Audio Mixers for Centralized VoIP Conferencing

submitted by

Zhihong Qian, B. Eng.

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

Master of Applied Science

Ottawa-Carleton Institute For Electrical and Computer Engineering

Faculty of Engineering

Department of Systems and Computer Engineering

Carleton University

Ottawa, Ontario, Canada K1S 5B6

May 2003

© Copyright
Zhihong Qian, 2003

National Library
of Canada

Bibliothèque nationale
du Canada

Acquisitions and
Bibliographic Services

Acquisitions et
services bibliographiques

395 Wellington Street
Ottawa ON K1A 0N4
Canada

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

ISBN: 0-612-93929-4

Our file Notre référence

ISBN: 0-612-93929-4

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

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

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

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

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this dissertation.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de ce manuscrit.

While these forms may be included in the document page count, their removal does not represent any loss of content from the dissertation.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

PAGINATION ERROR.

ERREUR DE PAGINATION.

TEXT COMPLETE.

LE TEXTE EST COMPLET.

Abstract

In centralized VoIP conferencing the audio mixer receives IP packets from all of the participants in the conference and combines them to generate the output signal. These packets may have different compression formats. At the present time, these packets are decompressed (decoded), mixed in the linear domain, then compressed back to their original formats (encoded), and sent to their destinations. Due to the packetization and compression, VoIP conferencing introduces unacceptable additional delays and distortion.

This thesis investigates the structure of the audio mixers used in centralized VoIP conferencing. It proposes simplified mixer structures reducing the delay and complexity while maintaining a good audio quality. The thesis addresses narrowband, wideband, and hybrid-bandwidth conferencing. Narrowband (G.729 and G.729A), wideband (G.722), and hybrid-bandwidth (G.711 and G.729(A) with G.722) audio mixers and transcoders are simplified by strong pitch domination, subband summation, high frequency suppression, and filter combination, respectively. Simulation shows the simplified mixers can maintain the audio quality, while reducing the complexity and delay.

Acknowledgements

First of all, I would like to express my sincere thanks to my advisor and co-supervisor, professor Rafik A. Goubran and Dr. Samy El-Hennawey, for all of their guidance, patience and encouragement for my research.

I greatly appreciate the financial support provided by Nortel Networks, National Capital Institute of Telecommunications (NCIT), and Carleton University during my graduate studies.

All the graduate students in the DSP lab are owed my thanks for their helpful discussions and friendship. Thanks also to Danny Lemay, Judy Bowman and other staff in SCE department for their assistance.

To my parents and my sister, I don't know how to express my thanks. Your love and encouragement support me all the time.

Finally, I want to present this thesis to my beloved husband, Xianjun. We know how hard the past two years have been for us, not only for the studies. Just want to thank you for your boundless love.

Contents

ABSTRACT	III
ACKNOWLEDGEMENTS	IV
INTRODUCTION	1
1.1 Overview	1
1.2 Problem Statement.....	2
1.3 Thesis Scope.....	3
1.4 Thesis Objectives and Contributions.....	4
1.5 Thesis Outline.....	6
LITERATURE REVIEW.....	8
2.1 VoIP Standards.....	9
2.1.1 Signalling Control	9
2.1.2 Gateway Control.....	13
2.1.3 Media Control.....	15
2.2 Mixing Options.....	16
2.3 Topology of VoIP Conferencing	18
2.3.1 Full Mesh Conferencing	18
2.3.2 Multicast Conferencing	20
2.3.3 Centralized Single-Speaker Conferencing.....	20

2.3.4	<i>Tandem Free Operation Conferencing (TFO)</i>	22
2.3.5	<i>Centralized Multi-Speaker Conferencing (CMS)</i>	22
2.3.6	<i>TFO vs CMS</i>	24
2.4	<i>Conversation Model</i>	26
2.5	<i>Speech Production System</i>	28
2.5.1	<i>Pitch</i>	30
2.5.2	<i>Formant</i>	32
2.5.3	<i>Linear Prediction Model</i>	33
2.6	<i>Narrowband vs Wideband</i>	34
2.7	<i>Factors Imposing on the QoS of VoIP Conferencing</i>	36
2.7.1	<i>Speech Coders</i>	36
2.7.2	<i>Delay</i>	39
2.7.3	<i>Packet Loss</i>	39
2.7.4	<i>Jitter</i>	40
2.8	<i>Several Involved Codecs</i>	40
2.8.1	<i>G.711</i>	40
2.8.2	<i>G.729</i>	41
2.8.3	<i>G.729A</i>	45
2.8.4	<i>G.722</i>	46
2.9	<i>Audio Quality Evaluation</i>	48
2.9.1	<i>Subjective Measurement Methods</i>	48
2.9.2	<i>Objective Measurement Methods</i>	49
<u>SIMULATION ENVIRONMENT</u>		52
3.1	<i>Speech Samples</i>	52
3.2	<i>Comparison Model</i>	54
3.3	<i>Other Issues</i>	56
<u>AUDIO MIXERS FOR NARROWBAND VOIP CONFERENCING.....</u>		57
4.1	<i>G.711 Mixer</i>	57
4.2	<i>Strong-Pitch-Domination Mixer</i>	58

4.2.1	<i>Algorithm Breakdown</i>	58
4.2.2	<i>Voiced vs Unvoiced</i>	59
4.2.3	<i>Simplified Mixer</i>	61
4.3	<i>Simulation and Results</i>	64
4.3.1	<i>Two-Speaker Conferencing</i>	65
4.3.2	<i>Three-Speaker Conferencing</i>	67
4.4	<i>Summary</i>	67
<u>AUDIO MIXERS FOR WIDEBAND VOIP CONFERENCING</u>		69
5.1	<i>Conventional G.722 Audio Mixer</i>	69
5.2	<i>Subband Summation Mixer</i>	70
5.3	<i>Simulation Results</i>	72
5.4	<i>Summary</i>	73
<u>AUDIO MIXERS FOR HYBRID-BANDWIDTH VOIP CONFERENCING</u>		75
6.1	<i>Conventional Transcoders for G.711 and G.722</i>	76
6.2	<i>Simplified G.722-to-G.711 Transcoder</i>	79
6.2.1	<i>High-Frequency-Suppression Transcoder</i>	79
6.2.2	<i>Simulation Results</i>	81
6.3	<i>Simplified Hybrid-Bandwidth-to-Narrowband Mixer</i>	84
6.3.1	<i>High-Frequency-Suppression Mixer</i>	84
6.3.2	<i>Simulation Results</i>	86
6.4	<i>Simplified G.711-to-G.722 Transcoder</i>	89
6.4.1	<i>Filter-Combination Transcoder</i>	89
6.4.2	<i>Simulation Results</i>	93
6.5	<i>Simplified Hybrid-Bandwidth-to-Wideband Mixer</i>	94
6.5.1	<i>Filter-Combination Mixer</i>	94
6.5.2	<i>Simulation Results</i>	96
6.6	<i>Summary</i>	99
<u>CONCLUSION</u>		100

7.1	Summary	100
7.2	Future Work	103
<u>REFERENCES.....</u>		<u>107</u>

List of Tables

Table 1-1 Audio Mixers and Transcoders Involved	5
Table 2-1 H.323: The Umbrella Specification.....	10
Table 2-2 Number of Operations in MCU and Terminals of Different Models.....	24
Table 2-3 Total Number of Operations of Different Models.....	25
Table 2-4 Worst-case Bandwidth of Different Models (kbps)	25
Table 2-5 Some Specifications of Several Codecs	38
Table 2-6 Comparison of Codec Parameters of G.729(A) and G.723.1	41
Table 2-7 MOS Rating.....	49
Table 4-1 Breakdown of the Codec Complexity (Worst Case) for G.729 and G.729A...	58
Table 4-2 Reduced Complexity of the Simplified G.729 Mixer in Terms of Operations	64
Table 4-3 The PESQ Scores of the Mixed Voices in the G.729(A) Mixer	66
Table 5-1 Simulation Results of the Conventional and Simplified G.722 Mixer.....	72
Table 6-1 Simulation Results of the G.722-to-G.711 Transcoders	83
Table 6-2 Simulation Results of the G.711 Input + G.722 Input -> G.711 Output Mixer	88
Table 6-3 Simulation Results of the G.711-to-G.722 Transcoders	93

Table 6-4 Simulation Results of the G.711 Input + G.722 Input ->G.722 Output Mixer	98
Table 7-1 Summaries of the Simplification Methods of Audio Mixers	101
Table 7-2 Summaries of the Simplification Methods of Audio Transcoders	101
Table 7-3 Features of G.722, G.722.1, and G.722.2.....	104
Table 7-4 Similarities and Differences of G.722.2 and G.729(A).....	104

List of Figures

Figure 2.2.1 Subtraction Mixer ($M = 3$)	17
Figure 2.2.2 Addition Mixer ($M = 3$).....	17
Figure 2.3.1 The Structure of Full Mesh Conferencing.....	19
Figure 2.3.2 The Structure of Multicast Conferencing.....	19
Figure 2.3.3 The Structure of Centralized Single-Speaker Conferencing	21
Figure 2.3.4 The Structure of Tandem Free Operation Conferencing.....	21
Figure 2.3.5 The Structure of Centralized Multi-Speaker Conferencing.....	23
Figure 2.5.1 “True” Discrete-Time Speech Production Model	29
Figure 2.5.2 Spectral View of Voiced Sound /a:/	31
Figure 2.5.3 Spectral View of Unvoiced Sound /t ʔ/.....	31
Figure 2.5.4 Estimated Speech Production Model Using Linear Prediction Analysis	33
Figure 2.6.1 Narrowband and Wideband of Voiced Sound /a:/	35
Figure 2.6.2 Narrowband and Wideband of Unvoiced Sound /t ʔ/	35
Figure 2.8.1 Code-Excited Linear Prediction Model.....	42
Figure 2.8.2 G.729 Encoder.....	44

Figure 2.8.3 G.729 Decoder.....	45
Figure 2.8.4 Block Diagram of G.722 Encoder.....	47
Figure 2.8.5 Block Diagram of G.722 Decoder.....	47
Figure 3.1.1 Waveforms of Three Input Channels and the Directly Mixed Signal	54
Figure 3.2.1 Comparison Mode	55
Figure 4.2.1 The Structure of the Strong-Pitch-Domination Mixer.....	62
Figure 4.2.2 Pitch Estimator	63
Figure 5.1.1 Block Diagram of the Conventional G.722 Mixer	70
Figure 5.2.1 Block Diagram of the Subband Summation G.722 Mixer	71
Figure 6.1.1 Frequency Response of the Lower Subband QMF.....	77
Figure 6.1.2 Frequency Response of the Higher Subband QMF	77
Figure 6.1.3 G.711-to-G.722 Transcoder Indicated by Rec. G.722.....	78
Figure 6.1.4 G.722-to-G.711 Transcoder Indicated by Rec. G.722.....	78
Figure 6.2.1 Frequency Response of the 16 kHz LPF	80
Figure 6.2.2 Frequency Response of the 8 kHz LPF	80
Figure 6.2.3 Block Diagram of the Simplified G.722-to-G.711 Transcoder	81
Figure 6.2.4 Spectral View of the Original Signal.....	82
Figure 6.2.5 Spectral View of the Fullband Processed Signal.....	82
Figure 6.2.6 Spectral View of the Lower Subband Processed Signal	82
Figure 6.3.1 Block Diagram of the Conventional G.711+G.722 -> G.711 Mixer	85
Figure 6.3.2 Block Diagram of the Simplified G.711+G.722-> G.711 Mixer	85

Figure 6.3.3 2D Reference, Simplified, Error Signal View Using PAMS Analysis (G.711+G.722 -> G.711 Mixer).....	87
Figure 6.3.4 Initial Spectrum View Using PAMS Analysis (G.711+G.722 -> G.711 Mixer)	87
Figure 6.4.1 Block Diagram of the Conventional G.711-to-G.722 Transcoder	89
Figure 6.4.2 Analysis of the Filter Combination (I)	90
Figure 6.4.3 Analysis of the Filter Combination (II).....	90
Figure 6.4.4 Analysis of the Filter Combination (III).....	90
Figure 6.4.5 Analysis of the Filter Combination (IV)	90
Figure 6.4.6 Frequency Response of the Combined Filter 0	92
Figure 6.4.7 Frequency Response of the Combined Filter 1	92
Figure 6.4.8 Block Diagram of the Simplified G.711-to-G.722 Transcoder	93
Figure 6.5.1 Block Diagram of the Conventional G.711+G.722 -> G.722 Mixer	94
Figure 6.5.2 Block Diagram of the Simplified G.711+G.722 -> G.722 Mixer	95
Figure 6.5.3 2D Reference, Simplified, Error Signal View Using PAMS Analysis (G.711+G.722 -> G.722 Mixer).....	97
Figure 6.5.4 Initial Spectrum View Using PAMS Analysis (G.711+G.722 -> G.722 Mixer)	97

List of Acronyms

3GPP	Third Generation Partnership Project
ACELP	Algebraic-Code-Excited Linear-Prediction
ACR	Absolute Category Rating
ADPCM	Adaptive Differential Pulse Code Modulation
AMR-WB	Adaptive Multi-rate Wideband Codec
AR	Autoregressive
BSD	Bark Spectral Distortion
CELP	Code Excited Linear Prediction
CNAME	Canonical Name
CNG	Comfortable Noise Generation
CMS	Centralized Multi-speaker
CS-ACELP	Conjugate-Structure Algebraic-Code-Excited Linear-Prediction
DSL	Digital Speech Level Analyzer
DSP	Digital Signal Processing
GK	Gatekeeper

GW	Gateway
HPF	Highpass Filter
HTTP	Hypertext Transfer Protocol
IETF	Internet Engineering Task Force
IP	Internet Protocol
ITU	International Telecommunication Union
ITU-T	ITU Telecommunication Standardization Sector
LAN	Local Area Network
LP	Linear Prediction
LPC	Linear Prediction Coding
LPF	Lowpass Filter
LSP	Line Spectral Pair
MC	Multipoint Controller
MCU	Multipoint Control Unit
MGCP	Media Gateway Control Protocol
MIPS	Millions of Instructions Per Second
MNB	Measuring Normalizing Blocks
MOPS	Millions of Operations Per Second
MOS	Mean Opinion Score
MP	Multipoint Processor
PAMS	Perceptual Analysis Measurement System

PCM	Pulse Code Modulation
PESQ	Perceptual Evaluation of Speech Quality
PSQM	Perceptual Speech Quality Measure
PSTN	Public Switched Telephone Network
QMF	Quadrature Mirror Filter
QoS	Quality of Service
RAM	Random Access Memory
ROM	Read Only Memory
RTCP	RTP Control Protocol
RTP	Real-Time Transport Protocol
SB-ADPCM	Sub-band Adaptive Differential Pulse Code Modulation
SIP	Session Initiation Protocol
SMTP	Simple Mail Transfer Protocol
SNR	Signal-to-Noise Ratio
TCP	Transport Control Protocol
TFO	Tandem Free Operation
UDP	User Datagram Protocol
VAD	Voice Activity Detection
VoIP	Voice over Internet Protocol
WMOPS	Weighted Millions of Operations Per Second

Chapter 1

Introduction

Voice transmission over packet networks using the Internet Protocol (VoIP) is gradually replacing conventional Public Switched Telephone Networks (PSTN) systems. In VoIP the speech is sampled, compressed, packetized, and transmitted over the network as IP packets. VoIP allows the service providers to integrate voice, data, and other multimedia traffic on the same digital network, offering potential flexibility and cost savings.

1.1 Overview

Voice transmission over IP networks is used in many applications. Some applications such as voice messaging and audio streaming can tolerate some delays, whereas other applications such as real-time telephony are very sensitive to delays [1][2].

This thesis deals with the latter class of applications. It focuses on the design of audio mixers suitable for centralized VoIP conferencing.

Traditional circuit-switched networks establish dedicated paths between the participants and the central conferencing unit. With the fixed bandwidth and short latency, they provide a satisfactory quality of service. However, the bandwidth is not used efficiently since the bandwidth is assigned whether the participant is talking or not.

In centralized VoIP conferencing the audio mixer receives IP packets from all of the participants in the conference. The active talkers are identified and their signals are combined to form the output signal.

1.2 Problem Statement

Conferencing in VoIP is a challenging problem. Voice packets originating from different participants are sent to a central conferencing unit to be combined. At the present time, these packets are decompressed (decoded), mixed in the linear domain, then compressed back to their original formats (encoded), and sent to their destinations. Due to the packetization and compression, VoIP conferencing introduces unacceptable additional delays and distortion. This is especially true when the different participants in the conference use different compression standards.

