

Declaration of Catherine Vassilkova

1. I, Catherine Vassilkova, am an employee of BEADOX LLC ("BEA").
2. I am over the age of 21 and have personal knowledge of the facts contained herein unless otherwise indicated.
3. I have worked in the field of information research and document retrieval for law firms since 2005.
4. I have been with BEA since I established the company in 2019.
5. BEA, based in Silver Spring, MD, conducts research at the District of Columbia federal and local courts, archives, government agencies, and libraries around the world. BEA's services include interlibrary loans, specialized research, article location and delivery, and other document retrieval. One of my responsibilities is to scan textbooks for requesters.
6. BEA specializes in the research and delivery of materials from the Library of Congress.
7. On September 26, 2025, I was asked by the requester Morrison and Foerster LLP to obtain a scan from the following book: Springer handbook of speech processing (2008).
8. On September 26, 2025, I obtained Springer handbook of speech processing (2008) from the Library of Congress, and scanned the requested excerpts including book cover, title page, copyright page date-stamped (December 11 2007), table of contents pages, and chapter sections 50.1 (p.1021-1023), 51.2.1 from Springer handbook of speech processing (2008). Those excerpts are included as Exhibit 1009.
9. I confirm that the copyright page of the scanned book has a stamped date of receipt of December 11 2007.
10. I declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true, and that these statements were made with knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under section 1001 of Title 18 of the United States Code.
11. Executed on the 29th of September, 2025, in Silver Spring, MD 20910.

Signed:

*C. Vassilkova*

# Springer Handbook of Speech Processing



Benesty  
Sondhi  
Huang  
Editors

From common consumer products such as cell phones and MP3 players to more sophisticated projects such as human-machine interfaces and responsive robots, speech technologies are now everywhere. Many think that it is just a matter of time before more applications of the science of speech become inescapable in our daily life. This Springer Handbook of Speech Processing is meant to play a fundamental role for sustainable progress in speech research and development. It targets three categories of readers: graduate students, professors and active researchers in academia and research labs, and engineers in industry who need to understand or implement some specific algorithms for their speech-related products. The handbook could also be used as a sourcebook for one or more graduate courses on signal processing for speech and different aspects of speech processing and applications. A quickly accessible source of application-oriented, authoritative and comprehensive information about these technologies, it combines the established knowledge derived from research in such fast evolving disciplines as Signal Processing and Communications, Acoustics, Computer Science, and Linguistics.

## Key Topics

- › Production, Perception, and Modeling of Speech
- › Signal Processing for Speech
- › Speech Coding
- › Text-to-Speech Synthesis
- › Speech and Speaker Recognition
- › Language Recognition
- › Speech Enhancement
- › Multichannel Speech Processing

## Features

- › Authoritative desk reference of one of tomorrow's breakthrough technologies
- › Provides quick access to applicable, reliable, and comprehensive knowledge
- › Contains over 450 two-color illustrations
- › Parts and chapters with summaries, detailed index and fully searchable DVD-ROM guarantee quick access to data and links to other sources
- › Delivers a wealth of up-to-date references and further reading



Springer  
Handbook  
of  
Speech  
Processing

TK 7882  
.S65  
.S6715  
2008  
Copy 1

ISBN 978-3-540-49125-5



springer.com

DVD containing all contents  
best viewed with Adobe Reader 8



## System Requirements

- Windows 95/98/ME 32 MB RAM, Windows NT 4/2000/XP 64 MB RAM, Windows Vista 512 MB RAM
- MAC OS 9 or higher, 64 MB RAM, 333 MHz
- LINUX Pentium I /166 MHz, 64 MB RAM

# Springer Handbook of

# Speech Processing



Benesty  
Sondhi  
Huang  
Editors

Springer

---

**Springer Handbook  
of Speech Processing**

**Springer Handbooks** provide a concise compilation of approved key information on methods of research, general principles, and functional relationships in physical sciences and engineering. The world's leading experts in the fields of physics and engineering will be assigned by one or several renowned editors to write the chapters comprising each volume. The content is selected by these experts from Springer sources (books, journals, online content) and other systematic and approved recent publications of physical and technical information.

The volumes are designed to be useful as readable desk reference books to give a fast and comprehensive overview and easy retrieval of essential reliable key information, including tables, graphs, and bibliographies. References to extensive sources are provided.

# Springer Handbook of Speech Processing

Jacob Benesty, M. Mohan Sondhi, Yiteng Huang  
(Eds.)

