INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

STRUCTURES AND INTERFACES FOR VOICE OVER IP

by

Trevor Yensen, 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 Electrical Engineering

Ottawa-Carleton Institute for Electrical and Computer Engineering
Faculty of Engineering
Department of Systems and Computer Engineering
Carleton University
Ottawa, Ontario, Canada
K1S 5B6

September 17, 1998

© Copyright 1998, Trevor Yensen

National Library
of Canada

Acquisitions and
Bibliographic Services

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale
du Canada

Acquisitions et
services bibliographiques

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

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

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

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

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

0-612-36903-X

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

"STRUCTURES AND INTERFACES FOR VOICE OVER IP"

submitted by Trevor Yensen, B.Eng.

in partial fulfillment of the requirements for

the degree of Masters of Engineering

Department Chair

Thesis Supervisor

Thesis Supervisor

CARLETON UNIVERSITY
Date 09/17/98

Abstract

Computer telephony integration (CTI) is a new communications technology that provides enhanced services not offered by traditional telephone networks. CTI integrates voice communication technology such as voice over Internet protocol (VoIP) with video conferencing and the ability to exchange data. This thesis explores innovative new structures and interfaces, which will make VoIP a more marketable communication technology for corporate Intranet use.

This thesis proposes a new synthetic stereo echo cancellation structure for multiple participant full-duplex hands-free conferencing. The synthetic stereo structure uses spatialization functions to acoustically separate multiple talkers and improve the user's ability to distinguish between talkers.

A new method to compute echo cancellation targets for digital VoIP networks is also proposed. The mean one-way propagation delay determines acoustic echo cancellation targets using psycho-acoustic data. The mean one-way propagation delay can be approximated by the H.323 compliant acoustic round trip delay empirical measurement method proposed in this thesis.

For this thesis, a VoIP platform was designed to demonstrate the synthetic stereo canceller and compute acoustic round trip delays.

Acknowledgements

I would like to thank my supervisors and research associates for their assistance during the course of my thesis: Dr. Rafik Goubran and Dr. John Lambadaris, for their guidance and support in developing this research project to its full potential; Marios Parperis, for his role in the development of the real-time transport protocol (RTP) layer of the Carleton University DSP multi-participant Intranet phone system (dMIPS) platform and his extensive participation in the development of the acoustic round trip delay empirical measurement method and theoretical model; Meliha Ferhatbegovic for her assistance in debugging the MIPS platform.

I would like to thank Jill Renaud for her many years of encouragement and for her hard work in editing many of my papers, projects and this thesis.

I would like to thank my family for their many years of support and encouragement during my university career.

The support staff at Carleton University deserves thanks for their assistance: Danny Lemay and the other lab technicians for their help in solving technical problems; the secretarial staff in the Department of Systems and Computer Engineering for all their help.

Finally, I would like to thank the Natural Sciences and Engineering Research Council of Canada, Nortel and Carleton University for their financial support.

Table of Contents

ABSTRACT	III
ACKNOWLEDGEMENTS	IV
TABLE OF CONTENTS.....	V
LIST OF FIGURES.....	XI
LIST OF TABLES.....	XIII
LIST OF SYMBOLS.....	XIV
1 INTRODUCTION.....	1
1.1 PROBLEM STATEMENT.....	1
1.1.1 <i>Hands-Free Full-Duplex Conferences</i>	1
1.1.2 <i>Acoustic Echo Cancellation</i>	2
1.1.3 <i>Stereophonic Acoustic Echo Cancellation</i>	3
1.2 CONTRIBUTIONS TO ENGINEERING KNOWLEDGE	3
1.2.1 <i>Synthetic Stereo Acoustic Echo Cancellation Structure</i>	3
1.2.2 <i>Psycho-Acoustic Determination of Echo Cancellation Targets</i>	4
1.2.3 <i>Acoustic Round Trip Delay Computation</i>	5
1.2.4 <i>VoIP Test Platform</i>	6
2 BACKGROUND INFORMATION.....	7
2.1 HALF-DUPLEX CONFERENCING.....	7
2.1.1 <i>Press-to-Talk</i>	7
2.1.2 <i>Electronically Switched</i>	8
2.1.3 <i>Switched-Loss</i>	8

2.2	FULL-DUPLEX CONFERENCING	9
2.2.1	<i>Acoustic Echo Cancellation</i>	9
2.2.1.1	Acoustic Echo Canceller Adaptation	10
2.2.1.2	Near-End Speech	11
2.2.1.3	ERLE Computation	11
2.2.1.4	Maximum ERLE Computation	12
2.2.1.5	ERLE Target Determination.....	13
2.2.1.6	Limitations of Acoustic Echo Cancellers.....	14
2.2.2	<i>Acoustic Echo Cancellation with Switched-Loss</i>	15
2.3	TCP/IP PROTOCOL SUITE.....	15
2.3.1	<i>User Datagram Protocol</i>	16
2.3.2	<i>Transmission Control Protocol</i>	17
2.3.3	<i>Internet Protocol</i>	17
2.4	ITU-T H.323 STANDARD	20
2.4.1	<i>Multimedia Services</i>	21
2.4.2	<i>H.32x Interoperability</i>	22
2.4.3	<i>Multipoint Conferences</i>	22
2.4.4	<i>Gateways</i>	24
2.4.5	<i>Gatekeepers</i>	24
2.4.6	<i>Call Setup</i>	25
2.4.7	<i>Logical Channels</i>	26
2.4.8	<i>Quality of Service</i>	29
2.5	DIGITAL SIGNAL PROCESSING ON A FIXED-POINT PROCESSOR	30
2.5.1	<i>Q.F Notation</i>	30
2.5.2	<i>Q.0 Notation</i>	31
2.5.3	<i>Q.x Operation Rules</i>	31
3	REVIEW OF THE STATE OF THE ART.....	33
3.1	STEREOPHONIC ACOUSTIC ECHO CANCELLATION STRUCTURES	33

3.1.1	<i>Stereophonic Acoustic Echo Cancellation Using a Multiple Canceller Structure</i>	33
3.1.2	<i>Stereophonic Acoustic Echo Cancellation Using a Single Canceller Structure</i>	37
3.1.3	<i>Stereophonic Acoustic Echo Cancellation Using a Subband Structure</i>	39
3.1.4	<i>Stereophonic Acoustic Echo Cancellation Structures with Pre-processing</i>	41
3.1.4.1	Pre-processing with the Addition of a Nonlinearity.....	42
3.1.4.2	Pre-processing with the Addition of a Nonlinearity and Comb Filtering	43
3.1.4.3	Pre-processing with Random Noise Controlled by Auditory Properties	43
3.1.4.4	Pre-processing with Adaptation Based Solely on the Uncorrelated Signals	44
3.1.4.5	Pre-processing with Periodically Time Delayed Input Signals	44
3.2	ACOUSTIC ECHO CANCELLATION ALGORITHMS.....	44
3.2.1	<i>The Least Mean Square Algorithm (LMS)</i>	45
3.2.2	<i>The Recursive Least Square Algorithm (RLS)</i>	46
3.2.3	<i>The Affine Projection Algorithm (APA)</i>	47
3.3	SPATIALIZATION TECHNIQUES.....	49
3.3.1	<i>ILD and ITD Techniques</i>	50
3.3.2	<i>Head Related Transfer Function Technique (HRTF)</i>	51
3.3.3	<i>Crosstalk Cancellation</i>	52
4	SYNTHETIC STEREO ECHO CANCELLATION STRUCTURE.....	54
4.1	INTRODUCTION TO THE SYNTHETIC STEREO STRUCTURE.....	54
4.2	STEREOPHONIC AND SYNTHETIC STEREO HIGH-LEVEL DIFFERENCES.....	55
4.3	EXPLANATION OF THE SYNTHETIC STEREO STRUCTURE.....	56
4.4	SYNTHETIC STRUCTURE FOR MULTIPLE ACTIVE TALKERS.....	59
4.5	SYNTHETIC STEREO IMPLEMENTATION ISSUES.....	61
4.6	SPATIALIZATION FUNCTIONS	62
4.7	SIMULATION	63
4.8	STEREOPHONIC AND SYNTHETIC STEREO PROPERTIES	64
5	ACOUSTIC ROUND TRIP DELAY	66

5.1	MEASURING THE QUALITY OF SERVICE WITH ACOUSTIC ROUND TRIP DELAY.....	66
5.2	SWITCHED-LOSS LEVEL ADJUSTMENT WITH ACOUSTIC ROUND TRIP DELAY.....	67
5.3	THEORETICAL MODEL FOR THE ACOUSTIC ROUND TRIP DELAY.....	69
5.3.1	<i>Assumptions for the Theoretical Model.....</i>	70
5.3.2	<i>Description of the Delay Components.....</i>	71
5.3.3	<i>Pipeline Analogy for the Theoretical Model.....</i>	73
5.3.4	<i>Equations for the Theoretical Model.....</i>	76
5.4	EMPIRICAL MEASUREMENT METHOD FOR THE ACOUSTIC ROUND TRIP DELAY.....	76
5.4.1	<i>Acoustic Round Trip Delay Measurement.....</i>	79
5.4.2	<i>DSP/PC Platform Results.....</i>	80
5.5	ACOUSTIC DOMAIN VERIFICATION TEST.....	82
5.5.1	<i>Acoustic Domain Verification Test Procedure.....</i>	83
5.5.2	<i>Acoustic Domain Verification Test Results.....</i>	84
6	REAL-TIME VOIP TEST PLATFORM.....	86
6.1	DSP/PC PLATFORM HARDWARE REQUIREMENTS.....	87
6.2	VOIP ACOUSTIC FRONT-END.....	87
6.3	VOIP SOFTWARE REQUIREMENTS.....	88
6.4	VOIP AUDIO DATA FORMAT.....	90
6.5	H.323 COMPATIBILITY.....	90
6.5.1	<i>Connectionless Service Requirement.....</i>	91
6.5.2	<i>Multipoint Support.....</i>	91
6.6	DSP IMPLEMENTATION OF AN NLMS ACOUSTIC ECHO CANCELLER.....	92
6.7	DSP IMPLEMENTATION OF AN FIR SPATIALIZATION FILTER.....	94
7	CONCLUSIONS.....	95
7.1	SYNTHETIC STEREO ACOUSTIC ECHO CANCELLATION STRUCTURE.....	95
7.1.1	<i>Future Research in Synthetic Stereo Cancellation.....</i>	95
7.2	PSYCHO-ACOUSTIC DETERMINATION OF ECHO CANCELLATION TARGETS.....	96

7.2.1	<i>Future Research in Psycho-Acoustic Echo Cancellation Targets</i>	96
7.3	ACOUSTIC ROUND TRIP DELAY COMPUTATION	97
7.3.1	<i>Future Research in Acoustic Round Trip Delay</i>	97
7.4	VOIP PLATFORM DESIGN	98
7.4.1	<i>Future Research in VoIP Platform Design</i>	98
8	REFERENCES	99
APPENDIX A - IMPULSE RESPONSE COMPUTATION		103
APPENDIX B - ADAPTIVE ALGORITHMS		106
B.1	FAST AFFINE PROJECTION ALGORITHM	106
B.2	RECURSIVE LEAST SQUARE ALGORITHM	108
APPENDIX C – SOFTWARE DESIGN		109
C.1	USE CASE MAP SYMBOLS	109
C.2	HIGH-LEVEL USE CASE MAP	110
C.3	CAPTURE DATA AND ACOUSTIC ECHO CANCELLATION USE CASE MAP	111
C.4	PLAYBACK DATA USE CASE MAP	113
C.5	JITTER BUFFER USE CASE MAP	115
C.6	SEND DATA USE CASE MAP	116
C.7	RECEIVE DATA USE CASE MAP	117
C.8	JITTER BUFFER STRUCTURE FOR MULTIPLE PARTICIPANTS	119
C.9	HOST PORT INTERFACE AND TRANSFER REGISTERS	120
C.10	PC CONFERENCE STATUS WORD	121
C.11	DSP CONFERENCE STATUS WORD	122
C.12	CONTROL INFORMATION AND SPATIALIZATION FUNCTIONS	123
C.13	CAPTURE AND JITTER BUFFER	124
C.14	ROUND TRIP DELAY INFORMATION	125
C.15	CONTROL USE CASE MAP	126

APPENDIX D - C54X CODE FOR NLMS ALGORITHM	127
APPENDIX E - C54X CODE FOR SPATIALIZATION ALGORITHM	132

List of Figures

FIGURE 1 - HALF-DUPLEX STRUCTURE.....	7
FIGURE 2 – BASIC ACOUSTIC ECHO CANCELLATION STRUCTURE.....	9
FIGURE 3 - TALKER ECHO TOLERANCE CURVES	13
FIGURE 4 - ACOUSTIC ECHO CANCELLATION STRUCTURE WITH SWITCHED-LOSS.....	15
FIGURE 5 - TCP/IP PROTOCOL STACK.....	16
FIGURE 6 - IP PACKET FORMAT	19
FIGURE 7 - H.323 TERMINAL EQUIPMENT.....	21
FIGURE 8 – STEREOPHONIC ACOUSTIC ECHO CANCELLATION WITH A MULTIPLE CANCELLER STRUCTURE...	36
FIGURE 9 - STEREOPHONIC ACOUSTIC ECHO CANCELLATION WITH A SINGLE CANCELLER STRUCTURE.....	38
FIGURE 10 - STEREOPHONIC ACOUSTIC ECHO CANCELLATION WITH A SUBBAND STRUCTURE	40
FIGURE 11 - STEREOPHONIC ACOUSTIC ECHO CANCELLATION STRUCTURE WITH PREPROCESSING	42
FIGURE 12 – TRANSAURAL MONITORING WITH CROSSTALK	53
FIGURE 13 - TRANSAURAL STEREO WITH HRTF SYNTHESIS AND CROSSTALK CANCELLATION.....	53
FIGURE 14 - MULTIPLE PARTICIPANT SPATIALIZATION	55
FIGURE 15 - HIGHLEVEL STEREOPHONIC STRUCTURE FOR VOIP	56
FIGURE 16 - HIGHLEVEL SYNTHETIC STEREO STRUCTURE FOR VOIP	57
FIGURE 17 - SYNTHETIC STEREO ACOUSTIC ECHO CANCELLATION STRUCTURE	58
FIGURE 18 - SYNTHETIC STEREO ACOUSTIC ECHO CANCELLATION STRUCTURE FOR TWO TALKERS	60
FIGURE 19 – TRANSFER FUNCTIONS FOR MEASURED ROOM.....	64
FIGURE 20 - SYNTHETIC STEREO CANCELLER CONVERGENCE	65
FIGURE 21 – SWITCHED-LOSS LEVEL COMPUTATION.....	68
FIGURE 22 - SOURCES OF DELAY IN A VOIP SYSTEM.....	70
FIGURE 23 – ACOUSTIC ROUND TRIP DELAY PIPELINE ANALOGY.....	75
FIGURE 24 – ACOUSTIC ROUND TRIP DELAY EMPIRICAL MEASUREMENT METHOD	77
FIGURE 25 – DSP/PC ACOUSTIC ROUND TRIP DELAY MEASUREMENTS FOR 20MS FRAME	81

FIGURE 26 – DSP/PC ACOUSTIC ROUND TRIP DELAY MEASUREMENTS FOR 40MS FRAME	81
FIGURE 27 - ACOUSTIC DOMAIN VERIFICATION TEST	83
FIGURE 28 - ACOUSTIC DOMAIN VERIFICATION TEST RESULTS	85
FIGURE 29 – HIGH-LEVEL VIEW OF VOIP PLATFORM	86
FIGURE 30 – MULTIMEDIA CONFIGURATION FOR VOIP PLATFORM	88
FIGURE 31 - ACOUSTIC FRONT-END FOR VOIP PLATFORM	89
FIGURE 32 - LEFT CHANNEL IMPULSE RESPONSE COMPUTATION	104
FIGURE 33 - MEASUREMENT OF SOUND CARD SKEW.....	105
FIGURE 34 - USE CASE MAP SYMBOLS	109
FIGURE 35 – HIGH-LEVEL USE CASE MAP FOR DSP/PC PLATFORM.....	110
FIGURE 36 – CAPTURE DATA AND ACOUSTIC ECHO CANCELLATION USE CASE MAP	111
FIGURE 37 – PLAYBACK DATA USE CASE MAP	113
FIGURE 38 - JITTER BUFFER USE CASE MAP	115
FIGURE 39 - SEND DATA USE CASE MAPS.....	116
FIGURE 40 - RECEIVE DATA USE CASE MAP.....	117
FIGURE 41 - JITTER BUFFER STRUCTURE.....	119
FIGURE 42 - HOST PORT INTERFACE AND TRANSFER REGISTERS.....	120
FIGURE 43 - PC CONFERENCE STATUS WORD.....	121
FIGURE 44 - DSP CONFERENCE STATUS WORD	122
FIGURE 45 - CONTROL INFORMATION AND SPATIALIZATION FUNCTIONS	123
FIGURE 46 - CAPTURE AND JITTER BUFFER	124
FIGURE 47 - ROUND TRIP DELAY INFORMATION	125
FIGURE 48 - CONTROL USE CASE MAP.....	126

List of Tables

TABLE 1 – LIMITATIONS OF ACOUSTIC ECHO CANCELLERS	14
TABLE 2 - CALL SETUP SIGNALING FOR H.323 CONFERENCE	26
TABLE 3 - DESCRIPTION OF ACOUSTIC ROUND TRIP DELAY COMPONENTS	72
TABLE 4 – CAPTURE DATA AND ACOUSTIC ECHO CANCELLATION ACTIVITY DESCRIPTIONS	112
TABLE 5 – PLAYBACK DATA ACTIVITY DESCRIPTIONS	114
TABLE 6 - JITTER BUFFER ACTIVITY DESCRIPTIONS	115
TABLE 7 - SEND DATA ACTIVITY DESCRIPTIONS	116
TABLE 8 - RECEIVE DATA ACTIVITY DESCRIPTIONS	118

List of Symbols

APA	Affine Projection Algorithm
ATM	Asynchronous Transfer Mode
BRLS	Block Based Recursive Least Square Algorithm
Codec	Coder/Decoder
CTI	Computer Telephony Integration
DMA	Direct Memory Access
dMIPS	DSP Multi-Participant Intranet Phone System (Carleton University Platform)
DNS	Domain Name Server
DSP	Digital Signal Processing
EGP	Exterior Gateway Protocol
ERLE	Echo Return Loss Enhancement
ETSI	European Telecommunications Standards Institute
FAP	Fast Affine Projection (Fast APA)
FIR	Finite Impulse Response
HRTF	Head Related Transfer Function
IETF	Internet Engineering Task Force
IGP	Interior Gateway Protocol
ILD	Interaural Level Difference
I/O	Input/Output
IP	Internet Protocol
IPv4	Internet Protocol version 4
IPv6	Internet Protocol version 6 (also Internet Protocol Next Generation, IPng)
ISDN	Integrated Services Digital Network
IsoEthernet	Isochronous Ethernet
ITD	Interaural Time Difference
ITU-T	International Telecommunications Union – Telephony Division
LAN	Local Area Network
LMS	Least Mean Square Algorithm
MC	Multipoint Controller

MCU	Multipoint Control Unit
MIPS	Multi-participant Intranet Phone System (Carleton University Platform)
MP	Multipoint Processor
NLMS	Normalized Least Mean Square Algorithm
NLP	Nonlinear Processor
PC	Personal Computer
PDU	Protocol Data Unit
PIO	Programmed Input/Output
PSTN	Public Switched Telephone Network
QoS	Quality-of-Service
RLS	Recursive Least Square Algorithm
RSVP	Reservation Protocol
RTCP	RTP Control Protocol
RTP	Real-Time Transport Protocol
RTSP	Real-Time Streaming Protocol
TCP	Transmission Control Protocol
TCP/IP	Transmission Control Protocol/Internet Protocol
TELR	Talker Echo Loudness Rating
TSAP	Transport Service Access Point
TTL	Time-To-Live
UDP	User Datagram Protocol
VoIP	Voice Over Internet Protocol

1 Introduction

1.1 Problem Statement

Multimedia network computers and digital telephone sets equipped with network interfaces are the future of communication systems [9]. An integrated medium, such as computer telephony integration (CTI), is required to provide voice communication, video conferencing and data exchange capabilities for these types of systems. This thesis examines new acoustic front-end designs for hands-free full-duplex stereo voice communication technologies such as voice over IP (VoIP) for corporate Intranet usage. Three inter-related technologies, required for front-end design, are proposed in this thesis.

