talking, listening and interaction qualities: Towards an objective model of the conversational quality. In *Proceedings of the Measurement of Speech and Audio Quality in Networks workshop (MESAQIN'05)*, Prague, Czech Republic, June 2005.
- [47] Olof Hagsand, Ian Marsh, and Kjell Hanson. Sicsophone: A Low-Delay Internet Telephony Tool. In *IEEE 29th Euromicro Conference*, pages 189–195, Belek, Turkey, September 2003.

- [48] Florian Hammer, Peter Reichl, and Alexander Raake. Elements of Interactivity in Telephone Conversations. In *8th International Conference on Spoken Language Processing (ICSLP/INTERSPEECH 2004)*, pages 1741–1744, Jeju Island, Korea, October 2004.
- [49] Florian Hammer, Peter Reichl, and Alexander Raake. The well-tempered conversation: Interactivity, delay and perceptual VoIP quality. In *IEEE ICC 2005*, Seoul, Korea, 2005.
- [50] Mark Handley and Van Jacobson. SDP: Session Description Protocol. *Request for Comments (Standards Track) RFC 2327*, Internet Engineering Task Force, April 1998.
- [51] Mark Handley, Henning Schulzrinne, Eve Schooler, and Jonathan Rosenberg. SIP: Session initiation protocol. *Request for Comments (Standards Track) RFC 1883*, Internet Engineering Task Force, March 1999.
- [52] Vicky Hardman, Angela Sasse, Mark Handley, and Anna Watson. Reliable Audio for Use over the Internet. In *Proceedings of INET'95*, Honolulu, Hawaii, June 1995.
- [53] Olivier Hersent, David Gurle, and Jean-Pierre Petit. *IP telephony*. Addison Wesley, Reading, Massachusetts, 2000.
- [54] Christian Hoene. *Internet Telephony over Wireless Links*. PhD thesis, Technical University of Berlin, Germany, September 2005.
- [55] Christian Hoene and Georg Carle. Umts networks and internet telephony. <http://net.informatik.uni-tuebingen.de/en/teaching/umts-voip/ss2007>.
- [56] Christian Hoene, B. Rathke, and Adam Wolisz. On the Importance of a VoIP Packet. In *Proceedings Of 1st ISCA Tutorial and Research Workshop On The Auditory Quality Of Systems*, Mont-Cenis, Germany, April 2003.
- [57] J. M. Holtzman. The interaction between queueing and voice quality in variable bit rate packet voice systems. In Minoru Akiyama, editor, *Eleventh International Teletraffic Congress*, pages 151–154, Kyoto, Japan, September 1985. Elsevier Science Publishers.
- [58] International Telecommunication Union. 7 kHz audio-coding within 64 kbit/s. *ITU-T Recommendation G.722*, November 1988.
- [59] International Telecommunication Union. Echo suppressors. *ITU-T Recommendation G.164*, November 1988.

- [60] International Telecommunication Union. Specification for an intermediate reference system. *ITU-T Recommendation P.48*, November 1988.
- [61] International Telecommunication Union. Annex B: A silence compression scheme for G.729 optimized for terminals conforming to recommendation V.70. *ITU-T Recommendation G.729 Annex B*, November 1996.
- [62] International Telecommunication Union. Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear-prediction (CS-ACELP). *ITU-T Recommendation G.729*, March 1996.
- [63] International Telecommunication Union. Modulated noise reference unit (MNRU). Technical Report ITU-T Recommendation P.810, Telecommunication Standardization Sector of ITU, 1996.
- [64] International Telecommunication Union. The E-model, a computational model for use in transmission planning. Recommendation G.107, Telecommunication Standardization Sector of ITU, Geneva, Switzerland, December 1998.
- [65] International Telecommunication Union. Pulse Code Modulation (PCM) of Voice Frequencies. *ITU-T Recommendation G.711*, November 1998.
- [66] International Telecommunication Union. Appendix I: A high quality low-complexity algorithm for packet loss concealment with G.711. *ITU-T Recommendation G.711, Appendix I*, September 1999.
- [67] International Telecommunication Union. Single ended method for objective speech quality assessment in narrow-band telephony applications. *ITU-T Recommendation P.563*, 2004.
- [68] IPTEL. SIP versus H.323. <http://www.iptel.org/info/trends/sip.html>.
- [69] ITU-T Recommendation H.323. Packet-based multimedia communications systems, July 2003.
- [70] Van Jacobson. Multimedia conferencing on the Internet. In *ACM Symposium on Communications Architectures and Protocols*, London, England, August 1994. Tutorial slides.
- [71] Van Jacobson and Steve McCanne. vat - LBNL Audio Conferencing Tool, July 1992. Available at <http://www-nrg.ee.lbl.gov/vat/>.
- [72] Nugehally S. Jayant and Susan W. Christensen. Effects of packet losses in waveform coded speech and improvements due to an odd-even