This thesis investigates the structure of the audio mixers used in centralized VoIP conferencing. Consider the decoding, audio mixing, and decoding functions described above. Some unnecessary steps can be skipped in order to reduce the complexity and delay without degrading the audio quality. For example the encoder may use some of the parameters already computed by the decoder, thus reducing the delay and complexity of the mixer.

In other situations, the various participants may be using different codecs, which requires transcoding and increases the complexity of the mixer. For example, some participants may be using narrowband codecs, while the others are using wideband codecs. The mixer decodes them respectively and encodes the mixed signal with some desired codec. The signal conversion of different bandwidths can be simplified by some filter transformation. In this thesis, such audio mixers and transcoders are investigated and simplified for hybrid-bandwidth VoIP Conferencing.

1.3 Thesis Scope

The compression algorithms of interest in this thesis are G.711, G.729(A) and G.722. G.711 and G.729(A) are currently applied in most VoIP platforms. G.711 is the mandatory algorithm required by H.323, and G.729A is the most popular algorithm nowadays due to its low delay (15 ms algorithmic delay) and complexity (about 10.5 millions of instructions per second on fixed-point Digital Signal Processor). On the other

hand, with the fast development and construction of packet-switched networks, the wideband application appears to dominate the future market. Recommendation G.722 is one of the wideband VoIP solutions for compressing. We also discuss the audio mixers and transcoders for inputs conforming to G.722 and G.711/G.729(A) in this thesis.

1.4 Thesis Objectives and Contributions

The objectives of the thesis are as follows:

- ✓ Simplifying and validating the G.729(A) audio mixers for narrowband VoIP conferencing.
- ✓ Simplifying and validating the G.722 audio mixers for wideband VoIP conferencing.
- ✓ Simplifying and validating the audio mixers and transcoders (G.711, G.729(A), and G.722) for hybrid-bandwidth VoIP conferencing.

Table 1-1 lists the audio mixers and transcoders involved. Those in red bold font are innovative approaches put forward by this thesis.

Table 1-1 Audio Mixers and Transcoders Involved

Algorithm	G.711	G.729(A)	G.722
G.711	M	M/T	M/T
G.729(A)	M/T	M	M/T
G.722	M/T	M/T	M
M: Mixer		T: Transcoder	
Note: Those in bold font are simplified and simulated in this thesis. Those in red are innovative approaches.			

The main contributions of the thesis involve three aspects.

The first simplification is for narrowband VoIP conferencing. The G.729(A) audio mixers are simplified by a scheme named Strong Pitch Domination, that is to say, replacing the open-loop pitch delay of the mixed signal by the pitch delay of the input signal with strongest pitch energy.

The second and the third are for hybrid-bandwidth VoIP conferencing. The audio transcoders from G.722 to G.711/G.729(A) are simplified by a scheme named High Frequency Suppression. In brief, it is not necessary to process the fullband signal conforming to G.722. Actually, the aliasing existing in the lower subband signal can be suppressed in order to obtain the desired narrowband signal. The audio mixers, whose inputs conform to G.711/G.729(A) and G.722 and output conforms to G.711/G.729(A), can be simplified by the same idea.

Last but not least, the audio transcoders from G.711/G.729(A) to G.722 are simplified by a scheme named Filter Combination. The audio mixers, whose inputs

conform to G.711/G.729(A) and G.722 and desired output format is G.722, can employ the same scheme to be simplified.

1.5 Thesis Outline

The rest of the thesis is organized as follows:

Chapter 2 gives some background information of the whole art of audio mixers. First, VoIP standards are discussed and the entities like multipoint control unit are defined. The different topologies of VoIP conferencing are discussed and compared. Secondly, the fundamental knowledge of the mixing options, conversation model, speech processing system, and narrowband and wideband are presented to give a better understanding of the whole topic and the methods to be utilized later. The factors imposing on the quality of service of VoIP conferencing, especially some particular codecs involved in this thesis, and how to evaluate the quality, are considered finally.

Chapter 3 states the simulation environment for the following chapters, such as the speech samples, the comparison model, and so on.

In chapter 4, the audio mixers for narrowband VoIP conferencing are investigated. Firstly, the G.711 mixer is discussed briefly. The focus is the G.729(A) mixer. Accordingly, the CS-ACELP algorithms are described. The idea of using pitch

substitution to simplify the mixer is elaborated next. Simulations of two-speaker and three-speaker cases are made to verify and assess the simplification.

The wideband case is presented in chapter 5. By looking into the G.722 algorithm, the subband summation mixer is investigated and verified to achieve competitive quality while reducing the complexity and delay.

Chapter 6 discusses the audio mixers and transcoders for hybrid-bandwidth VoIP conferencing. The aliasing existing in the lower and higher subbands of the signal conforming to G.722 prevents us from using the lower subband signal as the narrowband signal directly. The conventional transcoders between the narrowband signal and the 64 kbit/s G.722 signal are investigated. These prompt us to employ different schemes to simplify the transcoders and hence the audio mixers. First the lower subband of the G.722 signal is used only to obtain the narrowband signal by the idea of high frequency suppression. For the transcoding from narrowband to G.722 signal, some filter combinations are performed to generate the lower and higher subband respectively. The same idea to simplify the audio transcoders is deployed by the audio mixers and simulations show satisfactory results.

All the work is summarized in chapter 7. At the end of this chapter, the new wideband compression algorithm, G.722.2, is discussed and some heuristic methods to simplify the G.722.2 mixer and G.722.2 with G.729(A) mixer are sketched.

Chapter 2

Literature Review

This chapter gives background information relevant to the whole art of audio mixers. First, the VoIP standards are described, to provide an integral view of the VoIP systems, and define the multipoint control unit in H.323 and SIP, respectively. The five elementary topologies of VoIP conferencing are introduced consecutively. Centralized VoIP conferencing is compared with, and shown to surpass, the other conferencing topologies. However, the problems of centralized conferencing, that is, the tandem encoding, heavy complexity and extra delay, degrade the quality of service. How to mix the multiple channels and the conversation model are discussed to provide a better understanding of the audio mixer. The speech producing system is addressed to understand concepts such as formant, pitch, voiced and unvoiced etc. Furthermore, narrowband and wideband design are presented to give a better understanding of the whole topic and the methods to

be utilized later. The factors imposing on the quality of service of VoIP conferencing are discussed. Several codecs, like G.711, G.729(A), G.722 involved in this thesis are particularized. At the end of this chapter, the rating of audio quality and subjective and objective evaluation methods are introduced.

2.1 VoIP Standards

Standards are like traffic regulations. Traffic without any common rules has little chance to avoid accidents. It's also true for communication systems. The VoIP standards are discussed below, from signalling control, gateway control, and media control perspectives.

2.1.1 *Signalling Control*

From the moment a user dials a telephone number, signalling begins to function. Signalling protocols are usually rather complicated, to create and manage connections between terminals, especially when the VoIP users are on different networks. Interconnection between the packet networks and PSTN escalates the level of complexity further.

2.1.1.1 H.323

As the first protocol for real-time multimedia communication over IP, H.323 is the most mature and widely deployed solution for the market. The first version of H.323 was released by ITU-T in May 1996, designed for small or medium audio/video call or

conferencing. Its widespread adoption and deployment by the VoIP community drive the addition of more and more features to the pure VoIP system to support a large-scale distributed environment. In November 2000, version 4 of H.323 was ratified with high performance and rich features for carrier-class networks.

H.323 is an umbrella specification under which a set of general purpose and dedicated protocols work together harmoniously. It defines the system architecture and implementation guidelines covering the entire call set up, call control, and media used in the call.

Table 2-1 H.323: The Umbrella Specification

Audio/Video Applications		Terminal Control and Management				Data	
G.7x	H.26x	RTCP	Terminal to Gatekeeper Signalling RAS	H.225.0 Call Signalling	H.245	T.125	
						T.124	
RTP				Reliable Transport (TCP)		T.123	
Unreliable Transport (UDP)							
Network Layer (IP)							
Link Layer							
Physical Layer							

Table 2-1 shows the H.323 protocol architecture. H.323 itself covers all areas in white and it relies directly on the components in black and gray for application and transport layer service.

The entities defined in H.323 are Terminal, Gateway (GW), Gatekeeper (GK), and Multipoint Control Units (MCU) [3].

H.323 Terminals are the endpoints that can provide real-time two-way audio and optionally video and data communications capability with other H.323 entities.

Gateways are the endpoints that can provide the ability to interwork with other H-series terminals, GSTN or ISDN voice terminals, or GSTN or ISDN data terminals.

Gatekeepers are responsible for admission control and address translation services for H.323 terminals, GWs, and MCUs. They may also provide other services such as bandwidth management and locating gateways.

MCUs are the key devices for multipoint conferencing to permit multiple terminals and GWs to participate in a VoIP conference. The MCU consists of a mandatory Multipoint Controller (MC) and optional Multipoint Processors (MPs). The MC provides capability for negotiation, to coordinate all terminals for common levels of communications. In some situations, the MC may control conference resources. The MP is the component that performs centralized processing of the audio, video, or data streams. For a centralized audio conference, the MP is responsible for mixing, switching, and transcoding of the media streams. The audio mixer is one type of MCUs.

2.1.1.2 SIP

Session Initiation Protocol (SIP) is developed and maintained by the Internet Engineering Task Force (IETF). Unlike H.323, which is designed for telecommunications, SIP is Internet-oriented. It exploits the manageability of IP, and makes developing the IP

telephony relatively easier. In [4], the architecture of a VoIP system based on SIP and other IETF protocols is discussed.

SIP itself only defines creating, modifying, and terminating interactive sessions, while leaving all the other parts of the session to be covered by other protocols. The sessions include Internet telephone calls, multimedia distribution, and multimedia conferences [5]. SIP just provides primitives to be used by other IETF protocols to accomplish a multimedia session.

The entities defined in SIP are User Agents, Proxy Servers and Registrars.

User Agent is a logical entity that can act as User Agent Clients or User Agent Server. User Agent Clients can initiate a new SIP request, while User Agent Server can process a SIP request and generate a response.

Neither does SIP define any conferencing entities such as MCU in H.323, nor provide conferencing control service. However, the SIP user agents can model such a conferencing entity, and other conference control protocols can be used with SIP support[6].

Proxy Server is an intermediary entity that makes requests on behalf of other clients. Proxy servers are in charge of routing, proxy interpretation and so on.

Registrars are servers that process registration requests. They can update the location database with the contact information as specified in the request for the domain they handle.

SIP is a text-based protocol that is designed to work hand in hand with other core Internet protocols such as hypertext transfer protocol (HTTP) and simple mail transfer protocol (SMTP). Not as strictly defined as a complete system such as H.323, SIP leaves many aspects of its architecture open to interpretation. These features make SIP less complex, more scalable and flexible, compared with H.323. Meanwhile, H.323 surpasses SIP regarding its interoperability, network management, and call control.

As approaches of ITU-T and IETF respectively, H.323 and SIP both have strengths and weaknesses. However, they are much more than competitors. The two standard organizations are seeking the improvement and better interoperation of these two specifications. It may not take much time for the development of universal standards over the same network infrastructure, which offers a simple, manageable, flexible, and rich-featured VoIP system.

2.1.2 *Gateway Control*

As defined in H.323, gateways are the boundary devices, which operate to make the audio, video, or data streams in different formats understandable between and within PSTN, broadband, wireless networks and Internet. According to its functions, the

gateway can be decomposed into Media Gateway, Signalling Gateway, and Media Gateway Controller.

Typically, a Media Gateway provides conversion between the audio signals carried on telephone circuits and data packets carried over the packet networks. Trunking gateways, voice over ATM gateways, residential gateways, access gateways, business gateways, and network access servers can all be defined as gateways.

The Signalling Gateways function as the relays for signalling information between different networks.

The Media Gateway Controllers manage the call procedure and control the Media Gateway, using the gateway control protocols.

2.1.2.1 MGCP

Media Gateway Control Protocol (MGCP)[7] is the IETF approach for gateway control. It isolates the call control outside the Media Gateway and processes by external control components (Call Agents). The Call Agents are master devices, which handle call signalling, call processing, and send commands and responses to Media Gateways.

MGCP increases reliability and interoperability for the VoIP system, as well as reduces complexity. Further, the security is enhanced by processing the critical information at trusted servers.

2.1.2.2 MEGACO/H.248

MEGACO/H.248 is a collaborative effort of the ITU-T and IETF, following an agreement by both bodies to cooperate on a single unified protocol. All MGCP's functionality is contained in MEGACO. What's more, superior controls over analog telephone lines and the ability to transport multiple commands in a single packet are added into MEGACO.

2.1.3 *Media Control*

2.1.3.1 RTP

The Real-time Transport Protocol (RTP) defines end-to-end network transport services for real-time traffic support, such as audio, video transmission, and so on. The services include payload type identification, sequence numbering, timestamping, and delivery monitoring. The sequence number in RTP packets allows the received disordered packets to be reconstructed correctly.

RTP only defines how to carry data with real-time characteristics, without any guarantee of timely delivery or QoS (Quality of Service) itself [8]. RTP usually runs on top of User Datagram Protocol (UDP), relying on its multiplexing and checksum functions.

RFC 1890[9] helps to describe a profile for RTP and RTP Control Protocol (RTCP) for audio and video conferencing with minimal control. The formats of carrying audio

data are standardized in detail with regard to several audio codecs, for example, 1016, DVI4, G.721, G.722, G.728 etc.

2.1.3.2 RTCP

The RTP Control Protocol enhances RTP service in a manner scalable to a large conference. It periodically transmits control packets to all participants in the on-going session. RTCP augments the RTP session by providing feedback reports on the quality of data distribution, carrying a canonical name (CNAME) for an RTP source to keep track of each participant, controlling the RTCP packets sending rate, and optionally, conveying minimal session control information [8].

2.2 Mixing Options

In this section, the general mixing algorithms, subtraction mixing and addition mixing, are discussed. The number of mixing operations for both of them is calculated to be used later.

It's important to distinguish the speaker from the conferee. All the participants in the conference are conferees, and the speakers are referred to as those who talk at the investigated time. Assume the speaker selection scheme is M active speakers from N conferees ($M \leq N$). For each selected active speaker, the bridge only mixes the other $M-1$ speeches respectively. We have two choices to obtain the mixed signals.

One solution is first to mix the M active inputs together, and then for the M active channels, subtract the input with fixed gain from the mixed signal respectively. This is shown in Figure 2.2.1. The number of mixing operations is $(M - 1) + (M - 1) = 2(M - 1)$.

Figure 2.2.1 Subtraction Mixer ($M = 3$)

The other solution is more flexible. That is to add the active inputs for individual channels as Figure 2.2.2 shows. The number of mixing operations is $M \cdot (M - 2) + 1 = M^2 - 2M + 1$.

Figure 2.2.2 Addition Mixer ($M = 3$)

For common conferencing, the value of M usually is 2 or 3. If $M = 2$, the subtraction mixer gives the number of the mixing operation = 2, while the second mixer gives 1. If $M = 3$, both mixers give 4. So we prefer the second mixer to the first one.

2.3 Topology of VoIP Conferencing

Due to the connectionless nature of IP networks, VoIP conferencing system architectures are very flexible. According to the topology, we can simply classify the architecture into five elementary groups [10]: Full mesh, multicast, centralized single-speaker, tandem free operation, and centralized multi-speaker conferencing.

The last three topologies employ the Multipoint Control Unit in their framework. General MCU-based conferencing problems are addressed by D. Horn and A. Sharma in [11], regarding its architecture, implementation and so on.

2.3.1 *Full Mesh Conferencing*

Full mesh conferencing means that each participant is connected with each other by a simultaneous one-to-one connection, which is shown in Figure 2.3.1. Each pair has a common codec. The source speech is coded at the speaker side and sent to all other participants. Each terminal is equipped with an audio mixer to support the conference. It's suitable for small conferences with a few conferees, but not so easily scalable to accommodate a large number of participants.

Figure 2.3.1 The Structure of Full Mesh Conferencing

Figure 2.3.2 The Structure of Multicast Conferencing

2.3.2 *Multicast Conferencing*

Multicast conferencing topology is described in [2][12]. It needs the support of network-layer multicast. In multicast conferencing, each participant encodes its own voice and transmits it to the conference multicast address and receives $N-1$ streams in return. All the terminals must share a common codec. Similar to full mesh conferencing, the terminals have to mix the voices by themselves. The topology of multicast conferencing is shown in Figure 2.3.2.