The first technology is a multiple participant acoustic echo cancellation structure for hands-free full-duplex synthetic stereo conferences. The second technology is a method of determining acoustic echo cancellation targets through psycho-acoustic data. The third technology is an empirical method of computing the acoustic round trip delay.

1.1.1 Hands-Free Full-Duplex Conferences

Hands-free technology is important in conference systems because it frees the user from having to use a handset or headset. The user is free to perform other activities such as moving about freely in the conference area. VoIP systems should be capable of hands-free operation for user convenience.

Modern conferencing systems must be capable of hands-free full-duplex operation because conversation flow is more natural to users than in hands-free half-duplex systems. Full-duplex systems switch instantaneously between talkers, unlike half-duplex systems, which cut off speech and create switching noise. In half-duplex systems the loudest talker or an environmental noise could cause the system to switch to that end allowing them to dominate the conversation. There are many advantages to conferences that use full-duplex technology but they require acoustic echo cancellation for hands-free conferencing.

1.1.2 Acoustic Echo Cancellation

Hands-free full-duplex VoIP products are plagued by large amounts of far-end talker acoustic echo. These echoes become unbearable when combined with large Internet protocol (IP) network latencies. To achieve full-duplex conferencing in a hands-free environment, acoustic echo cancellers must be employed to cancel the feedback from the speakers to the microphone transducers. In the past, designers had to resort to half-duplex technology since electronic echo cancellation was not feasible. Today high-powered digital signal processors (DSPs) are available that are capable of handling the complex mathematical operations required to implement an acoustic echo canceller.

1.1.3 Stereophonic Acoustic Echo Cancellation

Stereophonic conferencing is desirable for VoIP systems because it provides localization information not available in monaural systems. The far-end talker employs stereo microphone transducers that make their spatial information available to the near-end user for peer-to-peer conferencing.

Stereophonic acoustic echo cancellation has been the focus of many recent papers. The main problem with stereophonic acoustic echo cancellation is cross-correlation between the primary input signals to the acoustic echo canceller. The structures detailed in sections 3.1.1, 3.1.2, 3.1.3, and 3.1.4 attempt to resolve the cross-correlation problem. However, these structures are not acceptable solutions to the stereophonic acoustic echo cancellation problem because they exhibit one or more of the following characteristics: slow convergence speed, canceller misconvergence, a lack of tracking ability, excessive computational complexity or they add noise objectionable to the user.

1.2 Contributions to Engineering Knowledge

1.2.1 Synthetic Stereo Acoustic Echo Cancellation Structure

This thesis proposes a new structure called the synthetic stereo acoustic echo cancellation structure, for multiple participant stereo conferencing. Monaural microphone transducers are used to eliminate far-end talker spatial information and static stereo images are generated instead. Spatialization functions are used to generate a stereo image with localization information for distinguishing between

talkers. The spatialization method chosen for the synthetic stereo canceller is the stereo pan/pot, which uses interaural level differences (ILDs), see section 3.3.1, to produce localization cues. Head related transfer functions provide superior localization cues, but there are problems with speaker de-correlation that must be resolved for transaural monitoring, see sections 3.3.2 and 3.3.3.

The multiple participant synthetic stereo acoustic echo cancellation structure proposed in this paper has rapid convergence, good tracking ability, low computational complexity and fully whitens the primary signals. The primary signals do not suffer from cross-correlation because only one acoustic echo canceller is used at a time. The widely-used normalized least mean square (NLMS) algorithm can be applied to the synthetic stereo structure for acoustic echo cancellation with low computational complexity.

1.2.2 Psycho-Acoustic Determination of Echo Cancellation Targets

High levels of acoustic echo cancellation are required in practical VoIP system implementations. A control system called the nonlinear processor (NLP) is required to generate switched-loss, which is applied to the signal path in addition to acoustic echo cancellation. The level of switched-loss must be minimized to maintain full-duplex conferencing without the negative half-duplex effects.

This thesis proposes a new algorithm to minimize half-duplex effects by determining echo return loss enhancement (ERLE) targets through psycho-acoustic data. Psycho-acoustic data shows that the ERLE target is based on the mean one-

way propagation delay for the conference. The mean one-way propagation delay is approximately equal to one half of the total acoustic round trip delay for the conference.

Public switched telephone networks (PSTNs) are incapable of computing the acoustic round trip delay, but digital VoIP networks can measure the acoustic round trip delay and adjust their ERLE targets.

1.2.3 Acoustic Round Trip Delay Computation

This thesis proposes a theoretical model for the acoustic round trip delay and an empirical method for measuring this delay. The acoustic round trip delay is the sum of all the delays from the time the near-end user's begins speaking until the time the near-end user's echo, which is generated at far-end, returns to their ear. The theoretical model was used to derive equations for the acoustic round trip delay of a VoIP system. The empirical measurement method was proposed to validate the equations for the acoustic round trip delay and for measuring delays in VoIP systems. The empirical measurement method was used on the Carleton University digital signal processing/personal computer (DSP/PC) VoIP platform and a multimedia PC platform to collect acoustic round trip delay data.

1.2.4 VoIP Test Platform

A Carleton University VoIP test platform was developed to demonstrate the synthetic stereo acoustic echo cancellation structure and to measure acoustic round trip delay through the acoustic round trip delay empirical measurement method. A custom platform was created because there were no commercially available systems capable of supporting the proposals of this thesis. The platform is called the DSP multi-participant Intranet phone system (dMIPS). dMIPS conforms to the ITU-T H.323 standard for compatibility with other VoIP implementations. The platform requires a digital signal processing (DSP) card because it provides the high level of processing power necessary for acoustic echo cancellation. The DSP card is also able to lower the round trip delay of the system more than platforms based on a PC sound card. A multimedia PC platform, multi-participant Intranet phone system (MIPS), was also developed at Carleton University to test acoustic round trip delay for sound card based CTI systems.

2 Background Information

2.1 Half-Duplex Conferencing

2.1.1 Press-to-Talk

The first generation of half-duplex hands-free conferencing systems used a press-to-talk methodology for echo suppression. Manually controlled switches changed these press-to-talk systems from reception to transmission mode. These switches stopped the loudspeaker's signal from echoing and returning through the microphone transducer. Figure 1 shows two control blocks, A and B, with switches that change between transmission and reception mode for half-duplex conferencing.

Figure 1 - Half-Duplex Structure

2.1.2 Electronically Switched

The second generation of half-duplex hands-free conferencing systems used electronic switches for echo suppression. Manually controlled switches were replaced with electronically controlled switches. These switches automatically switched direction based on the nonlinear processor's ability to detect whether the near-end signal was stronger than the far-end signal.

2.1.3 Switched-Loss

The third generation of half-duplex hands-free conferencing systems used attenuators for echo suppression. The switches from the press-to-talk and electronically switched generations, see sections 2.1.1 and 2.1.2, were replaced with attenuators shown in Figure 1. The level of attenuation in switched-loss systems is variable and gradual compared with the infinite and instantaneous attenuation provided by a switch. Switched-loss systems require more complex algorithms to evaluate the required amount of echo suppression and determine near-end talker activity.

Depending on background noise and talker levels, the required switched-loss attenuation level can cause performance between half-duplex and close to full-duplex.

2.2 Full-Duplex Conferencing

2.2.1 Acoustic Echo Cancellation

The purpose of an acoustic echo canceller is to break the feedback loop between the speaker and the microphone. The feedback loop causes the user's speech to echo and return to the far-end. A monaural acoustic echo cancellation structure with only one speaker and microphone is shown in Figure 2.

Figure 2 – Basic Acoustic Echo Cancellation Structure

The source signal, $s(t)$, can be speech, music or another acoustic source. The acoustic echo cancellation field, however, normally deals with speech signals. The source signal, $s(t)$, is the reference to the echo canceller.

The reception room transfer function, $H(f)$ or $h(t)$ in the time domain, is the transfer function between the speaker transducer and microphone transducer. The error signal, $e(t)$, shown after the adder in the system, is zero for perfect echo cancellation with no near-end speech. The near-end is the end of the conference

where the acoustic echo canceller being discussed is located. The acoustic echo canceller cancels the feedback between the speaker and microphone by computing an estimate of the reception room transfer function, $\bar{h}(t)$ and subtracting this estimate from the desired signal, $e(t) = s(t) * h(t) - s(t) * \bar{h}(t)$.

The primary signal to the canceller, $d(t)$, is also called the desired signal. The desired signal in Figure 2 is equal to $d(t) = s(t) * h(t)$. The echo estimate computed by the acoustic echo canceller is equal to $s(t) * \bar{h}(t)$.

2.2.1.1 Acoustic Echo Canceller Adaptation

The acoustic echo canceller uses an adaptive algorithm which is designed to minimize the error, $e(t)$. The error is the difference between the reference signal, $x(t)$, convolved with the reception room transfer function, $h(t)$, and the reference signal convolved with the reception room estimate, $\bar{h}(t)$.

$$e(t) = s(t) * h(t) - s(t) * \bar{h}(t) \quad (1)$$

The acoustic echo canceller builds an estimate of the reception room transfer function to minimize the error. The error is minimized when $\bar{h}(t) = h(t)$.

2.2.1.2 Near-End Speech

If the near-end talker is active then their speech passes through the acoustic echo cancellation structure. The error function for near-end speech is

$$e(t) = s(t) * \left[h(t) - \tilde{h}(t) \right] + v(t) \quad (2)$$

where $v(t)$ is the near-end talker's speech. If the acoustic echo canceller is in steady state, $h(t) - \tilde{h}(t) = 0$, then $e(t) = v(t)$. The far-end echo is completely cancelled and the near-end talker's speech passes to the far-end.

The acoustic echo canceller must not be allowed to adapt during periods where the near-end talker is active. In steady state, the reception room transfer function has already been correctly identified. The residual error, $e(t) = v(t)$, from the near-end speech will cause the adaptive algorithm to wrongly minimize the error resulting in canceller divergence. It is important for the acoustic echo canceller to detect whether near-end speech is present in the system and to freeze adaptation under these conditions. When system adaptation is frozen, acoustic echo cancellation is performed by convolving the reference signal with the most recent set of canceller coefficients.

2.2.1.3 ERLE Computation

The echo return loss enhancement (ERLE) is an important property used to characterize acoustic echo canceller performance. ERLE is a measure of the

amount of signal power cancelled by the acoustic echo canceller. The ERLE is computed across the adder in the structure shown in Figure 2. The ERLE is computed by the following equation

$$ERLE(dB) = 10 \log \frac{E[d^2(n)]}{E[e^2(n)]} \quad (3)$$

2.2.1.4 Maximum ERLE Computation

The maximum amount of acoustic echo cancellation for an adaptive FIR filter is limited by the number of filter taps. Acoustic echo cancellers must be designed with a balance between the number of adaptive filter taps and the convergence rate of the adaptive algorithm. The NLMS algorithm's convergence rate is inversely proportional to the number of adaptive filter taps in the system.

The maximum ERLE for an acoustic echo canceller can be evaluated by computing the total impulse power to tail power ratio [41]

$$\frac{TIP}{TP}(n, M)(dB) = 10 \log \frac{\sum_{i=0}^{M-1} h^2(i)}{\sum_{i=n}^{M-1} h^2(i)} \quad (4)$$

where $h(i)$ is the impulse response of the system, M is the length of the impulse response of the system, and n is the length of the adaptive filter in taps. In reality M is infinite but for practical reasons the impulse response of the system is truncated to a fixed value. For the maximum ERLE to be computed accurately M must be much larger than n , $M \gg n$.

2.2.1.5 ERLE Target Determination

Studies showing the amount of echo that can be tolerated for a given mean one-way propagation delay have been published in ITU-T G.131 [22] and ETSI ETR 250 [23]. The minimum tolerable acoustic echo cancellation is shown in Figure 3. The talker echo loudness rating (TELRL) curves are generated by

$$TELRL = A_0 - 6e^{-0.3T^2} + 40 \log \frac{1 + \frac{T}{10}}{1 + \frac{T}{150}} \quad (5)$$

where T is the one-way transmission time in milliseconds and A₀ is 14dB for the acceptable case, 1% objection, or A₀ is 7dB for the limiting case, 10% objection.

Figure 3 - Talker Echo Tolerance Curves

The acceptable case indicates the level of attenuation that must be present in the system for most users to perceive the conversation quality as acceptable. The limiting case indicates the level of attenuation for some users to find the quality acceptable. The acceptable case is used in the rest of this thesis as the minimum tolerable acoustic echo cancellation for a given delay.

2.2.1.6 Limitations of Acoustic Echo Cancellers

The maximum level of echo suppression that an acoustic echo canceller can provide is limited by the factors detailed in Table 1 [14], [41].

Limitation	Explanation
Noise	Acoustic noise from heating and cooling systems, computer fans, hum from lights and other noise sources. Thermal and impulsive circuit noise.
Truncation and Quantization Effects Under-modelling	Fixed point DSP processors suffer from truncation and quantization effects due to a fixed word length. A room's impulse response is infinite, but the number of taps in the acoustic echo canceller implementation must be fixed. The number of taps is chosen to represent the majority of the power in the impulse response, see section 2.2.1.4.
Enclosure Vibration	Enclosure vibration and resonances can limit the performance of the acoustic echo canceller.
Transducer Nonlinearities	Loudspeakers and microphones, to a lesser degree, produce nonlinear distortions that cannot be modeled by the linear adaptive FIR filter.
Algorithmic Performance	The ability to track changes in non-stationary environments is characterized by the algorithm's tracking performance. The speed at which an adaptive algorithm initially converges to the steady-state solution is characterized by the algorithm's convergence rate. The ability of the algorithm to correctly identify double-talk and near-end speech affects its performance.

Table 1 – Limitations of Acoustic Echo Cancellers

2.2.2 Acoustic Echo Cancellation with Switched-Loss

Practical full-duplex hands-free conferencing system implementations require the use of switched-loss in addition to acoustic echo cancellation. Switched-loss is required to reach the ERLE targets, which are normally higher than the acoustic echo canceller is able to obtain by itself. Figure 4 shows four attenuators, labeled L_1 , L_2 , L_3 and L_4 which provide switched-loss to reduce the amount of echo from the far-end. The nonlinear processor (NLP) computes the amount of switched-loss to be added to the system through the four attenuators.

Figure 4 - Acoustic Echo Cancellation Structure with Switched-Loss

2.3 TCP/IP Protocol Suite

The transmission control protocol/Internet protocol (TCP/IP) suite is the most widely used suite of protocols for non-guaranteed quality-of-service networks. The TCP/IP suite provides reliable packet transmission service through the transmission

control protocol (TCP) and unreliable best-try service through the user datagram protocol (UDP). The function of the IP layer is to route TCP packets or UDP datagrams. Figure 5 shows a hierarchical view of the TCP/IP protocol stack.

Figure 5 - TCP/IP Protocol Stack

2.3.1 User Datagram Protocol

The user datagram protocol (UDP) is a best-try, connectionless service. No prior connection needs to be established and packet overhead is lower than for connection-oriented service. UDP is used by H.323, see section 2.4, to send packets with the real-time transport protocol (RTP), H.225.0. The user data protocol is used since there is not usually enough time to retransmit packets and lower packet overhead is desirable for real-time traffic [43].

2.3.2 Transmission Control Protocol

The transmission control protocol (TCP) is a connection-oriented, reliable service. A TCP connection is established between two endpoints using a three-way handshaking protocol. The handshaking protocol exchanges sequence numbers and other information before a stream is transmitted. TCP establishes a full-duplex connection with sliding window flow control and go-back-N error correction. Real-time traffic does not normally use TCP since the packet overhead is much larger than for UDP. TCP is used when errors cannot be tolerated and preventative measures such as three-way handshaking and error correction are required.

2.3.3 Internet Protocol

The Internet protocol (IP) is responsible for packet fragmentation and reassembly, routing and error reporting. The IP protocol provides a best-try connectionless service. The version of IP currently in use on the Internet is version 4 (IPv4) although tests are being performed on IP version 6 (IPv6), also called IP next generation (IPng). The IPv4 packet format is shown in Figure 6.

Packet fragmentation and reassembly are resource-consuming operations that occur when an IP packet is too large to be transmitted over the desired link and must be split into several smaller packets. Avoiding fragmentation by intermediate systems is desirable for real-time traffic to avoid the extra overhead. Real-time streams should set the don't fragment bit in the IP header to ensure that the network chosen to route the IP packet can handle datagrams of that size.

A specific type of service can be requested in the IP header to influence route selection. A priority between 0-7, and flags to indicate low delay, high throughput or high reliability can be set to influence the type of service.

The time-to-live (TTL) for an IP datagram is specified in its header and is decremented at each router along its route. This quantity can be used to limit the time the datagram is in transit across the network, since it will be discarded when the TTL reaches zero. Real-time streams should limit the TTL since severely delayed datagrams cannot usually be recovered and should be discarded.

Gateways are used to route IP datagrams if they are not addressed to the local area network (LAN). There are two types of gateways that use different routing protocols. An interior gateway runs the interior gateway protocol (IGP). IGP maintains sufficient information to forward datagrams to hosts on connected LANs or to other interior gateways within the same autonomous system. (An autonomous system refers to a network that is separately managed and run.) An exterior gateway runs the exterior gateway protocol (EGP). EGP maintains sufficient information to forward datagrams to an interior gateway if the datagram is destined for the same autonomous system or another exterior gateway if the datagram is destined for another autonomous system.

Figure 6 - IP Packet Format

The new IP protocol, IPv6, expands routing and addressing capabilities. The address fields have increased from 32-bits to 128-bits to support the rapid growth of the Internet and its addressing hierarchy. Greater support for multicasting has been introduced by replacing the broadcast address class with a multicast address class.

A new address type called “anycast” has been defined to specify that a datagram is destined for one of a set of nodes. IPv6 defines specific values for its 4-bit priority field. A priority value between 0-7 indicates traffic for which the source will provide congestion control and 8-15 indicates traffic that does not back off in response to congestion such as real-time traffic at a constant bit rate. The IPv6 header format has been simplified by eliminating fields, which are less used. These simplifications were necessary to offset the increased header size due to the expanded address fields.

IPv6 supports new fields to specify the desired quality-of-service (QoS). Datagrams are grouped into traffic flows and the sender can request special handling for real-time service. IPv6 contains many enhancements, which should be used to enhance future CTI products.