With DVD-ROM, 456 Figures and 113 Tables

 Springer

Editors:

Jacob Benesty  
INRS-EMT, University of Quebec  
800 de la Gauchetiere Ouest, Suite 6900  
Montreal, Quebec, H5A 1K6, Canada  
benesty@emt.inrs.ca

M. Mohan Sondhi  
Avayalabs Research  
233 Mount Airy Road  
Basking Ridge, NJ 07920, USA  
mms@research.avayalabs.com

Yiteng Huang  
Bell Laboratories, Alcatel-Lucent  
600 Mountain Avenue  
Murray Hill, NJ 07974, USA  
arden\_huang@iccc.org

Library of Congress Control Number: 2007931999

ISBN: 978-3-540-49125-5 e-ISBN: 978-3-540-49127-9

This work is subject to copyright. All rights reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September, 9, 1965, in its current version, and permission for use must always be obtained from Springer-Verlag. Violations are liable for prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media  
springer.com

© Springer-Verlag Berlin Heidelberg 2008

The use of designations, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Product liability: The publisher cannot guarantee the accuracy of any information about dosage and application contained in this book. In every individual case the user must check such information by consulting the relevant literature.

Typesetting and production:  
LE- $\text{\TeX}$  Jelonek, Schmidt&Vöckler GbR, Leipzig  
Senior Manager Springer Handbook: Dr. W. Skolaut, Heidelberg  
Typography and layout: schreiberVIS, Seeheim  
Illustrations: Hippmann GbR, Schwarzenbruck  
Cover design: eStudio Calamar Steinen, Barcelona  
Cover production: WMXDesign GmbH, Heidelberg  
Printing and binding: Stürtz GmbH, Würzburg

Printed on acid free paper

SPIN 11544036 60/3180/YL 5 4 3 2 1 0

DEC 11 2007  
57E

## Foreword

Over the past three decades digital signal processing has emerged as a recognized discipline. Much of the impetus for this advance stems from research in representation, coding, transmission, storage and reproduction of speech and image information. In particular, interest in voice communication has stimulated central contributions to digital filtering and discrete-time spectral transforms.

This dynamic development was built upon the convergence of three then-evolving technologies: (i) sampled-data theory and representation of information signals (which led directly to digital telecommunication that provides signal quality independent of transmission distance); (ii) electronic binary computation (aided in early implementation by pulse-circuit techniques from radar design); and, (iii) invention of solid-state devices for exquisite control of electronic current (transistors – which now, through microelectronic materials, scale to systems of enormous size and complexity). This timely convergence was soon followed by optical fiber methods for broadband information transport.

These advances impact an important aspect of human activity – information exchange. And, over man's existence, speech has played a principal role in human communication. Now, speech is playing an increasing role in human interaction with complex information systems. Automatic services of great variety exploit the comfort of voice exchange, and, in the corporate sector, sophisticated audio/video teleconferencing is reducing the necessity of expensive, time-consuming business travel. In each instance an overarching target is a user environment that captures some of the naturalness and spatial realism of face-to-face communication. Again, speech is a core element, and new understanding from diverse research sectors can be brought to bear.

Editors-in-Chief Benesty, Sondhi and Huang have organized a timely engineering handbook to answer this need. They have assembled a remarkable compendium of current knowledge in speech processing. And, this accumulated understanding can be focused upon enlarging the human capacity to deal with a world ever increasing in complexity. Benesty, Sondhi and Huang are renowned researchers in their own right, and they have attracted an international cadre of over 80 fellow authors and collaborators who constitute a veritable *Who's Who* of world leaders in speech processing research. The resulting book provides under one cover authoritative treatments that commence with the basic physics and psychophysics of speech and hearing, and range through the related topics of computational tools, coding, synthesis, recognition, and signal enhancement, concluding with discussions on capture and projection of sound in enclosures. The book can be expected to become a valuable resource for researchers, engineers and speech scientists throughout the global community. It should equally serve teachers and students in human communication, especially delimiting knowledge frontiers where graduate thesis research may be appropriate.