2.3.3 *Centralized Single-Speaker Conferencing*

This is a simple and direct realization of the audio mixer. At the MCU, a primary speaker from N input streams is selected and sent to all the other participants. This makes the conference like a series of monologues.

The tandem encodings can be avoided by this means but it also severely limits the interaction of the participants. J. Forgie indicated in [13] that multi-talk accounts for 6~11% of the total conference time. Further, most conferencing control is handled by interruptions and overlapping speeches; it's inappropriate to permit only one speaker at any time [14]. Figure 2.3.3 shows the basic structure of centralized single-speaker conferencing.

Figure 2.3.3 The Structure of Centralized Single-Speaker Conferencing

Figure 2.3.4 The Structure of Tandem Free Operation Conferencing

2.3.4 *Tandem Free Operation Conferencing (TFO)*

The tandem free operation conferencing architecture is put forward and analyzed by R. Rabipour and P. J. Smith etc. [10] [15] and shown in Figure 2.3.4. A similar structure was presented by T.G. Champion in [14]. The speaker selection is centrally controlled by the bridge, and the other functions like decoding and mixing are accomplished by each terminal. Each source terminal sends the signal power together with the compressed voice in the RTP packets. Then the bridge can choose the active speakers without decoding. The selected active signals are forwarded to all the terminals to be mixed by each individually.

2.3.5 *Centralized Multi-Speaker Conferencing (CMS)*

The conference bridge in centralized multi-speaker conferencing has to make each participant hear the sum of all other talkers in the conference. So the speaker selection, decoding and mixing are performed by the bridge. The topology of centralized multi-speaker conferencing is shown in Figure 2.3.5. The audio mixer has to decode the inputs of all talkers and select the active speeches first. The selected inputs are mixed linearly. The aggregate signal is re-encoded and sent to all the participants respectively. (For each active speaker, the aggregate speech doesn't include his/her own voice.) The mixer is equipped with all the codecs required by the participants and accomplishes mixing or necessary transcoding in the desired format. The terminals can be blind, without caring about the requirements of the others. This feature makes the centralized multi-speaker

topology highly manageable and scalable. However, the speech quality is degraded by the tandem encodings and extra processing delay at the mixer. Furthermore, the VoIP vocoders always have heavy computational demands, so the tandem vocoders in the mixer will double the complexity. All these problems are taken into consideration with simplification of the audio mixers.

Figure 2.3.5 The Structure of Centralized Multi-Speaker Conferencing

Besides the five elementary topologies mentioned above, there are some other topologies, such as logical ring topology in [16], AD Hoc, or a mixture of some of the elementary topologies.

The problems of full mesh conferencing, multicast conferencing, and centralized single-speaker conferencing are very obvious. Concerning the tandem free operation and centralized multi-speaker conferencing, which one is better for VoIP conferencing

carriers and customers? That is to say, which one can offer us a highly manageable framework, cost-effective equipment, competitive quality, efficient bandwidth utilization, and convenient value-added services? There must be a tradeoff. It will be further discussed in the next section.

2.3.6 TFO vs CMS

Table 2-2 gives the number of operations of the tandem free operation bridge (TFO) and centralized multi-speaker bridge (CMS). The operations are encoding, decoding, and mixing.

Table 2-2 Number of Operations in MCU and Terminals of Different Models

Model	MCU			Terminal ($\times N$)		
	Encoding	Decoding	Mixing	Encoding	Decoding	Mixing
TFO	-	-	-	1	M	M-1
CMS	M+1	N	$M^2 - 2M + 1$	1	1	-

For the TFO, the MCU selects the M active speakers without decoding, and forwards them to the desired terminals. For the N terminals, each has one encoder for its own speech, and M decoders for the returned speeches, so the number of mixing operations is (M-1) at each terminal.

For the CMS, each terminal encodes its speech and decodes the returned mixed signal. They need not be concerned about the mixing issues. The MCU has to decode the N channels respectively, mix them, and encode the mixed signals for the M active

speakers and the other listeners. Thus, the number of mixing operations at MCU is $M^2 - 2M + 1$, as obtained in section 2.2.

The total numbers of encoding, decoding, and mixing of TFO and CMS are shown in Table 2-3.

Table 2-3 Total Number of Operations of Different Models

Model	TFO	CMS
Encoding	N	$(M+1)+N$
Decoding	MN	$N+N$
Mixing	$(M-1)N$	$M^2 - 2M + 1$

Table 2-4 shows the worst-case bandwidth requirement (in kbps). Assume that the bandwidth of a traditional voice packet is B , and the extra information about the signal power for the TFO requires another bandwidth of ϵ . Thus, the bandwidth of the new packet is $B + \epsilon$.

Table 2-4 Worst-case Bandwidth of Different Models (kbps)

Model	MCU		Terminal ($\times N$)	
	In	Out	In	Out
TFO	$N(B + \epsilon)$	$M(N-1)B$	MB	$B + \epsilon$
CB	NB	NB	B	B

We find that the TFO model can avoid tandem vocoders, thus, it reduces the delay and improves the overall audio quality, however, it distributes the complexity to all terminals, and an extra bandwidth is required. If the conference participants choose different codecs, then each terminal must support all these codecs. It's apparent that this

increases the cost for the terminals. All these considerations determine that Centralized VoIP conferencing topology is preferable as a carrier class solution.

2.4 Conversation Model

For a conference with N participants, the conferencing state can be silence, single speaker, two speakers, ..., or N speakers. As we mentioned in section 2.3.3, it's not appropriate to permit only one speaker at a time. Then why set 2 or 3 simultaneous speakers in most conferencing designs? What's the distribution time for each state?

To analyze conversation, we need to classify the total session into different states, such as silence, single talk, pause, interruption, overlapped talk, and so on. Voice Activity Detection (VAD) technology is used to accomplish the classification. It was found that the intervals of these states approximately yield exponential distribution. Brady's six-state model well approximates the two-way conversation behaviors [17]. ITU-T P.59 [18] adopts a four-state model based on simplification of Brady's model, which can be used to generate artificial two-way conversations with single-talk, double-talk, and mutual silence accounting for about 70.7%, 11.2%, and 18.1% of the total conversation time. It should be pointed out that the generally accepted length of double talk is about 6.58%, so the P.59 is not suitable for modelling double talk.

Although we can refer to the interval distributions of these models, they are, however, only valid for two-way conversation, not for multi-speaker conferencing.

Conferences with different objectives can have different models. Let's say, some conferences may ask all participants to play along with each other to solve problems, while some conferences are held to announce some decisions to all, and the interactions of such conferencing must be very limited.

An objective representation of conference dynamics is presented in [19] by research on several conferencing models for Problem Solving task and Role Play task, respectively. It's indicated that for the multi-speaker conferencing, setting $M=2$ improves the subjective rating, and setting $M=3$ can further ameliorate and achieve performance close to that of more than 3 speakers conferencing. An interesting scenario is that the participants always intend to be heard, instead of adapting to the provided system conditions. Thus the speech dynamics are not affected by conferencing modes and size.

Another factor that needs to be taken into account is the auditory perception. Human hearing has such features as simultaneous masking, and temporal masking [20]. As we can notice in our life, conversations tend to be dominated by those with higher volume.

In general, the multi-speaker interval accounts for 5~11% of the total conferencing time. Forgie classifies it into five categories [13]:

Reinforcement: Feedback to the primary speaker, usually short exclamations like "oh", "yes", and "really", or non-speech sound like harrumph, and chuckles.

Collision: Several participants try to talk at the same time. Generally speaking, when realizing a collision, one or more speakers stop talking and try after others finish.

Overlap: New speaker judges that the previous speaker is finishing from his content and intonation of the utterance, so starts talking right before the previous finishes.

Interruption: New speaker attempts to cut-off the primary speaker.

Noise: Such as coughs, door closing, and so on, which are not supposed to be heard by others.

2.5 Speech Production System

Speech production is such a complex system that it cannot be fully elaborated in several pages. Hereby, a brief introduction is given to educe some basic and important concepts involved in this thesis.

The gross human speech production system consists of lungs, trachea, larynx, pharyngeal cavity (throat), buccal cavity (mouth), and nasal cavity (nose). What we call the vocal tract refers to the throat and mouth, which starts from the output of the larynx, and ends at the input to the lips. The average length of the vocal tract for an adult male is about 17 cm, while for an adult female is about 14 cm.

The articulators, that is to say the vocal cords, soft palate, tongue, teeth, and lips, can produce various sounds by moving to different positions. They are critical to speech production. Vowels, the most prominent and central sound of a syllable, are created by the relatively free passage of airflow along the vocal tracts, for example, /i/, /u/, /e/. Consonants, however, are due to a partial or complete obstruction of the airflow, so usually they sound noisier than vowels. The examples are /v/, /z/, /b/, /p/, /f/, /h/, /m/ and so on.

Figure 2.5.1 “True” Discrete-Time Speech Production Model

The widely accepted Discrete-Time Speech Production Model is shown in Figure 2.5.1 [21]. The vocal tract is usually modelled as a filter $H(z)$, and the excitations to such filter are quasi-periodic pulses of air (voiced) or white noise (unvoiced). The air passed through the vocal tract filter is further processed by $R(z)$ considering the lip radiation effect. The model is further interpreted in the following sections, combining with several important concepts like pitch, formant.

2.5.1 *Pitch*

Pitch is the most important characteristic of voiced sound. As known, voiced sound is generated by forcing the airflow through the glottis, the opening between the vocal cords at the upper part of the larynx. The flow of air from the lungs to the vocal tract is interrupted with the vibrational opening and closing of the vocal cords, which produces quasi-periodic pulses of air as the excitation to the vocal tract. The rate of the opening and closing gives the pitch of the sound, which can be adjusted by varying the shape of, and the tension in, the vocal cords, and the pressure of the air behind them. For example, /a/, /e/, /i/, /u/, /v/, /d/, /z/, /t/, /l/, /b/, /d/, /g/ are all voiced sound.

In the spectral view, the voiced sound always has greater signal energy, pitch, and a spectrum biased towards the lower frequencies. Figure 2.5.2 shows the magnitude spectrum of voiced sound /a:/.

In order to generate unvoiced sound, a constriction is formed at some point along the vocal tract. Then the airflow is forced through the constriction to produce turbulence. /f/, /t/, /s/, /p/, /k/ are some examples of unvoiced sound.

The characteristics of unvoiced speech are relatively low signal energy, no pitch, and a frequency spectrum biased towards the higher frequencies of the audio band. Figure 2.5.3 is the spectral view of unvoiced sound /t ʃ /.

Figure 2.5.2 Spectral View of Voiced Sound /a:/

Figure 2.5.3 Spectral View of Unvoiced Sound /t f /

Besides the two elementary excitation types, voiced and unvoiced, there are other cases. By combining the voiced and unvoiced with silence, we can further obtain mixed, plosive, etc. [21][22].

Pitch can be used in many applications such as voiced/unvoiced detection. In chapter 4, for several channels to be mixed, the one with strongest pitch energy is selected to be the pitch of the summation signal by considering it dominates the excitation to the vocal tract.

Pitch and fundamental frequency are often used interchangeably. However, the pitch refers to the perceived fundamental frequency, not necessarily existing in the waveform. For example, G.729's pitch delay ranges from 19 to 143, and corresponds to 55.94 Hz ~ 421 Hz. But as is known, the G.729 encoder includes a highpass filtering with cut-off frequency of 140 Hz, so the fundamental frequency cannot be less than 140 Hz. If a signal has a pitch of 60 Hz, the corresponding fundamental frequency is 180 Hz (the least multiple of 60 Hz which is greater than 140 Hz).

2.5.2 *Formant*

If we have a look at any sound spectrum, there are some regions of relatively great intensity. These resonance-like frequency regions are named formants, indicating that they tend to form the overall spectrum.

The movement of articulators comparts the vocal tract cavities into several small acoustical cavities and subcavities. Such concatenation of acoustic organ pipes can be regarded as a filter, amplifying some frequencies, while attenuating some others. The shape and dimensions of the vocal tract determine the location of formants, in other words, the spectral envelope. Conversely, we can characterize the state of the vocal tract by formants.

Strictly speaking, there are infinite formants in human speech. However, human perception is sensitive to the first three formants, so usually we only take the first 3~5 into consideration.

2.5.3 *Linear Prediction Model*

Linear Prediction (LP) on speech analysis came up at the beginning of the 1970s, and now it is one of the most important techniques for speech processing.

Figure 2.5.4 Estimated Speech Production Model Using Linear Prediction Analysis

Figure 2.5.4 [21] shows the estimated speech production models used in Linear Prediction. In the LP model, a time-variable autoregressive (AR) filter, or all-pole filter, is applied to simulate the vocal tract. As is known, most phoneme classes, especially vowels, involve an acoustical resonant vocal tract, which can be represented by an all-pole filter. However, the all-pole model preserves all the acoustic information in the speech, but not the temporal relationships. Since the phase information in the speech has no effect on speech perception, it's not a problem for such applications as coding and re-

synthesis [21]. The estimated coefficients of the all-pole synthesis filter can be obtained by short-term analysis of the speech.

The excitation signal is different for voiced and unvoiced sounds. As indicated in section 2.5.1, for voiced sound, quasi-periodic pulses of air are produced as the excitation to the vocal tract. For voiced sound, the excitation signal can be obtained by examining the long-term correlation of the speech. For unvoiced sound, we use white noise as the excitation signal. In most Code Excited Linear Prediction (CELP) models, several discrete pulses are used instead.

2.6 Narrowband vs Wideband

Traditionally, narrowband speech is used in most speech transmission and processing systems. The bandwidth of speech is limited to 200 ~ 3400 Hz, and sampled at 8 kHz. This limitation no doubt imposes on the quality of speech. If we take the limitation of the organs for human speech production and auditory system into consideration, it's usually suitable to limit the speech to a bandwidth of 7 ~ 8 kHz [21].

To illustrate, the magnitude spectrums of voiced sound /a:/ and unvoiced sound /t f / are shown in Figure 2.6.1 and Figure 2.6.2 respectively.

Figure 2.6.1 Narrowband and Wideband of Voiced Sound /a:/

Figure 2.6.2 Narrowband and Wideband of Unvoiced Sound /t f /

From the figures, we can notice not only the unvoiced sounds, but also voiced sound, extend their energy to high frequency. So if we extend the bandwidth of speech into a wider scope, the communication quality can be definitely improved.

The wideband speech processing is bandlimited to 50 ~ 7000 Hz, and sampled at 16 kHz. From the perceptual point of view, the extended 50 ~ 200 Hz can enhance the sensation of naturalness, presence and comfort, and the extended 3400 ~ 7000 Hz enhances fricative differentiation and higher intelligibility [23].

More and more applications are requiring high-quality audio service. With the wideband technology introduced to VoIP conferencing, the high-quality audio option can lead the conferees to a virtual face-to-face conversation experience.

2.7 Factors Imposing on the QoS of VoIP Conferencing

Transferring audio signals through packet-based networks introduces challenges that don't exist in circuit-switched networks. The first factor imposing on the quality of service of VoIP conferencing is the compression algorithms of the speech. While reducing the bit rate of the audio stream using some codec, quality degradation is also experienced in some degree. Section 2.7.1 deals with some general problems of speech codecs, while some codecs involved in this thesis will be particularized in 2.8. Further, since the communication path is no longer a dedicated one with a constant transmission delay any more, the large delay variation that may be present in networks with limited or no QoS guarantees can lead to severe problems such as packet loss, jitter, and end-to-end delay.

2.7.1 *Speech Coders*

The most attractive feature of VoIP is that it's not necessary to occupy the 64 kbit/s channel all the time for an ongoing session. First, the Voice Activity Detector can differentiate voice from noise. It's natural that we can use a very low bit rate to send the background noise, or we even needn't send the noise at all, just inform the receiver side

and activate the Comfortable Noise Generation (CNG) function. Second, for the speech, we can employ speech coders (codecs) to compress the digital stream at a lower bit rate without significantly degrading the audio quality.

A low bit rate codec usually has higher complexity, longer delay, or poorer quality. So it doesn't mean the codec with lowest bit rate is the best. Then how to evaluate a codec? The attributes that should be taken into account are bit rate, complexity, delay, bandwidth and quality [24]. These attributes are not independent from each other. So choosing a good codec means a tradeoff.

If the participants are talking in a noisy environment, it's hard for the VAD to declare the noise and speech. The VAD may have a front-end clipping problem, if it fails to judge the onset of the speech, and this can seriously impair the interaction of the conference.