2.4 ITU-T H.323 Standard

H.323, [19] [27], is a ‘toolkit standard’ which details audio, video and data services transmission over networks which do not provide a guaranteed quality-of-service. H.323 is called a ‘toolkit standard’ since it combines several ITU-T standards, see Figure 7. The H.323 standard covers system and component descriptions, call model descriptions and control method formats. This standard closely parallels work being performed by the Internet Engineering Task Force (IETF) on standards such as the real-time transport protocol/RTP control protocol

(RTP/RTCP) [29][30]. H.323 is the dominant standard for CTI on the Internet and most products claim full compliance with this standard.

Figure 7 - H.323 Terminal Equipment

The shaded sections of Figure 7 show the building blocks that were implemented for the Carleton University dMIPS VoIP platform, see section 6.

An H.323 compliant network is required to provide reliable and unreliable delivery mechanisms. The reliable delivery mechanism in the TCP/IP protocol suite is TCP and the unreliable delivery mechanism is UDP.

2.4.1 Multimedia Services

There are three multimedia services that can be carried over H.323 CTI systems. Voice is a mandatory service that must be supported on all H.323 compliant

systems. Video and data exchange are optional multimedia services. The following types of conferences are supported by H.323 :

- voice only (voice is a required basic service)
- voice and video (video is optional)
- voice and data (data is optional)
- voice, video and data (video and data are optional)

H.323 is an open standard, which allows innovation by manufacturers and researchers. Terminals from the same manufacturer can support proprietary multimedia services. If a terminal does not support these proprietary services then the conference will fallback to required components specified by H.323.

2.4.2 H.32x Interoperability

A successful H.323 implementation must be inter-operable with H.320, for narrowband ISDN, H.321, for broadband ISDN, H.322, for isochronous Ethernet (IsoEthernet), H.324, for public service telephone network (PSTN), and H.310, for asynchronous transfer mode (ATM).

2.4.3 Multipoint Conferences

H.323 is capable of supporting the following types of conferences:

- point-to-point (two terminals with send and receive capability)
- multipoint (three or more terminals with send and receive capability)
- broadcast (one sender and multiple receivers)

Multipoint conferences are supported over TCP/IP by IP unicast or by using new IP multicast techniques. Multipoint conferencing by IP unicast requires a device called the multipoint control unit (MCU) that replicates copies of packets. The MCU can be incorporated into the terminal, gateway or gatekeeper. The MCU is composed of the multipoint processor (MP) and multipoint controller (MC). The MP's services include media processing, audio mixing, video mixing and switching. The MC provides conference control such as negotiating a common communications mode and establishing media channels.

Multipoint conferencing by IP multicast means that the network is capable of distributing the media streams to multiple points along the network. An example of a multicast enabled protocol is IP version 6 (IPv6) which replaces IPv4 broadcast addresses with multicast addresses. In IPv6 multicast, the network and terminal interfaces are aware of multicast addresses and are able to process them. The main advantage of a multicast enabled network is that an MCU is not required to replicate the packet stream. The efficiency of the terminal and network are greatly improved by multicast, as the number of participants in the conference increases, because packet replication is not required.

H.323 specifies three multipoint conference models that cover all possible combinations of unicast and multicast enabled networks. They are a centralized model, which is unicast based, a decentralized model, which is multicast based, and a hybrid conference, which combines centralized and decentralized networks. An MCU is required in the centralized model to replicate a packet stream using IP unicast, which increases traffic on a network and loads the MCU. IP multicast is

used in the decentralized model and packets are distributed without any packet replication required by the MCU. The hybrid model is reserved for incremental upgrades of network infrastructure where part of the network uses legacy IP unicast and the other uses IP multicast.

2.4.4 Gateways

A gateway provides interoperability between network segments with different configurations. It translates call signaling and control channel messages. The gateway also performs conversions between different multiplexing techniques. The gateway may contain the MCU to support multipoint conferencing between terminals on different network segments. An example would be an H.323 to H.310 gateway, or packet switched to ATM gateway.

2.4.5 Gatekeepers

The gatekeeper is used for bandwidth control, monitoring security through admissions control and providing address translation services in CTI systems. Gatekeeper usage is optional to the system designer, but a different call model is necessary if a gatekeeper is present.

The gatekeeper's admissions control function is used to regulate the amount of telephony traffic in a network. The network administrator can specify maximum levels for video, voice and data exchange traffic. Admissions controls prevent a

network from becoming congested by CTI traffic and being unusable for other networking requirements.

The gatekeeper provides an address translation service for aliases or other addressing conventions. An example is the translation of an E.164 telephone number such as '613-520-2600' into an IP transport service access point (TSAP) such as '134.117.4.1:2000'. Another example is the translation of an alias such as 'tyensen' into an IP TSAP such as '134.117.4.65:500'. The address translation service is similar to address lookup by the domain name server (DNS) in a TCP/IP implementation.

A system that does not have a gatekeeper uses the endpoint signaling call model, but systems that have a gatekeeper use the gatekeeper routed call signaling model. In endpoint signaling, the call is setup regardless of network conditions and without requesting permission from any third party. Permission is requested from the gatekeeper to setup a call in the gatekeeper routed call model. The only way to provide bandwidth management on a network is to use the gatekeeper routed call model, which monitors bandwidth usage through admissions control.

2.4.6 Call Setup

An H.323 connection is established using the H.225.0 call setup procedure. Call setup procedures for the gatekeeper routed call model and the endpoint signaling model are similar. Table 2 summarizes the call setup sequence for both call models, a flow graph representation can be found in [19].

Following call setup, the endpoints establish an H.245, [15], control channel. The H.245 control channel is used for capability exchange, master/slave determination and multiplex table signaling.

CALL MODEL	STEP	ACTION TAKEN BY H.225.0 PROTOCOL
Gatekeeper Routed	1	Originating terminal requests permission from gatekeeper using 'ARQ' message.
Gatekeeper Routed	2	Gatekeeper accepts call and 'ARJ' message is returned to originating terminal.
ALL	3	Originating terminal sends 'Setup' message to remote endpoint.
ALL	4	Remote endpoint responds with 'Call Proceeding' acknowledgment message.
Gatekeeper Routed	5	Remote endpoint requests permission from gatekeeper using 'ARQ' message.
Gatekeeper Routed	6	Gatekeeper accepts call and 'ARJ' message is returned to remote endpoint.
ALL	7	Remote sends 'Alerting' message to originating terminal to indicate that the remote user is being notified of the incoming call. (ie phone rings)
ALL	8	Remote sends 'Connect' message to originating terminal if the call is answered.

Table 2 - Call Setup Signaling for H.323 Conference

2.4.7 Logical Channels

After call setup is complete, the H.245 protocol is used to establish H.225.0 logical channels. Separate logical channels are established for each class of media present in the CTI conference. For example, a separate channel would be established for audio, video, and data. Multiple channels for any media class can be opened if the conference is multipoint. For example, if a three-way audio, video and data exchange conference was proceeding there would be three audio channels,

one for each talker, three video channels and three data channels established at each CTI terminal.

In addition to the logical media channels, one control channel is established per logical channel. The control channel reports on network conditions for the associated logical channel. Information such as packets lost, jitter and packets sent are sent through the RTP control protocol (RTCP) sender and receiver reports. These channels are distinguished through unique transport service access point (TSAP) addresses. In TCP/IP, a TSAP is composed of a port and a network IP address.

The H.225.0 protocol, which is also called the real-time transport protocol (RTP) was designed to carry traffic with real-time properties. RTP adds header information which contains the type of payload such as G.711 audio or H.261 video, timestamps and sequence numbers. This information is used to determine if the network has lost, reordered or delayed any packets. The receiver can estimate network statistics by using the timestamps and sequence number information. RTP protocol data units (PDUs) are transmitted using a connectionless service such as UDP, see section 2.3.1.

The RTP control protocol (RTCP) is used to monitor the quality-of-service (QoS) for a media stream and to report network conditions. Sender and receiver reports are used to monitor network conditions and the statistics they report can be used to provide congestion and flow control. These reports can be used to adaptively adjust the bit-rate for audio or video streams [43]. Multicast enabled networks also rely on

this information to diagnose faults in the distribution. Each participant in an H.323 session sends RTCP packets as often as possible, typically 5% of the logical channel bandwidth, to maximize the resolution of the statistics being collected. The statistics collected by RTCP are:

- sender's packet count: total number of RTP PDUs transmitted by the sender since transmission commenced
- sender's octet count: total number of payload octets, not including header, transmitted in RTP PDUs since transmission commenced
- fraction lost: fraction of RTP data packets lost since the previous receiver or sender report packet was sent
- cumulative number of packets lost: total number of RTP PDUs that have been lost since the beginning of reception
- interarrival jitter: estimate of the statistical variance of the RTP PDU interarrival time measured in timestamp units

The relative transmit time, D , is the difference in packet spacing at the receiver as compared to the sender, [29]

$$D(i-1, i) = (R_i - S_i) - (R_{i-1} - S_{i-1}) \quad (6)$$

where S_i is the sender's timestamp for the i th packet, R_i is the receiver's timestamp for the i th packet and S_{i-1} and R_{i-1} apply to the previous packet.

The interarrival jitter, J , can be computed by (7), [29]

$$J = J + (|D(i-1, i) - J|) / 16 \quad (7)$$

2.4.8 Quality of Service

The quality-of-service (QoS) for real-time traffic can be evaluated by measuring jitter, packet loss and network congestion. H.323 does not provide the system administrator with any mechanisms for guaranteeing the quality-of-service on a CTI network. H.323 does provide tools for assessing and controlling the quality-of-service, such as RTCP sender and receiver reports.

Methods for dealing with errors are required since the quality-of-service is not guaranteed. An RTP PDU that is lost or received in error may be retransmitted if jitter buffers are large and retransmission time is short. Real-time traffic does not usually favour retransmission. When errors occur the decoder must perform error concealment such as generating comfort noise or repeating the last packet received.

A network administrator can regulate the QoS of the network in terms of bandwidth allocation if a gatekeeper is implemented. The network administrator can assign an upper limit to the bandwidth that CTI traffic can occupy on a local area network (LAN) segment.

The reservation protocol (RSVP) will offer future QoS guarantees by reserving bandwidth at each gateway in the network. The RSVP protocol assumes that the bandwidth on any given segment of the network is unlimited, but the bandwidth between segments is limited. The RSVP protocol reserves bandwidth at each gateway such that CTI traffic is guaranteed a certain level of QoS.

2.5 Digital Signal Processing on a Fixed-Point Processor

The fixed-point digital signal processor (DSP) is frequently used instead of floating-point processors because the cost per million instructions per second is lower than on the fixed-point version. Fixed-point representation of fractional numbers must be used to program these devices.

2.5.1 Q.F Notation

The Q.F notation can be used to represent fractional numbers in two's complement sign extension mode ranging from

$$\text{Q.F range} \quad \left[-1, 1 - 2^{-F} \right] \quad (8)$$

where F represents the number of bits which are to the right of the implicit binary point.

For example, a 16-bit accumulator used to represent fractional numbers in Q.15 notation has a range between $\left[-1, 1 - 2^{-15} \right]$.

The following conversion rule applies to converting a Q.F number into a fraction

$$N = D \times 2^{-F} \quad (9)$$

where N is the number in fractional format between $[-1, 1 - 2^{-F}]$ and D is the Q.F number.

For example, if a 16-bit accumulator on a fixed point DSP contained the value 16384 then the corresponding fractional value would be

$$0.5 = 16384 \times 2^{-15} \quad (10)$$

2.5.2 Q.0 Notation

Normal twos complement notation on a DSP can also be called Q.0 notation. These values range from

$$\left[-2^{S-1}, 2^{S-1} - 1 \right] \quad (11)$$

where S is the size of the accumulator in bits.

2.5.3 Q.x Operation Rules

The following rules apply to Q.x operations.

$$\text{Multiplication} \quad Q.x \times Q.y = Q.(x + y) \quad (12)$$

$$\text{Division} \quad \frac{Q.x}{Q.y} = Q.(x - y) \quad (13)$$

$$\text{Arithmetic} \quad Q.x \pm Q.y, x \text{ and } y \text{ must be equal} \quad (14)$$

The Texas Instruments TMS320C54x DSP processors have a special flag, which can be set to automatically left shift the result from a multiplication operation by one

bit. This flag is called the fractional (FRCT) bit. A Q.15 fraction multiplied by another Q.15 fraction forms a Q.31 fraction when the FRCT bit is enabled. A store high word operation can then be used to store the result as a Q.15 fraction. The multiplication rule for Q.x notation with the FRCT bit enabled is

$$\text{Multiplication (FRCT bit enabled)} \quad Q.x \times Q.y = Q.(x + y + 1) \quad (15)$$

This reduces the number of cycles required to complete common operations such as finite impulse response (FIR) filtering and adaptive filtering.

3 Review of the State of the Art

3.1 Stereophonic Acoustic Echo Cancellation Structures

Stereophonic conferencing requires two microphone transducers and two speakers for each participant in a conference. This physical setup is intended for peer-to-peer conferencing only. This setup gives the near-end listener spatial information about the far-end talker. If the audio is coupled with video then the listener would see the far-end user move on the video monitor and be able to hear the user's localization cues change accordingly.

The main problem in stereophonic acoustic echo cancellation is the cross-correlation between the primary input signals. This thesis investigates steps taken by various authors to eliminate the cross-correlation problem in the following stereophonic acoustic echo cancellation structures:

- Stereophonic Acoustic Echo Cancellation Using a Multiple Canceller Structure
- Stereophonic Acoustic Echo Cancellation Using a Single Canceller Structure
- Stereophonic Acoustic Echo Cancellation Using a Subband Structure
- Stereophonic Acoustic Echo Cancellation With Pre-processing

3.1.1 Stereophonic Acoustic Echo Cancellation Using a Multiple Canceller Structure

The multiple canceller structure is the simplest stereophonic acoustic echo cancellation structure. It is a straightforward extension to the monaural acoustic echo cancellation structure. In monaural acoustic echo cancellation there is one

echo path between the speaker and microphone. In stereo there are four coupling paths between the speakers and the microphone. These paths are from the left microphone to the left speaker, from the left microphone to the right speaker, from the right microphone to the right speaker and from the right microphone to the left speaker. Four acoustic echo cancellers are required since there are four coupled paths that must be cancelled [8] [13] [34].

Half of a multiple canceller structure is shown in Figure 8. A stereo signal is generated at the far-end by stereo microphones. At the near-end only one microphone is shown for simplicity. A full multiple canceller structure would have two microphone inputs and four echo cancellers.

The remote talker in Figure 8 is labeled $s(t)$. The transmission room transfer functions are labeled $g_1(t)$ and $g_2(t)$. The transmission room transfer functions are dynamic. These transfer functions change when the talker moves or their environment changes. The correlation between the left and right references to the canceller are unknown because $g_1(t)$ and $g_2(t)$ are unknown. The correct adaptive filter solution for the echo paths is as follows

$$\tilde{h}_1(t) = h_1(t) \quad (16)$$

$$\tilde{h}_2(t) = h_2(t) \quad (17)$$

The reference signals to the cancellers are

$$x_1 = s(t) * g_1(t) \quad (18)$$

$$x_2 = s(t) * g_2(t) \quad (19)$$

The error signal for this structure is

$$\begin{aligned} e(t) &= s(t) * g_1(t) * h_1(t) + s(t) * g_2(t) * h_2(t) \\ &\quad - s(t) * g_1(t) * \bar{h}_1(t) - s(t) * g_2(t) * \bar{h}_2(t) \end{aligned} \quad (20)$$

$$e(t) = s(t) * g_1(t) * \left[h_1(t) - \bar{h}_1(t) \right] + s(t) * g_2(t) * \left[h_2(t) - \bar{h}_2(t) \right] \quad (21)$$

If the error signal is minimized by the acoustic echo canceller, $e(t) = 0$, then we obtain the following relationship

$$s(t) * g_1(t) * \left[h_1(t) - \bar{h}_1(t) \right] = -s(t) * g_2(t) * \left[h_2(t) - \bar{h}_2(t) \right] \quad (22)$$

From (22) it can be seen that there are an infinite number of solutions to the problem. It does not imply that $h_1(t) = \bar{h}_1(t)$ and $h_2(t) = \bar{h}_2(t)$ which is the case for perfect acoustic echo cancellation. The correct echo paths are, therefore, not identified correctly by the multiple canceller structure. The cancellers converge to multiple instantaneous solutions to the problem which do not identify the correct echo path, [5] [44].

Figure 8 – Stereophonic Acoustic Echo Cancellation with a Multiple Canceller Structure

True echo paths can be identified, in practice, even though cross-correlation significantly impedes the echo canceller's convergence. The following conditions cause the multiple canceller structure to converge to the true echo paths [8] [13] [34]

- 1) The two stereo input signals contain independent noise.
- 2) The length of the adaptive filters is shorter than the length of the impulse response in the transmission room. The truncated impulse response acts as independent noise in the system.
- 3) The cross-correlation between the stereo signals varies slightly even when the talker does not move their body or head while speaking.

The multiple canceller structure uses the slight variations in cross-correlation to converge to the true echo path solutions. The convergence rates for this structure, however, are very slow, which makes this structure impractical. This structure must track changes in two transfer functions, the transmission and the reception room transfer functions [34]. The two levels of adaptation make rapid convergence and tracking rates essential.

3.1.2 *Stereophonic Acoustic Echo Cancellation Using a Single Canceller Structure*

The multiple canceller structure was unable to correctly identify the echo paths and filter coefficients due to redundancy in canceller structure [36] [44]. Eliminating one of the acoustic echo cancellers makes it possible for the remaining acoustic echo canceller to converge to a unique solution for the correct echo path. A single acoustic echo canceller can model the signal path that generates the echo [44]. A single canceller structure is shown in Figure 9.

The primary signal at the input to the single canceller structure is the same as for the multiple canceller structure

$$d(t) = s(t) * g_1(t) * h_1(t) + s(t) * g_2(t) * h_2(t) \quad (23)$$

If one echo canceller is removed then the primary signal can be rewritten as

$$d(t) = s(t) * g_1(t) * [h_1(t) + g_1^{-1}(t) * g_2(t) * h_2(t)] \quad (24)$$

The remaining acoustic echo canceller converges to the following transfer function, [44]

$$\bar{h}(t) = h_1(t) + g_1^{-1}(t) * g_2(t) * h_2(t) \quad (25)$$

Figure 9 - Stereophonic Acoustic Echo Cancellation with a Single Canceller Structure

The reference signal that is ahead in phase must be identified to select which echo canceller to use. There are two possible choices for the acoustic echo canceller shown in Figure 9. A reliable means of determining which signal is ahead in phase is proposed by Hirano and Sugiyama [44]. This reference signal, which is ahead in phase, must be identified in order to satisfy causality between the reference and echo signals.

This structure is capable of solving the cross-correlation problem by eliminating one of the acoustic echo cancellers. There is a unique echo path solution since only one echo canceller is used at a time. This structure is still unsuitable for practical implementation because of slow tracking and instability in the echo path

solution. The inverse of the transmission room transfer function is often unstable, [34] causing instability in the echo path solution given by (25). There is also additional complexity in implementing a method to determine which reference signal is ahead in phase.

3.1.3 *Stereophonic Acoustic Echo Cancellation Using a Subband Structure*

The multiple canceller structure was shown to converge, in practice, due to the slight variation in cross-correlation between the primary signals. The subband structure was proposed to further exploit these variations in cross-correlation. Figure 10 shows a typical subband structure.

The input signals are bandlimited and downsampled by the subband structure which increases the variation in cross-correlation between the signals. The variation in cross-correlation increases because the time interval between successive samples increases due to downsampling [12]. Increasing the variation acts to further de-correlate the reference and primary signals.