Warren, New Jersey  
October 2007

Jim Flanagan



**J. L. Flanagan**  
Professor Emeritus  
Electrical and Computer  
Engineering  
Rutgers University

Editors:

Jacob Benesty  
INRS-EMT, University of Quebec  
800 de la Gauchetière Ouest, Suite 6900  
Montreal, Quebec, H5A 1K6, Canada  
benesty@emt.inrs.ca

M. Mohan Sondhi  
Avayalabs Research  
233 Mount Airy Road  
Basking Ridge, NJ 07920, USA  
mms@research.avayalabs.com

Yiteng Huang  
Bell Laboratories, Alcatel-Lucent  
600 Mountain Avenue  
Murray Hill, NJ 07974, USA  
arden\_huang@ieee.org

Library of Congress Control Number: 2007931999

ISBN: 978-3-540-49125-5 e-ISBN: 978-3-540-49127-9

This work is subject to copyright. All rights reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September, 9, 1965, in its current version, and permission for use must always be obtained from Springer-Verlag. Violations are liable for prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media  
springer.com

© Springer-Verlag Berlin Heidelberg 2008

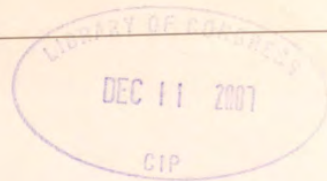
The use of designations, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Product liability: The publisher cannot guarantee the accuracy of any information about dosage and application contained in this book. In every individual case the user must check such information by consulting the relevant literature.

Typesetting and production:  
LE-TeX Jelonek, Schmidt&Vöckler GbR, Leipzig  
Senior Manager Springer Handbook: Dr. W. Skolaut, Heidelberg  
Typography and layout: schreiberVIS, Seeheim  
Illustrations: Hippmann GbR, Schwarzenbruck  
Cover design: eStudio Calamar Steinen, Barcelona  
Cover production: WMXDesign GmbH, Heidelberg  
Printing and binding: Stürtz GmbH, Würzburg

Printed on acid free paper

SPIN 11544036 60/3180/YL 5 4 3 2 1 0



## Foreword

Over the past three decades digital signal processing has emerged as a recognized discipline. Much of the impetus for this advance stems from research in representation, coding, transmission, storage and reproduction of speech and image information. In particular, interest in voice communication has stimulated central contributions to digital filtering and discrete-time spectral transforms.

This dynamic development was built upon the convergence of three then-evolving technologies: (i) sampled-data theory and representation of information signals (which led directly to digital telecommunication that provides signal quality independent of transmission distance); (ii) electronic binary computation (aided in early implementation by pulse-circuit techniques from radar design); and, (iii) invention of solid-state devices for exquisite control of electronic current (transistors – which now, through microelectronic materials, scale to systems of enormous size and complexity). This timely convergence was soon followed by optical fiber methods for broadband information transport.

These advances impact an important aspect of human activity – information exchange. And, over man's existence, speech has played a principal role in human communication. Now, speech is playing an increasing role in human interaction with complex information systems. Automatic services of great variety exploit the comfort of voice exchange, and, in the corporate sector, sophisticated audio/video teleconferencing is reducing the necessity of expensive, time-consuming business travel. In each instance an overarching target is a user environment that captures some of the naturalness and spatial realism of face-to-face communication. Again, speech is a core element, and new understanding from diverse research sectors can be brought to bear.

Editors-in-Chief Benesty, Sondhi and Huang have organized a timely engineering handbook to answer this need. They have assembled a remarkable compendium of current knowledge in speech processing. And, this accumulated understanding can be focused upon enlarging the human capacity to deal with a world ever increasing in complexity. Benesty, Sondhi and Huang are renowned researchers in their own right, and they have attracted an international cadre of over 80 fellow authors and collaborators who constitute a veritable *Who's Who* of world leaders in speech processing research. The resulting book provides under one cover authoritative treatments that commence with the basic physics and psychophysics of speech and hearing, and range through the related topics of computational tools, coding, synthesis, recognition, and signal enhancement, concluding with discussions on capture and projection of sound in enclosures. The book can be expected to become a valuable resource for researchers, engineers and speech scientists throughout the global community. It should equally serve teachers and students in human communication, especially delimiting knowledge frontiers where graduate thesis research may be appropriate.