Codecs are typical DSP (Digital Signal Processing) algorithms, and usually they are implemented on DSP chips. The complexity of a codec means how many resources are needed to carry out the algorithm. There are many standards to evaluate the complexity of codecs, such as MIPS (Millions of Instructions Per Second), MOPS (Millions of Operations Per Second), RAM (Random Access Memory) usage, ROM (Read Only Memory) usage and so on. Currently, the codecs requiring less than 15 MIPS are

considered as low-complexity, while those requiring more than 30 MIPS are deemed as high-complexity [24].

If we exclude the transmission delay, the delay introduced by the codec includes algorithmic delay and processing delay. Most codecs process speech data in frame format, so there is a buffer to collect frames. Further, to achieve better analysis, some codecs may need extra information from a future data stream (lookahead). The buffer delay and lookahead compose the algorithmic delay. The processing delay is easy to understand. The codecs consist of many filters. Transmission of the data stream through all the filters introduces the processing delay.

A candidate codec is expected to provide a competitive quality. In section 2.9, how to evaluate the audio quality will be particularized. Table 2-5 gives some specifications of several codecs [25][26].

Table 2-5 Some Specifications of Several Codecs

Compression Method	Bit Rate (kbps)	MOS Score	Algorithm Delay (ms)
G.711 PCM	64	4.3	0.125
G.726 ADPCM	32	4.1	1
G.728 LD-CELP	16	4.0	3 to 5
G.722 SB-ADPCM	48, 56, 64	4.5 (64 kbit/s)	1.625
G.729 CS-ACELP	8	3.92	15
G.729 x 2 Encodings	8	3.27	-
G.729 x 3 Encodings	8	2.68	-
G.729A CS-ACELP	8	3.76	15
G.729A x 2 Encodings	8	3.13	-
G.729A x 3 Encodings	8	2.51	-
G.723.1 MP-MLQ	6.3	3.9	37.5
G.723.1 ACELP	5.3	3.65	37.5

We can see that the single G.729 vocoder has a Mean Opinion Score (MOS) of 3.92, while the tandem of two G.729 vocoders has a MOS of 3.27, and the tandem of three G.729 vocoders has a MOS of 2.68, which is highly degraded [27]. The speech quality can be further reduced when multi-speech speech is involved.

2.7.2 *Delay*

As mentioned before, the packet-based networks are not designed for real-time applications. The links are not dedicated to each session but shared with other applications instead. Network behaviors, such as queuing, or multitasking, can introduce unpredictable delay during the transmission. The transmission delay, together with the codec delay, incurs sudden stops and starts in the voice reproduction. Significant delays can render such time-sensitive applications as audio conferencing into a mess.

2.7.3 *Packet Loss*

If the congestion of the networks is severe, and the buffered data exceed the queue capacity, the subsequent packets are dropped, so the listeners suffer from packet loss.

Many technologies have been developed to recover the lost packets, for example, waveform substitution and packet repetition [28]. However, all of them introduce more delay and complexity to the system. For overloaded networks, the quality of voice may be poor even with packet recovery support.

2.7.4 Jitter

Since packet-based networks don't guarantee the delivery time of data packets (or their order), the transmission delay over the networks is indeterminate and variable. At the receiver side, the audio packets arrive at very inconsistent rates. The variation in inter-packet arrival rate is referred to as jitter. The delay jitter with different traffic loads and its effect on the speech quality are examined in [29].

In order to remove jitter, a jitter buffer is used to collect packets and store them long enough to permit the slowest packets to arrive in time to be played in the correct sequence. Apparently, jitter buffers increase the overall delay.

2.8 Several Involved Codecs

2.8.1 G.711

G.711 is the international standard for encoding telephone audio and the mandatory codec of a VoIP session. In North America and Japan, G.711 is known as u-law, and in Europe it is known as A-law. G.711 u-law and A-law compress 14-bit linear Pulse Code Modulation (PCM) samples and 13-bit linear PCM samples, respectively, into 8-bit logarithmic samples, producing a bit rate of 64 kbps [30]. G.711 is a high-quality codec (MOS 4.3) with low delay and complexity.

2.8.2 G.729

At the request of applications like personal communication systems, digital satellite systems, packetized speech transmission and circuit multiplexing equipment, Study Group 15 of the ITU-T has been involved in a standardization process for a speech codec at 8 kbit/s since 1990. The desired codec should be able to produce equivalent speech quality to G.726 under most operating conditions[31][32][33]. In November 1995, G.729[34] was accepted for ratification, as a joint effort of France Telecom CNET/University of Sherbrooke, NTT, and AT&T.

Compared with G.723.1[35], another VoIP compression algorithm, G.729 has shorter delay, but higher complexity. G.729A coming up, which reduces the complexity of G.729 while keeping almost the same quality, surpasses G.723.1 to be the preferred choice for VoIP compressing. Table 2-6 gives some codec parameters of G.729(A) and G.723.1.

Table 2-6 Comparison of Codec Parameters of G.729(A) and G.723.1

Parameter	G.729	G.729A	G.723.1 MP-MLQ	G.723.1 ACELP
Bit rate (kb/s)	8	8	6.3	5.3
Frame size (ms)	10	10	30	30
Subframe (ms)	5	5	7.5	7.5
Algorithmic delay (ms)	15	15	37.5	37.5
MIPS (fixed-point DSP)	20	10.5	14.6	16
RAM (16-bit Words)	2.7 k	2 k	2.2 k	2.2 k

The conjugate-structure algebraic-code-excited linear-predictive (CS-ACELP) coding concept is adopted in G.729(A). The excitation to the all-pole synthesis filter is

constructed by adding the adaptive and fixed-codebook vectors scaled by their respective gains. The speech is reconstructed by filtering the excitation through the LP synthesis filter, as shown in Figure 2.8.1. Compared with other parametric vocoders, CELP-type vocoders have some capability of representing multiple speakers[14].

Figure 2.8.1 Code-Excited Linear Prediction Model

The G.729 coder operates on a speech frame of 10 ms corresponding to 80 samples at a sampling rate of 8000 samples per second. For each frame, it extracts the parameters of the CELP model - line spectrum pairs (LSP, for the short-term correlations in the speech signal, determine the spectral envelope), adaptive-codebook parameters (pitch parameters, for the long-term correlations in the speech signal, determine the spectral fine structure), and the fixed-codebook parameters (to refine the excitation signal) [24]. These parameters are encoded and transmitted. The signal flow of the G.729 encoder is shown in Figure 2.8.2. Some functions are processed once per frame (10 ms), like the pre-processing by a highpass filter (cut-off at 140 Hz), the LP analysis, and the open-loop pitch search. The open-loop pitch search analyzes the perceptual weighted signal to find a

rough open-loop pitch delay T_{op} . The consequent closed-loop pitch search seeks to find the optimal pitch delay for the current subframe (5 ms), so it is rather complex. By limiting the search range around T_{op} , the complexity of the closed-loop search is greatly reduced. As well as the closed-loop pitch analysis, the algebraic codebook search, adaptive-codebook gain (pitch gain) and algebraic-codebook gain quantization, and memory update are all performed once per subframe.

Figure 2.8.2 G.729 Encoder

At the decoder, all the parameters are used to retrieve the excitation and synthesis filter parameters, and the speech is reconstructed by filtering this excitation through the short-term synthesis filter, and further enhanced by a postfilter [34]. The signal flow of the G.729 Decoder is shown in Figure 2.8.3.

Figure 2.8.3 G.729 Decoder

2.8.3 G.729A

G.729A is a reduced complexity version of G.729, and it's interoperable with G.729. It slightly degrades the speech quality, but achieves a 50% drop in complexity.

The major algorithm changes of G.729A include [36]: Perceptual Weighting Filter, open-loop pitch delay (using decimation for the correlation calculation), computation of the impulse response of the weighted synthesis filter, target signal, and updating the filter states, adaptive codebook search, fixed codebook search and harmonic postfilter at the decoder.

The single G.729A vocoder gives a MOS of 3.76, and the tandem of two G.729A vocoders gives 3.13, while the tandem of three gives 2.51 [25], which are a little worse than G.729's corresponding performances, which are 3.92, 3.27, and 2.68 respectively.

2.8.4 G.722

G.722 is the first wideband audio codec of ITU-T, released in 1988. It uses the sub-band adaptive differential pulse code modulation (SB-ADPCM) within a bit rate of 64 kbit/s. The frequency band is split into higher and lower subbands and each subband signal is encoded using ADPCM technology. To support an auxiliary data channel, the system has three modes of operation corresponding to the bit rates used for 7 kHz audio coding: 64, 56, and 48 kbit/s. The latter two allow an auxiliary data channel of 8 and 16 kbit/s respectively by reducing the bit rate of the lower subband.

The structure of the G.722 encoder is shown in Figure 2.8.4 [37]. The wideband signal sampled at 16 kHz is split into the lower subband (0 ~ 4000 Hz) and higher subband (4000 ~ 8000 Hz) by the Quadrature Mirror Filter (QMF). The two subbands are

quantized using Adaptive Differential Pulse Code Modulation (ADPCM) respectively. Since the lower subband is perceptually rather more important than the higher subband, more bits (6 over 8 bits) are given to the lower subband. The 48 kbit/s lower subband stream and 16 kbit/s higher subband stream are combined into a composite 64 kbit/s signal.

Figure 2.8.4 Block Diagram of G.722 Encoder

Figure 2.8.5 Block Diagram of G.722 Decoder

The G.722 decoder is shown in Figure 2.8.5 [37]. Symmetric to the G.722 encoder, the received 64 kbit/s ADPCM stream is separated into lower and higher subbands and decoded respectively. The decoded lower and higher subband signals are processed by the receive QMF to obtain the original wideband signal.

G.722 can achieve good performance with its three modes. In ITU-T Recommendation G.722, the speech quality of the 64 kbit/s (7 kHz) audio coding system is quantified in terms of Q_w values, where Q_w is a measure of the signal-to-correlated noise ratio of the wideband system, measured in dB. The corresponding MOS of a single G.722 64 kbit/s is about 4.5, and a tandem of two G.722 is about 4.1 [23].

The algorithmic delay of G.722 is 1.625 ms, which is composed of a 0.125 ms frame size and 1.5 ms lookahead (for the 24-point QMF filtering).

2.9 Audio Quality Evaluation

One of the most important aspects of a communication system is the audio quality. Audio quality is evaluated by the listener, so it is subjective. ITU-T Recommendation P-series deal with methods for objective and subjective assessment of audio quality, trying to represent the perception by some rating scales.

2.9.1 *Subjective Measurement Methods*

The Absolute Category Rating (ACR), defined in ITU-T Recommendation P.800[38], is widely used for subjective measurement of audio quality. Actually, it's accepted as the norm for comparative rating of different systems. In the test, a statistically valid mixed panel of men and women listens to a collection of voice samples under

controlled conditions, and votes on their quality using an integer opinion scale. The rating results are averaged to produce a MOS score.

MOS ranges from 1 to 5, with 5 presenting excellent speech quality, and 4 or higher presenting toll quality. The MOS rating, the corresponding listening quality and listening effort are shown in Table 2-7.

Table 2-7 MOS Rating

MOS Rating	Listening Quality (LQ)	Listening Effort (LE)
5	<i>Excellent</i>	<i>Complete relaxation possible; no effort required</i>
4	<i>Good</i>	<i>Attention necessary; no appreciable effort required</i>
3	<i>Fair</i>	<i>Moderate effort required</i>
2	<i>Poor</i>	<i>Considerable effort required</i>
1	<i>Bad</i>	<i>No meaning understood with any feasible effort</i>

2.9.2 Objective Measurement Methods

Determining the subjective speech quality of a transmission system has always been an expensive and laborious process, thus it's inappropriate for many situations. Objective measurement methods come up.

Most objective measurement methods are signal-based, that is, they use the reference signal and degraded signal as inputs, and identify the audible distortions based on the perceptual domain representation of two signals incorporating human auditory models. This scheme is employed by Bark Spectral Distortion (BSD), Perceptual Speech

Quality Measure (PSQM), Measuring Normalizing Blocks (MNB), Perceptual Analysis Measurement System (PAMS), and Perceptual Evaluation of Speech Quality (PESQ).

PAMS and PESQ are utilized for evaluation in the simulation of this thesis. Let's have a brief look at these two methods.

2.9.2.1 Perceptual Analysis Measurement System

PAMS was developed by the PsyTechnics group at British Telecommunications in 1998. The current version 3.0, widely accepted in Europe and North America, was released in December 1999 [39]. Unlike the former models, PAMS is the first model to focus on end-to-end behavior, including the filtering and variable delay. The PAMS score is mapped to a MOS-like scale, i.e., 0 to 5. The ratings, the corresponding listening quality and listening effort are similar as MOS.

2.9.2.2 Perceptual Evaluation of Speech Quality

Perceptual Evaluation of Speech Quality (PESQ) is the current ITU objective speech quality standard, defined in ITU-T P.862 [40]. It combines the merits of PAMS and PSQM (ITU-T P.861, Feb. 1998). In order to take such effects as filtering, variable delay, coding distortions and channel errors into account, it employs transfer function equalization, time alignment, and a new algorithm for averaging distortions over time[41], so it's preferable to evaluate the speech quality of VoIP.

The PESQ score is mapped to a MOS-like scale, a single number in the range of – 0.5 to 4.5, although for most cases the output range will be between 1.0 and 4.5, the normal range of MOS values found in an ACR listening quality experiment.

In our simulation, the multi-speeches generated by the audio mixer have to be assessed. Since none of the objective measurement tools are validated for multiple simultaneous talkers, subjective measurement no doubt is the best choice. However, the impracticality of setting up subjective measurement environment drives us to accept PESQ and PAMS, with the aid of informal subjective measurement. The results may not be accurate enough, but it does illustrate the audio quality improved or degraded.

In IEEE P.862, it's recommended that PESQ be used for speech quality assessment of 3.1 kHz (narrowband) handset telephony and narrowband speech codecs. Currently, PESQ has not been fully characterized for wideband telephony applications, and may need parts of the model to be changed. Two delayed contributions of ITU-T SG12 group [42][43] indicated this and put forward some modification advice. Thus, in evaluating the wideband audio signals, we prefer PAMS as the objective speech quality measurement tool. For the wideband and hybrid-bandwidth cases, the PESQ scores are tested together with the PAMS scores, as a reference. For the narrowband case and the transcoders for hybrid-bandwidth codecs, PESQ does work well, so it's adopted for these situations.

Chapter 3

Simulation Environment

3.1 Speech Samples

The speech files used in our simulation are obtained from TIMIT Acoustic-Phonetic Continuous Speech Corpus, which was jointly accomplished by the Massachusetts Institute of Technology (MIT), SRI International (SRI) and Texas Instruments, Inc. (TI) from January to May 1986. It provides speech data for acoustic-phonetic studies and for the development and evaluation of automatic speech recognition systems. TIMIT contains broadband recordings of 630 speakers of 8 major dialects of American English, each reading 10 phonetically rich sentences. The detailed information about TIMIT can be found in [44].

Since all the samples from TIMIT are sampled at 16 kHz, in order to obtain the narrowband signals, the sampling rate of the samples is converted from 16 kHz to 8 kHz using Cool Edit Pro. The only exception is, when generating the desired 8 kHz reference signal for the G.711+G.722-> G.711 mixer, we use the same scheme to convert the sampling rate as the transcoder does.

TIMIT has some naming rules for the speakers and sentences. The initial 'f' represents a female voice, while 'm' represents a male voice. In chapters 4, 5, and 6, 24 two-speaker and one three-speaker conferences are simulated. As for the two-speaker conference, eight samples from seven females and eight samples from seven males with different accents are chosen and grouped. To verify the algorithms, the speakers are all overlapping with each other. The length of the samples used ranges from 1.97 seconds to 4.41 seconds. Seven of the 24 cases are female talking with female, seven are male with male, and the remaining 10 are female talking with male.

The three-speaker conference uses different speeches from those in the two-speaker conference. The speakers are active now and then, like a normal conversation. For demonstration purposes, the overlaps of the talks are longer than the normal case, and there are periods of two and three simultaneous active talks. The waveforms of the three channels and the mixed signal are shown in Figure 3.1.1. The first channel is a male speaker, which is active during 0~4.60 seconds, and 7.60~11.80 seconds, the second one is from a lady and active during 4~7.03 seconds, and 9.03~13.35 seconds,

and the third one is male speak again and active during 6~10.54 seconds. All the three channels are active from 9.03 seconds to 10.54 seconds.

Figure 3.1.1 Waveforms of Three Input Channels and the Directly Mixed Signal

3.2 Comparison Model

Figure 3.2.1 shows the comparison model for the simulations in the following three chapters. The reference signal is obtained by mixing the inputs linearly.