Bandlimiting and decimating the signals reduces the length and complexity of the adaptive filters. Dividing the signals into narrower frequency subbands reduces the eigenvalue spread of the input signals and increases the convergence speed [12] [24].

The variation in cross-correlation between the input signals can be further exploited when the affine projection algorithm (APA) or fast affine projection algorithm (FAP) is used. The APA can de-correlate or fully whiten the primary input with a small projection order because the number of filter taps in each subband is reduced by the downsampling [12] [24]. The Gram-Schmidt process states that a

large input vector space cannot be whitened by a small projection order. A small input vector space, however, can be whitened by a small projection order [24]. The convergence rate for speech can be doubled with a projection order of two and even further rate increases are possible with larger projection orders [24].

Figure 10 - Stereophonic Acoustic Echo Cancellation with a Subband Structure

The main drawback of the subband structure is that it suffers from delays introduced by the subband synthesis and analysis filter banks. Narrower subband filter banks cause even further increases in complexity and delay. Narrowing the subbands, however, increases the whitening effect of the subband structure. The subband filter bank width is a compromise between the delay and the desired whitening effect.

The problems with the subband stereophonic acoustic echo cancellation structure are a high degree of computational complexity and large delays resulting from its analysis and synthesis filter banks. The structure is able to achieve reasonable convergence rates but its complexity and delay are sufficient to discourage its use.

3.1.4 Stereophonic Acoustic Echo Cancellation Structures with Pre-processing

The pre-processed stereophonic acoustic echo cancellation structure attempts to de-correlate the primary signals through direct manipulation of the reference signals in the system. Several pre-processed stereophonic acoustic echo cancellation structures have been proposed:

- adding a nonlinear transformation in the pre-processor [8] [13]
- adding a nonlinear transformation in the pre-processor with comb filtering [3]
- adding random noise controlled by auditory properties in the pre-processor [5]
- adapting acoustic echo canceller coefficients based on uncorrelated signals generated in the pre-processor [6]
- generating periodically time delayed input signals in the pre-processor [4] [7]

A simple pre-processing structure is shown in Figure 11 that adds a nonlinear transformation in the pre-processor. Pre-processed structures are currently the focus of research in the area of stereophonic acoustic echo cancellation.

All the pre-processed structures generate de-correlated stereo reference signals by modifying them in a manner that the user cannot perceive. Research shows that

adding independent noise to reference signals can be annoying to subjects but does cause correct convergence [8] [13]. The goal of the pre-processed structures is to devise a pre-processing scheme that is the least offensive to the users.

Figure 11 - Stereophonic Acoustic Echo Cancellation Structure with Preprocessing

3.1.4.1 Pre-processing with the Addition of a Nonlinearity

This structure reduces interchannel cross-correlation by adding a small nonlinearity into each reference channel. This nonlinearity is not noticeable for speech signals due to self-masking [13]. The reference signal is generated by the following function

$$x'_i(n) = x_i(n) + \alpha f[x_i(n)], i = 1, 2, \dots \quad (26)$$

A simple nonlinear function that can be used in (26) is the half-wave rectifier [13]

$$\tilde{x} = f(x) = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad (27)$$

The stereo localization cues are not affected by the addition of this nonlinear function and the distortion is barely audible to the user.

3.1.4.2 Pre-processing with the Addition of a Nonlinearity and Comb Filtering

This structure uses psycho-acoustic data which states that the stereo effect is mostly due to energy below 1 kHz and comb filtering above 1 kHz does not degrade the localization ability of the listener [3]. The signals are divided into two frequency bands. The low frequency band contains the stereo localization cues and the high frequency band contains information not required for localization. A nonlinear transformation is used on the low frequency band and complementary comb filtering is used on the upper frequency band to de-correlate the input. Comb filtering cannot be used on the fullband signal since it would destroy the stereo localization cues.

3.1.4.3 Pre-processing with Random Noise Controlled by Auditory Properties

This structure uses random noise whose shape is controlled by auditory masking rules [5]. Adding masked noise to the signals, even if the noise is inaudible to the user, can destroy spatial cues. To preserve the spatial cues, the masked noise must be adjusted by a correction coefficient. The maximum level of masked noise that can be generated before destroying the stereo spatial cues or annoying the user limits the amount of de-correlation for this approach.

3.1.4.4 Pre-processing with Adaptation Based Solely on the Uncorrelated Signals

This structure adapts coefficients based on the uncorrelated portion of the signals generated through a nonlinear transformation [6]. Four adaptive echo cancellation structures are used per channel. Two cancellers are used for the nonlinear pre-processed portion of the signal and two for the unmodified portion of the signal. The canceller coefficients from these four cancellers are not used to cancel the acoustic echo. The canceller coefficients from the two nonlinear cancellers are transferred to fixed filter structures when the coefficients are judged to be in good condition. The cross-correlation is minimized by this structure since convergence is based on the uncorrelated signals generated in the pre-processor.

3.1.4.5 Pre-processing with Periodically Time Delayed Input Signals

This structure periodically time delays one of the input signals using a time-varying filter in the pre-processor [4]. In [7], a unique time-varying all-pass filter is applied to each channel in order to produce a just noticeable inter-aural delay. The just noticeable inter-aural delay varies between $30\mu\text{sec}$ to $200\mu\text{sec}$ [7]. The time-varying filter maintains spatial cues and is able to de-correlate the signals.

3.2 Acoustic Echo Cancellation Algorithms

There are many different adaptive algorithms, which can be used for acoustic echo cancellation. A brief discussion of three algorithms follows, for a more detailed

description of these algorithms refer to [32][37][39][45][47][49]. The following adaptive algorithms are discussed:

- The Least Mean Squares Algorithm (LMS)
- The Recursive Least Squares Algorithm (RLS)
- The Affine Projection Algorithm (APA)

3.2.1 The Least Mean Square Algorithm (LMS)

The least mean square algorithm (LMS) is important to adaptive signal processing because of its simplicity and low computational complexity. The LMS algorithm is simple because it does not require any off-line gradient estimations or data repetition [49]. The LMS algorithm is

$$\tilde{\underline{h}}(n+1) = \tilde{\underline{h}}(n) + \mu e(n) \underline{x}(n) \quad (28)$$

$$e(n) = d(n) - \tilde{\underline{h}}^T(n) \underline{x}(n) \quad (29)$$

where $\tilde{\underline{h}} = [\tilde{h}_1, \tilde{h}_2, \dots, \tilde{h}_N]$ is the adaptive filter vector of length N , $\underline{x} = [x_1, x_2, \dots, x_N]$ is the reference vector of length N , $e(n)$ is the error signal and μ is the step-size.

The normalized least mean square (NLMS) algorithm is a widely used variant of the LMS algorithm. The step-size, μ , is normalized to the input power of the system. The standard LMS algorithm uses a fixed step-size. The step-size for the NLMS algorithm is computed at each iteration by (30)

$$\mu = \frac{\alpha}{L \cdot E\{x^2(n)\}} \quad (30)$$

The NLMS algorithm is the most commonly used variant of the LMS algorithm since its convergence rate and steady state performance are invariant to the input signal characteristics.

The LMS algorithm is not appropriate for multi-channel acoustic echo cancellation because it does not explore the variations in cross-correlation between the channels [34].

3.2.2 The Recursive Least Square Algorithm (RLS)

The recursive least-square (RLS) algorithm has the fastest convergence rate of all the adaptive algorithms but at the cost of the highest computational complexity [31][33]. The convergence rate of the RLS algorithm is fast because it is independent of the spectral characteristics of the speech signal at the system's input. The computational complexity of the RLS algorithm is $2N^2+4N$ multiplications. The RLS algorithm is stated in [47] and Appendix B.2.

The fast two-channel RLS algorithm solves the stereophonic acoustic echo cancellation problem in an optimal way. This algorithm is able to rapidly converge to the true echo paths [25][31][35]. The fast two-channel RLS algorithm fully whitens the stereo signal, however the computational cost is very high. The computational cost is very high since four fast RLS filters of size L need to be employed [35].

3.2.3 The Affine Projection Algorithm (APA)

The affine projection algorithm (APA) was introduced to make better use of input signal data to improve the adaptive behavior of the NLMS algorithm [31]. The affine projection algorithm uses a multi-dimensional projection for each adaptive filter tap [31]. The NLMS algorithm is equivalent to an APA with a projection order of 1 since each adaptive filter tap is one-dimensional.

The affine projection algorithm is capable of canceling P a posteriori errors such that $1 \leq P \leq N$ where P is the projection order and N is the length of the filter vector. If the projection order is 1, $P=1$, then the algorithm performs identically to the NLMS, but if the projection order is N , $P=N$, then the algorithm performs identically to the block based recursive least square (BRLS) algorithm [31]. The filter coefficients are adjusted in the direction of the plane produced by $x(k)$, $x(k-1)$, ..., $x(k-P+1)$, where k is the time index [32]. The input is whitened according to the projection order, P [24].

The convergence and tracking performance of the affine projection algorithm is very good. Increasing the projection order increases the convergence speed of the APA adaptive filter coefficients and the computational complexity [37]. APA tracking ability can outperform the LMS and RLS algorithms [39].

The affine projection algorithm is particularly effective in identifying the true echo path impulse response in stereophonic acoustic echo cancellation. The APA emphasizes the slightly varying cross-correlation between the stereo signals [12]. The APA is applied to stereophonic acoustic echo cancellation for a projection order of two in [25][28] and in general in [26][38].

The affine projection algorithm is

$$\underline{e}_n = \underline{s}_n - \underline{X}_n^T \underline{h}_{n-1} \quad (31)$$

$$\underline{\varepsilon}_n = [\underline{X}_n^T \underline{X}_n + \sigma \underline{I}] \underline{e}_n \quad (32)$$

$$\underline{h}_n = \underline{h}_{n-1} + \mu \underline{X}_n \underline{\varepsilon}_n \quad (33)$$

with excitation signal matrix

$$\underline{X}_n = [\underline{x}_n, \underline{x}_{n-1}, \dots, \underline{x}_{n-(N-1)}] \quad (34)$$

where $\underline{x}_n = [x_n, \dots, x_{n-P+1}]^T$, $\underline{h}_n = [h_{0,n}, \dots, h_{P-1,n}]^T$, $\underline{e}_n$ is the error vector and σ is the regularization parameter.

The computational complexity of the affine projection algorithm is $2PN + K_{\text{inv}}N^2$ multiplies per sample period, where K_{inv} is a constant depending on the inversion method used. (K_{inv} for Levinson Durbin is about 7.)

The fast affine projection algorithm (FAP) is the most important advancement for the APA. The fast affine projection algorithm has LMS like complexity and memory requirements, and RLS like convergence for speech signals. The computational complexity of FAP is roughly $2P + 20N$ multiplications per sample period [37]. FAP does not delay the input or output signals [37]. If the relaxation parameter is set to one, or smaller, the computational complexity becomes $2P + 14N$ multiplications per sample period. The FAP Algorithm is stated in [37] and Appendix 2.1.

The affine projection algorithm and the subband structure, see section 3.1.3, can be combined to emphasize the variation in cross-correlation between the stereo

signals. The APA and FAP algorithms are powerful algorithms for de-correlating input signals and are important for stereophonic acoustic echo cancellation.

3.3 Spatialization Techniques

Spatial images for stereo output transducers can be generated by several methods such as interaural level differences (ILDs) and interaural time differences (ITDs) or head related transfer functions (HRTFs). Localization is described as a subject's ability to perceive distance, azimuth and elevation characteristics from the sound [18]. Spatial cues can be reproduced by binaural and transaural monitoring.

Binaural recording transcribes signals that represent what the ear hears. Binaural recordings can be produced by two methods. In the first method an artificial head fitted with microphones or a human head with microphone probes is used. In the second method, called binaural synthesis, free-space microphones are used and the transmission of sound around the human head to the ears is simulated [46].

Listening to stereo signals through headphones or earphone transducers is called binaural monitoring. The signals require equalization to account for the frequency and phase characteristics of the headphones or earphones.

Listening to stereo signals from loudspeakers is called transaural monitoring. The signals require loudspeaker equalization and crosstalk cancellation. The dMIPS VoIP platform that was built for this thesis is a stereo hands-free platform, which means that the stereo signals are heard through transaural monitoring.

3.3.1 ILD and ITD Techniques

Human localization ability is a function of the interaural level differences (ILDs) and interaural time differences (ITDs). These interaural differences are produced by the acoustic shadow of the human head [42]. In three-dimensional space these differences are ambiguous, which creates an effect known as the cone of confusion. This thesis, however, is restricted to localization in the horizontal plane and the front listener stage.

The amplitude level between the two ears varies with the incident angle of the auditory event. If the source is closer to one ear then the level at that ear will be higher than the level at the other ear. The stereophonic law of sines can be used to compute interaural level differences for reproduction over stereo loudspeakers [48]

$$\sin \Theta = \frac{L - R}{L + R} \sin \Theta_0 \quad (35)$$

where $-\Theta_0 \leq \Theta \leq \Theta_0$, L is the left speaker level, R is the right speaker level and Θ is the off-centre angle between the two loudspeakers and the listener's head. The ILDs computed by this methodology lead to consistent imaging over loudspeakers which is the basis for recommending this technique in section 4.6.

The time delay between the two ears varies with the incident angle of the auditory event. These time delays are called interaural time differences ITDs and for inter-loudspeaker delays of $\pm 3\text{ms}$, it is possible to localize an impulsive source over a

stereo sound stage [48]. For non-impulsive sound sources the ITDs do not lead to consistent localization imaging over the stereo sound stage [18]. ITDs can lead to 'phasiness' effects that do not contribute to source localization [48].

3.3.2 Head Related Transfer Function Technique (HRTF)

Head related transfer functions (HRTFs) are measurements of the transfer functions between an auditory source and subject's eardrums. It is assumed that the external ears superimpose linear distortions on the source signals necessary for the subject's determination of azimuth, elevation and distance [18]. These linear distortions create the spatial information that is encoded into the signals received by the eardrum. The head related transfer function containing spatial information is found through determination of the transfer function between the auditory source and the subject's eardrums.

The head and shoulders are physical characteristics that vary regardless of whether the subject is a dummy or human. Research is being performed to determine if there is an ideal dummy or human subject or if personalized HRTFs need to be used for proper localization ability. A complete set of HRTF measurements has been performed by MIT on a KEMAR dummy and is available in the public domain [40].

3.3.3 Crosstalk Cancellation

Crosstalk cancellation or speaker de-correlation is required for transaural monitoring of a stereo source. The crosstalk must be cancelled in order to preserve spatial information. Crosstalk cancellation inverts the effects of the crosstalk between the left loudspeaker and right ear and right loudspeaker and left ear.

Transaural monitoring is shown in Figure 12. The transfer function between the left ear and left speaker is labeled S , and the crosstalk between the right speaker and left ear is labeled A . If symmetry is assumed, then A also represents the crosstalk between the left speaker and right ear and S represents the transfer function between the right speaker and right ear.

A crosstalk cancellation filter is used to cancel the stereo crosstalk. The equation for a crosstalk cancellation filter in shuffler form is derived in [21], [46]

$$Y = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} \frac{1}{2(S+A)} & 0 \\ 0 & \frac{1}{2(S-A)} \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \quad (36)$$

Transaural monitoring of a stereo source generated by HRTFs requires the implementation of the shuffler filter for crosstalk cancellation. Figure 13 shows HRTF synthesis and crosstalk cancellation in a structural diagram. Loudspeaker equalization is required to preserve spatial cues, although it is not shown.

The crosstalk cancellation filter is dependent on the listener's position and loudspeaker setup. Any movement of the listener's head or change in loudspeaker configuration requires new crosstalk cancellation filters. The listener cannot be

expected to remain perfectly still and to have an ideal loudspeaker setup. A simpler technique consisting of ILDs and some ITDs is recommended in section 4.6 for spatialization.

Figure 12 – Transaural Monitoring with Crosstalk

Figure 13 - Transaural Stereo with HRTF Synthesis and Crosstalk Cancellation

4 Synthetic Stereo Echo Cancellation Structure

4.1 Introduction to the Synthetic Stereo Structure

The synthetic stereo acoustic echo cancellation structure provides full-duplex hands-free CTI conferencing for multiple participants. The system uses static spatialization functions which give the near-end talker localization cues to distinguish between far-end talkers in the conference. This is different than true stereophony, which gives the near-end listener localization cues for the far-end talker's location and movement. Synthetic stereo does not provide information about the far-end talker's movement. The far-end talker's spatial cues remain constant. Figure 14 shows localization cues for the near-end listener to distinguish among far-end talkers. The synthetic stereo structure is discussed further in [2].

The synthetic stereo structure is compatible with the standard multimedia computer configuration. The standard multimedia computer configuration consists of a monaural microphone and a set of stereo speakers. If a stereo microphone was used in the conference, strict microphone placement guidelines would be required to produce effective binaural recordings. Microphone equalization would also be required to preserve the spatial cues. It is difficult to support the wide range of microphones, loudspeakers and physical configurations to produce effective binaural recordings. The synthetic stereo structure does not require equalization or strict microphone placement guidelines for monaural recordings.

Figure 14 - Multiple Participant Spatialization

4.2 Stereophonic and Synthetic Stereo High-level Differences

The high-level structure of a typical stereophonic conference is shown in Figure 15. There are two microphones per talker and two loudspeakers per talker. The proposed high-level synthetic stereo structure differs from the stereophonic structure because there is only one microphone per talker and the stereo output is produced by static spatialization functions. The stereophonic acoustic echo cancellation structure, such as the one in Figure 8, employed dynamic transmission room transfer functions. These transmission room transfer functions are replaced by static

spatialization functions, in the synthetic stereo structure, which are unique to each participant. The high-level synthetic stereo structure is shown in Figure 16.

Figure 15 - Highlevel Stereophonic Structure for VoIP

4.3 Explanation of the Synthetic Stereo Structure

The high-level synthetic stereo structure in Figure 16 is used to derive the acoustic echo cancellation structure shown in Figure 17. The structure shown in Figure 17 represents the blocks from Figure 16 that are labeled left channel

spatialization, right channel spatialization and acoustic echo canceller. The synthetic stereo acoustic echo cancellation structure is shown with only one active talker.

Figure 16 - Highlevel Synthetic Stereo Structure for VoIP

The remote talker's speech is shown as $s(t)$ in Figure 17. The left channel is formed by convolving the monaural source by the left channel spatialization function, $g_1(t)$

$$left(t) = s(t) * g_1(t) \quad (37)$$

Figure 17 - Synthetic Stereo Acoustic Echo Cancellation Structure

The right channel is formed by convolving the monaural source, $s(t)$, by the right channel spatialization function, $g_2(t)$

$$right(t) = s(t) * g_2(t) \quad (38)$$

The primary signal received by the microphone is

$$d(t) = s(t) * g_1(t) * h_1(t) + s(t) * g_2(t) * h_2(t) \quad (39)$$

where $h_1(t)$ is the left channel transmission room transfer function and $h_2(t)$ is the right channel transmission room transfer function.

The reference, $s(t)$, to the acoustic echo canceller is taken before the spatialization functions. The canceller will, therefore, converge to

$$\bar{h}(t) = g_1(t) * h_1(t) + g_2(t) * h_2(t) \quad (40)$$

Synthetic stereo cancellers must only track changes in the reception room, since the spatialization functions are static. Stereophonic cancellers must track changes in both the transmission room, $g_1(t)$ and $g_2(t)$, and the reception room transfer functions, $h_1(t)$ and $h_2(t)$. The adaptive filter length in synthetic stereo is much shorter than for stereophonic acoustic echo cancellation. The adaptive filter length is shorter because the FIR representation for the spatialization functions is much shorter than for the transmission room transfer functions.