Warren, New Jersey  
October 2007

Jim Flanagan



**J. L. Flanagan**  
Professor Emeritus  
Electrical and Computer  
Engineering  
Rutgers University

**Stephanie Seneff**

Massachusetts Institute of Technology  
 Computer Science and Artificial  
 Intelligence Laboratory  
 32 Vassar Street  
 Cambridge, MA 02139, USA  
 e-mail: [senef@csail.mit.edu](mailto:senef@csail.mit.edu)

**Wade Shen**

Massachusetts Institute of Technology  
 Communication Systems, Information Systems  
 Technology, Lincoln Laboratory  
 244 Wood Street  
 Lexington, MA 02420-9108, USA  
 e-mail: [swade@ll.mit.edu](mailto:swade@ll.mit.edu)

**Elliot Singer**

Massachusetts Institute of Technology  
 Information Systems Technology Group, Lincoln  
 Laboratory  
 244 Wood Street  
 Lexington, MA 02420-9108, USA  
 e-mail: [es@ll.mit.edu](mailto:es@ll.mit.edu)

**Jan Skoglund**

Global IP Solutions  
 301 Brannan Street  
 San Francisco, CA 94107, USA  
 e-mail: [jan.skoglund@gipscorp.com](mailto:jan.skoglund@gipscorp.com)

**M. Mohan Sondhi**

Avayalabs Research  
 233 Mount Airy Road  
 Basking Ridge, NJ 07920, USA  
 e-mail: [mms@research.avayalabs.com](mailto:mms@research.avayalabs.com)

**Sascha Spors**

Deutsche Telekom AG, Laboratories  
 Ernst-Reuter-Platz 7  
 10587 Berlin, Germany  
 e-mail: [Sascha.Spors@telekom.de](mailto:Sascha.Spors@telekom.de)

**Ann Spriet**

ESAT-SCD/SISTA, K.U. Leuven  
 Department of Electrical Engineering  
 Kasteelpark Arenberg 10  
 3001 Leuven, Belgium  
 e-mail: [ann.spriet@esat.kuleuven.be](mailto:ann.spriet@esat.kuleuven.be)

**Richard Sproat**

University of Illinois at Urbana-Champaign  
 Department of Linguistics  
 Urbana, IL 61801, USA  
 e-mail: [rws@uiuc.edu](mailto:rws@uiuc.edu)

**Yannis Stylianou**

Institute of Computer Science  
 Heraklion, Crete 700 13, Greece  
 e-mail: [yannis@csd.uoc.gr](mailto:yannis@csd.uoc.gr)

**Jes Thyssen**

Broadcom Corporation  
 5300 California Avenue  
 Irvine, CA 92617, USA  
 e-mail: [jthyssen@broadcom.com](mailto:jthyssen@broadcom.com)

**Jay Wilpon**

Research AT&T Labs  
 Voice and IP Services  
 Florham Park, NJ 07932, USA  
 e-mail: [jgw@research.att.com](mailto:jgw@research.att.com)

**Jan Wouters**

ExpORL, Department of Neurosciences, K.U. Leuven  
 O.& N2, Herestraat 49  
 3000 Leuven, Belgium  
 e-mail: [jan.wouters@med.kuleuven.be](mailto:jan.wouters@med.kuleuven.be)

**Arie Yeredor**

Tel-Aviv University  
 Electrical Engineering - Systems  
 Tel-Aviv 69978, Israel  
 e-mail: [arie@eng.tau.ac.il](mailto:arie@eng.tau.ac.il)

**Steve Young**

Cambridge University Engineering Dept  
 Cambridge, CB21PZ, UK  
 e-mail: [sjy@eng.cam.ac.uk](mailto:sjy@eng.cam.ac.uk)

**Victor Zue**

Massachusetts Institute of Technology  
 CSAI Laboratory  
 32 Vassar Street  
 Cambridge, MA 02139, USA  
 e-mail: [zue@csail.mit.edu](mailto:zue@csail.mit.edu)

**Contents**

<b>List of Abbreviations</b> .....	XXXI
<b>1 Introduction to Speech Processing</b>	
<i>J. Benesty, M. M. Sondhi, Y. Huang</i> .....	1
1.1 A Brief History of Speech Processing .....	1
1.2 Applications of Speech Processing .....	2
1.3 Organization of the Handbook .....	4
<b>References</b> .....	4
<b>Part A Production, Perception, and Modeling of Speech</b>	
<b>2 Physiological Processes of Speech Production</b>	
<i>K. Honda</i> .....	7
2.1 Overview of Speech Apparatus .....	7
2.2 Voice Production Mechanisms .....	8
2.3 Articulatory Mechanisms .....	14
2.4 Summary .....	24
<b>References</b> .....	25
<b>3 Nonlinear Cochlear Signal Processing and Masking in Speech Perception</b>	
<i>J. B. Allen</i> .....	27
3.1 Basics .....	27
3.2 The Nonlinear Cochlea .....	35
3.3 Neural Masking .....	45
3.4 Discussion and Summary .....	55
<b>References</b> .....	56
<b>4 Perception of Speech and Sound</b>	
<i>B. Kollmeier, T. Brand, B. Meyer</i> .....	61
4.1 Basic Psychoacoustic Quantities .....	62
4.2 Acoustical Information Required for Speech Perception .....	70
4.3 Speech Feature Perception .....	74
<b>References</b> .....	81
<b>5 Speech Quality Assessment</b>	
<i>V. Grancharov, W. B. Kleijn</i> .....	83
5.1 Degradation Factors Affecting Speech Quality .....	84
5.2 Subjective Tests .....	85
5.3 Objective Measures .....	90
5.4 Conclusions .....	95
<b>References</b> .....	96