Result 1 is obtained by comparing the decoded signal at the receiver using the traditional mixer, with the reference signal, while result 2 is obtained by comparing the decoded signal at the receiver using the simplified mixer, with the reference signal.

Figure 3.2.1 Comparison Mode

If all inputs are narrowband signals or all wideband signals, then the reference signal is of the same type as the input. However, for the hybrid bandwidth inputs, whether the reference signal is a narrowband or wideband signal, depends on what the mixed signal is desired to be. If the receiver expects a narrowband mixed signal, then a narrowband reference signal is required too, and vice versa. Hence, the reference signal is a “best-quality” signal.

3.3 Other Issues

It's necessary to point out that the other functions besides decoding, mixing and coding are not considered in this thesis and are assumed to exist already, such as Voice Energy Monitor, Automatic Gain Control and so on. The simplification of the mixer shouldn't damage their normal function.

In the following chapters, two-speaker and three-speaker conferencing are simulated for each mixer. The PESQ scores are obtained by using the P.862 software provided by ITU-T, and the PAMS scores are obtained by using the Digital Speech Level Analyzer (DSLAs), a measurement tool manufactured by Malden Electronics Ltd., U.K.

In the simulation, the mixer sums all active speeches for the listeners. For the speakers themselves, the returned signal is the sum of the other speeches, not including his/her own. This case is not discussed, since it's the same design with one less channel mixing.

Chapter 4

Audio Mixers for Narrowband VoIP Conferencing

4.1 G.711 Mixer

If several inputs conforming to G.711 need to be mixed, first they have to be transformed to linear format, then summed together, and passed by a compander to transform back to the desired A-law or u-law format.

There is another idea, that is, to design a lookup table for addition of G.711 inputs. Since G.711 is a low-complexity and low-delay codec, this design is not preferable in terms of the memory usage.

The rest of this chapter will focus on the G.729/G.729A audio mixers, simplifying them by a scheme named strong pitch domination.

4.2 Strong-Pitch-Domination Mixer

4.2.1 Algorithm Breakdown

Table 4-1 shows the breakdown of the encoding and decoding steps of G.729 and G.729A [45], measured by Weighted Millions of Operations Per Second (WMOPS) and Texas Instruments (TI) c50 DSP MIPS.

Table 4-1 Breakdown of the Codec Complexity (Worst Case) for G.729 and G.729A

Function	WMOPS		C50MIPS	
	G.729	G.729A	G.729	G.729A
Pre-processing	0.2	0.2	0.226	0.226
LP Analysis	1.63	1.28	1.957	1.696
LSP Quantization and Interpolation	0.95	0.95	1.39	1.39
LSP to A(z) and Weighting	0.3	0.12	0.461	0.173
Open-loop Pitch	1.45	0.82	1.563	0.955
Closed-loop Pitch	2.83	1.55	3.453	1.778
Algebraic Codebook	6.35	1.86	8.406	3.046
Quantization of Gains	0.46	0.46	0.643	0.643
Find Excitation and Memory Update	0.21	0.08	0.278	0.112
Total (Coder)	14.38	7.32	18.377	10.019
Decoder	0.68	0.68	1.133	1.133
Postfilter	2.13	0.73	2.539	1
Post-processing	0.22	0.22	0.266	0.266
Total (Decoder)	3.03	1.63	3.938	2.399
Total (Duplex)	17.41	8.95	22.315	12.418

The open-loop pitch analysis is of interest to us. It is done once per frame, to find a candidate pitch delay T_{op} to be used in the closed-loop pitch analysis. Generally speaking, T_{op} is not the true pitch delay of the current frame, however, it can narrow down the search range of the closed-loop pitch search, which employs a more complicated method

to find a more accurate pitch delay for each subframe. From Table 4-2, we can see that for G.729, the open-loop pitch analysis takes about $1.45/14.38 = 10.08\%$ in encoder complexity only, and $1.45/17.41 = 8.32\%$ in total complexity, and for G.729A, it takes $0.82/7.32 = 11.20\%$ in encoder complexity only, and $0.82/8.95 = 9.16\%$ in total complexity. As for the mixer, which involves two decoders and one encoder, the open-loop pitch search takes $1.45/(3.03*2 + 14.38) = 7.09\%$ in the entire mixer's complexity for G.729, and $0.82/(1.63*2 + 7.32) = 7.75\%$ in the entire mixer's complexity for G.729A.

4.2.2 *Voiced vs Unvoiced*

As we know, the tandem vocoders do the same pitch analysis, and the pitch represents the long-term correlation in the speech. At the bridge, the decoder has all the pitch information of all inputs, which represents the long-term correlation of the input signals. The long-term correlation of each component must be related to the long-term behavior of the summation signal. So it's natural to consider making use of the pitch information from the decoder for the G.729 encoding of the summation signal.

The G.729 vocoder processes unvoiced speech like voiced speech; that is, it extracts the pitch parameters for the unvoiced speech too. Since unvoiced speech exhibits no pitch, actually, the adaptive-codeword component contributes little to the synthesis of unvoiced signal.

In contrast, the pitch information is significant for the voiced speech. This doesn't mean the fixed-codebook part contributes little, since the fixed-codebook is used to refine the excitation signal of the LP model.

This means that if the inputs are the sum of single or multiple unvoiced speeches and single voiced speech, the mixed voice must have a similar pitch as that of the voiced speech. Further, if there is more than one voiced speech input, the long-term correlation is dominated by the one with strong pitch energy.

Based on the above, a smart audio mixer can be constructed. That is, compare the pitch energy of each input after the decoder, choose the strongest one and send its pitch delay directly to the encoder. The encoder of the mixer can use this value as the candidate delay T_{op} , which used to be obtained from the open-loop pitch analysis, and this reduces the complexity of the mixer by more than 7%.

As is known, the coder uses the open-loop pitch delay T_{op} to find the optimal delay for each subframe. The optimal delay is determined using closed-loop analysis that minimizes the weighted mean-squared error. In the first subframe, the delay T_1 is found by searching a small range (6 samples) of delay values around the open-loop delay T_{op} . For the second subframe, closed-loop pitch analysis is done around the pitch selected in the first subframe to find the optimal delay T_2 . Therefore, if the pitch delay from the

decoder doesn't significantly differ from the true delay pitch, we need not worry that the pitch delay from the decoder is not the true pitch delay for the mixed speech. The vocoder will adjust it in some range and the fixed-codebook will further refine the excitation signal.

It must be noted that in some situations the pitch delay may not be accurate enough, for example the plosive case. But if such situation doesn't occur frequently, the total speech quality cannot be degraded noticeably.

4.2.3 *Simplified Mixer*

The structure of the smart Strong-Pitch-Domination mixer is shown in Figure 4.2.1, which is the two-speaker case. Compared with the conventional design, the Pitch Estimator (green block) is added, while the Open Loop Pitch Search (yellow block) is eliminated.

Figure 4.2.1 The Structure of the Strong-Pitch-Domination Mixer

For the simplified mixer, the pitch delay (second subframe), pitch gain, and frame energy need to be extracted from each decoder, as the input to the pitch estimator. The pitch estimator compares the pitch energies of all inputs, and chooses the one with strongest pitch energy as the open-loop pitch delay of the summation signal. The estimated pitch delay is consequently sent to the encoder, as the candidate pitch delay for the closed-loop pitch analysis. By this means, the open-loop pitch analysis can be eliminated.

The structure of the pitch estimator is shown in Figure 4.2.2. The pitch of the summation signal is selected by considering three terms. First is the pitch gain (adaptive-codebook gain), which is normalized by the signal energy. The second term is the signal energy, which can be obtained from the decoder directly. The third term is optional, and

that is to take the pitch variation into consideration. The default assumption is the pitch delay shouldn't greatly vary for the voiced speech. Thus if the pitch delay differs from that of the last frame by 20, we deem it the unvoiced or the transient period of the voiced and unvoiced speech, so a smaller weight (0.5 is used in the simulation) is given to this case.

Figure 4.2.2 Pitch Estimator

Experiments are made to evaluate speech quality of the mixed speech of different cases. We will show that the quality of the G.729(A) + Simplified G.729(A) can be comparable with the common case (G.729(A) + G.729(A)).

We already indicated that the complexity reduction achieved by employing the Strong-Pitch-Domination idea is more than 7% of the entire mixer's complexity. Due the different realizations of algorithms on different DSP chips, it's usually hard to give exactly how much complexity or processing delay is reduced. So sometimes such operations like multiplication or addition are investigated in order to evaluate the conventional and simplified design. It may not be accurate enough, but it does illustrate the complexity in some sense.

Table 4-2 can provide a better idea of the reduced complexity and new load in terms of operations. The new load is different for different pitch energy comparator designs, such as introducing a moving average of the pitch gain and energy, or considering the pitch variation, and these can improve the quality a little more while only adding slightly more complexity.

Table 4-2 Reduced Complexity of the Simplified G.729 Mixer in Terms of Operations

Operations per frame (10ms)	Multiplication	Addition	SQRT	Division
Speech Signal Weighting	>40*40	>21	0	0
Open-loop Pitch Search	>80*124 + 1	>200	1	1
Total Reduced	>11640	>220	1	1
New Load	4	1~7	0	0

4.3 Simulation and Results

The two-speaker and three-speaker conferencing as indicated in chapter 3 are simulated for the comparison of conventional and simplified G.729(A) mixers.

It should be indicated that the conferences are not restricted to two or three conferees. The simplified mixer can accommodate any number of conferees and any number of active speakers.

4.3.1 *Two-Speaker Conferencing*

Simulation results are shown in Table 4-3. The pitch energy comparator only takes pitch gain and energy of current frame into consideration, so it's the simplest design.

The 1st and 2nd columns are the two inputs. The reference signal is obtained by mixing them in the linear format.

The 3rd column is the mixed speech received by the listeners, which is processed by the tandem G.729 procedure (encoded at the sender end, decoded at the mixer, mixed and re-encoded at the mixer, and re-decoded at the receiver end).

The 4th column is similar to the 3rd column, but the second G.729 encoder makes use of the pitch information from the decoder, which reduces the algorithm complexity by more than 7% of the complexity of the entire mixer.

The 5th column is obtained by subtracting the 3rd column from the 4th column. A negative value means quality degradation, while a positive value means improvement.

The 6th, 7th and 8th columns are the corresponding results using G.729A instead of G.729.

Table 4-3 The PESQ Scores of the Mixed Voices in the G.729(A) Mixer

File 1	File 2	G.729			G.729A		
		Conventional Mixer	Simplified Mixer	Degradation	Conventional Mixer	Simplified Mixer	Degradation
faks0sil573_8kHz	fasw0sa1_8kHz	3.437	3.392	-0.045	3.243	3.05	-0.193
faks0sil573_8kHz	fbjl0sx292_8kHz	3.17	3.178	0.008	3.089	2.86	-0.229
fasw0sx200_8kHz	fdac1x394_8kHz	3.613	3.614	0.001	3.473	3.387	-0.086
fbjl0sx292_8kHz	fdac1x394_8kHz	3.427	3.366	-0.061	3.285	3.157	-0.128
fbmh0sa2_8kHz	fcdr1sx106_8kHz	3.311	3.364	0.053	3.252	3.076	-0.176
fbmh0sa2_8kHz	fcjff0sx397_8kHz	3.417	3.412	-0.005	3.212	3.181	-0.031
fcdr1sx106_8kHz	fcjff0sx397_8kHz	3.22	3.264	0.044	3.157	3.067	-0.09
faks0sil573_8kHz	mahh0si1924_8kHz	3.261	3.23	-0.031	3.119	2.98	-0.139
fasw0sa1_8kHz	mdlr1si669_8kHz	3.577	3.596	0.019	3.466	3.427	-0.039
fasw0sx200_8kHz	mcmp0sa1_8kHz	3.525	3.512	-0.013	3.391	3.369	-0.022
fbjl0sx292_8kHz	mdac2sx279_8kHz	3.643	3.575	-0.068	3.446	3.448	0.002
fbmh0sa2_8kHz	mctt0sx298_8kHz	3.599	3.631	0.032	3.554	3.399	-0.155
fbmh0sa2_8kHz	medr0sx114_8kHz	3.522	3.514	-0.008	3.45	3.37	-0.08
fcdr1sx106_8kHz	mcmp0sa1_8kHz	3.408	3.402	-0.006	3.386	3.24	-0.146
fcjff0sx397_8kHz	mdac2si999_8kHz	3.56	3.555	-0.005	3.418	3.378	-0.04
fcjff0sx397_8kHz	mchl0sx87_8kHz	3.671	3.695	0.024	3.537	3.395	-0.142
fdac1x394_8kHz	mdlr1si669_8kHz	3.458	3.47	0.012	3.404	3.217	-0.187
mahh0si1924_8kHz	mcmp0sa1_8kHz	3.554	3.591	0.037	3.512	3.468	-0.044
mahh0si1924_8kHz	mdac2sx279_8kHz	3.651	3.752	0.101	3.498	3.44	-0.058
mchl0sx87_8kHz	medr0sx114_8kHz	3.585	3.623	0.038	3.45	3.401	-0.049
mchl0sx87_8kHz	mdac2si999_8kHz	3.557	3.553	-0.004	3.394	3.413	0.019
mcmp0sa1_8kHz	mdac2sx279_8kHz	3.685	3.663	-0.022	3.558	3.435	-0.123
mctt0sx298_8kHz	medr0sx114_8kHz	3.656	3.583	-0.073	3.713	3.533	-0.18
mdac2sx279_8kHz	mdlr1si669_8kHz	3.685	3.63	-0.055	3.532	3.412	-0.12
Average		3.508	3.507	-0.001	3.397	3.296	-0.102

We can see that by using the simplified Strong-Pitch-Domination mixer, the quality of the mixed speech only decreases about 0.001 for G.729 and 0.102 for G.729A, which is insignificant. Experiment shows that if we consider the moving average of the pitch gain and pitch variation in the pitch energy comparator, the quality can be better.

The simulation results prove that the open-loop pitch analysis in the encoder of the mixer can be skipped without degrading the performance if the estimated pitch delay is replaced by the pitch delay of the speaker with strongest pitch energy.

4.3.2 *Three-Speaker Conferencing*

The simplification is valid not only for two active speakers. We can have a look at a three-speaker conference compressed with G.729 and G.729A. The simplification of the audio mixer is similar to the two-speaker case. For each channel, the pitch delay of the second subframe, pitch gain, and frame energy are extracted once per frame. The pitch estimator compares the pitch energies of the three channels and selects the one with strongest pitch energy as the candidate pitch delay for closed-loop pitch delay of the summation signal.

The simulation results are as satisfactory as the two-speaker case. The conventional tandem vocoder gives a PESQ value of 3.602 for G.729 and 3.476 for G.729A respectively, while the simplified mixer gives a PESQ value of 3.559 and 3.394, degraded by about 0.04 and 0.08. By informal subjective assessment, the simplified mixer provides a comparable quality to the conventional design.

4.4 Summary

In this chapter, the G.729(A) mixer is simplified by the scheme of Strong-Pitch-Domination. The pitch energies of all active speakers are compared and the pitch delay

with the strongest pitch energy is selected to substitute for the candidate pitch delay of the open-loop pitch search for the mixed signal. The simulation shows that the simplified mixer can acquire comparable quality with the conventional G.729(A) mixer, while reducing the computation complexity by more than 7%. The processing delay is reduced too. The multi-speaker conferencing of inputs conforming to G.711 and G.729(A) codecs are not discussed in this thesis.

Chapter 5

Audio Mixers for Wideband VoIP Conferencing

5.1 Conventional G.722 Audio Mixer

The conventional design is to decode the inputs conforming to G.722, then mix them as linear signals, and re-encode the summation signal to a 64 kbit/s ADPCM bit stream again. The signal flow is shown in Figure 5.1.1.

It's noticed that the receive QMFs combine the lower and higher subband signals into wideband signals, while the wideband summation signal is split into lower and higher subband again by the transmit QMF again. So why not do the summation in the subband level?

Figure 5.1.1 Block Diagram of the Conventional G.722 Mixer

5.2 Subband Summation Mixer

The idea is that the signals conforming to G.722 can be summed in the subband domain. It is already indicated in the ITU-T Recommendation G.722 [37], which says:

Signal summing at the subband uniform PCM level is preferred for the following reasons:

1. The hardware is minimized in the MCU by eliminating the need for quadrature mirror filters,
2. Signal quality is maximized and algorithmic delay is reduced by avoiding additional filtering,
3. Echo control is likely to be simpler to perform at the subband level.