The synthetic stereo acoustic echo canceller identifies the echo path solution without the stereophonic cross-correlation problem. The synthetic stereo structure is comparable to a monaural acoustic echo cancellation structure in terms of canceller convergence, Figure 2. The well-known NLMS algorithm may be used for the synthetic stereo structure. The tracking performance and misadjustment are identical for both the synthetic stereo canceller and standard monaural echo canceller using the NLMS algorithm.

4.4 Synthetic Structure for Multiple Active Talkers

The synthetic stereo structure can converge, in practice, when multiple talkers are active. The multiple active talker structure, shown in Figure 18 for two talkers, is

similar to the stereophonic multiple canceller structure, see section 3.1.1. One acoustic echo canceller is used per talker.

Figure 18 - Synthetic Stereo Acoustic Echo Cancellation Structure for Two Talkers

The correct echo paths for the two active talkers are

$$\tilde{h}_1(t) = g_1(t) * h_1(t) + g_3(t) * h_2(t) \quad (41)$$

$$\tilde{h}_2(t) = g_2(t) * h_1(t) + g_4(t) * h_2(t) \quad (42)$$

The error function, $e(t)$, for this system is

$$\begin{aligned}
e(t) = & s_1(t) * [g_1(t) * h_1(t) + g_3(t) * h_2(t)] \\
& + s_2(t) * [g_2(t) * h_1(t) + g_4(t) * h_2(t)] - s_1(t) * \tilde{h}_1(t) - s_2(t) * \tilde{h}_2(t)
\end{aligned} \tag{43}$$

If the error signal is minimized by the canceller, $e(t) = 0$, then

$$\begin{aligned}
s_1(t) * [g_1(t) * h_1(t) + g_3(t) * h_2(t) - \tilde{h}_1(t)] = \\
- s_2(t) * [g_2(t) * h_1(t) + g_4(t) * h_2(t) - \tilde{h}_2(t)]
\end{aligned} \tag{44}$$

From (44) it can be seen that there are multiple solutions to the problem. The same cross-correlation problem that existed for the stereophonic multiple canceller structure, see section 3.1.1, is encountered when multiple talkers are active.

System adaptation must be frozen when more than one talker is active to avoid the problem of cross-correlation between echo cancellers. In practice, the multiple active talker acoustic echo cancellation structure converges but the convergence rate and tracking ability are very poor relative to the single active talker version.

4.5 Synthetic Stereo Implementation Issues

The multiple participant synthetic stereo structure requires a separate set of adaptive filter coefficients for each participant. The memory requirements for P participants, where P includes the near-end listener, is

$$MEM = 2N*(P-1) \tag{45}$$

The coefficients must be swapped whenever the active participant in the conference changes.

The echo canceller must be allocated to the participant producing the echo signal with the highest signal power since only one echo canceller can be active at a time. Active participants that are not allocated the echo canceller must still be convolved with their frozen echo paths.

The synthetic stereo structure requires speech detectors and a method for dealing with double-talk. The speech detector is used to determine which far-end talkers are active. If double-talk is detected the echo canceller must be frozen for all the participants and all active talkers will still be convolved with their frozen echo paths.

4.6 Spatialization Functions

Multiple talker localization for the far-end talkers is performed by spatialization functions. The two spatialization methods discussed in this thesis are the head related transfer function and interaural level and time differences.

The head related transfer function (HRTF) is well suited to binaural applications, such as those involving headphones [17]. Transaural monitoring of HRTF stereo signals can be performed through convolution with the appropriate loudspeaker de-correlation functions [21], see section 3.3.3. The speaker de-correlation functions

are very sensitive to listener and loudspeaker position. Any variation in the physical setup or movement of the listener's head will cause listener confusion.

The solution is to use interaural level differences (ILDs) and interaural time differences (ITDs) [18][48]. The stereophonic law of sines can be used to compute ILDs, see section 3.3.1. ITDs can be used to further enhance the listener's sound stage but should not exceed $\pm 3\text{ms}$.

4.7 Simulation

Room transfer functions were measured for the CTI setup shown in Figure 30. The left channel impulse response, $h_1(t)$, was measured from the left speaker to the microphone. The right channel impulse response, $h_2(t)$, was measured from the right speaker to the microphone. Further description for these measurements can be found in Appendix A. The room transfer functions were measured for a sound pressure level of 100dB and are shown in Figure 19.

The spatialization functions are based on intensity stereo techniques, ILDs, and small amounts of ITD. The spatialization functions are 24-tap finite impulse response (FIR) filters, which correspond to a maximum ITD of $\pm 3\text{ ms}$ at an 8 kHz sampling rate.

The synthetic stereo canceller structure simulations were performed using the room transfer functions and the 24-tap FIR spatialization functions. The simulations were performed using one active participant. Figure 20 shows the convergence rate

for the synthetic stereo canceller with a 500-tap adaptive filter using the NLMS filter. The maximum theoretical ERLE for the 500-tap adaptive filter is 25.5 dB which the filter is shown to converge to in Figure 20.

Figure 19 – Transfer Functions for Measured Room

4.8 Stereophonic and Synthetic Stereo Properties

The basic stereophonic acoustic echo cancellation structure has many problems with convergence to the true echo paths. The difficulties are as follows:

- cross-correlation between the two primary signals
- echo paths are not uniquely identified (system misconverges) [44]
- the algorithm must track changes in the transmission and reception room transfer functions (two levels of tracking) [34]

Figure 20 - Synthetic Stereo Cancellation Convergence

The advantages offered by the synthetic stereo structure for multiple participant conferences are as follows:

- no cross-correlation between the primary signals
- echo paths are uniquely identified
- static spatialization functions (one level of tracking for reception room)

5 Acoustic Round Trip Delay

The acoustic round trip delay of a VoIP system involves a large number of system components that must interact in the real-time conference. Current round trip delay methods address the computation of the network round trip delay, but do not address the complete acoustic round trip delay as perceived by the user [9][11][43]. Further information concerning the determination of acoustic round trip delay can be found in [1].

This thesis proposes a theoretical model that illustrates the delays in VoIP implementations. An empirical measurement method is also proposed to perform actual acoustic round trip delay measurements on VoIP platforms. The empirical measurement method is H.323 [27] compliant, which is important for maintaining compatibility with the CTI standard. The empirical method was applied to the DSP/PC platform, dMIPS, and an additional platform called the multimedia PC platform, MIPS. The MIPS platform was developed in order to apply our research in the area of acoustic round trip delay to platforms based on a PC equipped with a sound card. Finally an acoustic domain verification test was performed to ensure that the theoretical model and empirical measurement method are accurate.

5.1 Measuring the Quality of Service with Acoustic Round Trip Delay

The acoustic round trip delay that the user perceives partially characterizes the quality-of-service (QoS) of the conference system front-end. Studies have shown that increasingly long latencies cause talker confusion and difficulties in maintaining

normal conversation flow [51]. A conference system with long delays causes users to engage in double talk and mutual silence more often than one with short delays. As the delay increases, the likelihood of a participant interrupting the conversation increases, the likelihood of the user terminating the interruption decreases and the likelihood that the user will respond to questions or pauses in the conversation decreases [51]. These effects reflect an overall change in normal conversation flow. Comparing acoustic round trip delay data with these psychological effects allows an evaluation of the QoS that the user perceives.

5.2 Switched-Loss Level Adjustment with Acoustic Round Trip Delay

All practical implementations of full-duplex hands-free conference sets require switched-loss in addition to acoustic echo cancellation. The nonlinear processor (NLP) sets the amount of switched-loss that is required, see section 2.2.2. PSTN networks do not have the ability to determine the acoustic round trip delay. CTI systems that employ the acoustic round trip delay empirical measurement method, however, have the ability to determine this delay. The acoustic round trip delay can be used by the NLP to set the level of switched-loss required by the conference.

Psycho-acoustic data [22][23] relates the level of Talker Echo Loudness Rating (TELRL) to the mean one-way propagation delay, see Figure 3. The mean one-way propagation delay can be calculated by dividing the acoustic round trip delay by two. The amount of switched-loss can be computed for a given TELRL

$$TELRL = ERLE_c + A_s \quad (46)$$

where $ERLE_C$ is the maximum ERLE for a specific acoustic echo canceller, see 2.2.1.4, and A_S is the amount of switched-loss required by the system.

Figure 21 shows an example for a maximum $ERLE_C$ of 40dB. Delays under 50ms require no switched-loss, however delays over 50ms require the addition of switched-loss.

Figure 21 – Switched-Loss Level Computation

The amount of switched-loss added to a conference should be minimized to maintain the full-duplex nature of the conference and to avoid adverse half-duplex effects, see 1.1.

No excessive amounts of half-duplex need to be added to the conference if the precise level of switched-loss is known. Previously, the NLP determined the appropriate level of switched-loss based on the worst case scenario. If the upper bound on the mean one-way propagation delay for the conference is 300ms then approximately 15dB of switched-loss attenuation is required for the example shown in Figure 21. If the acoustic round trip delay empirical measurement method is used to determine that the delay is actually 100ms then only 5dB of switched-loss attenuation is required. The combination of the empirical measurement method with psycho-acoustic data minimizes switched-loss attenuation by 10dB.

5.3 Theoretical Model for the Acoustic Round Trip Delay

The acoustic round trip delay is the time taken for the near-end user's voice to echo at the far-end in a full-duplex hands-free system and return to their ears. These delays are shown in Figure 22.

The multimedia PC platform, MIPS and the DSP/PC platform, dMIPS, are both affected by the same delays, except that the multimedia PC platform employs an audio double buffer for playback and the DSP platform requires only a single buffer for playback. The DSP/PC platform is able to service requests for new samples on a sample-by-sample basis at the sampling rate of the audio playback codec. The multimedia PC platform services requests for new audio frames on a frame-by-frame basis and must maintain an audio double buffer for continuous playback.

Figure 22 - Sources of Delay in a VoIP System

5.3.1 Assumptions for the Theoretical Model

Several assumptions concerning the VoIP conference are made in order to model the acoustic round trip delay of the system. The theoretical model and empirical measurement method use these assumptions.

1. All observations are made with reference to the start of a recorded audio frame.
2. Both ends are identical in hardware and software. (Same CPU power, operating system, memory configuration, etc...) Therefore $\Delta p_x = \Delta p_x'$, $\Delta p_y = \Delta p_y'$ and $\Delta p_z = \Delta p_z'$. (see Table 3)
3. Recording, encapsulation and network events are independent and they occur simultaneously.
4. The delayed playback jitter buffer threshold is equal at both stations.

5. A new audio frame may be queued in the playback audio buffer without interrupting playback.
6. The playback audio buffer queue is a double buffer for the multimedia PC platform and a single buffer for the PC/DSP platform.
7. Both participants have reached steady state. (Conference has started and proceeded for some time.)
8. The room impulse responses for both ends are static and identical. ($t_a = t_a'$)
9. $(t_r - t_a)$ is constant for a station at steady-state.

5.3.2 Description of the Delay Components

Delay	Multimedia PC and DSP/PC Platforms
t_f	The length of an audio frame in seconds.
Δp_x	Processing time taken to code the audio frame with the codec and to encapsulate the frame with RTP, UDP and IP.
Δt_n	DSP to PC communication time (DSP only). Network one-way transmission and propagation delay.
Δp_y	Processing time to decapsulate the frame, generate RTP statistics, generate an RTCP packet (if required) and enqueue the audio frame in the jitter buffer <ul style="list-style-type: none"> • IP, UDP and RTP decapsulation of network audio packet to form audio frame • Generate RTP statistics such as jitter and total received bytes • Generate RTCP packet at set interval to inform conference participants of pending network congestion and other statistics
Jitter Buffer	PC to DSP communication time (DSP only). The expected jitter buffer size is $E[JBS]$.
Queue G/D/1	The jitter buffer can be realized as a single server - single queue with general arrival and deterministic service time.
Audio Buffer	The expected audio buffer size is $E[ABS]$.

D/D/1	<p>The audio buffer can be realized as a single server – single queue with deterministic arrival and service distributions.</p> <p>$ABS_{max} = 2$ on the multimedia PC platform and $ABS_{max} = 1$ on the DSP/PC platform.</p> <p>Processing time to decode the audio frame with the voice codec.</p>
Δp_z	Audio propagation time
t_a	<ul style="list-style-type: none"> Includes propagation time through the microphone and speakers Includes the direct acoustic path from the speaker(s) to the microphone Measured as $t = d/v$ where d is the distance between the speaker and the microphone, v is the velocity of sound waves (330m/s) at room temperature OR measured by the room impulse response
$t_r - t_a$	<p>Self-synchronization parameter with audio delay</p> <ul style="list-style-type: none"> The amount of time remaining in the recording of an audio frame after a frame begins playback at the same station Accounts for the lack of synchronization between record and playback events in a station The empirical measurement model does not include the acoustic delay in the synchronization parameter, t_r $E[t_r] = 1/2t_f$
$t_o + t_a = t_r - t_r + t_a$	<p>Complementary self-synchronization parameter with audio delay</p> <ul style="list-style-type: none"> The time delay to reach the beginning of the echo recorded by the other station in the audio frame being played (includes acoustic delay from other station) The empirical measurement model does not include the acoustic delay in the complementary synchronization parameter, $t_o = t_r - t_r$
t_v	<p>Cross-synchronization parameter</p> <ul style="list-style-type: none"> Amount of time remaining in the playback of an audio frame when an incoming frame is placed in the jitter buffer $E[t_v] = 1/2t_f$

Table 3 - Description of Acoustic Round Trip Delay Components

The lack of synchronization between the record and playback events at a station requires compensation by the self-synchronization parameter, t_r . This parameter represents the difference between the start of a recorded frame and the start of a playback frame. This lack of synchronization is caused by the random startup times

for participants in the conference since all stations operate concurrently. The self-synchronization parameter is distributed between 0 and t_f

$$0 \leq t_\gamma \leq t_f \quad (47)$$

The self-synchronization parameter with acoustic delay, $t_\gamma - t_a$, adjusts the amount of time remaining in the recorded frame during the timestamp loopback to account for acoustic echo propagation. The complementary self-synchronization parameter with acoustic delay, $t_\gamma + t_a$, represents the offset time delay in the audio frame before the echo is played.

The lack of synchronization between stations in the conference requires the introduction of the cross-synchronization parameter t_v . This parameter represents the amount of time remaining in playback when a new frame is placed in the jitter buffer. It is distributed between 0 and t_f

$$0 \leq t_v \leq t_f \quad (48)$$

5.3.3 Pipeline Analogy for the Theoretical Model

A pipeline delay analogy is proposed to explain the concurrent nature of a VoIP system. Many activities in the VoIP system must occur concurrently to maximize the system's efficiency and model hardware operation. The acoustic round trip delay pipeline analogy model is shown in Figure 23.

The maximum jitter buffer size, JBS_{max} , and the maximum audio buffer size, ABS_{max} , are values critical to system performance. If a new frame does not arrive during the time the last queued frame spends in the jitter buffer, then the jitter buffer will starve by one frame. If the jitter buffer is allowed to starve by too many frames then there will be audible gaps in the system playback. An audible gap is characterized by the system shutting down the playback and re-buffering enough frames in the jitter buffer to start the delayed playback again. The audio buffer size is also important to the real-time nature of the conference. If the $E[ABS]$ becomes less than zero then the conference system is not able to servicing the audio buffer of the sound card or codec of the DSP card in real-time.

The pipeline analogy, Figure 23, starts with the first packet labeled, Packet 1, and proceeds with subsequent packets in increasing order. Each packet undergoes the same sequence of delays that were shown in Figure 22.

Figure 23 – Acoustic Round Trip Delay Pipeline Analogy

5.3.4 Equations for the Theoretical Model

The major points of interest on the time axis of Figure 23 are shown below.

$$\textbf{Event A: First frame queued} \quad T_A = t_f + \Delta p_{x1} + \Delta t_{n1} + \Delta p_{y1} \quad (49)$$

$$\textbf{Event B: Second frame queued} \quad T_B = t_f + \Delta p_{x2} + \Delta t_{n2} + \Delta p_{y2} \quad (50)$$

$$\textbf{Event C: Third frame queued} \quad T_C = t_f + \Delta p_{x3} + \Delta t_{n3} + \Delta p_{y3} \quad (51)$$

$$\textbf{Event D: First frame begins playback} \quad T_D = t_f + \Delta p_{x1} + \Delta t_{n1} + \Delta p_{y1} + E[JBS]t_f + \Delta p_{z1} + [E[ABS] - 1]t_f + t_v \quad (52)$$

$$\textbf{Event E: Second frame begins playback} \quad T_E = t_f + \Delta p_{x2} + \Delta t_{n2} + \Delta p_{y2} + E[JBS]t_f + \Delta p_{z2} + [E[ABS] - 1]t_f + t_v \quad (53)$$

$$\textbf{Event F: Third frame begins playback} \quad T_F = t_f + \Delta p_{x3} + \Delta t_{n3} + \Delta p_{y3} + E[JBS]t_f + \Delta p_{z3} + [E[ABS] - 1]t_f + t_v \quad (54)$$

$$\begin{aligned} \textbf{Event Z: First frame round trip delay} \quad T_Z = & t_f + 2\Delta p_{x1} + 2\Delta t_{n1} + 2\Delta p_{y1} \\ & + 2E[JBS]t_f + 2\Delta p_{z1} + 2[E[ABS] - 1]t_f \\ & + t_v + t_v' + t_a + t_\gamma - t_a + t_f - t_\gamma + t_a + t_a \end{aligned} \quad (55)$$

$$\begin{aligned} \textbf{Round Trip Delay} \quad ARTD = & 2t_f + 2\Delta p_x + 2\Delta p_y + 2\Delta p_z + 2\Delta t_n \\ & + 2E[JBS]t_f + 2[E[ABS] - 1]t_f + t_v + t_v' + 2t_a \end{aligned} \quad (56)$$

5.4 Empirical Measurement Method for the Acoustic Round Trip Delay

The acoustic round trip delay can be measured by adding timestamp information through an RTP header extension. This method is fully H.323 and RTP compliant since clients not supporting the acoustic round trip delay measurement method would ignore the header extension.

Figure 24 shows the progression of the timestamp information in the empirical measurement method. The shaded blocks are included in the acoustic round trip delay empirical computation method.

Figure 24 – Acoustic Round Trip Delay Empirical Measurement Method

The initial recording delay, t_f , is not captured by the empirical measurement method. After the initial recording delay every frame is timestamped at transmission time by the RTP layer of station A.

When the frame reaches station B, the timestamp information from station A is looped back through the RTP transmission layer of station B. The timestamp information from station A is piggybacked onto the next frame that completes recording as a proprietary RTP extension. The acoustic delay, t_a , is not captured by the empirical method.

The effective playback audio buffer delay includes the cross-synchronization parameter, t_v . The stations are not synchronized in their producer-consumer relationship and a new frame can arrive at any time during playback. The maximum

audio buffer delay is $EABD = [E[ABS]-1]t_f + t_v$. The reference for the empirical model is the beginning of a recorded frame, therefore, one frame must be subtracted from the $E[ABS]$ since the audio information is at the start of the frame.

