1997 IEEE ICASSP VOL. 1

GOOGLE EXHIBIT 1005

b89059970699a

KURT F. WENDT LIBRARY
COLLEGE OF ENGINEERING

UW-MADISON, WI 53706

1997 IEEE

International Conference on Acoustics, Speech, and Signal Processing

Volume I

KURT F. WENDT LIBRARY
COLLEGE OF ENGINEERING

UW-MADISON, WI 53706

1997 IEEE International Conference on Acoustics, Speech, and Signal Processing

Volume I of V

Plenary, Expert Summaries, Special, Audio, Underwater Acoustics, VLSI

April 21-24, 1997
Munich, Germany

Sponsored by

The Institute of Electrical and Electronics Engineers
Signal Processing Society

Co-Sponsored by

DPG, GI, ITG, and TUM

IEEE Computer Society Press
Los Alamitos, California
Washington • Brussels • Tokyo

Kurt F. Wendt Library
University of Wisconsin - Madison
215 N. Randall Avenue
Madison, WI 53703-1688

IEEE Computer Society Press
10662 Los Vaqueros Circle
P.O.Box 3014
Los Alamitos, CA 90720-1264

Copyright ©1997 by The Institute of Electrical and Electronics Engineers, Inc.
All rights reserved.

Copyright and Reprint Permissions: Abstracting is permitted with credit to the source. Libraries may photocopy beyond the limits of US copyright law, for private use of patrons, those articles in this volume that carry a code at the bottom of the first page, provided that the per-copy fee indicated in the code is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923.

Other copying, reprint, or republication requests should be addressed to: IEEE Copyrights Manager, IEEE Service Center, 445 Hoes Lane, P.O. Box 1331, Piscataway, NJ 08855-1331.

The papers in this book comprise the proceedings of the meeting mentioned on the cover and title page. They reflect the authors' opinions and, in the interests of timely dissemination, are published as presented and without change. Their inclusion in this publication does not necessarily constitute endorsement by the editors, the IEEE Computer Society Press, or the Institute of Electrical and Electronics Engineers, Inc.

IEEE Computer Society Press Order Number PR07919
ISBN 0-8186-7919-0
ISBN 0-8186-7920-4 (case)
ISBN 0-8186-7921-2 (microfiche)
IEEE Order Plan Catalog Number 97CB36052
Library of Congress Number 96-80328

Additional copies may be ordered from:

IEEE Computer Society Press
Customer Service Center
10662 Los Vaqueros Circle
P.O. Box 3014
Los Alamitos, CA 90720-1314
Tel: +1-714-821-8380
Fax: +1-714-821-4641
Email: cs.books@computer.org

IEEE Service Center
445 Hoes Lane
P.O. Box 1331
Piscataway, NJ 08855-1331
Tel: +1-908-981-1393
Fax: +1-908-981-9667
misc.custserv@computer.org

IEEE Computer Society
13, Avenue de l'Aquilon
B-1200 Brussels
BELGIUM
Tel: +32-2-770-2198
Fax: +32-2-770-8505
euro.ofc@computer.org

IEEE Computer Society
Ooshima Building
2-19-1 Minami-Aoyama
Minato-ku, Tokyo 107
JAPAN
Tel: +81-3-3408-3118
Fax: +81-3-3408-3553
tokyo.ofc@computer.org

Editorial production by Bob Werner
Printed in the United States of America by Sony Electronic Publishing Services

The Institute of Electrical and Electronics Engineers, Inc.

Wanot
TA
365
I102
V.1997
1

AWC 976

Table of Contents

Volume I
Plenary, Expert Summaries, Special, Audio, Underwater Acoustics, VLSI

Conference Committee xvi
IEEE Signal Processing Society xvii
ICASSP 1998 in Seattle xviii
Call for Papers xix
ICASSP-98 Paper Cover Sheet xx
IEEE Copyright Form xxi

Plenary Lectures and Expert Summaries

Space-Time Processing for Wireless Communications 1
A. Paulraj

Variability of Performance in Video Coding 5
D. Pearson

Expert Summaries 9
R. Chellapa, R. de Mori, G. Giannakis, H. Musmann, H. Ney, M. Smith, J. Treichler and M. Zoltowski

Signal Processing for Education

Expanding Team Experiences in DSP Education 11
D. Etter and G. Orsak

Interactive Classroom for DSP/Communication Courses 15
H. Abut and Y. Öztürk

Experiences in Teaching DSP First in the ECE Curriculum 19
J. McClellan

Analog Signal Processing: A Replacement for the Sophomore-Level Circuit Analysis Course 23
D. Munson

Re-Engineering the Electrical Engineering Curriculum 27
S. Mitra

New Methods for Design of SP-Algorithms

Structural Subband Decomposition: A New Concept in Digital Signal Processing 31
S. Mitra

A New Algorithm for the Generalized Eigenvalue Problem 35
K. Hüper and U. Helmke

A Lattice Structure for Perfect Reconstruction Linear Time Varying Filter Banks with All Pass Analysis Banks 39
S. Dasgupta, C. Schwarz and M. Fu

Algorithm Design for Structured Systems: Application to Pole Placement 43
S. Paul and J. Nossek

Actions of Noncompact Groups and Algorithm Design: A Case Study 47
K. Diepold and R. Pauli

Discretization Issues for the Design of Optimal Blind Algorithms 51
R. Kennedy, D. Borah and Z. Ding

Continuous-Time Envelope-Constrained Filter Design via Laguerre Filters and H(Inf) Optimization Methods 55
Z. Zang, A. Cantoni and K. Teo

Local Adaptive Algorithms for Information Maximization in Neural Networks, and Application to Source Separation 59
J. Dehaene and N. Twum-Danso

Quick Aggregation of Markov Chain Functionals via Stochastic Complementation 63
K. Dogançay and V. Krishnamurthy

A Rank Preserving Flow Algorithm for Quadratic Optimization Problems Subject to Quadratic Equality Constraints 67
J. Moore and D. Jiang