**Part B Signal Processing for Speech**

<b>6 Wiener and Adaptive Filters</b>	
<i>J. Benesty, Y. Huang, J. Chen</i> .....	103
6.1 Overview.....	103
6.2 Signal Models.....	104
6.3 Derivation of the Wiener Filter.....	106
6.4 Impulse Response Tail Effect.....	107
6.5 Condition Number.....	108
6.6 Adaptive Algorithms.....	110
6.7 MIMO Wiener Filter.....	116
6.8 Conclusions.....	119
<b>References</b> .....	120
<b>7 Linear Prediction</b>	
<i>J. Benesty, J. Chen, Y. Huang</i> .....	121
7.1 Fundamentals.....	121
7.2 Forward Linear Prediction.....	122
7.3 Backward Linear Prediction.....	123
7.4 Levinson–Durbin Algorithm.....	124
7.5 Lattice Predictor.....	126
7.6 Spectral Representation.....	127
7.7 Linear Interpolation.....	128
7.8 Line Spectrum Pair Representation.....	129
7.9 Multichannel Linear Prediction.....	130
7.10 Conclusions.....	133
<b>References</b> .....	133
<b>8 The Kalman Filter</b>	
<i>S. Gannot, A. Yeredor</i> .....	135
8.1 Derivation of the Kalman Filter.....	136
8.2 Examples: Estimation of Parametric Stochastic Process from Noisy Observations.....	141
8.3 Extensions of the Kalman Filter.....	144
8.4 The Application of the Kalman Filter to Speech Processing.....	149
8.5 Summary.....	157
<b>References</b> .....	157
<b>9 Homomorphic Systems and Cepstrum Analysis of Speech</b>	
<i>R. W. Schafer</i> .....	161
9.1 Definitions.....	161
9.2 Z-Transform Analysis.....	164
9.3 Discrete-Time Model for Speech Production.....	165
9.4 The Cepstrum of Speech.....	166
9.5 Relation to LPC.....	169
9.6 Application to Pitch Detection.....	171

9.7 Applications to Analysis/Synthesis Coding.....	172
9.8 Applications to Speech Pattern Recognition.....	176
9.9 Summary.....	180
<b>References</b> .....	180
<b>10 Pitch and Voicing Determination of Speech with an Extension Toward Music Signals</b>	
<i>W. J. Hess</i> .....	181
10.1 Pitch in Time-Variant Quasiperiodic Acoustic Signals.....	182
10.2 Short-Term Analysis PDAs.....	185
10.3 Selected Time-Domain Methods.....	192
10.4 A Short Look into Voicing Determination.....	195
10.5 Evaluation and Postprocessing.....	197
10.6 Applications in Speech and Music.....	201
10.7 Some New Challenges and Developments.....	203
10.8 Concluding Remarks.....	207
<b>References</b> .....	208
<b>11 Formant Estimation and Tracking</b>	
<i>D. O'Shaughnessy</i> .....	213
11.1 Historical.....	213
11.2 Vocal Tract Resonances.....	215
11.3 Speech Production.....	216
11.4 Acoustics of the Vocal Tract.....	218
11.5 Short-Time Speech Analysis.....	221
11.6 Formant Estimation.....	223
11.7 Summary.....	226
<b>References</b> .....	226
<b>12 The STFT, Sinusoidal Models, and Speech Modification</b>	
<i>M. M. Goodwin</i> .....	229
12.1 The Short-Time Fourier Transform.....	230
12.2 Sinusoidal Models.....	242
12.3 Speech Modification.....	253
<b>References</b> .....	256
<b>13 Adaptive Blind Multichannel Identification</b>	
<i>Y. Huang, J. Benesty, J. Chen</i> .....	259
13.1 Overview.....	259
13.2 Signal Model and Problem Formulation.....	260
13.3 Identifiability and Principle.....	261
13.4 Constrained Time-Domain Multichannel LMS and Newton Algorithms.....	262
13.5 Unconstrained Multichannel LMS Algorithm with Optimal Step-Size Control.....	266
13.6 Frequency-Domain Blind Multichannel Identification Algorithms.....	268
13.7 Adaptive Multichannel Exponentiated Gradient Algorithm.....	276