Figure 5.2.1 shows the structure of the simplified G.722 subband summation mixer. Note in this structure, both the two receive and one transmit QMFs are eliminated (6 12-tap filters working at 8 kHz).

Figure 5.2.1 Block Diagram of the Subband Summation G.722 Mixer

The subband summation mixer saves the multiplication operation $6 \times 12 \times 8000 = 576000$ times/sec, and the addition operation $6 \times 13 \times 8000 = 624000$ times/sec. This represents a reduction of the entire mixer complexity of about 36.9% for the multiplications, and 45.3% for the additions. This measurement may not be accurate enough, since there are some other operations such as table-lookup, but it gives us some idea of the complexity reduction.

The delay is reduced too. Firstly, the elimination of transmit and receive QMFs reduces the algorithmic delay by 3 ms (2×1.5 ms). Besides, the processing delay is reduced too.

A simulation is performed to compare the audio quality with conventional and simplified mixers.

5.3 Simulation Results

As before, first the two-speaker case is investigated. Table 5-1 shows results for various cases involving two active channels.

Table 5-1 Simulation Results of the Conventional and Simplified G.722 Mixer

File 1 (G.722)	File 2 (G.722)	PESQ			PAMS		
		Conventional Mixer	Simplified Mixer	Degradation	Conventional Mixer	Simplified Mixer	Degradation
faks0sil573_16kHz	fasw0sa1_16kHz	4.352	4.362	0.01	4.69	4.71	0.02
faks0sil573_16kHz	fbjl0sx292_16kHz	4.35	4.35	0	4.59	4.61	0.02
fasw0sx200_16kHz	fdac1x394_16kHz	4.359	4.357	-0.002	4.54	4.56	0.02
fbjl0sx292_16kHz	fdac1x394_16kHz	4.418	4.404	-0.014	4.5	4.5	0
fbmh0sa2_16kHz	fedr1sx106_16kHz	3.835	4.034	0.199	4.69	4.69	0
fbmh0sa2_16kHz	fcjf0sx397_16kHz	4.39	4.416	0.026	4.65	4.65	0
fedr1sx106_16kHz	fcjf0sx397_16kHz	4.392	4.406	0.014	4.59	4.61	0.02
faks0sil573_16kHz	mahh0si1924_16kHz	4.255	4.322	0.067	4.7	4.7	0
fasw0sa1_16kHz	mdlr1si669_16kHz	4.422	4.417	-0.005	4.63	4.65	0.02
fasw0sx200_16kHz	mcamp0sa1_16kHz	4.357	4.359	0.002	4.59	4.58	-0.01
fbjl0sx292_16kHz	mdac2sx279_16kHz	4.444	4.457	0.013	4.68	4.69	0.01
fbmh0sa2_16kHz	mett0sx298_16kHz	4.388	4.394	0.006	4.67	4.66	-0.01
fbmh0sa2_16kHz	medr0sx114_16kHz	4.307	4.305	-0.002	4.64	4.65	0.01
fedr1sx106_16kHz	mcamp0sa1_16kHz	4.307	4.305	-0.002	4.55	4.59	0.04
fcjf0sx397_16kHz	mdac2si999_16kHz	4.43	4.425	-0.005	4.62	4.66	0.04

fcjt0sx397_16kHz	mchl0sx87_16kHz	4.398	4.393	-0.005	4.58	4.6	0.02
fdac1x394_16kHz	mdlr1si669_16kHz	4.42	4.432	0.012	4.49	4.5	0.01
mabh0si1924_16kHz	mcmp0sal_16kHz	4.349	4.33	-0.019	4.48	4.5	0.02
mabh0si1924_16kHz	mdac2sx279_16kHz	4.469	4.465	-0.004	4.7	4.7	0
mchl0sx87_16kHz	medr0sx114_16kHz	4.373	4.388	0.015	4.59	4.6	0.01
mchl0sx87_16kHz	mdac2si999_16kHz	4.448	4.447	-0.001	4.69	4.7	0.01
mcmp0sal_16kHz	mdac2sx279_16kHz	4.387	4.362	-0.025	4.71	4.71	0
mcht0sx298_16kHz	medr0sx114_16kHz	4.361	4.376	0.015	4.53	4.55	0.02
mdac2sx279_16kHz	mdlr1si669_16kHz	4.442	4.426	-0.016	4.65	4.67	0.02
Average		4.361	4.372	0.012	4.615	4.627	0.012

It can be concluded that the subband summation solution does work, as shown by the experiments. It's also noticed that by eliminating the unnecessary filtering, the audio quality is slightly improved by using the subband summation (improved by 0.012, measuring in both PESQ and PAMS).

For the three-speaker case, the same scheme is employed. Simulation shows that the conventional mixer can achieve 4.354 in PESQ and 4.63 in PAMS, while the subband summation mixer can achieve 4.367 in PESQ and 4.64 in PAMS. The audio quality is improved using the simplified design, too.

5.4 Summary

Since G.722 employs subband processing, we can make use of the subband summation for the G.722 mixer. The lower and higher subband signals are mixed respectively and re-encoded using ADPCM. The receive QMF in the two decoders and transmit QMF in the encoder are all eliminated, which reduces the algorithmic delay by 3

ms. The audio quality in the simulations shows the subband summation mixer achieves a better performance, compared with the conventional design.

Chapter 6

Audio Mixers for Hybrid-Bandwidth VoIP Conferencing

Sometimes, we may have users residing in different networks or using different compression algorithms. Some may reside in narrowband networks, while some may reside in wideband networks. Some may use an 8 kHz sampling rate and low bit-rate codecs, while some may prefer the 16 kHz sampling rate and higher bit-rate codecs. Some may use waveform codecs, while some may use parametric codecs. The mixer has to deal with the different request of terminals, i.e., decoding inputs with different compression formats, and encoding the mixed signal according to the specified codec.

6.1 Conventional Transcoders for G.711 and G.722

The question is, while mixing the narrowband signal with the 7 kHz wideband G.722 signal, is it possible to use the lower sub-band only, ignoring the higher subband?

The answer is no. The reason is the quadrature mirror filters. The transmit and receive QMFs are identical two-channel subband coding filter banks with power complementary frequency response.

[37] gives the coefficients of the QMFs

```
hl = [ 0.0004 -0.0013 -0.0013 0.0065 0.0015 -0.0190 0.0039 0.0442 -0.0256 -0.0983 0.1161 0.4731
       0.4731 0.1161 -0.0983 -0.0256 0.0442 0.0039 -0.0190 0.0015 0.0065 -0.0013 -0.0013 0.0004 ]
hh = [ 0.0004 0.0013 -0.0013 -0.0065 0.0015 0.0190 0.0039 -0.0442 -0.0256 0.0983 0.1161 -0.4731
       0.4731 -0.1161 -0.0983 0.0256 0.0442 -0.0039 -0.0190 -0.0015 0.0065 0.0013 -0.0013 -0.0004 ]
```

The frequency response of the lower subband QMF is shown in Figure 6.1.1. We can see, for the lower subband QMF, the passband frequency is about 4 kHz, and the stopband frequency is about 5.6 kHz. The non-zero frequency response above 4 KHz introduces aliasing to lower subband signal. As for the higher subband QMF, it has a symmetric frequency response to that of the lower subband, so the aliasing also exists in the higher subband signal. The frequency response of the higher subband QMF is shown in Figure 6.1.2. The aliasing existing in the lower and higher subband signals can be cancelled completely by the receive QMF. It's apparent that the quantization error introduced in the ADPCM coding of the lower and higher subband cannot be eliminated completely by the receive QMF [46].

Figure 6.1.1 Frequency Response of the Lower Subband QMF **Figure 6.1.2** Frequency Response of the Higher Subband QMF

Figure 6.1.3 and Figure 6.1.4 are the transcoding structures between the G.722 signal and the 64 kbps PCM indicated by Recommendation G.722 [37]. The G.711-to-G.722 transcoder first converts the A-law/u-law signal into uniform PCM, then interpolates it from 8 kHz to 16 kHz, and processes the wideband uniform PCM signal with the G.722 encoder. On the other hand, the G.722-to-G.711 transcoder first decodes the input signal conforming to G.722, downsamples it to 8 kHz, and finally converts it into an A-law or u-law signal.

Figure 6.1.3 G.711-to-G.722 Transcoder Indicated by Rec. G.722

Figure 6.1.4 G.722-to-G.711 Transcoder Indicated by Rec. G.722

The lowpass filter (LPF) is required to sample at 16 kHz and not significantly modify the baseband frequency response up to 3.4 kHz and attenuates the frequency components above 4.6 kHz. The recommendation doesn't give the other specifications of the LPF, except the cutoff frequencies.

6.2 Simplified G.722-to-G.711 Transcoder

6.2.1 *High-Frequency-Suppression Transcoder*

The conventional G.722-to-G.711 transcoder includes a G.722 decoder, a decimator and a compander. Our idea is that the lower subband G.722 decoded signal can be used merely to obtain the PCM signal. Considering the LPF, which attenuates the frequency components above 4.6 kHz, we can add the order of the LPF to get a narrower transition band to remove the high frequency aliasing.

To illustrate the question, the LPF is constructed using the Remez FIR filter design, which requires:

Passband ripple $r_p = 3\text{dB}$,
 Stopband ripple $r_s = 40\text{dB}$,
 Transition bands are [3400, 4600] and [3400, 4000] for two different designs.
 Note: $f_{s1} = 16\text{ kHz}$ and $f_{s2} = 8\text{ kHz}$

The ripples for passband and stopband can be other reasonable values. Now they are set to 3 dB and 40 dB for illustration purpose. Remez design is used since it can minimize the maximum error between the desired frequency response and the actual frequency response. Other filter design can be used instead, however.

The Remez design gives two filters to satisfy the two specifications above and the orders of both filters are 14.

```
lpf1 = [ 0.0132  0.0651  0.0725 -0.0184 -0.1013  0.0208  0.3160  0.4783  0.3160  0.0208
        -0.1013 -0.0184  0.0725  0.0651  0.0132]
lpf2 = [ 0.0527 -0.0402  0.0523 -0.0640  0.0742 -0.0824  0.0875  0.9108  0.0875 -0.0824
        0.0742 -0.0640  0.0523 -0.0402  0.0527]
```

The frequency responses of the two filters are shown in Figure 6.2.1 and Figure 6.2.2.

Figure 6.2.1 Frequency Response of the 16 kHz LPF

Figure 6.2.2 Frequency Response of the 8 kHz LPF

Thus the structure of the simplified transcoder is shown in Figure 6.2.3.

Figure 6.2.3 Block Diagram of the Simplified G.722-to-G.711 Transcoder

Compared with the old structure as shown in Figure 6.1.4,

the advantages of the new structure are:

1. Higher Subband Decoder and Receive QMF are eliminated,
2. LPF works at 8 kHz instead of 16 kHz.

It's clear that by eliminating the receive QMF, we reduce the algorithmic delay by 1.5 ms.

6.2.2 Simulation Results

Some experiments are made to verify our design. Figure 6.2.4, Figure 6.2.5 and Figure 6.2.6 are the spectral views of the original signal, fullband processed signal and lower subband only processed signal. Also note that the sampling frequency of the original signal is 16 kHz, while that of the other two is 8 kHz.

Figure 6.2.4 Spectral View of the Original Signal **Figure 6.2.5 Spectral View of the Fullband Processed Signal** **Figure 6.2.6 Spectral View of the Lower Subband Processed Signal**

By using the lower subband component only, we lose some information in the higher frequency. This is the cost of not eliminating the high frequency aliasing in the lower subband signal perfectly. Since narrowband telecommunication systems usually limit the signals to a bandwidth of 3.4 kHz, the frequency suppression above 3.4 kHz will not degrade the integral quality too much.

Table 6-1 shows the PESQ and PAMS results of the conventional and simplified G.722-to-G.711 transcoders. By using the high frequency suppression in the transcoder, the audio quality is reduced by about 0.02 in PESQ, or about 0.09 in PAMS.

Table 6-1 Simulation Results of the G.722-to-G.711 Transcoders

Input File (G.722)	PESQ			PAMS		
	Conventional Transcoder	Simplified Transcoder	Degradation	Conventional Transcoder	Simplified Transcoder	Degradation
faks0sil573_16kHz	4.211	4.174	-0.037	4.56	4.41	-0.15
fasw0sa1_16kHz	4.293	4.226	-0.067	4.56	4.49	-0.07
fasw0sx200_16kHz	4.288	4.197	-0.091	4.61	4.57	-0.04
fbpl0sx292_16kHz	4.256	4.248	-0.008	4.45	4.36	-0.09
fbmh0sa2_16kHz	4.335	4.31	-0.025	4.49	4.3	-0.19
fedr1sx106_16kHz	4.224	4.247	0.023	4.2	4.21	0.01
fcj10sx397_16kHz	4.246	4.234	-0.012	4.37	4.17	-0.2
fdac1x394_16kHz	4.311	4.299	-0.012	4.54	4.51	-0.03
mabb0si1924_16kHz	4.239	4.228	-0.011	4.28	4.19	-0.09
mcmp0sa1_16kHz	4.249	4.262	0.013	4.37	4.25	-0.12
mchl0sx87_16kHz	4.352	4.359	0.007	4.64	4.74	0.1
mcu0sx298_16kHz	4.36	4.38	0.02	4.53	4.55	0.02
mdac2si999_16kHz	4.304	4.202	-0.102	4.61	4.42	-0.19
mdac2sx279_16kHz	4.336	4.267	-0.069	4.62	4.45	-0.17
mdlr1si669_16kHz	4.275	4.281	0.006	4.54	4.32	-0.22
medr0sx114_16kHz	4.217	4.225	0.008	4.42	4.37	-0.05
Average	4.286	4.264	-0.021	4.482	4.393	-0.089

It should be noted that the structure could be used not only for G.722 to G.711, but also G.722 to any narrowband codec. The only difference is to replace the compander by the desired encoder.

6.3 Simplified Hybrid-Bandwidth-to-Narrowband Mixer

6.3.1 *High-Frequency-Suppression Mixer*

Since the lower subband transcoder works properly, it can be applied into the G.711 + G.722 $\rightarrow$ G.711 mixer. Figure 6.3.1 and Figure 6.3.2 give the conventional and simplified block diagrams. The advantages of the simplified mixer are the same as the G.722-to-G.711 transcoder. So, the higher subband decoder and receive QMF are eliminated, and the LPF works at 8 kHz instead of 16 kHz. The algorithmic delay is reduced by 1.5 ms, too.

Figure 6.3.1 Block Diagram of the Conventional G.711+G.722 -> G.711 Mixer

Figure 6.3.2 Block Diagram of the Simplified G.711+G.722-> G.711 Mixer

Here it's necessary to point out that the G.711 encoder and decoder can be replaced by any narrowband codec, such as G.729(A) and so on. It will be verified in the simulation of three-speaker conferencing.

6.3.2 *Simulation Results*

Experiments are made to compare the performance of the conventional and simplified structures. For two-speaker conferencing, only the G.711 is tested as the narrowband codec.

The reference is the summation signal of the original signal of input 1 (8 kHz sampling rate, to be companded as u-law), and the downsampled 8 kHz signal of input 2 (16 kHz, to be coded by G.722). The downsampling of the second channel adopts the same lowpass filter as that in the conventional mixer.

An example is given below, in which we use the fullband-processed signal as reference, to evaluate the lower-subband-only processed signal. Figure 6.3.3 and Figure 6.3.4 are the waveform view and spectral view respectively, showing that the mixed signals from the conventional and simplified mixers are very close to each other.

Reference Signal:

mctt0sx298_8kHz_medr0sx114_16kHz_mix_fullband_tandem.wav

Degraded Signal:

mctt0sx298_8kHz_medr0sx114_16kHz_mix_lowerband_tandem.wav

Figure 6.3.3 2D Reference, Simplified, Error Signal View Using PAMS Analysis (G.711+G.722 -> G.711 Mixer)

Figure 6.3.4 Initial Spectrum View Using PAMS Analysis (G.711+G.722 -> G.711 Mixer)

All the experiment results of two-speaker conferencing are listed in Table 6-2, showing that using the simplified structure does not significantly reduce the audio quality.