Speech-to-Speech Translation

Verbmobil: The Combination of Deep and Shallow Processing for Spontaneous Speech Translation 71
T. Bub, W. Wahlster and A. Waibel

Prosodic Processing and its Use in Verbmobil 75
H. Niemann, E. Nöth, A. Kießling, R. Kompe and A. Batliner

The Language Components in Verbmobil 79
H. Block

The Karlsruhe-Verbmobil Speech Recognition Engine 83
M. Finke, P. Geutner, H. Hild, T. Kemp, K. Ries and M. Westphal

An Experiment on Korean-to-English and Korean-to-Japanese Spoken Language Translation 87
J. Yang and J. Park

Multilingual Person to Person Communication at IRST 91
B. Angelini, M. Cettolo, A. Corazza, D. Falavigna and G. Lazzari

Fast Word-Graph Generation for Spontaneous Conversational Speech Translation 95
T. Shimizu, H. Singer and Y. Sagisaka

JANUS-III: Speech-to-Speech Translation in Multiple Languages 99
A. Lavie, A. Waibel, L. Levin, M. Finke, D. Gates, M. Gavalda, T. Zeppenfeld and P. Zhan

State-Transition Cost Functions and an Application to Language Translation 103
H. Alshawi and A. Buchsbaum

Hybrid Language Processing in the Spoken Language Translator 107
M. Rayner and D. Carter

Finite-State Speech-to-Speech Translation 111
E. Vidal

An Experimental Bidirectional Japanese/English Interpreting Video Phone System Using Internet 115
S. Hiraoka, M. Hoshimi, K. Matsui and J. Junqua

Advanced Neural Applications

From Neural Networks to Neural Strategies 119
C. Goerick, B. Sendhoff and W. von Seelen

Neural and Traditional Techniques in Diagnostic ECG Classification 123
R. Silipo and G. Bortolan

Unsupervised Learning for Blind Source Separation: An Information-Theoretic Approach 127
D. Obradovic and G. Deco

Applications of Neural Blind Separation to Signal and Image Processing 131
J. Karhunen, A. Hyvärinen, R. Vigário, J. Hurri and E. Oja

Communications and Neural Networks: Theory and Practice 135
M. Plumbley

Robust Vector Quantization by Competitive Learning 139
J. Buhmann and T. Hofmann

Recognizing Faces from a New Viewpoint 143
T. Vetter

Hybrid Optimization of Feedforward Neural Networks for Handwritten Character Recognition 147
W. Utschick and J. Nossek

Reading Checks with Multilayer Graph Transformer Networks 151
Y. LeCun, L. Bottou and Y. Bengio

Neural Networks for Process Control in Steel Manufacturing 155
M. Schlang, E. Broese, B. Feldkeller, O. Gramckow, M. Jansen, T. Poppe, C. Schaeffner and G. Soergel

A Neuro-Dynamic Programming Approach to Admission Control in ATM Networks: The Single Link Case 159
P. Marbach and J. Tsitsiklis

Signal Processing Technology for Multi-Media Human-Machine Interaction

Issues in Measuring the Benefits of Multimodal Interfaces 163
J. Flanagan and I. Marsic

Multimodal Interfaces for Multimedia Information Agents 167
A. Waibel, B. Suhm, M. Vo and J. Yang

Smart Rooms, Desks, and Clothes 171
A. Pentland

Human Machine Interaction by Voice and Gesture 175
N. Jayant

Audio-Visual Interaction in Multimedia Communication 179
T. Chen and R. Rao

LIP Motion Modeling and Speech Driven Estimation 183
F. Lavagetto, S. Lepsoy, C. Braccini and S. Curinga

Voice Source Localization for Automatic Camera Pointing System in Videoconferencing 187
H. Wang and P. Chu

Video Interface for Spatiotemporal Interactions Based on Multi-Dimensional Video Computing 191
A. Akutsu, Y. Tonomura and H. Hamada

Indexing and Search of Multimodal Information 195
A. Hauptmann and H. Wactlar

Acoustic Indexing for Multimedia Retrieval and Browsing 199
S. Young, J. Foote, G. Jones, K. Sparck-Jones and M. Brown

Broadcast News Transcription 203
F. Kubala, H. Jin, L. Nguyen, R. Schwartz and S. Matsoukas

Image/Speech Processing that Adopts an Artistic Approach Toward Integration of Art and Technology 207
R. Nakatsu

Microphone Array Signal Processing

Noise Cancelling for Microphone Arrays 211
J. Meyer and C. Sydow

A Microphone Array System for Speech Recognition 215
K. Kiyohara, Y. Kaneda, S. Takahashi, H. Nomura and J. Kojima

Strategies for Combining Acoustic Echo Cancellation and Adaptive Beamforming Microphone Arrays 219
W. Kellermann

A Steerable and Variable First-Order Differential Microphone Array 223
G. Elko and A. Pong

Microphone Array Based Speech Recognition with Different Talker-Array Positions 227
M. Omologo, M. Matassoni, P. Svaizer and D. Giuliani

Acoustic Source Location in a Three-Dimensional Space Using Crosspower Spectrum Phase 231
P. Svaizer, M. Matassoni and M. Omologo

Superdirective Microphone Array for a Set-Top Videoconferencing System 235
P. Chu

Simultaneous Echo Cancellation and Car Noise Suppression Employing a Microphone Array 239
M. Dahl, I. Claesson and S. Nordebo

Analytical Evaluation of a Self-Calibrating Microphone Array 243
S. Nordholm and I. Claesson

Microphone Array Response to Speaker Movements 247
Y. Grenier and S. Affes

A Digital Processing System for Source Location and Sound Capture by Large Microphone Arrays 251
H. Silverman, W. Patterson, J. Flanagan and D. Rabinkin

DSP for Mobile Communication

3-D Unitary ESPRIT for Joint 2-D Angle and Carrier Estimation 255
M. Haardt and J. Nossek