- sample-interpolation procedure. *IEEE Transactions on Communications*, 29(2):101–109, February 1981.
- [73] Wenyu Jiang and Henning Schulzrinne. Analysis of On-Off Patterns in VoIP and Their Effect on Voice Traffic Aggregation. In *9th IEEE International Conference on Computer Communication Networks*, Las Vegas, Nevada, October 2000.
- [74] Wenyu Jiang and Henning Schulzrinne. Modeling of Packet Loss and Delay and their Effect on Real-Time Multimedia Service Quality. In *Proceedings International Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV)*, June 2000.
- [75] Wenyu Jiang and Henning Schulzrinne. Comparison and Optimization of Packet Loss Repair Methods on VoIP Perceived Quality under Bursty Loss. In *Proceedings International Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV)*, Miami Beach, Florida, May 2002.
- [76] Wenyu Jiang and Henning Schulzrinne. Speech Recognition Performance as an Effective Perceived Quality Predictor. In *IWQoS*, Miami Beach, May 2002.
- [77] Jonathan Rosenberg and Henning Schulzrinne. SIP: Comparison of SIP and H.323. In *Proceedings of NOSSDAV*, Cambridge, UK, July 1998.
- [78] Jonathan Rosenberg and R. Mahy and P. Matthews and D. Wing. Session Traversal Utilities for (NAT) (STUN), July 2008.
- [79] Jonathan Rosenburg. Internet Telephony: A (Partial) Research Agenda, October 1997.
- [80] Lars-Erik Jonsson. RObust Header Compression (ROHC): The ROHC Architecture. *IETF Internet Draft, draft-jonsson-rohc-architecture-00.txt*, December 2002.
- [81] Takehiro Kawata, S. Shin, Andrea G. Forte, and Henning Schulzrinne. Using dynamic PCF to improve the capacity for VoIP traffic in IEEE 802.11 networks. In *IEEE WCNC*, March 2005.
- [82] Moo Young Kim and W. Bastiaan Kleijn. Rate-Distortion comparisons between FEC and MDC based on Gilbert channel model. In *Proc. IEEE Int. Conf. on Networks (ICON)*, pages 495–500, Sydney, 2003.
- [83] Nobuhiko Kitawaki, M. Honda, and K. Itoh. Speech-quality assessment methods for speech-coding systems. *IEEE Communications Magazine*, pages 26–32, October 1984.

- [84] Bastiaan Kleijn and Kuldip. K. Paliwal. *Speech Coding and Synthesis*. Amsterdam: Elsevier, 1995.
- [85] Vineet Kumar, Markku Korpi, and Senthil Sengodan. *IP Telephony with H.323: Architectures for Unified Networks and Integrated Services*. Wiley, March 2001.
- [86] H. H. Lee and C. K. Un. A study of on-off characteristics of conversational speech. *IEEE Transactions on Communication*, 34(6):630–637, June 1986.
- [87] Fengyi Li. Measurements of Voice over IP Quality. Master’s thesis, KTH, Royal Institute of Technology, Sweden, 2002.
- [88] Anders Lindgren, Andreas Almquist, and Olov Schelén. Quality of Service Schemes for IEEE 802.11 Wireless LANs - An Evaluation. In *the Journal on Special Topics in Mobile Networking and Applications (MONET) on Performance Evaluation of Qos Architectures in Mobile Networks*, 8:223–235, June 2003.
- [89] M. R. Schroeder and B. S. Atal. Code-excited linear prediction (CELP): high-quality speech at very low bit rates. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '85)*, pages 937–940, April 1985.
- [90] Andrew J. Mackie, Salah E. Aidarous, Samy A. Mahmoud, and J. Spruce Riordon. Design and performance evaluation of a packet voice system. *IEEE Transactions on Vehicular Technology*, VT-32(2):158–168, May 1983.
- [91] D. T. Magill. Adaptive speech compression for packet communication systems. In *IEEE National Telecommunications Conference*, pages 29D–1–29D–5, 1973.
- [92] Ian Marsh, Björn Grönvall, and Florian Hammer. The design and implementation of a quality-based handover trigger. In *Proceedings Of The 5th IFIP-TC6 Networking Conference*, Coimbra, Portugal, May 2005.
- [93] Piet Van Mieghem. A lower bound for the end-to-end delay in networks: Application to voice over IP. In *IEEE Globecom*, pages 2508–2513, Sydney, Australia, November 1998.
- [94] Warren A. Montgomery. Techniques for Packet Voice Synchronization. *IEEE Journal on Selected Areas in Communications*, SAC-1(6):1022–1028, December 1983.