After playback at station A the acoustic round trip delay is measured as the difference between the current system time and the old time contained in the RTP header extension. The final acoustic delay, t_a' , is not included in the measurement method.

$$ARTD_{Measured} = T_{current} - T_{old} \quad (57)$$

$$ARTD_{Measured} = 2[\Delta p_x + \Delta p_y + \Delta p_z + \Delta t_n] + \langle 2[E[JBS] + E[ABS] - 1] \rangle t_f + t_v + t_v' + t_f \quad (58)$$

The empirical measurement method does not include the initial recording delay, t_f , or the acoustic delays for either participant. The missing delays are included in the correction factor, C_F

$$C_F = ARTD_{Missing} = t_f + t_a + t_a' \quad (59)$$

The actual acoustic round trip delay is the sum of the measured round trip delay and the correction factor

$$ARTD = ARTD_{Measured} + C_F \quad (60)$$

$$ARTD = 2[\Delta p_x + \Delta p_y + \Delta p_z + \Delta t_n] + 2t_f + 2t_a + \langle 2[E[JBS] + E[ABS] - 1] \rangle t_f + t_v + t_v' \quad (61)$$

Equation (61) is the same as the acoustic round trip delay for the theoretical model given in (56).

The dominant factors in the acoustic round trip delay are the audio frame size, the jitter buffer size, the playback buffer size and the voice coder delay. The user or system designer sets the audio frame size, based on the platform, to optimize the system's performance. The optimal jitter buffer size can be determined analytically, [50]. The audio playback buffer delay is platform dependent. The multimedia PC platform requires an audio double buffer for playback that causes increased delay over the DSP/PC platform. The DSP/PC platform requires only a single buffer for playback. The voice coder delay can vary widely dependent on the coding scheme in use. A typical voice coder delay can range from 30 to 100ms [16].

The acoustic delay, t_a , is dependent on the room in use and the physical setup of the speakers and microphone. The acoustic delays are negligible compared with the network delays, jitter and audio buffer delays and voice coder delays, therefore

$$ARTD \equiv ARTD_{Meas} + t_f \quad (62)$$

5.4.1 Acoustic Round Trip Delay Measurement

Several acoustic round trip delay measurements were completed using empirical measurement method on the dMIPS VoIP platform. The results were taken for a local LAN CTI conference.

The DSP/PC platform gives the best results since the DSP is dedicated to running the dMIPS software only. The DSP/PC platform shows very little variation since the DSP processor generates timestamp information and performs the final round trip delay computation without operating system intervention.

Large spikes in the acoustic round trip delay data are caused by lost packets in the network. Our filtering method computes an upper boundary for the acoustic round trip delay based on a priori knowledge of the system parameters and eliminates these spikes from the round trip delay data.

5.4.2 DSP/PC Platform Results

The results shown in Figure 25 and Figure 26 verify (58). Each total delayed playback threshold (TDPT) setting was ensemble averaged over a sequence of five independent runs. The TDPT is the maximum number of jitter and audio buffers for both stations in the conference.

$$TDPT = 2(JBS_{\max} + ABS_{\max}) \quad (63)$$

Figure 25 – DSP/PC Acoustic Round Trip Delay Measurements for 20ms Frame

Figure 26 – DSP/PC Acoustic Round Trip Delay Measurements for 40ms Frame

The graphs show that as the TDPT setting increases by two there is a difference of two audio frames in the acoustic round trip delay measurement. In Figure 25 a difference of 40ms is shown for each increase of two in TDPT, since the frame size is 20ms. In Figure 26 a difference of 80ms is shown for each increase of two in TDPT.

The results from the acoustic round trip delay empirical measurement method verify that the theoretical model is in accordance with the empirical measurement method.

5.5 Acoustic Domain Verification Test

The acoustic domain verification test was used as a sanity check to measure the acoustic round trip delay that the user perceives. The acoustic domain verification test can include all the system delays. The verification test is not appropriate for measuring acoustic round trip delay on a VoIP platform, however, because of its complexity and white Gaussian noise (WGN) reference signal requirements. These requirements make acoustic domain verification testing appropriate for off-line use only. The white Gaussian noise reference signal required to perform the cross-correlation would make conversation very difficult for the VoIP users. Also, the immense computational complexity required for cross-correlation with long system delays makes real-time signal processing impractical.

The results obtained for the acoustic domain verification test will show that the empirical measurement method is accurate and that there is no need to perform acoustic domain verification testing for VoIP implementations.

5.5.1 Acoustic Domain Verification Test Procedure

The procedure for the acoustic domain verification test is to play a white Gaussian noise reference signal through a multi-track digital audio tape (DAT) recording and record the far-end acoustic echo. The reference signal and the echo are then cross-correlated to obtain the acoustic round trip delay.

Figure 27 - Acoustic Domain Verification Test

Figure 27 shows the setup for the acoustic domain verification test. The DAT plays white Gaussian noise through the microphone input to the DSP card and records the echo through the speaker output from the DSP card. Station A is engaged in a steady-state VoIP conference with station B. The acoustic loopback at

station B is simulated by directly cabling the speaker output from the DSP card to the microphone input to the DSP card.

The acoustic domain verification test does not include the audio propagation times, t_a and t_a' , for comparison with the empirical measurement method, which does not include them either. The acoustic round trip delay measured by the acoustic domain verification test is

$$ARTD_{Acoustic} = ARTD_{Measured} + t_f \quad (64)$$

5.5.2 Acoustic Domain Verification Test Results

The results shown in Figure 28 show that the acoustic round trip delay measured by the empirical measurement method is one audio frame less than the delay measured by the acoustic domain verification test. These test results verify (64) and prove that the empirical measurement method is an accurate way to determine the acoustic round trip delay that the user perceives.

Figure 28 - Acoustic Domain Verification Test Results

6 Real-Time VoIP Test Platform

The Carleton University real-time VoIP platform was designed to test the synthetic stereo acoustic echo cancellation structure and measure acoustic round trip delay through the empirical measurement method. The VoIP platform is called the DSP multi-participant Intranet phone system (dMIPS). The high-level view of the system's building blocks is shown in Figure 29.

Figure 29 – High-Level View of VoIP Platform

6.1 DSP/PC Platform Hardware Requirements

The hardware and software requirements for the DSP/PC platform, dMIPS, are as follows:

- Intel Pentium Processor based PC 166 Mhz or Greater
- Windows NT Workstation 4.0 or Windows '95
- Visual C++ v5.0 32-bit (development only)
- PCI 10Mb/s Ethernet Card
- Winsock 2.0 for Windows '95 (included for Windows NT Workstation)
- Tetrydyne Systems' DriverX Software v3.3 or Greater
- Stereo Multimedia Speakers with Internal Amplifier
- Monaural Cardioid Microphone (ex. Audio Technica ATM—87R boundary mic)
- Microphone Pre-Amp with Line Level Output to DSP Card
- Go-DSP's Code Composer Software (required to load DSP code)
- Tiger 542/PC DSP Board or Tiger 548/PC DSP Board Based on the TMS320C542 or TMS320C548, respectively
 - Jumper settings are: OVL off (connect pins 2-3), TBCOFF no jumper, ALN no jumper, ROM no jumper, IRQ select IRQ 10, S1 select IO port 2A0h ON-OFF-ON, MIC 2Vrms JP16 and JP18 jumper installed, OGAIN don't care

6.2 VoIP Acoustic Front-End

The acoustic front-end consists of a set of amplified stereo loudspeakers and a microphone, which is placed on the user's desk or mounted on their computer. The microphone is connected to a microphone pre-amp and the pre-amp output is connected to the line level input to the DSP board's codec. The loudspeakers are

connected directly to the DSP board's codec output. The acoustic front-end is shown in Figure 30 and the DSP board connections are shown in Figure 31.

Figure 30 – Multimedia Configuration for VoIP Platform

6.3 VoIP Software Requirements

The DSP card software performs the acoustic echo cancellation, spatialization, jitter buffering, and voice coding and decoding required by the platform. The PC software sends and receives packets over the Ethernet using UDP/IP, performs RTP encapsulation and decapsulation, and transfers data frames to and from the DSP card. The PC and DSP software communicate through transfer registers on the DSP card, through interrupts to the DSP and PC, and shared memory on the DSP card, called the host port interface memory. For the software design use case maps and activity descriptions see Appendix C.

Figure 31 - Acoustic Front-End for VoIP Platform

The DSP board generates an interrupt to the PC whenever captured data is ready for transmission over the network by the PC. The PC generates an interrupt to the DSP whenever valid data frames have been received by the network and are ready for playback. The data is transferred by programmed input/output (PIO) mode instead of direct memory access (DMA) mode because the Tiger 542/PC does not support DMA. Using a DSP board capable of DMA transfers would reduce the data transfer overhead.

6.4 VoIP Audio Data Format

The dMIPS software supports a simple audio frame format. The DSP card codec produces linear PCM 16-bit data at an 8 kHz fixed sampling rate. The frames transmitted by dMIPS are not coded or decoded in the current version. A PSTN produces analog audio that is equivalent to digital audio with linear PCM 8-bit data at an 8 kHz sampling rate. The audio quality produced by dMIPS is better than the audio quality of a PSTN since enhanced echo cancellation depth and sound quality were desired.

6.5 H.323 Compatibility

Real-time data transmission over the network conforms to the recommendations in ITU-T H.323, [19], concerning the usage of RTP/RTCP. Call setup and logical channel setup are not implemented by the Carleton University VoIP platform. The user manually specifies the destination IP address and RTP port to setup the logical channel. The logical channel consists of an RTP channel for real-time data exchange and an RTCP channel to report network statistics.

The following ITU-T H.225.0, [20], or RTP/RTCP functions are implemented by the VoIP platform:

- end-to-end real-time data delivery
- payload type identification
- sequence numbering
- timestamping

- delivery monitoring
- quality-of-service (QoS) monitoring through RTCP

6.5.1 Connectionless Service Requirement

The real-time transport protocol (RTP) and the RTP control protocol (RTCP) frames are transmitted by UDP/IP. Connectionless service, such as UDP, is required by H.323 for real-time traffic since UDP has lower overhead and does not perform automatic packet recovery, see section 2.3.1.

6.5.2 Multipoint Support

Multiple participant conferences are supported by the platform to test the synthetic stereo echo cancellation algorithm. Two multipoint distribution schemes were considered, see section 2.4.3, for the IPv4 network used by dMIPS:

1. using an Ethernet broadcast to distribute packets to all stations on the local LAN (IPv4 broadcasts will not pass through gateways, routers, switches or bridges)
2. replicate packets at the host and direct them to multiple IP addresses

The first scheme is preferable because it requires less processing power and uses less network bandwidth. Only one copy of each packet needs to be sent to the participants in the conference. The second scheme requires packet replication by the MCU, see section 2.4.3, inside the PC terminal for IP unicast. The Carleton

University VoIP platform uses the first scheme for multipoint conferencing by IP broadcast.

6.6 DSP Implementation of an NLMS Acoustic Echo Canceller

Several TMS320C54x features were used to implement the normalized least mean square (NLMS) algorithm on the DSP board. The DSP features used to implement the algorithm are circular addressing, free shifts by powers of 2, the multiply and accumulate (MAC) instruction for FIR filtering and a fast unsigned division algorithm. The NLMS DSP code is contained in Appendix D.

Circular addressing is a powerful feature of the C54x architecture. The reference taps for an FIR filter can be stored in a buffer which uses circular addressing to access the filter taps. The circular addressing operator, +0%, increments the address in the specified auxiliary register by the value stored in auxiliary register 0. The circular addressing operator automatically restarts the address at the beginning of the buffer after the end is reached.

The MAC instruction is used by the NLMS algorithm to convolve the adaptive filter coefficients with the reference taps. When the MAC instruction is combined with the repeat instruction (RPT) and circular buffering it can execute in one cycle. This combination of instructions can be used to convolve an N tap FIR filter in N cycles. The MAC instruction, the RPT instruction and circular addressing are used in the following example:

```

RPT #(N-1)                ; Repeat for N taps
MAC *COEF_PTR+, *REF_PTR+0%, A ; multiple, accumulate and round

```

The C54x architecture can perform free shifts by powers of 2 when they are combined with other operations. The C54x architecture has a 40-bit barrel shifter that can be programmed to shift results from the arithmetic logic unit or data memory. Free shifts by powers of 2 are performed by 'STM #(16+ALPHASFT),T', 'SUB LTPOWER, TS, A', 'ADD POWER0, TS, A' and 'ADD EPSILON, TS, A' in this routine to compute the long term power:

```

; Update long-time reference power estimate (LTPOWER)
STM #(16+ALPHASFT), T      ; for dynamic shift load ALPHA = 2^ALPHASFT
; Shift arguments into high bit by 16+ALPHASFT which results in a 2^ALPHASFT
; division in the high part of the accumulator
DLD LTPOWER, A             ; double precision load old LTPOWER Q.31
; LTPOWER(n) = LTPOWER(n-1) - ALPHA*LTPOWER(n-1) + ALPHA*x(n)^2
SUB LTPOWER, TS, A         ; A = A - LTPOWER*2^ALPHASFT
ADD POWER0, TS, A          ; A = A + POWER0*2^ALPHASFT
DST A, LTPOWER             ; double precision store new LTPOWER
ADD EPSILON, TS, A         ; A = A + EPSILON*2^ALPHASFT
STH A, ALTPOWER            ; single precision store adjusted LTPOWER

```

The NLMS algorithm requires unsigned division to normalize results by the long-term signal power. The following algorithm divides two Q.15 operands to obtain a Q.0 result in very few cycles. The following unsigned division algorithm computes $RESULT = 1/OPERAND$:

```

LD #32767, A
RPT #(16-1)
SUBC OPERAND, A
STL A, RESULT

```

6.7 DSP Implementation of an FIR Spatialization Filter

The DSP implementation of the finite impulse response (FIR) spatialization algorithm uses several C54x features. The FIR filtering is performed by the multiply and accumulate (MAC) instruction combined with the repeat (RPT) instruction and circular addressing, see section 6.6. The multiply, accumulate and delay (MACD) instruction is normally used for FIR filtering in Texas Instrument's fixed-point processors. The MACD instruction could not be used for the spatialization filter since each participant's coefficients start at a different memory location. The MACD instruction requires that the filter coefficient's start at a constant address. The DSP code for the spatialization algorithm is contained in Appendix E.

7 Conclusions

7.1 Synthetic Stereo Acoustic Echo Cancellation Structure

The proposed synthetic stereo acoustic echo cancellation structure solves the multiple participant full-duplex hands-free stereo conferencing problem. The synthetic stereo structure completely de-correlates the primary signals through the use of static spatialization functions. The near-end listener can distinguish between far-end talkers through localization cues generated by spatialization functions.

The synthetic stereo structure simulation shows that the convergence rate and misadjustment for the synthetic stereo acoustic echo cancellation and monaural echo cancellation structure are equivalent when the NLMS algorithm is used. Monaural echo cancellation structure research is also applicable to the synthetic stereo structure because the structures are equivalent.

The synthetic stereo acoustic echo cancellation structure was implemented in the Carleton University dMIPS VoIP platform for demonstration purposes.

7.1.1 *Future Research in Synthetic Stereo Cancellation*

A subjective analysis should be undertaken to determine the maximum number of participants that users can distinguish between using ILD spatialization techniques. Enhanced virtual sound techniques appropriate for use as VoIP spatialization functions should be investigated. The maximum number of participants that can be represented by these enhanced spatialization functions should also be determined.

Further research into speaker de-correlation functions appropriate for VoIP conferencing should be investigated. A generic speaker de-correlation function that can be used for a wide range of listener positions is required for transaural monitoring of head related transfer functions (HRTFs). Speaker de-correlation functions capable of tracking the user's position and loudspeaker setup are also important for research in transaural monitoring.

7.2 Psycho-Acoustic Determination of Echo Cancellation Targets

This thesis shows that psycho-acoustic data can be used to determine the echo return loss enhancement (ERLE) target for an acoustic echo canceller. The ERLE target is determined by the mean one-way propagation delay. ERLE targets can be used by the nonlinear processor (NLP) to minimize the level of switched-loss in a conference.

7.2.1 *Future Research in Psycho-Acoustic Echo Cancellation Targets*

The method for determining echo cancellation targets based on psycho-acoustic data should be implemented on the Carleton University dMIPS platform for subjective testing. Conventional NLP designs should be compared against the proposed psycho-acoustic NLP design.

7.3 Acoustic Round Trip Delay Computation

A theoretical model for the acoustic round trip delay in a VoIP conference was proposed. An empirical measurement method was also proposed to measure acoustic round trip delay data. Both the theoretical model and empirical measurement method were checked by the acoustic domain verification test.

The empirical measurement method was implemented on the Carleton University dMIPS platform to collect acoustic round trip delay data. Ensemble averages were performed on the acoustic round trip delay data to obtain a plot showing the delay for various buffer sizes. It was shown that increasing the buffer sizes created a corresponding increase in the acoustic round trip delay.

Acoustic domain verification testing showed that the delays measured by the empirical measurement method were in accordance with what the VoIP user would perceive to be the conference delay.

7.3.1 Future Research in Acoustic Round Trip Delay

The effect of the operating system and system device drivers on VoIP system performance should be investigated. This research could be used to derive statistical models for various parameters in the acoustic round trip delay theoretical model.

7.4 VoIP Platform Design

A DSP/PC VoIP platform called dMIPS was design and implemented for demonstrating the synthetic stereo acoustic echo cancellation structure and measuring acoustic round trip delay. A multimedia PC platform called MIPS was also designed and implemented to measure acoustic round trip delay data. These platforms were built because there were no commercially available systems that could be customized for VoIP research.

7.4.1 *Future Research in VoIP Platform Design*

The MIPS platform should be used to measure acoustic round trip delay data to verify that the theoretical model and empirical method are valid for the multimedia PC configuration. The spatialization functions and synthetic stereo acoustic echo cancellation structure could also be implemented on the multimedia PC platform if there is enough processing power.

The impact of Internet protocol version 6 (IPv6) and the reservation protocol (RSVP) on CTI design should be studied. The combination of these protocols should reduce network latency and enhance the quality-of-service (QoS) for CTI conferencing. Various RSVP bandwidth management schemes and their impact on CTI design should also be studied.

8 References

- [1] T. Yensen, M. Parperis, I. Lambadaris, and R. Goubran, "Determining Acoustic Round Trip Delay for VoIP Conferences," MMSP98, Dec 1998.
- [2] T. Yensen, R. Goubran and I. Lambadaris, "A Multiple Participant CTI Acoustic Echo Cancellation Structure," NORSIG98, pp. 193-196, June 1998.
- [3] J. Benesty, D. Morgan and M. Sondhi, "Stereophonic Acoustic Echo Cancellation Using Nonlinear Transformations and Comb Filtering," ICASSP98, pp. 3673-3676, May 1998.
- [4] Y. Joncour and A. Sugiyama, "A Stereo Echo Canceller with Pre-Processing for Correct Echo-Path Identification," ICASSP98, pp. 3677-3680, May 1998.
- [5] A. Gilloire and V. Turbin, "Using Auditory Properties to Improve the Behaviour of Stereophonic Acoustic Echo Cancellers," ICASSP98, pp. 3681-3684, May 1998.
- [6] S. Shimauchi, Y. Haneda, S. Makino and Y. Kaneda, "New Configuration for a Stereo Echo Canceller with Nonlinear Pre-Processing," ICASSP98, pp. 3685-3688, May 1998.
- [7] M. Ali, "Stereophonic Acoustic Echo Cancellation System Using Time-Varying All-Pass Filtering for Signal Decorrelation," ICASSP98, pp. 3689-3692, May 1998.
- [8] J. Benesty, D. Morgan and M. Sondhi, "A Better Understanding and an Improved Solution to the Specific Problems of Stereophonic Acoustic Echo Cancellation," IEEE Trans. on Speech and Audio Proc., vol. 6, no. 2, pp. 156-165, March 1998.
- [9] L. Wolf, C. Griwodz and R. Steinmetz, "Multimedia Communication," Proc. IEEE, vol. 85, no. 12, pp. 1915-1933, Dec. 1997.
- [10] P. White and J. Crowcroft, "The Integrated Services in the Internet: State of the Art," Proc. IEEE, vol. 85, no. 12, pp. 1934-1946, Dec. 1997.
- [11] K. Van Der Wal, M. Mandjes and H. Bastiaansen, "Delay Performance Analysis of the New Internet Services with Guaranteed QoS," Proc. IEEE, vol. 85, no. 12, pp. 1947-1957, Dec. 1997.
- [12] S. Makino, K. Strauss, S. Shimauchi, Y. Haneda and A. Nakagawa, "Subband Stereo Echo Canceller using the Projection Algorithm with Fast Convergence to the True Echo Path," Proc. ICASSP97, pp. 299-302, April 1997.
- [13] J. Benesty, D. Morgan and M. Sondhi, "A Better Understanding and an Improved Solution to the Problems of Stereophonic Acoustic Echo Cancellation," Proc. ICASSP97, pp. 303-306, April 1997.
- [14] N. Birkett, Nonlinear Adaptive Filtering with Application to Acoustic Echo Cancellation, Ph.D Thesis, Carleton University, Ottawa, Canada, April 1997.