Quality Enhancement of Coded and Corrupted Speeches in GSM Mobile Systems Using Residual Redundancy 259
T. Hindelang, W. Xu and C. Erben

Pilot Assisted Coherent DS-CDMA Reverse-Link Communications with Optimal Robust Channel Estimation 263
F. Ling

A New Frequency Estimator Applied to Burst Transmission 267
C. Bergogne, M. Bousquet and P. Sehier

Unified Specification of Control and Data Flow 271
T. Groetker, R. Schoenen and H. Meyr

Reconfigurable Processing: The Solution to Low-Power Programmable DSP 275
J. Rabaey

DSP Cores for Mobile Communications: Where Are We Going? 279
G. Fettweis

DSPs in Mobile Communication in the United States 283
S. Kasturia and C. Warwick

FRIDGE: An Interactive Code Generation Environment for HW/SW CoDesign 287
M. Willems, V. Bürsgens, T. Grötker and H. Meyr

Staying Ahead of the Game in Silicon for Digital Mobile Communications 291
R. Subramanian, M. Barberis and H. Dawid

Echo Cancellation

Approximation of Optimal Step Size Control for Acoustic Echo Cancellation 295
C. Antweiler, J. Grunwald and H. Quack

Subband Stereo Echo Canceller Using the Projection Algorithm with Fast Convergence to the True Echo Path 299
S. Makino, K. Strauss, S. Shimauchi, Y. Haneda and A. Nakagawa

A Better Understanding and an Improved Solution to the Problems of Stereophonic Acoustic Echo Cancellation 303
J. Benesty, D. Morgan and M. Sondhi

Comparison of Three Post-Filtering Algorithms for Residual Acoustic Echo Reduction 307
V. Turbin, A. Gilloire and P Scalart

Audio Coding & Transducer

Audio Coding Using Sinusoidal Excitation Representation 311
W. Chang, D. Wang and L. Wang

Optimum Bit Allocation and Decomposition for High Quality Audio Coding 315
X. Wei, M. Shaw and M. Varley

The D_5 Lattice Quantization for a 64 Kbit/S Low-Delay Subband Audio Coder with a 15 Khz Bandwidth 319
K. Hay, S. Saoudi and L. Mainard

An Experimental Audio Codec Based on Warped Linear Prediction of Complex Valued Signals 323
A. Härmä, U. Laine and M. Karjalainen

High Quality Low Complexity Scalable Wavelet Audio Coding 327
W. Dobson, J. Yang, K. Smart and F. Guo

An Efficient Tonal Component Coding Algorithm for MPEG-2 Audio NBC 331
Y. Takamizawa, M. Iwadare and A. Sugiyama

Spectral Amplitude Warping (SAW) for Noise Spectrum Shaping in Audio Coding 335
R. Lefebvre and C. Laflamme

A Fast Noise-Scaling Algorithm for Uniform Quantization in Audio Coding Schemes 339
C. Serantes, A. Pena and N. Prelcic

Pyramid Vector Coding for High Quality Audio Compression 343
D. Cadel and G. Parladori

Subband Audio Coding with Synthesis Filters Minimizing a Perceptual Distortion 347
K. Gosse, F. Moreau de Saint-Martin, X. Durot, P. Duhamel and J. Rault

New Results in Low Bitrate Audio Coding Using a Combined Harmonic-Wavelet Representation 351
S. Boland and M. Deriche

Adaptive Inverse Control of Weakly Nonlinear Systems 355
W. Klippel

Microphone Array & Active Noise Control

Broadband Beamforming with Adaptive Postfiltering for Speech Acquisition in Noisy Environments 359
S. Fischer and K. Kammeyer

Near-Field Beamforming for Microphone Arrays 363
J. Ryan and R. Goubran

A Robust Adaptive Microphone Array with Improved Spatial Selectivity and its Evaluation in a Real Environment 367
O. Hoshuyama, A. Sugiyama and A. Hirano

Tracking Multiple Talkers Using Microphone-Array Measurements 371
D. Sturim, H. Silverman and M. Brandstein

A Robust Method for Speech Signal Time-Delay Estimation in Reverberant Rooms 375
M. Brandstein and H. Silverman

A Model-Based Approach to Active Noise Cancellation Using Loudspeaker Array 379
J. Gu and S. Yau

Reverberant Sound Field Analysis Using a Microphone Array 383
W. Täger and Y. Mahieux

Minimisation of the Maximum Error Signal in Active Control 387
A. Gonzalez, A. Albiol and S. Elliott

Subband Active Noise Control Algorithm Based on a Delayless Subband Adaptive Filter Architecture 391
J. Yun, D. Youn and Y. Park

Nonlinear Active Noise Control in a Linear Duct 395
P. Strauch and B. Mulgrew

Fast Exact Filtered-X LMS and LMS Algorithms for Multichannel Active Noise Control 399
S. Douglas

A Novel Frequency Domain Filtered-X LMS Algorithm for Active Noise Reduction 403
T. Kosakat, S. Elliott and C. Boucher

Hearing Aids and Computer Music

Practical Supergain Head Sized Arrays 407
D. Masmoudi, D. Dallet and J. Dom

A Multichannel Compression Strategy for a Digital Hearing Aid 411
T. Schneider and R. Brennan

Multi-Microphone Sub-Band Adaptive Signal Processing for Improvement of Hearing Aid Performance: Preliminary Results Using Normal Hearing Volunteers 415
P. Shields and D. Campbell

Environmental Noise Reduction Based on Speech/Non-Speech Identification for Hearing Aids 419
K. Itoh and M. Mizushima

Blind Separation of Multiple Speakers in a Multipath Environment 423
R. Lambert and A. Bell

A Single-Chip 1,200 Sinusoid Real-Time Generator for Additive Synthesis of Musical Signals 427
F. De Bernardinis, R. Roncella, R. Saletti, P. Terreni and G. Bertini

A Generalized Musical-Tone Generator with Application to Sound Compression and Synthesis 431
C. Drioli and D. Rocchesso