- [95] Sue B. Moon, Jim Kurose, and Don Towsley. Packet Audio Playout Delay Adjustment Algorithms: Performance Bounds and Algorithms. Research report, Department of Computer Science, University of Massachusetts, Amherst, Massachusetts, August 1995.
- [96] Nanying Yin and Thomas E. Stern and Song Li. Performance Analysis of a Priority-Oriented Packet Voice System. In *IEEE Infocom*, pages 856–863, San Francisco, California, April 1987.
- [97] William Edward Naylor. A status report on the real-time speech transmission work at UCLA. NSC Note 52, December 1974.
- [98] William Edward Naylor. *Stream traffic communication in packet switched networks*. PhD thesis, UCLA, August 1977.
- [99] William Edward Naylor and Leneord Kleinrock. Stream traffic communication in packet switched networks: destination buffering considerations. *IEEE Transactions on Communications*, 30:2527–2534, 1982.
- [100] Anders Gunnar (n’ee Andersson). Capacity Study of Statistical Multiplexing for IP Telephony. Master’s thesis, Uppsala University, 2000.
- [101] Olivier Hersent and Jean-Pierre Petit and David Gurle. *Deploying Voice-over-IP Protocols*. Wiley, 2005.
- [102] Jorg Ott and E. Carrara. Extended RTP Profile for RTCP-based Feedback (RTP/AVPF). *IETF Request for comments 4585*, July 2006.
- [103] Packetizer. H.323 versus SIP: A Comparison. http://www.packetizer.com/iptel/h323_vs_sip.
- [104] Charlie E. Perkins, Orion Hodson, and Vicky J. Hardman. A Survey of Packet Loss Recovery Techniques for Streaming Audio. *IEEE Network*, 12(5):40–48, September 1998.
- [105] Mathew Podolsky, Cynthia Romer, and Steve McCanne. Simulation of FEC-Based Error Control for Packet Audio on the Internet. In *Proceedings of the Conference on Computer Communications (IEEE Infocom)*, volume 2, pages 505–512, San Francisco, California, March 1998.
- [106] Alexander Raake. *Speech Quality of VoIP: Assessment and Prediction*. John Wiley & Sons, 2006.
- [107] Peter Reichl and Florian Hammer. Hot discussion or frosty dialogue? Towards a temperature metric for conversational interactivity. In *8th International Conference on Spoken Language Processing (ICSLP/INTERSPEECH 2004)*, Jeju Island, Korea, October 2004.

- [108] Antony W. Rix. Comparison between subjective listening quality and P.862 PESQ score. In *Proc. Measurement of Speech and Audio Quality in Networks (MESAQIN'03)*, Prague, Czech Republic, May 2003.
- [109] Rodeo group. <http://www-sop.inria.fr/rodeo/fphone/>, 1999.
- [110] J. Rosenberg and H. Schulzrinne. An RTP payload format for generic forward error correction. *Request for Comments (Standards Track) RFC 2733, Internet Engineering Task Force*, December 1999.
- [111] Jonathan Rosenberg, R. Mahy, and Christian Huitema. Traversal Using Relay NAT (TURN).
- [112] Jonathan Rosenberg, Lili Qiu, and Henning Schulzrinne. Integrating Packet FEC into Adaptive Voice Playout Buffer Algorithms on the Internet. In *Proceedings of the Conference on Computer Communications (IEEE Infocom)*, pages 1705–1714, Tel Aviv, Israel, March 2000.
- [113] Jonathan Rosenberg, Henning Schulzrinne, Gonzalo Camarillo, A. Johnston, J. Peterson, R. Sparks, Mark Handley, and Eve Schooler. RFC 3261 - SIP: Session initiation protocol. Technical report, Internet Engineering Task Force, June 2002.
- [114] Jonathan Rosenberg, J. Weinberger, Christian Huitema, and R. Mahy. STUN – Simple Traversal of User Datagram Protocol (UDP) through Network Address Translators (NATs). Internet Engineering Task Force: RFC 3489, March 2003.
- [115] Benny Sällberg, Henrik Åkesson, Nils Westerlund, Mattias Dahl, and Ingvar Claesson. Analog circuit implementation for speech enhancement purposes. In *Proc. 38th Asilomar Conference on Signals, Systems, and Computers*, volume 2, pages 2285–9, CA, USA, 2004.
- [116] Benny Sällberg and Mattias Dahl. Speech enhancement implementations in the digital, analog, and hybrid domain. In *Swedish System-on-chip Conference*, Tammsvik, Stockholm, 2005.
- [117] Salman A. Baset and Henning Schulzrinne. An Analysis of the Skype Peer-to-Peer Internet Telephony Protocol. In *Proceedings of the IEEE Infocom Conference*, Barcelona, Spain, April 2006.
- [118] Henning Sanneck. Fehlerverschleierungsverfahren für Sprachübertragung mit Paketverlust. Master's thesis, Telecommunications Department, University of Erlangen-Nuremberg, Germany, June 1995.