- [15] ITU-T Series H., "H.245 – Line Transmission of Non-Telephone Signals – Control Protocol for Multimedia Communication," March 1997.
- [16] S. Ahuja and K. Muti, "Packet Telephony," Bell Labs Technical Journal, vol. 2, no. 2, Spring 1997.
- [17] W. Gardner, "Head Tracked 3-D Audio Using Loudspeakers," Proc. IEEE Workshop on App. Signal Proc. To Audio and Acoustics, New Paltz, NY, 1997.
- [18] J. Blauert, Spatial Hearing: The Psychophysics of Human Sound Localization, revised ed., MIT Press, Cambridge, MA, 1997.
- [19] G. Thom, "H.323: The Multimedia Communications Standard for Local Area Networks," IEEE Communications Magazine, pp. 52-56, Dec. 1996.
- [20] ITU-T Series H., "H.225.0, Version 2 – Line Transmission of Non-Telephone Signals – Media Stream Packetization and Synchronization on Non-Guaranteed Quality of Service LANs," Dec. 1996.
- [21] J. Bauck and D. Cooper, "Generalized Transaural Stereo and Applications," J. Audio Eng. Soc., vol. 34, no. 9, pp. 716-744, Sept. 1996.
- [22] ITU-T Series G., "G.131 – Transmission Systems and Media – International Telephone Connections and Circuits – General Characteristics of the 4-wire Chain Formed by the International Circuits and National Extension Circuits – Control of Talker Echo," Aug. 1996.
- [23] ETSI Technical Report, "ETR 250 – Transmission and Multiplexing (TM); Speech Communication Quality from Mouth to Ear for 3,1kHz Handset Telephony Across Networks," July 1996.
- [24] S. Makino, J. Noebauer, Y. Haneda and A. Nakagawa, "SSB Subband Echo Canceller using Low-Order Projection Algorithm," Proc. ICASSP96, pp. 945-948, May 1996.
- [25] F. Amand, J. Benesty, A. Gilloire and Y. Grenier, "A Fast Two-Channel Projection Algorithm for Stereophonic Echo Cancellation," Proc. ICASSP96, pp. 949-952, May 1996.
- [26] S. Shimauchi and S. Makino, "Stereo Echo Cancellation Algorithm using Imaginary Input-Output Relationships," Proc. ICASSP96, pp. 941-944, May 1996.
- [27] ITU-T Series H., "H.323 – Line Transmission of Non-Telephone Signals – Visual Telephone Systems and Equipment for Local Area Networks which Provide a Non-Guaranteed Quality of Service," May 1996.
- [28] J. Benesty, P. Duhamel and Y. Grenier, "A Multichannel Affine Projection Algorithm with Applications to Multichannel Acoustic Echo Cancellation," IEEE Signal Processing Letters, vol. 3, no. 2, February 1996.
- [29] H. Schulzrinne, S. Casner, R. Frederick and V. Jacobson, IETF Audio-Video Transport Working Group RFC1889, "RTP: A Transport Protocol for Real-Time Applications," Jan. 1996.

- [30] H. Schulzrinne, IETF Audio-Video Transport Working Group RFC1890, "RTP Profile for Audio and Video Conferences with Minimal Control," Jan. 1996.
- [31] A. Gilloire, "Recent Advances in Adaptive Filtering Algorithms for Acoustic Echo Cancellation," Proc. International Workshop on Acoustic Echo and Noise Control, pp. 115-134, June 1995.
- [32] S. Makino, "Relationship between the "ES family" Algorithms and Conventional Adaptive Algorithms," Proc. International Workshop on Acoustic Echo and Noise Control, pp. 11-14, June 1995.
- [33] F. Amand, J. Benesty, A. Gilloire, Y. Grenier, "Multi-channel acoustic echo cancellation," Proc. International Workshop on Acoustic Echo and Noise Control, pp. 57-60, June 1995.
- [34] M. Sondhi, D. Morgan and J. Hall, "Stereophonic Acoustic Echo Cancellation – An Overview of the Fundamental Problem," IEEE Signal Processing Letters, vol. 2, no. 8, August 1995.
- [35] J. Benesty, F. Amand, A. Gilloire and Y. Grenier, "Adaptive Filtering Algorithms for Stereophonic Acoustic Echo Cancellation," Proc. ICASSP95, pp. 3099-3102, May 1995.
- [36] S. Minami, "A Stereophonic Echo Canceller Using Single Adaptive Filter," Proc. ICASSP95, pp. 3027-3030, May 1995.
- [37] S. Gay and S. Tavathia, "The Fast Affine Projection Algorithm," Proc. ICASSP95, pp. 3023-3026, May 1995.
- [38] S. Shimauchi and S. Makino, "Stereo Projection Echo Canceller with True Echo Path Estimation," Proc. ICASSP95, pp. 3059-3062, May 1995.
- [39] M. Montazeri and P. Duhamel, "A Set of Algorithms Linking NLMS and Block RLS Algorithms," IEEE Trans. on Signal Processing, pp. 444-453, Feb 1995.
- [40] W. Gardner and D. Martin, "HRTF Measurements of a KEMAR," J. Acoust. Soc. Am., vol. 97, no. 6, pp. 3907-3908, 1995.
- [41] M. Knappe and R. Goubran, "Steady State Performance Limitations of Full-Band Acoustic Echo Cancellers," Proc. ICASSP94, pp. 2073-2077, April 1994.
- [42] C. Mac Cabe and D. Furlong, "Virtual Imaging Capabilities of Surround Sound Systems," J. Audio Eng. Soc., vol. 42, no. 1/2, Jan/Feb 1994.
- [43] K. Jeffray, D. Stone and F. Smith, "Transport and display mechanisms for multimedia conferencing across packet-switched networks," Computer Networks and ISDN Systems, vol. 26, no. 10, pp. 1281-1304, 1994.
- [44] A. Hirano and A. Sugiyama, "A Compact Multi-Channel Echo Canceller with a Single Adaptive Filter per Channel," Proc. ISCAS92, pp. 1922-1925, May 1992.
- [45] S. Haykin, Adaptive Filter Theory, 2nd ed., Prentice-Hall, Englewood Cliffs, N.J., 1991.
- [46] D. Cooper and J. Bauck, "Prospects for Transaural Recording," J. Audio Eng. Soc., vol. 37, no. 1/2, Jan/Feb 1989.

- [47] J. Treichler, C. Johnson and M. Larimore, Theory and Design of Adaptive Filters, John Wiley and Son, Wiley, New York, NY, 1987.
- [48] S. Lipshitz, "Stereo Microphone Techniques... Are the Purists Wrong?," J. Audio Eng. Soc., vol. 34, no. 9, pp. 716-744, Sept. 1986.
- [49] B. Widrow and S. Stearns, Adaptive Signal Processing, Prentice-Hall, Englewood Cliffs, NJ, 1985.
- [50] G. Barberis, "Buffer sizing of a packet voice receiver," IEEE Transactions on Communications, pp. 152-156, vol. 29, no. 2, Feb. 1981.
- [51] P. T. Brady, "Effects of Transmission Delay on Conversational Behavior on Echo-Free Telephone Circuits," Bell Labs Technical Journal, pp. 115-134, v. 50, 1971.

Appendix A - Impulse Response Computation

Impulse responses for the left and right channels are required for simulation of the synthetic stereo echo cancellation structure. The left channel impulse response is the transfer function between the left loudspeaker and the microphone. The right channel impulse response is the transfer function between the right loudspeaker and the microphone. The structure used to compute the left channel impulse response is shown in Figure 32.

The left channel impulse response is determined in the following manner:

1. White Gaussian noise is played through the left loudspeaker with the right channel inactive. The microphone records the signal from the left loudspeaker.
2. An offline version of the NLMS algorithm is run with the left channel white Gaussian noise source as the reference and the microphone recording as the primary input signal. The algorithm is run until it converges.
3. The left channel impulse response equal to the adaptive filter's coefficients.

Room impulse responses were measured by software written for the PC. The software uses a sound card to measure the room impulse responses. This method is desirable because the actual hardware that is used in the VoIP system is used to measure the room impulse responses.

Figure 32 - Left Channel Impulse Response Computation

The sound card does not perfectly align recording and playback events when used in full-duplex. The method for determining the lag between the start of the recording and playback events, called the full-duplex skew, follows:

1. Play and record a test waveform simultaneously on the sound card.
2. Compute the cross-correlation for the playback and recorded signals.
3. The peak in the cross-correlation results is the full-duplex sound card skew.

The full-duplex skew is the sum of the delays from the sound card playback codec, any line delays and the delays from the record codec. The soundcard must be connected by a cable directly from its line output to mic input. The delays through

the speaker and microphone are eliminated by this coupling method. The delays in the sound card full-duplex skew are shown in Figure 33. The sum of the full-duplex skew delays is $n_1+n_2+n_3$ samples.

Figure 33 - Measurement of Sound Card Skew

The full-duplex skew must be eliminated from the recordings taken to compute the room impulse responses before running the NLMS algorithm.

Appendix B - Adaptive Algorithms

B.1 Fast Affine Projection Algorithm

1. Initialization: $E_{a,n} = E_{b,n} = \sigma$, $\underline{a}_o = [1, \underline{0}']'$, $\underline{b}_o = [\underline{0}', 1]'$
2. Use Sliding Windowed FRLS to update $E_{a,n}$, $E_{b,n}$, $\underline{a}_n$, $\underline{b}_n$
3. $\tilde{r}_{xx,n} = \tilde{r}_{xx,n-1} + x_n \underline{\tilde{a}}_n - x_{n-P} \underline{\tilde{a}}_{n-P}$
4. $\hat{e}_n = s_n - \underline{x}_n^t \hat{\underline{h}}_{n-1}$
5. $e_n = \hat{e}_n - \mu \tilde{r}_{xx,n} \overline{E}_{n-1}$
6. $\underline{e}_n = \begin{bmatrix} e_n \\ (1-\mu)\underline{e}_{n-1} \end{bmatrix}$
7. $\underline{\varepsilon}_n = \begin{bmatrix} 0 \\ \tilde{\underline{\varepsilon}} \end{bmatrix} + \frac{1}{E_{a,n}} \underline{a}_n \underline{a}_n^t \underline{e}_n$
8. $\begin{bmatrix} \underline{\varepsilon} \\ 0 \end{bmatrix} = \underline{\varepsilon}_n - \frac{1}{E_{b,n}} \underline{b}_n \underline{b}_n^t \underline{e}_n$
9. $\underline{E}_n = \begin{bmatrix} 0 \\ \overline{E}_{n-1} \end{bmatrix} + \underline{\varepsilon}_n$
10. $\hat{\underline{h}}_n = \hat{\underline{h}}_{n-1} + \mu \underline{x}_{n-(N-1)} E_{N-1,n}$
11. $\tilde{\underline{\varepsilon}}_{n+1} = (1-\mu)\tilde{\underline{\varepsilon}}_n$

where

$\underline{a}_n$, $\underline{b}_n$ are optimum forward and backward linear predictors for $\underline{R}_n$

$E_{a,n}$, $E_{b,n}$ are forward and backward prediction energies

$\underline{e}_n$ is the a priori error vector

$\underline{x}_n$ is the reference signal

s_n is the primary signal

$\hat{\underline{h}}_n$ is the alternate coefficient vector

$$\bar{\underline{\alpha}}_n = [x_{n-1}, \dots, x_{n-N+1}]$$

$\underline{\varepsilon}$ is the normalized residual echo vector

$$\underline{E}_n = \begin{bmatrix} \varepsilon_{0,n} \\ \varepsilon_{1,n} + \varepsilon_{0,n-1} \\ \vdots \\ \varepsilon_{N-1,n} + \varepsilon_{N-2,n-1} + \dots + \varepsilon_{0,n-(N-1)} \end{bmatrix}$$

$\bar{\underline{E}}$ is the N-1 length vector consisting of the upper most N-1 elements of $\underline{E}_n$

B.2 Recursive Least Square Algorithm

1. Accept new samples $x(k)$ and $d(k)$ to form new reference vector $\underline{X}(k)$.
2. Compute a priori output $y(k) = \bar{h}_k^T \underline{X}(k)$
3. Compute a priori error $e(k) = d(k) - y(k)$
4. Compute filtered information vector $\underline{Z}_k = \underline{R}_k^{-1} \underline{X}(k)$
5. Compute normalized error power $q = \underline{X}^T(k) \underline{Z}_k$
6. Compute gain constant $\nu = \frac{1}{1+q}$
7. Compute normalized filtered information vector $\bar{\underline{Z}}_k = \nu \cdot \underline{Z}_k$
8. Update optimal weight vector $\bar{h}_{k+1} = \bar{h}_k + e(k) \bar{\underline{Z}}_k$
9. Update inverse correlation matrix $\underline{R}_{k+1}^{-1} = \underline{R}_k^{-1} - \bar{\underline{Z}}_k \bar{\underline{Z}}_k^T$

Initialize inverse correlation matrix on algorithm start to $\underline{R}_{N-1}^{-1} = \eta I_N$ where η is a large positive constant and I_N is the identity matrix.

Appendix C – Software Design

C.1 Use Case Map Symbols

The VoIP software design is illustrated through use case maps. Use case maps show the interaction between concurrent system segments that execute in real-time. The symbols and terminology, used in use case maps, are shown in Figure 34.

Figure 34 - Use Case Map Symbols

C.3 Capture Data and Acoustic Echo Cancellation Use Case Map

This use case map, Figure 36, shows data capture from the acoustic interface, echo cancellation for the data, capture data buffering and data transfer to the PC when the capture buffer is full.

Figure 36 – Capture Data and Acoustic Echo Cancellation Use Case Map

Activity	Description	Activity Performed by
Capture Data	Data is transferred from the DSP card codec. This action will occur when the DSP card interrupts program flow to indicate the codec requires capture buffer servicing.	DSP Card Software
Detect Near End Speech and Multiple Talkers	Near end speech is detected to determine whether the acoustic echo canceller should be frozen. The status of the participants is checked to determine if multiple participants are active for echo cancellation.	DSP Card Software
Acoustic Echo Cancellation	The acoustic echo canceller is run using the reference store, the filter taps store and the current codec capture value. The error from the echo canceller is stored in the error store.	DSP Card Software
Add Data to DSP Card Capture Buffer	Transfer the acoustic echo canceller error value to the DSP card capture buffer. This data is now ready for retrieval by the PC program to send to the remote system(s)	DSP Card Software
Copy Completed Frame to Shared Capture Buffer (HPI)	The capture buffer threshold has been reached and the completed frame is sent to the PC for transmission to the remote system(s). The PC is alerted that there is new data via the HINT interrupt.	DSP Card Software

Table 4 – Capture Data and Acoustic Echo Cancellation Activity Descriptions

C.4 Playback Data Use Case Map

The sequence of events for data playback is shown in Figure 37. If the participant is active, data is retrieved from the jitter buffer. The data is convolved by the appropriate spatialization function, mixed with other talker data and is then played out on the codec.

Figure 37 – Playback Data Use Case Map

Activity	Description	Activity Performed by
Check for Active Participants <code>is_active[x] == TRUE; x={0..n-1}</code>	Check the participants to determine which ones are marked as active. Only the participants that are marked as active will have their audio frames played. This action will occur when an interrupt occurs at the codec to request playback servicing.	DSP Card Software
Convolve Active Talkers with Spatialization Functions	The audio frames from the active talkers are convolved with the appropriate spatialization functions. The audio frames for the active talkers must first be retrieved from the jitter buffer before the spatialization.	DSP Card Software
Mix Talkers	The audio frames from the active talkers are mixed together to form the sample that will be played out.	DSP Card Software
Play Data	Then synthetic stereo output data is transferred to DSP card codec and playback is initiated. The data is stored as the current reference value for use with the acoustic echo canceller.	DSP Card Software

Table 5 – Playback Data Activity Descriptions

C.5 Jitter Buffer Use Case Map

The jitter buffer use case map, Figure 38, shows data copy from the host port interface shared memory to the DSP jitter buffer when the PC generates a jitter buffer interrupt.

Figure 38 - Jitter Buffer Use Case Map

Activity	Description	Activity Performed by
Copy Data from Shared Rx Buffer to Jitter Buffer	This action occurs when the PC interrupts the DSP to indicate that incoming data for a participant is resident in the host port interface shared rx buffer. The DSP must copy this memory to the appropriate participant's jitter buffer.	DSP Card Software

Table 6 - Jitter Buffer Activity Descriptions

C.6 Send Data Use Case Map

The send data use case map, Figure 39, shows transmission of a captured audio frame through encapsulation by RTP and sending over UDP/IP.

Figure 39 - Send Data Use Case Maps

Activity	Description	Activity Performed by
Receive Data from DSP Card via PIO	Data is transferred from the DSP shared card capture buffer to the PC using the programmed I/O transfer mode. (PIO)	PC Software
Add RTP Header	Data received from the DSP card is encapsulated with a valid RTP header to form an RTP frame.	PC Software
Send UDP Datagram	The RTP frame is sent to the remote system(s) via the UDP/IP using IP multicasting.	PC Software

Table 7 - Send Data Activity Descriptions

C.7 Receive Data Use Case Map

The receive data use case map, Figure 40, shows incoming RTP packet decapsulation, RTP packet validation, sending valid audio frames to the DSP and the generation of RTCP statistics.

Figure 40 - Receive Data Use Case Map

Activity	Description	Activity Performed by
Receive UDP Datagram	An RTP frame is received from the network via the UDP protocol with IP multicast.	PC Software
Validate Frame and Collect Stats	The payload type is verified and the sequence number is checked to determine whether the packet is current or expired. The jitter is calculated, the total packet count is accumulated, and other statistics are computed.	PC Software
Remove RTP Header	The RTP frame header is verified and the data is removed from the RTP structure.	PC Software
Delete Frame	The RTP data that has been receive is non-current or corrupt and is deleted.	PC Software
Add to Jitter Buffer, Delete RTP Frame	The RTP data that has been removed from the frame is copied to the shared rx buffer in the host port interface memory. The DSP software will be responsible for adding it to the correct jitter buffer for the participant. The RTP frame is then deleted.	PC Software
Send RTCP Frame	Send an RTCP Sender Report (SR) frame to the designated participant when the percentage bandwidth is exceeded.	PC Software

Table 8 - Receive Data Activity Descriptions

C.8 Jitter Buffer Structure for Multiple Participants

The jitter buffer structure for multiple participants, Figure 41, has space to store incoming audio frames for n participants. The VoIP platform uses a delayed playback jitter buffer scheme to smooth network jitter by buffering several frames before playback begins. The number of frames buffered before playback is indicated by Δ in Figure 41. The participant's activity status and IP address must be logged by the jitter buffer for data playback and synchronization with the appropriate spatialization function.