A Singing Voice Synthesis System Based on Sinusoidal Modeling 435
M. Macon, L. Jensen-Link, J. Oliverio, M. Clements and E. George

Time-Scale Modification of Audio Signals with Combined Harmonic and Wavelet Representations 439
K. Hamdy, A. Tewfik, S. Takagi and T. Chen

A Waveguide Model for Slapbass Synthesis 443
E. Rank and G. Kubin

Minimum Perceptual Spectral Distance FIR Filter Design 447
S. Wu and W. Putnam

A Phase Interpolation Algorithm for Sinusoidal Model Based Music Synthesis 451
X. Qian and Y. Ding

Analytical Approximations of Fractional Delays: Lagrange Interpolators and Allpass Filters 455
S. Tassart and P. Depalle

Improved Discrete-Time Modeling of Multi-Dimensional Wave Propagation Using the Interpolated Digital Waveguide Mesh 459
L. Savioja and V. Välimäki

Matched Field Processing

Generalized Likelihood Ratio Test for Selecting a Geo-Acoustic Environmental Model 463
C. Mecklenbräuker, P. Gerstoft, P. Chung and J. Böhme

Tuning Genetic Algorithms for Underwater Acoustics Using a priori Statistical Information 467
M. Rendas and G. Bienvenu

Robust Beamformer Weight Design for Broadband Matched-Field Processing 471
K. Harmanci and J. Krolik

FASTMAP: A Fast, Approximate Maximum a posteriori Probability Parameter Estimator with Application to Robust Matched-Field Processing 475
B. Harrison, R. Vaccaro and D. Tufts

Electromagnetic Matched Field Processing for Source Localization 479
D. Gingras, P. Gerstoft, N. Gerr and C. Mecklenbräuker

Detection, Classification and Localisation

Power-Law Processors for Detecting Unknown in Signals in Colored Noise 483
I. Kirsteins, S. Mehta and J. Fay

Multitarget Detection/tracking of Echoes with Known Waveform: Algorithm and Applications 487
V. Rampa and U. Spagnolini

Detection of Gaussian Bandpass Transients Under Impulsive Noise: A Wavelet Transform Approach 491
F. Garcia and I. Lourtie

Maximum Likelihood Estimator for Magneto-Acoustic Localisation 495
G. Dassot, R. Blanpain and C. Jauffret

Barankin Bound for Source Localization in Shallow Water 499
J. Tabrikian and J. Krolik

Underwater Transient Signal Processing: Marine Mammal Identification, Localization, and Source Signal Deconvolution 503
Z. Michalopoulou

Numerical Optimization of Non-Adaptive Microphone Arrays 507
A. Goldin

Joint Direction-of-Arrival and Array Shape Tracking for Multiple Moving Targets 511
J. Goldberg, A. Perez-Neira and M. Lagunas

Comparison of Probabilistic Least Squares and Probabilistic Multi-Hypothesis Tracking Algorithms for Multi-Sensor Tracking 515
M. Krieg and D. Gray

Direction Finding with Imperfect Wavefront Coherence: A Matrix Fitting Approach Using Genetic Algorithm ..... 51[illegible]
A. Gershman, C. Mecklenbräuker and J. Böhme

Design of an Optimum Wideband Active Sonar Array with Robustness 523
S. Abeysekera and Y. Leung

Ocean Signal Processing

Multipath Time-Delay Estimation 527
J. Fuchs

Fast Maximum Likelihood Estimation with Multiple Signal Initialization 531
R. MacLeod and R. Vaccaro

An Algorithm for Detecting Closely Spaced Delay/Doppler Components 535
A. Habboosh, R. Vaccaro and S. Kay

Improvement of TDOA Measurement Using Wavelet Denoising with a Novel Thresholding Technique 539
S. Wu, H. So and P. Ching

A Short-Time Wiener Filter for Noise Removal in Underwater Acoustic Data 543
C. Therrien, K. Frack and N. Fontes

Fast Approximate Subspace Tracking (FAST) 547
D. Tufts, E. Real and J. Cooley

A Subspace Framework for Fast Parameter Estimation with Known Waveforms 551
B. Freburger, D. Tufts and T. Palka

Terrain Classification in Polarimetric SAR Using Wavelet Packets 555
N. Keshava and J. Moura

Electromagnetic Matched-Field Processing for Target Height Finding with Over-the-Horizon Radar 559
M. Papazoglou and J. Krolik

Time-Frequency Classification Using a Multiple Hypotheses Test: An Application to the Classification of Humpback Whale Signals 563
G. Roberts, A. Zoubir and B. Boashash

Source Classification Using Pole Method of AR Model 567
J. Huang, J. Zhao and Y. Xie

The LMMSE Estimate-Based Multiuser Detector: Performance Analyses and Adaptive Implementation 571
H. Ge

Improved Doppler Tracking and Correction for Underwater Acoustic Communications 575
M. Johnson, L. Freitag and M. Stojanovic

A Blind Multichannel Combiner for Long Range Underwater Communications 579
B. Sharif, J. Neasham, D. Thompson, O. Hinton and A. Adams

DSP Processors

Very Long Instruction Word Architectures for Digital Signal Processing 583
J. Mellott and F. Taylor

A Novel 32 Bit RISC Architecture Unifying RISC and DSP 587
C. Baumhof, F. Müller, O. Müller and M. Schlett

A Dual-Issue RISC Processor for Multimedia Signal Processing 591
H. Sato, E. Holmann, T. Yoshida, M. Matsuo and T. Kengaku

A Processor-Coprocessor Architecture for High End Video Applications 595
E. Maas, D. Herrmann, R. Ernst, P. Rüffer, S. Hasenzahl and M. Seitz

An MPEG-2 Encoder Architecture Based on a Single-Chip Dedicated LSI with a Control MPU 599
Y. Ooi, O. Ohnishi, Y. Yokoyama, Y. Katayama, M. Mizuno, M. Yamashina, H. Takano, N. Hayashi and I. Tamitani