Table 6-2 Simulation Results of the G.711 Input + G.722 Input -> G.711 Output Mixer

File 1 (G.711)	File 2 (G.722)	PESQ			PAMS		
		Conventional Mixer	Simplified Mixer	Degradation	Conventional Mixer	Simplified Mixer	Degradation
faks0si1573_8kHz	fasw0sa1_16kHz	3.673	3.612	-0.061	4.11	4.04	-0.07
faks0si1573_8kHz	fbjl0sx292_16kHz	3.823	3.846	0.023	4.11	4.05	-0.06
fasw0sx200_8kHz	fdac1x394_16kHz	3.667	3.793	0.126	3.97	3.96	-0.01
fbjl0sx292_8kHz	fdac1x394_16kHz	3.802	3.742	-0.06	3.97	3.84	-0.13
fbmh0sa2_8kHz	fedr1sx106_16kHz	3.429	3.37	-0.059	3.93	3.93	0
fbmh0sa2_8kHz	fcjfl0sx397_16kHz	3.638	3.569	-0.069	3.96	3.87	-0.09
fedr1sx106_8kHz	fcjfl0sx397_16kHz	3.825	3.763	-0.062	4.15	4.01	-0.14
faks0si1573_8kHz	mahh0si1924_16kHz	3.765	3.766	0.001	4.22	4.08	-0.14
fasw0sa1_8kHz	mdlr1si669_16kHz	3.631	3.736	0.105	4.02	3.95	-0.07
fasw0sx200_8kHz	mcmp0sa1_16kHz	3.633	3.731	0.098	3.99	3.96	-0.03
fbjl0sx292_8kHz	mdac2sx279_16kHz	4.005	3.864	-0.141	4.16	4.02	-0.14
fbmh0sa2_8kHz	mctt0sx298_16kHz	3.663	3.657	-0.006	4.19	4.19	0
medr0sx114_8kHz	fbmh0sa2_16kHz	3.708	3.629	-0.079	4.1	4.03	-0.07
mcmp0sa1_8kHz	fedr1sx106_16kHz	3.604	3.659	0.055	3.94	3.91	-0.03
mdac2si999_8kHz	fcjfl0sx397_16kHz	3.834	3.865	0.031	4.07	4.05	-0.02
mchl0sx87_8kHz	fcjfl0sx397_16kHz	3.86	3.819	-0.041	4.35	4.33	-0.02
mdlr1si669_8kHz	fdac1x394_16kHz	3.717	3.732	0.015	4.02	3.94	-0.08
mahh0si1924_8kHz	mcmp0sa1_16kHz	3.754	3.712	-0.042	3.85	3.78	-0.07
mahh0si1924_8kHz	mdac2sx279_16kHz	3.867	3.745	-0.122	4.19	4.09	-0.1
mchl0sx87_8kHz	medr0sx114_16kHz	3.693	3.646	-0.047	4.2	4.19	-0.01
mchl0sx87_8kHz	mdac2si999_16kHz	3.825	3.736	-0.089	4.14	3.97	-0.17
mcmp0sa1_8kHz	mdac2sx279_16kHz	3.959	3.856	-0.103	4.27	4.19	-0.08
mctt0sx298_8kHz	medr0sx114_16kHz	3.693	3.724	0.031	4.26	4.23	-0.03
mdac2sx279_8kHz	mdlr1si669_16kHz	3.63	3.643	0.013	4.05	3.93	-0.12
Average		3.737	3.717	-0.020	4.093	4.023	-0.070

The three-speaker conferencing is simulated. The first channel conforms to G.711, the second channel conforms to G.722, and the third channel conforms to G.729A. The conventional mixer gives a PESQ score of 3.802, or a PAMS score of 3.94, while the simplified mixer gives 3.801 and 3.94 in contrast.

So the hybrid-bandwidth-to-narrowband mixer is simplified by applying high frequency suppression. The audio quality is degraded a little, however, the complexity and the delay are reduced also. In terms of the multiplication and addition operations, the simplified mixer can reduce multiplication operations by about 66% of that of the entire mixer, and addition operations by about 67% of that of the entire mixer. These values can be deemed as a reference, but not the accurate measurement of the real complexity.

6.4 Simplified G.711-to-G.722 Transcoder

6.4.1 Filter-Combination Transcoder

The transcoder indicated by Recommendation G.722 has been shown in Figure 6.1.3. The G.722 encoder includes a transmit QMF, and lower and higher subband ADPCM coders. So Figure 6.4.1 shows the same transcoder structure as Figure 6.1.3.

Figure 6.4.1 Block Diagram of the Conventional G.711-to-G.722 Transcoder

Can we use the 4 kHz input only to obtain the aliased lower and higher subband signals for the G.722 encoder? We note that the QMF is actually a 2-channel polyphase filter [47], so the interpolator can be combined with the QMF. The transformation is shown in Figure 6.4.2, Figure 6.4.3, Figure 6.4.4, and Figure 6.4.5, step by step.

Figure 6.4.2 Analysis of the Filter Combination (I)

The QMF is a lowpass filter + downsampler and a highpass filter (HPF) + downsampler.

Figure 6.4.3 Analysis of the Filter Combination (II)

Let's combine the LPF and H_L , LPF and H_H respectively.

Figure 6.4.4 Analysis of the Filter Combination (III)

Using the polyphase structure again, we can eliminate the upsampler and downsampler. The filters

$$H_{c0,0} = \text{even coefficients of } H_{c0}$$

$$H_{c1,0} = \text{even coefficients of } H_{c1}$$

Figure 6.4.5 Analysis of the Filter Combination (IV)

The coefficients of the two combined filters are

$$\begin{aligned} H_{c0,0} &= [0.0000 \quad -0.0001 \quad 0.0003 \quad -0.0010 \quad 0.0022 \quad -0.0053 \quad 0.0437 \quad 0.0315 \quad -0.0655 \\ &\quad 0.4274 \quad 0.1687 \quad -0.0795 \quad 0.0831 \quad -0.0004 \quad 0.0031 \quad -0.0015 \quad 0.0005 \quad -0.0001 \quad 0.0000]; \\ H_{c1,0} &= [0.0000 \quad 0.0001 \quad -0.0006 \quad 0.0018 \quad -0.0032 \quad 0.0038 \quad -0.0055 \quad -0.0032 \quad 0.0131 \\ &\quad -0.0233 \quad 0.0296 \quad -0.0296 \quad 0.0229 \quad -0.0180 \quad 0.0078 \quad -0.0023 \quad 0.0002 \quad 0.0001 \quad -0.0000]; \end{aligned}$$

The frequency responses are shown in Figure 6.4.6 and Figure 6.4.7.

Figure 6.4.6 Frequency Response of the Combined Filter 0

Figure 6.4.7 Frequency Response of the Combined Filter 1

The simplified transcoder is shown in Figure 6.4.8.

Figure 6.4.8 Block Diagram of the Simplified G.711-to-G.722 Transcoder

6.4.2 Simulation Results

Note that this simplification is equivalent to the original design, so no quality degradation is expected. Table 6-3 lists the simulation results of the conventional and simplified G.711-to-G.722 transcoder, which match our expectation well. The reference signal is the original samples from TIMIT, before being downsampled to 8 kHz.

Table 6-3 Simulation Results of the G.711-to-G.722 Transcoders

Input File (G.711)	PESQ			PAMS		
	Conventional Transcoder	Simplified Transcoder	Degradation	Conventional Transcoder	Simplified Transcoder	Degradation
faks0s1573_8kHz	4.16	4.177	0.017	4.35	4.36	0.01
fasw0sa1_8kHz	4.22	4.24	0.02	4.19	4.2	0.01
fasw0sx200_8kHz	4.247	4.263	0.016	4.23	4.23	0
fbjl0sx292_8kHz	4.224	4.226	0.002	4.3	4.31	0.01
fbmh0sa2_8kHz	4.276	4.252	-0.024	4.31	4.32	0.01
fedr1sx106_8kHz	4.178	4.158	-0.02	4.07	4.04	-0.03
fejt0sx397_8kHz	4.139	4.169	0.03	4.32	4.31	-0.01
fdac1x394_8kHz	4.236	4.244	0.008	4.19	4.2	0.01
muhh0si1924_8kHz	4.159	4.14	-0.019	4.21	4.19	-0.02
mcomp0sa1_8kHz	4.126	4.133	0.007	4.19	4.21	0.02
mchl0sx87_8kHz	4.36	4.297	-0.063	4.45	4.45	0
mctt0sx298_8kHz	4.341	4.356	0.015	4.39	4.33	-0.06
mdac2si999_8kHz	4.252	4.255	0.003	4.21	4.25	0.04

mdac2sx279_8kHz	4.316	4.315	-0.001	4.46	4.49	0.03
mdlr1si669_8kHz	4.231	4.259	0.028	4.31	4.33	0.02
medr0sx114_8kHz	4.233	4.217	-0.016	4.15	4.13	-0.02
Average	4.236	4.235	-0.001	4.265	4.266	0.001

As for the G.722-to-G.711 transcoder, the above design can be applied to any other narrowband codec to G.722 transcoding. The only difference is replacing the G.711 decoder by the desired decoder.

6.5 Simplified Hybrid-Bandwidth-to-Wideband Mixer

6.5.1 Filter-Combination Mixer

The conventional design of the G.711 input + G.722 input to G.722 mixer is as Figure 6.5.1 shows.

Figure 6.5.1 Block Diagram of the Conventional G.711+G.722 -> G.722 Mixer

We can make use of the simplified G.711-to-G.722 transcoder in the G.711+G.722->G.722 mixer. The simplified mixer is shown in Figure 6.5.2.

Figure 6.5.2 Block Diagram of the Simplified G.711+G.722 -> G.722 Mixer

The advantages of the simplification are:

1. Decimator, Interpolator and the Receive QMF of the G.722 Decoder are eliminated,
2. LPF works at 8 kHz instead of 16 kHz.

Thus, $(2 \cdot 12 \cdot 8000 + 15 \cdot 16000) - (2 \cdot 19 \cdot 8000) = 128000$ times multiplication operations and $(2 \cdot 13 \cdot 8000 + 14 \cdot 16000) - (2 \cdot 18 \cdot 8000) = 144000$ times addition operations are saved per second. As for the complexity of the entire mixer, the multiplication operations drop about 21%, and the addition operations drop about 25%. These values are not an accurate measurement of the reduced complexity, and are just for reference purposes. Due to the elimination of the receive QMF, the algorithmic delay is reduced by 1.5 ms.

6.5.2 *Simulation Results*

Some experiments are made to verify the simplified structure. In order to find the reference signal, we need the original 16 kHz signal of the 8 kHz input (to be companded as u-law). The reference signal is obtained by mixing all the original 16 kHz inputs together, so it's the best quality signal. Due to the gain introduced by the filters, the 16 kHz signals are scaled according to the ratio of the amplitude of the signals mixed at the conventional mixer.

Below is an example. The reference is the fullband-processed signal, while the degraded signal is the lower-subband-only processed signal. Figure 6.5.3 and Figure 6.5.4 show the waveform view and spectral view.

Reference Signal:

mdac2sx279_8kHz_mix_md1r1si669_16kHz_G722_tandem_fullband.wav

Degraded (Simplified) Signal:

mdac2sx279_8kHz_mix_md1r1si669_16kHz_G722_tandem_subband.wav

Figure 6.5.3 2D Reference, Simplified, Error Signal View Using PAMS Analysis (G.711+G.722 -> G.722 Mixer)

Figure 6.5.4 Initial Spectrum View Using PAMS Analysis (G.711+G.722 -> G.722 Mixer)

We notice that the two signals are very similar, whether in the waveform view or spectral view. Table 6-4 lists the simulation results of the conventional and simplified G.711+G.722->G.722 mixers in two-speaker conferencing. By using the subband summation, the simplified mixer can provide a better audio quality, compared with the conventional design (improved by 0.161 in PESQ or 0.095 in PAMS).

Table 6-4 Simulation Results of the G.711 Input + G.722 Input ->G.722 Output Mixer

File 1 (G.711)	File 2 (G.722)	PESQ			PAMS		
		Conventional Mixer	Simplified Mixer	Degradation	Conventional Mixer	Simplified Mixer	Degradation
faks0sil573_8kHz	fasw0sa1_16kHz	3.757	3.885	0.128	4.18	4.21	0.03
faks0sil573_8kHz	fbjl0sx292_16kHz	3.804	3.919	0.115	3.97	4.03	0.06
fasw0sa200_8kHz	fdac1x394_16kHz	3.642	3.782	0.14	3.99	4.02	0.03
fbjl0sx292_8kHz	fdac1x394_16kHz	3.777	3.999	0.222	3.93	3.99	0.06
fbmh0sa2_8kHz	fedr1sx106_16kHz	3.134	3.67	0.536	3.72	3.97	0.25
fbmh0sa2_8kHz	fcjl0sx397_16kHz	3.537	3.939	0.402	3.89	4.06	0.17
fedr1sx106_8kHz	fcjl0sx397_16kHz	3.737	3.995	0.258	4.11	4.13	0.02
faks0sil573_8kHz	mahh0si1924_16kHz	3.957	3.723	-0.234	3.96	3.99	0.03
fasw0sa1_8kHz	mdlr1si669_16kHz	3.776	3.91	0.134	4.07	4.16	0.09
fasw0sa200_8kHz	mcmp0sa1_16kHz	3.525	3.705	0.18	3.78	4.02	0.24
fbjl0sx292_8kHz	mdac2sx279_16kHz	3.765	3.895	0.13	4.12	4.16	0.04
fbmh0sa2_8kHz	mctu0sx298_16kHz	3.693	3.861	0.168	4.26	4.37	0.11
medr0sx114_8kHz	fbmh0sa2_16kHz	3.788	3.913	0.125	4.17	4.2	0.03
mcmp0sa1_8kHz	fedr1sx106_16kHz	3.702	3.839	0.137	3.99	4.12	0.13
mdac2si999_8kHz	fcjl0sx397_16kHz	3.812	3.822	0.01	3.68	3.76	0.08
mchl0sx87_8kHz	fcjl0sx397_16kHz	3.732	3.964	0.232	4.02	4.13	0.11
mdlr1si669_8kHz	fdac1x394_16kHz	3.584	3.92	0.336	3.86	4.06	0.2
mahh0si1924_8kHz	mcmp0sa1_16kHz	3.68	3.793	0.113	4.07	4.14	0.07
mahh0si1924_8kHz	mdac2sx279_16kHz	3.937	4	0.063	4.41	4.41	0
mchl0sx87_8kHz	medr0sx114_16kHz	3.531	3.73	0.199	4.06	4.16	0.1
mchl0sx87_8kHz	mdac2si999_16kHz	3.756	3.871	0.115	4.06	4.1	0.04

mcomp0sa1_8kHz	mdac2sx279_16kHz	4.04	4.093	0.053	4.33	4.37	0.04
mcu0sx298_8kHz	medr0sx114_16kHz	3.513	3.678	0.165	3.98	4.1	0.12
mdac2sx279_8kHz	mdlr1si669_16kHz	3.638	3.784	0.146	3.98	4.2	0.22
Average		3.701	3.862	0.161	4.025	4.119	0.095

The three-speaker conferencing is simulated too. The first channel conforms to G.711, the second conforms to G.722, and the third conforms to G.729A. The conventional and simplified mixers give 3.806 and 3.793 in PESQ, and 4.02 and 4.03 in PAMS respectively.

6.6 Summary

The transcoders and mixers in hybrid-bandwidth VoIP conferencing are investigated. The conventional G.711 and G.722 transcoders are discussed and simplified. For the transcoding from G.722 to narrowband codec, the narrowband signal is obtained from the G.722 lower subband signal only by suppressing the high frequency band. For the transcoding from narrowband codec to G.722, the G.722 higher and lower subband signals are obtained by combining the polyphase filters. The simplified transcoders are utilized in the corresponding mixers. The simulation results show the simplified transcoders and mixers provide competitive audio quality while reducing the complexity and delay.

Chapter 7

Conclusion

7.1 Summary

Audio mixers for narrowband, wideband, and hybrid-bandwidth conferencing are investigated in this thesis. Specifically, the mixers with input conforming to such compression algorithms as G.711, G.729(A), and G.722 are analyzed and simplified with innovative methods. The transcoders of narrowband codec and G.722 are simplified too. The simulation results show that these simplification methods can achieve satisfactory performance. The simplified mixers and transcoders are summarized in Table 7-1 and Table 7-2.