- [119] Henning Schulzrinne. Voice Communication Across the Internet: A Network Voice Terminal. Technical Report TR 92-50, Dept. of Computer Science, University of Massachusetts, Amherst, Massachusetts, July 1992.
- [120] Henning Schulzrinne. *Reducing and characterizing packet loss for high-speed computer networks with real-time services*. PhD thesis, University of Massachusetts, Amherst, Massachusetts, May 1993.
- [121] Henning Schulzrinne et al. RTP: A transport protocol for real-time applications. *Request for Comments (Standards Track) RFC 3550*, Internet Engineering Task Force, July 2003.
- [122] Henning Schulzrinne, A. Rao, and R. Lanphier. Real time streaming protocol (RTSP). *Request for Comments (Standards Track) RFC 2326*, Internet Engineering Task Force, April 1998.
- [123] Scott Shenker. Fundamental design issues for the future Internet. *IEEE Journal on Selected Areas in Communications*, 13(7):1176–1188, September 1995.
- [124] Christian Sieckmeyer. Bewertung von adaptiven Ausspielalgorithmen für paketvermittelte Audiodaten Evaluation of adaptive playout algorithms for packet audio - in German. Studienarbeit, Dept. of Electrical Engineering, TU Berlin, Berlin, Germany, October 1995.
- [125] Henry Sinnreich and Alan B. Johnston. *Internet Communications Using SIP: Delivering VoIP and Multimedia Services with Session Initiation Protocol (Networking Council)*. Wiley, July 2006.
- [126] Skype Communications S.A. Skype Explained, 2007.
- [127] Kotikalapudi Sriram and Ward Whitt. Characterizing superposition arrival processes in packet multiplexers for voice and data. *IEEE Journal on Selected Areas in Communications*, SAC-4(6):833–846, September 1986.
- [128] Thomas E. Stern. A queueing analysis of packet voice. In *Proceedings of the IEEE Conference on Global Communications (GLOBECOM)*, pages 2.5.1–6, San Diego, California, November/December 1983.
- [129] Donald L Stone and S. Jeffay. An Empirical Study Of Delay Jitter Management Policies, 1995.
- [130] Tatsuya Suda, Yechiam Yemini, Hideo Miyahara, and Toshiharu Hasegawa. Performance Evaluation of a Packetized Voice System - A Simulation Study. In *IEEE ICC*, pages 749–753, Boston, Massachusetts, June 1983.

- [131] Kyoungwon Suh, Daniel R. Figueiredo, Jim Kurose, and Don Towsley. Characterizing and Detecting Skype-Relayed Traffic. In *IEEE INFOCOM 2006 - The Conference on Computer Communications*, 2006.
- [132] Krister Svanbro. Lower layer guidelines for robust RTP/UDP/IP header compression. *IETF Internet Draft (work in progress)*, *draft-ietf-rohc-rtp-lower-layer-guidelines-03.txt*, December 2001.
- [133] Thomas E. Tremain. The government standard linear predictive coding algorithm: LPC-10. *Speech Technology*, 1:40–49, April 1982.
- [134] International Telecommunication Union. Transmission Systems and Media, General Recommendation on the Transmission Quality for an Entire International Telephone Connection; One-Way Transmission Time. *G.114*, March 1993.
- [135] 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*, February 2001.
- [136] Martín Varela. *Pseudo-Subjective Quality Assessment of Multimedia Streams and its Applications in Control*. PhD thesis, University of Rennes, November 2005.
- [137] Martín Varela. Studying the Effects of FEC on Voice Traffic Using PSQA (extended abstract). In *IEEE INFOCOM 2005 Student Workshop*, Miami, USA, March 2005.
- [138] Dinesh C. Verma, Hui Zhang, and Domenico Ferrari. Delay jitter control for real-time communication in a packet switching network. Technical Report TR-91-007, University of California, Berkeley, CA, 1991.
- [139] M. A. Visser and Magda El Zarki. Voice and data transmission over an 802.11 wireless network. In *Proceedings of IEEE PIMRC'95*, pages 648–652, Toronto, Canada, September 1995.
- [140] Clifford J. Weinstein and James W. Forgie. Experience with Speech Communication in Packet Networks. *IEEE Journal on Selected Areas in Communications*, SAC-1(6):963–980, December 1983.

Appendix: Included articles