An Efficient and Reconfigurable VLSI Architecture for Different Block Matching Motion Estimation Algorithms 603
X. Zhang and C. Tsui

An Operation-Saving VLSI Geometry Engine Core 607
K. Karagianni, G. Diamantakos, V. Paliouras and T. Stouraitis

The FFT Butterfly Operation in 4 Processor Cycles on a 24 Bit Fixed-Point DSP with a Pipelined Multiplier 611
M. Grajcar and B. Sick

New Unified VLSI Architectures for Computing DFT and Other Transforms 615
S. Hsiao and C. Yen

Half-Rate GSM Vocoder Implementation on a Dual Mac Digital Signal Processor 619
M. Prasad, P. D'Arcy, A. Gupta, M. Diamondstein and H. Srinivas

VLSI Implementation of an Area-Efficient Architecture for the Viterbi Algorithm 623
C. Cabrera, M. Boo and J. Bruguera

DSP Building Blocks & Arithmetic

Low-Area Dual Basis Divider Over GF(2^M) 627
L. Song and K. Parhi

VLSI Architecture for Datapath Integration of Arithmetic Over GF(2^M) on Digital Signal Processors 631
W. Drescher, K. Bachmann and G. Fettweis

A Fast Direction Sequence Generation Method for CORDIC Processors 635
S. Nahm and W. Sung

A Radix-4 Redundant Cordic Algorithm with Fast On-Line Variable Scale Factor Compensation 639
C. Li and S. Chen

Pipelining of Cordic Based IIR Digital Filters 643
J. Ma, K. Parhi and E. Deprettere

An Asynchronous Implementation of the MAXLIST Algorithm 647
C. Myers and H. Zheng

A Novel Systematic Mapping Approach for Highly Efficient Multiplexed FIR-Filter Architectures 651
W. Wilhelm and T. Noll

Design Automation for DSP

An Upper Bound of the Throughput of Multirate Multiprocessor Schedules 655
R. Schoenen, V. Zivojnovic and H. Meyr

Minimizing the Number of Operations in DSP Computations 659
I. Hong and M. Potkonjak

BEEHIVE: An Adaptive, Distributed, Embedded Signal Processing Environment 663
S. Famorzadeh, V. Madisetti, T. Egolf and T. Nguyen

On Objective Function Selection in List Scheduling Algorithms for Digital Signal Processing Applications 667
J. Jonsson and J. Vasell

VLSI High Level Synthesis of Fast Exact Least Mean Square Algorithms Based on Fast FIR Filters 671
J. Diguet, O. Sentieys, D. Chillet and J. Philippe

Hierarchical VHDL Libraries for DSP ASIC Design 675
J. McCanny, D. Ridge, Y. Hu and J. Hunter

DSP Quant: Design, Validation, and Applications of DSP Hard Real-Time Benchmark 679
C. Lee, D. Kirovski, I. Hong and M. Potkonjak

Constructing Memory Layouts for Address Generation Units Supporting Offset 2 Access 683
B. Wess and M. Gotschlich

Modulo-Addressing Utilization in Automatic Software Synthesis for Digital Signal Processors 687
M. Willems, H. Keding, V. Zivojnovic and H. Meyr

Cooperative Register Assignment and Code Compaction for Digital Signal Processors with Irregular Datapaths 691
W. Kreuzer and B. Wess

Optimization of Embedded DSP Programs Using Post-Pass Data-Flow Analysis 695
A. Sudarsanam, S. Malik, S. Tjiang and S. Liao

Code Positioning to Reduce Instruction Cache Misses in Signal Processing Applications on Multimedia RISC Processors 699
H. Stolberg, M. Ikekawa and I. Kuroda

Code Generation by Using Integer-Controlled Dataflow Graph 703
T. Miyazaki and E. Lee

Fixed-Point C Compiler for TMS320C50 Digital Signal Processor 707
J. Kang and W. Sung

Author Index A1

This material may be protected by Copyright law (Title 17 U.S. Code)

MULTIMODAL INTERFACES FOR MULTIMEDIA INFORMATION AGENTS

Alex Waibel, Bernhard Suhm, Minh Tue Vo and Jie Yang

Interactive Systems Laboratories
Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh PA 15217 (USA)
{waibel,bsuhm,tue,yang+}@cs.cmu.edu

ABSTRACT

When humans communicate they take advantage of a rich spectrum of cues. Some are verbal and acoustic. Some are non-verbal and non-acoustic. Signal processing technology has devoted much attention to the recognition of speech, as a single human communication signal. Most other complementary communication cues, however, remain unexplored and unused in human-computer interaction. In this paper we show that the addition of non-acoustic or non-verbal cues can significantly enhance robustness, flexibility, naturalness and performance of human-computer interaction. We demonstrate computer agents that use speech, gesture, handwriting, pointing, spelling *jointly* for more robust, natural and flexible human-computer interaction in the various tasks of an information worker: information creation, access, manipulation or dissemination.

1. INTRODUCTION

Human-computer interfaces today are limited and inflexible and do not take advantage of the many communication channels humans use to communicate verbal and non-verbal ways. Humans speak, point, gesture, write, fixate, use facial expressions, head-motion, eye-contact, etc. to express ideas, intentions and feelings. To build computer systems that operate more flexibly and robustly, and that are more intuitive to use by anyone, computer interfaces must be able to understand and process this multiplicity of human communication cues. Recently many research projects have been conducted to address this problem (e.g. [1], [2], [3], [4]). The need for more intuitive interfaces becomes all the more pressing as multimedia presentation, that is output, is becoming commonly available.

In this paper we describe information agents that use multimodal interfaces to access, manipulate, create and disseminate information. In particular, we seek effective methods by which human users can easily interact with a computer by speaking, pointing, drawing, spelling, etc. We show how cross-modal cues can be used effectively to recover from errors or miscommunications, since confusions in one modality usually are unambiguous in another. We describe how multiple modalities can cooperate and complement each other in various phases of the preparation of multimedia documents. Moreover, the added flexibility and robustness significantly increases reliability and naturalness, and with it the usability and acceptance of information systems. We present QuickDoc as a prototypical application that combines our multimodal subsystems in a simple, but powerful way.