Table 7-1 Summaries of the Simplification Methods of Audio Mixers

Mixer			Simplification Method	Complexity Reduced	Delay Reduced	Quality Improved or Degraded (Simulation Results)
Input 1	Input 2	Output				
G.729(A)	G.729(A)	G.729(A)	Strong Pitch Domination	- More than 7%	Algo. Delay: Same Processing Delay: Reduced	G.729: -0.001 G.729A: <-0.10
G.722	G.722	G.722	Subband Summation	- 3 QMF are eliminated	Algo. Delay: 3 ms Processing Delay: Reduced	+0.01
G.711 G.729(A)	G.722	G.711 G.729(A)	High Frequency Suppression	- Higher Subband Decoder and Receive QMF are eliminated - LPF works at 8 kHz instead of 16 kHz	Algo. Delay: 1.5 ms Processing Delay: Reduced	-0.02
G.711 G.729(A)	G.722	G.722	Filter Combination	- Decimator, Interpolator and the Receive QMF of the G.722 Decoder are eliminated - LPF works at 8 kHz instead of 16 kHz.	Algo. Delay: 1.5 ms Processing Delay: Reduced	+0.10

Table 7-2 Summaries of the Simplification Methods of Audio Transcoders

Transcoder		Simplification Method	Complexity Reduced	Delay Reduced	Quality Improved or Degraded (Simulation Results in PESQ)
Input	Output				
G.722	G.711 G.729(A)	High Frequency Suppression	- Higher Subband Decoder and Receive QMF are eliminated - LPF works at 8 kHz instead of 16 kHz	Algo. Delay: 1.5 ms Processing Delay: Reduced	-0.02
G.711 G.729(A)	G.722	Filter Combination	- Decimator, Interpolator and the Receive QMF of the G.722 Decoder are eliminated - LPF works at 8 kHz instead of 16 kHz.	Algo. Delay: same Processing Delay: Reduced	-0.001

Table 7-1 shows four types of audio mixers investigated and simplified for narrowband, wideband, and hybrid-bandwidth VoIP conferencing, respectively.

First, Strong Pitch Domination is employed to simplify the G.729(A) audio mixers, which skips the open-loop pitch search of the encoder for the mixed signal and replaces the open-loop pitch delay with the pitch delay of the input signal with strongest pitch energy.

Second, Subband Summation is employed by the G.722 mixer, which achieves the mixing in subband level, and hence the receive and transmit QMFs are eliminated.

Third, High Frequency Suppression idea is employed by the hybrid-bandwidth (G.711/G.729(A), and G.722) to narrowband (G.711/G.729(A)) mixer, which avoids the fullband processing, and processes the lower subband signal only to obtain the narrowband signal.

The hybrid-bandwidth (G.711/G.729(A), and G.722) to wideband (G.722) mixer employs the idea of Filter Combination; that is, process the narrowband signal by two combined filters to obtain the lower and higher subband signal directly.

Comparing with the conventional mixers, the simplified design can reduce the complexity, delay, while maintain or slightly degrade the audio quality. The first, third, and last designs are innovative approaches.

Corresponding to the last two mixers, the audio transcoders between G.711/G.729(A) and G.722 are simplified by High frequency Suppression and Filter Combination, respectively, as shown in Table 7-2.

7.2 Future Work

A new wideband compression algorithm, G.722.2 (Adaptive Multi-rate Wideband Codec, or AMR-WB) [48], was adopted by ITU-T in Jan. 2002. It's the common wideband codec of Third Generation Partnership Project (3GPP) and ITU-T. It has 9 modes and the bit rates are 23.85, 23.05, 19.85, 18.25, 15.85, 14.25, 12.65, 8.85, 6.6 kbps respectively. The algorithmic delay is 20ms frame size plus 5ms lookahead. The complexity is about 38 WMOPS and it has good performance at rates 12.65 kbps and higher. What's more, G.722.2 has a built-in silence suppression mechanism including a VAD function, and CNG function in the decoder. Table 7-3 illustrates the features of G.722, G.722.1, and G.722.2.

Table 7-3 Features of G.722, G.722.1, and G.722.2

Standard		G.722	G.722.1	G.722.2
Date		1988	1999	2002
Bit Rate (kbps)		48,56,64	24,32	23.85, 23.05, 19.85, 18.25, 15.85, 14.25, 12.65, 8.85, 6.6
Type		Subband ADPCM	Transform Coding	Algebraic Code Excited Linear Prediction (ACELP)
Delay	Frame Size	0.125 ms	20 ms	20 ms
	Lookahead	1.5 ms	20 ms	5 ms
Quality		Complimentary (at 64 kbps)	Poor speech performance in some operation conditions; scope of standard limited to hands-free and low packet loss rates Good music performance	Good speech performance at rates 12.65 kbps and higher 15.85 >= G.722 @ 56 23.05 >= G.722 @ 64 Poor music performance
Complexity		10 MIPS	<15 MIPS	38 MOPS
RAM		1 K	2 K	5.3 K
VAD/DTX/CNG		None	None	Exists
Main Application		ISDN; Video Conferencing	Same + VoPN	3G Wireless + Same as G.722.1

Table 7-4 Similarities and Differences of G.722.2 and G.729(A)

Algorithm	G.722.2	G.729(A)
Input Bandwidth	50 Hz ~ 7 kHz	200 Hz ~ 3400 Hz
Sampling Rate	16 kHz (12.8 kHz Analysis)	8 kHz
Frame	20 ms	10 ms
Subframe	5 ms	5 ms
LP Filter	16-th order (30 ms window, 5 ms lookahead)	10-th order (30 ms window, 5 ms lookahead)
Open-loop Pitch Analysis	10 ms (20 ms for 6.6 kbps mode)	10ms
Pitch Range	34 ~ 231 (Different resolution for different modes)	19 ~ 143 (resolution $\frac{1}{3}$ in [19, 84], 1 in [85, 143])
Fixed Codebook	24/16/12/10/8/4/2 pulses	4 pulses

G.722.2 adopts the ACELP model, thus it performs the analysis of LPC, adaptive and fixed codebook parameters as G.729 does. Table 7-4 lists some similarities and differences of G.722.2 and G.729(A).

Since G.722.2 deploys the ACELP model too, the same Strong-Pitch-Domination idea for the G.729(A) mixer can be applied to the G.722.2 wideband mixer.

For the mixer of inputs conforming to G.722.2 and G.729(A) respectively, is it possible to map the LP coefficients, adaptive and fixed codebooks?

We know that G.729(A) models the vocal tract with a 10-th all-pole filter working at 8 kHz, while G.722.2 applies a 16-th all-pole filter working at 12.8 kHz. This means we need to map the 16 poles of G.722.2 codec with the 10 poles of G.729(A) codec. So we should find 10 poles in $-0.625\pi \sim 0.625\pi$. The poles have a direct relationship with the formants. Since human perception is sensitive to the first 3 formants, the imprecision of the other formants' positions will not degrade the quality much. We can make use of this perceptual mask feature to design the formant mapping.

For the adaptive codebook mapping, we can have a look at the pitch range. G.729's pitch delay ranges from 19 to 143, so it corresponds to 55.94 Hz $\sim$ 421 Hz. G.722.2's pitch delay ranges from 34 to 231, so it corresponds to 55.41 Hz to 376 Hz. The G.722.2 has better resolution in the lower frequency, and for the unvoiced speech towards higher

frequency, it refines the LPC excitation signal with more pulses. Thus, it's not a big problem to map the pitches with the fixed codebook search assistance.

The fixed codebook mapping requires the transformation of the 4 pulses of G.729(A) and the 24/16/12/10/8/4/2 pulses of G.722.2. Due to different models employed by the two algorithms, it seems hard to skip the algebraic codebook search.

All these need further investigation and proof.

References

- [1] P. C. Mehta, and S. Udani, "Overview of Voice over IP," technical report, University of Pennsylvania, <http://citeseer.nj.nec.com/467089.html>, Feb. 2001.
- [2] M. Handley, J. Crowcroft, C. Bormann, "The Internet Multimedia Conferencing Architecture," Internet Draft, Internet Engineering Task Force, Feb. 1996.
- [3] ITU-T Rec. H.323, "Packet-based Multimedia Communications Systems," Nov. 2000.
- [4] H. Schulzrinne, J. Rosenberg, "Internet Telephony: Architecture and Protocols an IETF Perspective," <http://citeseer.nj.nec.com/schulzrinne98internet.html>, July 1998.
- [5] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, and E. Schooler, "SIP: Session Initiation Protocol," RFC 3261, Internet Engineering Task Force, June 2002.
- [6] K. Singh, G. Nair, and H. Schulzrinne, "Centralized Conferencing Using SIP," Proc. 2nd IP-Telephony Workshop (IPTel 2001), New York, NY, Apr. 2001.

-
- [7] F. Andreassen, and B. Foster, "Media Gateway Control Protocol (MGCP)," RFC 3435, Internet Engineering Task Force, Jan. 2003.
 - [8] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications," RFC 1889, Internet Engineering Task Force, Jan. 1996.
 - [9] H. Schulzrinne, "RTP Profile for Audio and Video Conferences with Minimal Control," RFC 1890, Internet Engineering Task Force, Jan. 1996.
 - [10] P. J. Smith, P. Kabal, M. Blostein, and R. Rabipour, "Tandem Free Operation for VoIP Conference Bridges," Internal memo, Nortel Networks, Canada, 2002.
 - [11] D. N. Horn, and A. Sharma, "A Versatile Audio Bridge for Multimedia Conferencing," IEEE International Conference on Communications, vol. 3, pp. 1754-1762, May 1994.
 - [12] T. Braun, and L. Liu, "Multicast for Small Conferences", Proceedings of Sixth IEEE Symposium on Computers and Communications, pp. 145-150, 2001.
 - [13] J. Forgie, C. Feehrer, and P. Weene, "Voice Conferencing Technology Final Report," Tech. Rep. DDC AD-A074498, M.I.T. Lincoln Lab., Lexington, MA, Mar. 1979.
 - [14] T. G. Champion, "Multi-speaker Conferencing over Narrowband Channels," Proceedings of IEEE Military Communications Conference, Washington, D.C., pp. 1220-1223, Nov. 1991.
-

-
- [15] P. J. Smith, "Voice Conferencing over IP Networks", <http://www.tsp.ece.mcgill.ca/Theses/2002/SmithT2002.pdf>, Jan. 2002.
- [16] C. Ziegler, G. Weiss, and E. Friedman, "Implementation Mechanisms for Packet Switched Voice Conferencing," *IEEE Journal on Selected Areas in Communications*, vol. 7, pp. 698-706, June 1989.
- [17] P.T. Brady, "A Model for Generating On-Off Speech Patterns in Two-way Conversation," *Bell System Technical Journal*, vol. 48, pp. 2445-2471, Sept. 1969.
- [18] ITU-T Rec. P.59, "Artificial Conversational Speech," Mar. 1993.
- [19] J. D. Tardelli, P. D. Gatewood, E. W. Kreamer, and P. A. La Fol-lette, "The Benefits of Multi-speaker Conferencing and the Design of Conference Bridge Control Algorithms," *Proceedings of IEEE International Conference on Acoustics, Speech, Signal Processing*, Minneapolis, MN, vol. 2, pp. 435-438, Apr. 1993.
- [20] P. Noll, "Digital Audio Coding for Visual Communications," *Proceedings of the IEEE*, vol. 83, pp.925-943, June 1995.
- [21] J. R. Deller, J.G. Proakis, and J. H.L. Hansen, "Discrete-Time Processing of Speech Signals," New York, NY, IEEE Press and John Wiley & Sons Inc., pp. 99-348, 1993.
- [22] J.A. Marks, "Real Time Speech Classification and Pitch Detection," *Proceedings of IEEE Southern African Conference on Communications and Signal Processing*, pp.1-6, June 1988.
-

-
- [23] "Wideband Speech Coding Standards and Applications", White Paper, VoiceAge Corporation, Montreal, Canada, Sept. 2001.
- [24] R. V. Cox and P. Kroon, "Low Bit-Rate Speech Coders for Multimedia Communication," IEEE Communications Magazine, vol. 34, pp. 34–41, Dec. 1996.
- [25] R. Salami, C. Laflamme, B. Bessette, and J.-P. Adoul, "Description of ITU-T Recommendation G.729 Annex A: Reduced Complexity 8 kbit/s CS-ACELP Codec," IEEE International Conference on Acoustics, Speech, and Signal Processing, vol.2, pp. 775-778, Apr. 1997.
- [26] ITU-T SG16 Liaison Statement, "Liaison to Multiple SDOs Requesting Input for "Media Coding Summary Database" Project," www.3gpp.org/ftp/tsg_sa/TSG_SA/TSGS_18/Docs/pdf/SP-020635.pdf, Geneva, Oct. 2002
- [27] M.E. Perkins, K. Evans, D. Pascal, and L.A. Thorpe, "Characterizing the Subjective Performance of the ITU-T 8 kb/s Speech Coding Algorithm - ITU-T G.729," IEEE Communications Magazine, vol. 35, pp.74-81, Sept. 1997.
- [28] C. Perkins, O. Hodson, and V. Hardman, "A Survey of Packet Loss Recovery Techniques for Streaming Audio," IEEE Network, vol. 12, no. 5, pp. 40-48, Sept./Oct. 1998.
- [29] L. Zheng, L. Zhang, and D. Xu, "Characteristics of Network Delay and Delay Jitter and Its Effect on Voice over IP (VoIP)," IEEE International Conference on Communications, vol. 1, pp.122-126, 2001.
-

-
- [30] ITU-T Rec. G.711, "Pulse Code Modulation (PCM) of Voice Frequencies," Nov. 1972.
- [31] R. Salami, C. Laflamme, J.-P. Adoul, A. Kataoka, S. Hayashi, T. Moriya, C. Lamblin, D. Massaloux, S. Proust, P. Kroon, and Y. Shoham, "Design and Description of CS-ACELP: A Toll Quality 8 kb/s Speech Coder," *IEEE Transactions on Speech and Audio Processing*, vol. 6, pp. 116-130, March 1998.
- [32] A. Kataoka, J.-P. Adoul, P. Combescure, and P. Kroon, "ITU-T 8-kbit/s Standard Speech Codec for Personal Communication Services," *Record of Fourth IEEE International Conference on Universal Personal Communications*, pp. 6-10, Nov. 1995.
- [33] R. Salami, C. Laflamme, J.-P. Adoul, and D. Massaloux, "A Toll Quality 8 kb/s Speech Codec for the Personal Communications Systems (PCS) ," *IEEE Transaction on Vehicular Technology*, vol. 43, pp. 808-816, Aug. 1994.
- [34] ITU-T Rec. G.729, "Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP)," March 1996.
- [35] ITU-T Rec. G.723.1, "Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 kbit/s," March 1996.
- [36] ITU-T Rec. G.729 Annex A, "Reduced Complexity 8 kbit/s CS-ACELP Speech Codec," Nov. 1996.
- [37] ITU-T Rec. G.722, "7 kHz Audio – Coding within 64 kbit/s," Nov. 1988.
-

-
- [38] ITU-T Rec. P.800, "Methods for Subjective Determination of Transmission Quality," Aug. 1996.
- [39] PsyTechnics Group, British Telecommunications, "PAMS Usage Guidelines," Feb. 2000.
- [40] ITU-T Rec. P.862, "Perceptual Evaluation of Speech Quality (PESQ), An Objective Method for End-to-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs," Feb. 2001.
- [41] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra, "Perceptual Evaluation of Speech Quality (PESQ) - A New Method for Speech Quality Assessment of Telephone Networks and Codecs," Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, pp. 749 -752, 2001.
- [42] ITU-T SG12 Delayed Contribution [1], "Results of Quality Assessment of Wideband Speech Using PAMS," received from British Telecom (UK) on 2001-02-07.
- [43] ITU-T SG12 Delayed Contribution [7], "Proposed Modification to Draft P.862 to Allow PESQ to be Used for Quality Assessment of Wideband Speech," received from British Telecom (UK) and KPN (the Netherlands) on 2001-02-07.
- [44] TIMIT Acoustic-Phonetic Continuous Speech Corpus,
<http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1>.
-

-
- [45] R. Salami, C. Laflamme, B. Bessette, and J.-P. Adoul, "ITU-T G.729 Annex A: Reduced Complexity 8 kb/s CS-ACELP Codec for Digital Simultaneous Voice and Data," *IEEE Communications Magazine*, vol. 35, pp. 56–63, Sept. 1997.
- [46] P. Mermelstein, "G.722, A New CCITT Coding Standard for Digital Transmission of Wideband Audio Signals", *IEEE Communications Magazine*, vol. 26, pp. 8-15, Jan. 1988.
- [47] N.J. Fliege, "Multirate Digital Signal Processing: Multirate Systems, Filter Banks, and Wavelets," New York, NY, John Wiley & Sons Inc., pp. 141-152, 1994.
- [48] ITU-T Rec. G.722.2, "Wideband Coding of Speech at Around 16 kbit/s Using Adaptive Multi-rate Wideband (AMR-WB)," Jan. 2002.
-