Figure 41 - Jitter Buffer Structure

C.9 Host Port Interface and Transfer Registers

The audio frames and control information are transferred from the PC to the DSP and vice versa through the host port interface (HPI) shared memory. The high-level memory allocations for the host port interface shared memory are shown in Figure 42. The transfer registers and their addresses in PC I/O space and DSP I/O space are also shown.

Figure 42 - Host Port Interface and Transfer Registers

C.10 PC Conference Status Word

The PC conference status word, Figure 43, is used to send initial setup or update parameters to the DSP.

Figure 43 - PC Conference Status Word

C.11 DSP Conference Status Word

The DSP conference status word, Figure 44, is used by the DSP to acknowledge changes in conference parameters.

Figure 44 - DSP Conference Status Word

C.12 Control Information and Spatialization Functions

The control information and spatialization functions memory allocations in the host port interface shared memory are shown in Figure 45. The control information consists of conference parameters. The space allocated for spatialization functions is used to update the spatialization functions when a new participant is added.

Host Port Interface Memory		<u>Constant Name</u>
0x1000	Number of Participants, N_p	NUMPAR
0x1001	Spatialization Length, T_p	SPATLEN
0x1002	Frame Size, T_f	FRAMELEN
0x1003	Jitter Buffer Threshold, T_j	JITTERTH
0x1004	ERLE Target	ERLETAR
⋮		
0x100A	Spatial Participant Id	SPARTIC
0x100B	Left Spatialization	LSPATIAL
0x100B + T_p-1	Function [-3ms, 3ms]	
0x1024	Right Spatialization	RSPATIAL
0x1024 + T_p-1	Function [-3ms, 3ms]	

Figure 45 - Control Information and Spatialization Functions

C.13 Capture and Jitter Buffer

The capture and jitter buffer host port interface shared memory allocations are shown in Figure 46.

Figure 46 - Capture and Jitter Buffer

C.14 Round Trip Delay Information

The information required for acoustic round trip delay computation is stored in the host port interface shared memory, Figure 47.

		<u>Constant Name</u>
0x143E	Jitter Buffer Size	JBUFSIZE
0x143F	Jitter Buffer Restarts	JRESTARTS
0x1440	PC Cur. Timestamp, Ts(n)	PCURTS
0x1442	PC Old Timestamp, Ts(n-1)	POLDTS
0x1444	PC Current Seq #, Sn(n)	PCURSN
0x1445	PC Old Seq #, Sn(n-1)	POLDSEN
0x1446	Round Trip Delay	RTD
0x1448	Sequence # Delta	DSN
0x1449	DSP Old Seq. #, Sn(n-1)	DOLDSEN
0x144A	DSP Cur. Timestamp, Ts(n)	DCURTS
0x144C	DSP Old Timestamp, Ts(n-1)	DOLDTS
0x144E	Unused[946]	BLOCK
0x17FF		

Figure 47 - Round Trip Delay Information

C.15 Control Use Case Map

A use case map for conference control is shown in Figure 48. This use case map shows the PC issue a conference startup request to the DSP. The DSP must acknowledge that the startup request and any control information have been received.

Control Use Case Map

(Startup Example)

Figure 48 - Control Use Case Map

Appendix D - C54x Code for NLMS Algorithm

```
;-----
; NLMS ALGORITHM (C54x)
; By: Trevor Yensen
; May 22, 1998
;-----

        .mmregs

;-----
; CONSTANTS
;-----

LMS      .set 1          ; LMS mode define
NLMS     .set 2          ; NLMS mode define
mode     .set NLMS      ; set mode to NLMS

N        .set 127        ; the LMS adaptive filter will have N taps

ALPHASFT .set -6         ; ALPHA shift used as IIR pole in power calculation
; ALPHA = 1/16, or ALPHA = 2^ALPHASFT
EPSILONVAL.set 64        ; EPSILONVAL used as power control constant (Q15)
; value to cancel pole and have a resulting  $\sigma^2(n) = 1 \Rightarrow \text{ALPHA} * \text{EPSILON}$ 
; (for case were EPSILON overrides refvalues)
;  $\sigma^2(n) = (1 - 2^{\text{ALPHASFT}}) \sigma^2(n-1) + 2^{\text{ALPHASFT}} * \text{new\_val}$ 
MU_SFT   .set -9         ; MU shift used as gain parameter in calculation
; of step-size for use in NLMS
; MU = 2 * convergence factor / L
; MU < 1/L

;-----
; ASSIGNMENTS
;-----

.asg     AR3, DES_PTR    ; Desired Value Pointer (D)
.asg     AR4, COEF_PTR   ; Coefficient Vector Pointer (W)
.asg     AR5, REF_PTR    ; Reference Vector Pointer (X)

;-----
; VARIABLES
;-----

REFERENCE .usect "REFERENC", N ; allocate space for ref buffer
COEFFICIENTS .usect "LMSCOEF", N ; allocate space for adaptive filter

; LMS Algorithm Variables
        .bss     ERROR, 1     ; Error (Q.15)
```

```

.bss    GRADIENT, 1 ; LMS Gradient = MU * ERROR(n-1) LMS
                                ; NLMS Gradient = MU * IPOWER * ERROR(n-1)
                                ; (Q.15)
.bss    DESIRED, 1 ; desired value (Q.15)
.bss    Y, 1       ; Convolution Output (Q.15)

; C Interfacing Variables
.bss    REFVAR, 1  ; reference (Q.15)
.bss    PRIVAR, 1  ; primary (Q.15)
.bss    OLDREF, 1  ; old reference vector pointer (Q.15)

; Power Variables
.bss    LTPOWER, 2,1,1 ; long-time reference power estimate (double word)
                                ; (Q.31)
.bss    POWER0, 1     ; value of refvalue^2 (Q.15)
.bss    ALTPOWER, 1   ; adj. long-time ref power est. (includes EPSILON)
                                ; (Q.15)
.bss    IPOWER, 1     ; 1/(LTPOWER+ALPHA*EPSILON) (Q.0)
.bss    EPSILON, 1    ; power control, EPSILON (Q.15)

;-----
; PROGRAM SECTION
;-----
.text

;-----
; void initNLMS()
;-----
_initNLMS:
.global _initNLMS

; Setup OLDREF (Holds current value of REF_PTR between LMS iterations)
ST #REFERENCE, *(OLDREF)

; Zero Coefficients
STM #COEFFICIENTS, COEF_PTR
RPTZ A, #(N-1)          ; Zero Accumulator and repeat next N times
STL A, *COEF_PTR+

; Zero Reference
STM #REFERENCE, REF_PTR
RPTZ A, #(N-1)          ; Zero Accumulator and repeat next N times
STL A, *REF_PTR+

; Setup Epsilon

```

```
ST #EPSILONVAL, *(EPSILON)
```

```
; Setup LTPOWER, ERROR
LD #0, A
DST A, *(LTPOWER)
STL A, *(ERROR)
RET
```

```
;
```

```
;
```

```
; int runNLMS(int ref, int pri)
```

```
;
```

```
PC      .set    0      ; location of program counter on stack
DES     .set    1      ; new desired value for iteration
```

```
_runNLMS:
```

```
.global _runNLMS
```

```
; *****
```

```
; Initializations
```

```
; *****
```

```
LD *SP(DES), B      ; Load Desired Value (in SP mode)
RSBX CPL            ; Change to DP mode
LD #ERROR, DP       ; Setup Data Page
SSBX SXM            ; Sign Extension Enabled
SSBX FRCT           ; FRCT Bit Enabled (Q.15 * Q.15 = Q.31)
RSBX OVM            ; Overflow mode enabled (no wrap around)
LD #0, ASM          ; ASM = 0 (accumulator shift)
STL A, REFVAR        ; Store function new reference value argument
STL B, PRIVAR        ; Store function desired argument
STM #N, BK          ; Circular Buffer Size
STM #1, AR0          ; Index Increment Value (AR0 inc reg)
MVDM OLDREF, REF_PTR ; Retrieve value of REF_PTR
STM #DESIRED, DES_PTR ; Primary Variable Pointer
```

```
; *****
```

```
; STEP I: UPDATE REFERENCE VECTOR
```

```
; *****
```

```
LD REFVAR, 16, A
STH A, *REF_PTR+0%      ; new x(n)
```

```
; *****
```

```
; STEP II: COMPUTE LONG-TERM POWER AND INVERSE
```

```
; *****
```

```
; Compute new reference value squared (AH must be loaded with REFVAR)
```

```

; at the end of step I)
SQR A, A                ; Q.15 * Q.15 = Q.31 (FRCT on)
STH A, POWER0           ; POWER0 = new_ref^2 (Q.15 again)

; Update long-time reference power estimate (LTPower)
STM #(16+ALPHASFT), T    ; for dynamic shift load ALPHA = 2^ALPHASFT
; Shift arguments into high bit by 16+ALPHASFT which results in a 2^ALPHASFT
; division in the high part of the accumulator
DLD LTPower, A           ; double precision load old LTPower Q.31
; LTPower(n) = LTPower(n-1) - ALPHA*LTPower(n-1) + ALPHA*x(n)^2
SUB LTPower, TS, A       ; A = A - LTPower*2^ALPHASFT
ADD POWER0, TS, A        ; A = A + POWER0*2^ALPHASFT
DST A, LTPower           ; double precision store new LTPower
ADD EPSILON, TS, A       ; A = A + EPSILON*2^ALPHASFT
STH A, ALTPOWER          ; single precision store adjusted LTPower

; Invert long-time reference power 1/(LTPower+ALPHA*EPSILON)
; This algorithm is intended for unsigned division, so LTPower cannot be
; negative, which it never is anyway.
LD #32767, A             ; numerator = 1 in Q.15
RPT #(16-1)
SUBC ALTPOWER, A
STL A, IPOWER            ; IPOWER = 1/(LTPower+ALPHA*EPSILON)
; NOTE: IPOWER is now in Q.0 notation since IPOWER = 1 / fraction

; *****
; STEP IV: UPDATE PRIMARY VALUE (DESIRED VALUE)
; *****
LD PRIVAR, A
STL A, *DES_PTR          ; new d(n)

; *****
; STEP III: PERFORM CONVOLUTION AND COMPUTE ERROR
; *****
STM #COEFFICIENTS, COEF_PTR ; Setup Coefficient Vector COEF_PTR
LD #0, A
RPT #(N-1)              ; Repeat for N taps
MAC *COEF_PTR+, *REF_PTR+0%, A ; multiple, accumulate and round
STH A, Y                 ; Output of Convolution (not required)
SUB *DES_PTR, 16, A      ; A = DESIRED<<16 - A, new error
NEG A
STH A, ERROR             ; Error

; *****
; STEP IV: UPDATE GRADIENT

```

```

; *****
LD IPOWER, T          ; Load IPOWER(n) into T (Q.0)
MPY ERROR, A          ; A = ERROR(n)*IPOWER (Q.0*Q.15=Q.16) FRCT on
SFTA A, #(15+MU_SFT)  ; adjust decimal place to give Q.31 and
                      ; multiply by MU, where MU = 2^MU_SFT
STH A, GRADIENT       ; GRADIENT = Error(n) * IPOWER(n) * MU

; *****
; STEP V: TAP UPDATE
; *****
STM #COEFFICIENTS, COEF_PTR ; Setup Coefficient Vector COEF_PTR
STM #(N-1), BRC             ; Repeat for N taps
LD GRADIENT, T              ; T = GRADIENT(n)
RPTB UPD_END-1

MPY *REF_PTR+0%, A          ; A = GRADIENT(n) * x(n,0) (Q.15)
ADD *COEF_PTR, 16, A        ; A = A + w(n-1) = w(n-1) + GRADIENT(n) * x(n)
STH A, *COEF_PTR+
UPD_END:

; *****
; Cleanup
; *****
MVMD REF_PTR, OLDREF
LD ERROR, A                ; Return LMS error in A reg
NOP
NOP

SSBX CPL                   ; change to SP mode
RSBX FRCT                  ; turn off FRCT mode

RET
;-----

```

Appendix E - C54x Code for Spatialization Algorithm

```
; Filter Coefficients Section
; By: Trevor Yensen
; Last Update: July 22nd, 1998

        .mmregs

;-----
;  REGISTERS
.asg     AR3, REF_PTR      ; Reference Vector Pointer
.asg     AR4, COEF_PTR     ; Coefficient Pointer
.asg     AR5, OLD_PTR      ; Pointer to old ref pointer
;-----
;  CONSTANTS
MAXPARTIC .set    3        ; Maximum number of participants allowed
                                ; in the conference
MAXTAPS    .set    25      ; Maximum taps possible
                                ; NOTE: For circular buffering used
                                ; with MAC instruction, MAXTAPS must
                                ; be one less than a power of 2

;-----
        .sect "spatial"
;  FILTER TAPS (MAXIMUM OF 25 TAPS)
_m_spatLen .global _m_spatLen
        .word 0              ; Spatialization Function Length (m_spatLen is
                                ; SPATLEN - 1, to accomodate RPT instruction
                                ; which repeats blocks of size + 1)

;  NEW VALUE
newval    .word 0
;  PARTICIPANT BEING PROCESSED
ppartic   .word 0
;  RESULT
result    .word 0

;  OLD REFERENCE VECTOR POINTERS
old_lref  .space (16*MAXPARTIC)
                                ; Reserve space for MAXPARTIC WORDS
old_rref  .space (16*MAXPARTIC)
                                ; Reserve space for MAXPARTIC WORDS

;  FILTER COEFFICIENTS (MAXIMUM OF 25 COEFFICIENTS)
;  NOTE: Using Symmetric FIR Filter so coefficient order doesn't matter
_m_lSpat: .global _m_lSpat
        .space (16*MAXTAPS*MAXPARTIC)
```

```

; Reserve MAXTAPS*MAXPARTIC WORDS (16-bits)
; NOTE: Last filter coefficient is first in memory

_m_rSpat: .global _m_rSpat
         .space (16*MAXTAPS*MAXPARTIC)
; Reserve MAXTAPS*MAXPARTIC WORDS (16-bits)
; NOTE: Last filter coefficient is first in memory

;-----
;  LEFT CHANNEL TAPS
l_taps   .usect "LEFTREF",32*MAXPARTIC
; Reserve 32*MAXPARTIC WORDS (16-bits)
; For 25 taps we must be on a 25<2^N boundary

;  RIGHT CHANNEL TAPS
r_taps   .usect "RIGHTREF",32*MAXPARTIC
; Reserve 32*MAXPARTIC WORDS (16-bits)
; For 25 taps we must be on a 25<2^N boundary

.text
;-----
;  int _spatInit();
;  SETUP LAST POSITION IN REFERENCE VECTOR
_spatInit:
    .global _spatInit

; INITIALIZE OLD REFERENCE PTRs FOR LEFT CHANNEL(s) SO THEY ARE
; ON 32-bit BOUNDARIES
STM #old_lref, AR0      ; AR0 = old_lref
LD #l_taps, A           ; A = l_taps
STM #(MAXPARTIC-1), BRC ; Repeat for MAXPARTIC iterations
RPTB spatBlok1-1
STL A, *AR0+            ; old_lref(count) = l_taps+count*25
ADD #32, A
spatBlok1:

; Initialize old reference ptrs for right channel(s) so they are
; on 32-bit boundaries
STM #old_rref, AR0      ; AR0 = old_rref
LD #r_taps, A           ; A = r_taps
STM #(MAXPARTIC-1), BRC ; Repeat for MAXPARTIC iterations
RPTB spatBlok1-1
STL A, *AR0+            ; old_lref(count) = l_taps+count*25
ADD #32, A
spatBlok1:

```

```
; CLEAR LEFT TAPS TO ALL ZERO
STM #l_taps, AR0
RPTZ A, #(32*MAXPARTIC-1)
STL A, *AR0+
```

```
; CLEAR RIGHT TAPS TO ALL ZERO
STM #r_taps, AR0
RPTZ A, #(32*MAXPARTIC-1)
STL A, *AR0+
```

```
RET
```

```
;
; int _spatLeft(int new_tap, int participant);
; LEFT CHANNEL SPATIALIZATION FUNCTION.
; NOTE: CPL bit is set by C compiler so indirect absolute addressing
; must be used, or reset CPL to 0 and then set to 1 before exit.
; NOTE: First argument is in accumulator A, return value in accumulator A.
PC .set 0 ; location of program counter on stack
lpartic .set 1 ; participant #
```

```
_spatLeft:
```

```
.global _spatLeft
```

```
LD *SP(lpartic), B ; Load participant # into B (SP mode)
RSBX CPL ; Change to DP mode
LD #_m_lSpat, DP ; Setup Data Page
SSBX SXM ; Sign Extension Enabled
SSBX FRCT ; FRCT Bit Enabled (Q.15 * Q.15 = Q.31)
RSBX OVM ; Overflow Mode Enabled
STL A, newval ; Store new value
STL B, ppartic ; Store Participant being processed
LD _m_spatLen, A ; A = m_spatLen
ADD #1, A ; A = m_spatLen + 1
STLM A, BK ; Block Size Register
STM #1, AR0 ; index increment register
```

```
; *****
```

```
; STEP I: SET COEF_PTR TO POINT TO COEFFICIENTS
```

```
; *****
```

```
LD #(_m_lSpat-MAXTAPS), A ; A = _m_lSpat - 25
RPT ppartic
ADD #MAXTAPS, A ; A = A + 25*(ppartic+1) = _m_lSpat+25*ppartic
STLM A, COEF_PTR ; COEF_PTR = A (2 cycle latency)
```



```

; *****
; STEP II: RESTORE REF_PTR INDEX INTO REFERENCE VECTOR
; *****
LD #(old_lref), A          ; A=&old_lref
ADD ppartic, A             ; A=A+ppartic=&old_ref+ppartic
NOP
STLM A, OLD_PTR           ; OLD_PTR = A (2 cycle latency)
NOP
NOP
LD *OLD_PTR, A             ; A = *AR0
STLM A, REF_PTR           ; (2 cycle latency)
NOP

; *****
; STEP III: UPDATE REFERENCE VECTOR WITH NEW VALUE
; *****
LD newval, A               ; A = new_tap
STL A, *REF_PTR+0%

; *****
; STEP IV: PERFORM FILTER CONVOLUTION
; *****
LD #0, A                  ; A = 0
RPT _m_spatLen            ; Repeat for m_spatLen (which is SPATLEN-1)
MAC *COEF_PTR+, *REF_PTR+0%, A ; multiply, accumulate
STH A, result             ; store result
LDM AH, A

; *****
; STEP V: CLEANUP
; *****
; Restore pointer to old reference vector position
LDM REF_PTR, B            ; B = REF_PTR
STL B, *OLD_PTR           ; OLD_PTR = REF_PTR

SSBX CPL                  ; Change to SP Mode (C mode)
RSBX FRCT                 ; No FRCT mode (Q15*Q15=Q30)
NOP

RET

; Right Channel Spatialization Similar to Left Channel (Not Shown)
.end                      ; EOF

```

IMAGE EVALUATION TEST TARGET (QA-3)

APPLIED IMAGE, Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989

© 1993, Applied Image, Inc., All Rights Reserved