2. INTERPRETATION OF MULTIMODAL INPUT

2.1. Multimodal Components

Our labs have implemented recognizers and processing modules for various input modalities, most notably speech, pen input, and face tracking. These modality processors constitute the basic components of all our multimodal systems.

Speech

Our speech recognition subsystem is based on the recognition front-end of the JANUS speech translation system [5] which is capable of processing speaker-independent, spontaneous speech. The recognizer can be adapted to any task domain by retraining the language models and possibly tuning the acoustic models if the domain involves special vocabulary. We also have a high-performance, real-time continuous spelling recognizer for large lists of 100,000 or more names.

Gesture

Our approach to pen-based gesture recognition is to decompose pen strokes into sequences of basic shapes such as line, arc, arrow, circle, cross...[6] The same gesture shape may mean different things depending on the surrounding context, hence each gesture component is augmented by *gesture contexts* indicating spatial relationships between the gesture and nearby objects in the user interface.

Handwriting

Our MS-TDNN-based handwriting recognizer [7] is capable of processing writer-independent, continuous (cursive) handwriting at a recognition rate of 94% on a 20,000-word vocabulary. We employ simple heuristics to decide when to invoke handwriting recognition on pen input, e.g., when the gesture recognizer cannot identify the input strokes as basic shapes.

Face Tracking

We base our face-tracking subsystem on face color clustering and motion detection [8]. The face tracker can control a pan-tilt-zoom camera to follow a freely moving person, producing a constant-sized image of the face area in real time. Another software layer built on top of this face tracker can identify the eyes and other face features, build a model of the face, and estimate the gaze direction to track the head pose [9].

2.2. Joint Interpretation

In order to make sense of input from all available sources, we need a multimodal interpreter capable of producing an interpreta-

0-8186-7919-0/97 $10.00 © 1997 IEEE

tion of user intent (e.g., a command to execute in the application interface) from the output of the modality processors.

In our joint interpretation scheme, the user intent is represented by a frame consisting of slots specifying pieces of information such as the action to carry out or the parameters for that action. Recognition output from the modality processors are parsed into partially filled frames that are merged together to produce the combined interpretation as described in [6]. This technique leads to uniform handling of high-level information from all input sources, which is very important for modularity and extensibility. To add another input modality we need only provide a module to convert low-level recognizer output to a partially filled frame to be merged with others. In addition, context information can be retained across input events by merging with previous interpretation frames.

3. MULTIMODAL ERROR CORRECTION

Intensive research over recent years has boosted the performance of speech recognition technology significantly. Nevertheless it is widely believed that recognition performance will remain limited, at least for the foreseeable future. The potential for error with any speech interface will be balanced by the redundant and alternate ability to enter information or queries by spelled, handwritten or typed input. Rather than fighting a particular recognition error the user can therefore quickly circumvent the problem by choosing a different modality. Such alternates not only eliminate user frustration with potentially recurring errors, but also provide an avenue for background learning and adapting.

3.1. Multimodal Interactive Error Repair

Our approach is to have the user collaborate with the system in recovering from interpretation errors [10]. First, errors have to be identified, either initiated by the system, for instance based on some confidence measure, or by the user. If a graphical user interface (GUI) is available, the user can simply highlight erroneous words (the reparandum) in the recognition hypothesis displayed. For error correction, the user provides additional input, choosing among different correction methods: repair by repeating the reparandum by respeaking, spelling out loud, or handwriting; repair by paraphrasing the reparandum; repair by pen gestures (e.g. to delete or insert), in addition to the standard repair by typing or selecting among N-best alternatives.

The rationale for this multimodal approach to error recovery is twofold. First, it exploits the fact that different input modalities are orthogonal: words which are confusable in one modality, can be disambiguated in a different modality (e.g. "road" and "rote" spoken versus spelled). Secondly, recent studies [11] show that switching modality after system misinterpretation alleviates user frustration.

3.2. Error Repair for Multimedia Information Agents

The design of speech user interfaces is constrained by the context of the application, and so is the design of error repair. The application context varies along several dimensions: available modalities (speech only versus GUI), what is repaired (isolated words, phrases, sentences), goal of repair is (get verbatim every word versus get the intended action), dialogue metaphor (command controls versus conversational). For multimedia information agents, we see the following two application contexts as particularly relevant:

- (Mixed initiative) Spoken dialogue, for example in user queries and command control of the application. In this context, a GUI may be available, repair will typically be performed at the level of phrases, the goal of repair is to initiate the intended action (semantic repair), and a conversational dialogue is desirable.
- Dictation, for example to entry in fields, or to dictate reports. In this context, a GUI is very appropriate, repair will occur from isolated word to sentence level, the goal of repair is to get each word correct (verbatim repair).

We have built prototypical speech user interfaces with error repair capabilities for dictation and form filling tasks. In this application context, requiring the user to identify errors by highlighting them is appropriate, since the user naturally focuses on the actual recognition hypothesis. For error correction, we offer to repeat the input (potentially switching to another modality), and pen gestures to indicate where to insert words, or which word to delete. In form filling tasks, knowledge of the current field can provide powerful constraints for recognition by restricting vocabulary and language model accordingly.

3.3. Evaluating Interactive Error Repair

Given a set of error recovery methods feasible with current technology in a given application context, a crucial issue is to predict which methods a user will prefer. Based on the assumption of a rational user [13], we use the time to complete some input, including the time needed to correct errors, as an objective and easily quantifiable measure for the effectiveness of different error recovery methods, and main predictor of user preference. By assuming stochastic independence of the various repair attempts, the expected accuracy after a certain number of repair attempts can be estimated as a geometric series. Given a high level of accuracy sufficient for the application (e.g. 99%), the number of repair attempts can be estimated, and thus the total time required to input including repair [12].

A pilot evaluation on a form filling task [12] suggests that given current technology, repair by spelling and handwriting can be very effective, and significantly better than standard choice from N-best alternatives. Using the context of a repair, e.g. in rescoring N-best lists obtained by decoding repair input, can substantially improve the accuracy of repair. Although results are still preliminary, they show that our multimodal approach to interactive error recovery is very promising.

4. MULTIMODAL INFORMATION AGENTS

The objective for multmodal information agents is to quickly access, create, manipulate, and disseminate multimedia information. A multimodal agent can label multimedia information with appropriate classifications by voice or/and gesture annotations. All annotations are incorporated with original information. The agent will automatically attach hotlinks on the world-wide-web to broaden information sources. The agent then will generate an HTML format report. The agent can search all the information in the report including voice and gesture annotations and voice mails. The agent can repair errors and add more information. Finally, the agent can disseminate the report in electronic form or printed form.

4.1. Information Access

Multimodal information agents can speed up navigation of Information space and queries of a database. A city map, for example, can be queried by a combination of gesture, handwriting and speech. A query such as: "Show me all hospitals in this <circling-gesture> area," or "How far is from here to there <tracing-gesture>?" will access information that would otherwise have to be assembled and combined painstakingly in a series of steps. Multimodal agents will recognize gesture and speech concepts and map them onto a semantic representation. The semantic representation can be filled interchangeably by both gestures and verbal phrases using pattern recognition and chart parsing techniques. The parsing result is then sent to query the database. If interpretation errors occur, the user corrects them using multimodal correction techniques. The whole process is transparent to the user.

Searching multimedia data can be viewed as natural extension of multimodal database queries. For instance, voice annotations or video clips added during user interaction can be interpreted and indexed using multimodal interpretation components, and then accessed using database queries techniques.

4.2. Information Creation

A user often needs to add information to a database or modify existed information. Multimodal agents are also useful in such a process. For example, a doctor needs to mark his/her diagnoses on X-ray, CT, MR images and make comments on the case. He/she can use speech, gesture, and handwriting to do this. Similarly iconic gestures may be used on a touch sensitive display or writing pad, to quickly enter resources or objects in a database, such as a map, in non-verbal ways. Iconic gestures can be defined to represent objects that would be entered in the database. In the same way, handwriting symbols can be used to mark or tag objects on a screen. Multimodal input also need not always be interpreted, but can often be stored for later review. Thus a voice annotation, or a circling gesture can be simply stored to go along with an image for later review and retrieval. The efficiency of multimodal agents for information creation can illustrated by the following example. The scenario is an image analyst sitting in the front of a computer and classifying various targets. Images can be popped up quickly and in rapid succession and the analyst provides a quick classification. The image and its classification are entered into a database along with the analyst's spoken and gestured annotation. With the human in the loop to make decisions on intelligence relevance, but with the availability of a better (multimodal) interface, to do so quickly, the job of data creation and manipulation can be carried out with greater reliability and efficiency compared to a fully automated system.

4.3. Information Manipulation

Information will not only be requested to viewed, but will also need to be updated. A multimodal multimedia information system can offer new ways to maintain the underlying databases: the user can use speech, gesture, writing, spelling to update database information efficiently. manipulating visual representations on the display rather than having to learn and use a separate database manipulation system. Gestures play an important role in specifying object parameters (e.g. different iconic shapes to represent different object types) and spatial constraints (e.g., location and extent of objects), while speech is useful in specifying parameters not easily expressed visually. For example he/she might circle a bridge on a map and say "this bridge has been destroyed", or might cross out an object without words, etc.

In some situations, the answer for a query cannot be directly obtained from the database and the information from the database has to be future processed. For example, the query "Show me the nearest restaurant" requires not only database access but also other computations.

4.4. Information Dissemination

In addition to information access, creation and manipulation, the product has to be packaged and readied for later viewing. The world wide web can serve as a natural medium to disseminate multimedia information. The multimodal input signals are interpreted and compiled, and an HTML format report is automatically generated on the fly. It is then automatically available for viewing together with the original aligned speech waveform attached to it as hotlinks. Multimodal interaction can also be used to manipulate objects in a report, such as positioning objects on a page appropriately, or attaching appropriate hotlinks.

To quickly generate a multimedia report, speech dictation can be combined with point and click actions to arrange reports on the fly. Such a multimodal report generation would be faster than typing and can generate a hierarchy of facts and notions. The generation thus would consist of some dictation or typing, some multimodal object manipulation ("move this <point> here <point> and link it as an explanation for that <click>") and some handwriting or drawings that are to be attached. The result of these communicative acts will compile into HTML code, that is ready to be transferred. Time savings are possible in several ways: the report generation can be done more efficiently by speech, gesture, or handwriting, and by using references to already existing multimedia stickies, factoids, video footage or explanations, and the recipient of the report can probe and browse a hierarchical multimedia report in a non-linear fashion more efficiently, calling up aspects of the assessment as needed.

4.5. Controlling the Interface

Although the goal of a multimodal interface is to provide as transparent an access to functionality as possible, interaction to control various aspects of the system can't be eliminated.

For example for navigation within visually presented image data, the user typically modifies the view by operations such as zooming and panning. The integration of speech and gesture increases flexibility by allowing different attributes of an operation to be specified in whatever modality is more appropriate for each attribute; for instance, the effect of a spoken "zoom in to this <circle> area" accompanied by a circle drawn around the desired area is much more difficult to achieve using a single input modality

Additionally, multiple modalities can be used to enhance the reliability. For example, a gaze tracker can be used to detect the user's focus of attention. When the user is looking at a window on a screen, the window will be highlighted. No action is taken unless the user uses a voice command to confirm the selection. The voice commands could be ``select this window" or ``close window", etc. This can reduce unintended actions caused by the user randomly looking around.

5. THE QUICKDOC APPLICATION

QuickDoc is an example of an application that combines our multimodal subsystems in a simple but powerful way, letting the user perform a repetitive task with speed and convenience. The task is for a doctor to go through a series of images such as X-rays or computer-aided tomography scans, quickly identify an anomalous

area, label the area with the name of a disease or condition, and attach relevant comments. The end product is an HTML report that summarizes the doctor's findings in a compact table listing the annotated images, the corresponding preliminary diagnoses, and automatically generated hotlinks to relevant sites based on the diagnoses.

The QuickDoc user labels each presented image by circling an area on the image (drawing directly on a touch-sensitive screen) and speaking a disease name as well as a percentage representing the confidence level of the diagnosis. The gesture recognizer identifies the circle and generates an area marker on the image, attaching other parts of the gesture as written annotations (a future version may run the handwriting recognizer to turn those into text). The user can issue a spoken command to initiate a voice annotation attachment; the recorded audio file is also run through the speech recognizer to produce an automatic transcription. When all the images have been processed, another spoken command causes the HTML report to be generated. Optionally the collected data can also be edited in a multimodal dictation and repair facility before being processed into HTML.

The generated report contains thumbnails of the annotated images accompanied by disease label, confidence level, location in the image (automatically extracted from the circling gesture), and voice annotation playback icon if applicable. The thumbnail is linked to a more detailed page containing the full-size image and the automatic voice transcription. The name of the disease is looked up in a database and turned into a hotlink to a page listing Web sites relevant to that disease. A keyword search mode allows to search spoken keywords in the text of the report and the automatic voice transcriptions; entries containing hits are then highlighted in the report.

QuickDoc thus embodies multimodal creation, manipulation, dissemination, and access of multimedia information in a simple yet surprisingly useful application.

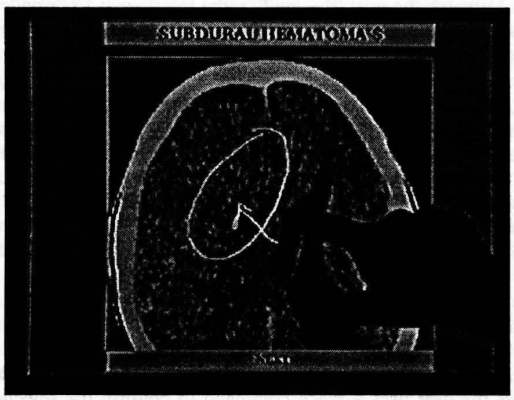

FIGURE 1: Multimodal information creation in QuickDoc: labeling some area by a circle gesture and speech ("subdural hematoma")

6. CONCLUSIONS

In this paper we have described systems that combine speech, gesture, handwriting, pointing, spelling, and other communication modalities into interfaces that are robust, flexible, and intuitive to use. We described how multimodal error recovery can ease the problem of unreliable automatic interpretation of communication modalities, in particular speech, thus removing a major obstacle in making multimodal interfaces truly usable. We demonstrated how these concepts can be combined for a more user friendly multimedia document production technology.

7. ACKNOWLEDGEMENTS

This research was sponsored by the DARPA under the Department of the Navy, Naval Research Office under grant number N00014-93-1-0806. Views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Navy or the U.S. Government.

8. REFERENCES

[1] H. Ando, Y. Kitahara and N. Hataoka: "Evaluation of Multimodal Interface Using Spoken Language and Pointing Gesture on Interior Design System", *Proc. ICSLP*, 1994, Vol. 2, pp. 567-570.

[2] T. Nishimoto, N. Shida, T. Kobayashi, K. Shirai: "Multimodal Drawing Tool Using Speech, Mouse and Keyboard", *Proc. ICSLP*, 1994, Vol. 3, pp. 1287-1290.

[3] J. Wang: "Integration of Eye-gaze, Voice and Manual Response in Multimodal User Interface", *Proc. ICSMC*, 1995, Vol. 5, pp. 3938-42.

[4] A. Waibel, M.T. Vo, P. Duchnowski, and S. Manke: "Multimodal Interfaces," *Artificial Intelligence Review, Special Volume on Integration of Natural Language and Vision Processing*, McKevitt, P. (Ed.), Vol. 10, Nos. 3-4, 1995.

[5] B. Suhm, P. Geutner, T. Kemp, A. Lavie, L.J. Mayfield, A. McNair, I. Rogina, T. Schultz, T. Sloboda, W. Ward, M. Woszczyna and A. Waibel: "JANUS: Towards Multilingual Spoken Language Translation," *Proc. ARPA SLT Workshop 95* (Austin, Texas).

[6] M.T. Vo and C. Wood: "Building and application framework for speech and pen input integration in multimodal learning interfaces," *Proc. ICASSP'96* (Atlanta, GA).

[7] S. Manke, M. Finke and A. Waibel: "The Use of Dynamic Writing Information in a Connectionist On-Line Cursive Handwriting Recognition System," *Advances in Neural Information Processing Systems 6*, Morgan Kaufmann, 1994.

[8] J. Yang and A. Waibel: "A real-time face tracker," *Proc. WACV'96*, pp. 142-147.

[9] R. Stiefelhagen, J. Yang and A. Waibel: "A Modelbased Gaze Tracking System," *Proc. IEEE International Joint Symposia on Intelligence and Systems - Image, Speech & Natural Language Systems*, pp. 304-310, 1996.

[10] A.E. McNair and A. Waibel: "Improving Recognizer Acceptance through Robust, Natural Speech Repair", *Proc. ICSLP*, 1994, Vol. 3, pp. 1299-1302.

[11] S.L. Oviatt and R. VanGent: "Error Resolution during Multimodal Human-Computer Interaction", *Proc. ICSLP*, 1996.

[12] B. Suhm, B. Myers and A. Waibel: "Interactive Recovery from Speech Recognition Errors in Speech User Interfaces", *Proceedings of the International Conference on Spoken Language Processing - ICSLP*, 1996.

[13] A. Newell, S. Card and P. Moran: "The Psychology of Human Computer Interaction", Lawrence Earlbaum Associates, 1983.