13.8 Summary .....	279
References .....	279
<b>Part C Speech Coding</b>	
<b>14 Principles of Speech Coding</b>	
<i>W. B. Kleijn</i> .....	283
14.1 The Objective of Speech Coding .....	283
14.2 Speech Coder Attributes .....	284
14.3 A Universal Coder for Speech .....	286
14.4 Coding with Autoregressive Models .....	293
14.5 Distortion Measures and Coding Architecture .....	296
14.6 Summary .....	302
References .....	303
<b>15 Voice over IP: Speech Transmission over Packet Networks</b>	
<i>J. Skoglund, E. Kozica, J. Linden, R. Hagen, W. B. Kleijn</i> .....	307
15.1 Voice Communication .....	307
15.2 Properties of the Network .....	308
15.3 Outline of a VoIP System .....	313
15.4 Robust Encoding .....	317
15.5 Packet Loss Concealment .....	326
15.6 Conclusion .....	327
References .....	328
<b>16 Low-Bit-Rate Speech Coding</b>	
<i>A. V. McCree</i> .....	331
16.1 Speech Coding .....	331
16.2 Fundamentals: Parametric Modeling of Speech Signals .....	332
16.3 Flexible Parametric Models .....	337
16.4 Efficient Quantization of Model Parameters .....	344
16.5 Low-Rate Speech Coding Standards .....	345
16.6 Summary .....	347
References .....	347
<b>17 Analysis-by-Synthesis Speech Coding</b>	
<i>J.-H. Chen, J. Thyssen</i> .....	351
17.1 Overview .....	352
17.2 Basic Concepts of Analysis-by-Synthesis Coding .....	353
17.3 Overview of Prominent Analysis-by-Synthesis Speech Coders .....	357
17.4 Multipulse Linear Predictive Coding (MPLPC) .....	360
17.5 Regular-Pulse Excitation with Long-Term Prediction (RPE-LTP) .....	362
17.6 The Original Code Excited Linear Prediction (CELP) Coder .....	363
17.7 US Federal Standard FS1016 CELP .....	367
17.8 Vector Sum Excited Linear Prediction (VSELP) .....	368
17.9 Low-Delay CELP (LD-CELP) .....	370

17.10 Pitch Synchronous Innovation CELP (PSI-CELP) .....	371
17.11 Algebraic CELP (ACELP) .....	371
17.12 Conjugate Structure CELP (CS-CELP) and CS-ACELP .....	377
17.13 Relaxed CELP (RCELP) – Generalized Analysis by Synthesis .....	378
17.14 eX-CELP .....	381
17.15 iLBC .....	382
17.16 TSNFC .....	383
17.17 Embedded CELP .....	386
17.18 Summary of Analysis-by-Synthesis Speech Coders .....	388
17.19 Conclusion .....	390
References .....	390
<b>18 Perceptual Audio Coding of Speech Signals</b>	
<i>J. Herre, M. Lutzky</i> .....	393
18.1 History of Audio Coding .....	393
18.2 Fundamentals of Perceptual Audio Coding .....	394
18.3 Some Successful Standardized Audio Coders .....	396
18.4 Perceptual Audio Coding for Real-Time Communication .....	398
18.5 Hybrid/Crossover Coders .....	403
18.6 Summary .....	409
References .....	409
<b>Part D Text-to-Speech Synthesis</b>	
<b>19 Basic Principles of Speech Synthesis</b>	
<i>J. Schroeter</i> .....	413
19.1 The Basic Components of a TTS System .....	413
19.2 Speech Representations and Signal Processing for Concatenative Synthesis .....	421
19.3 Speech Signal Transformation Principles .....	423
19.4 Speech Synthesis Evaluation .....	425
19.5 Conclusions .....	426
References .....	426
<b>20 Rule-Based Speech Synthesis</b>	
<i>R. Carlson, B. Granström</i> .....	429
20.1 Background .....	429
20.2 Terminal Analog .....	429
20.3 Controlling the Synthesizer .....	432
20.4 Special Applications of Rule-Based Parametric Synthesis .....	434
20.5 Concluding Remarks .....	434
References .....	434
<b>21 Corpus-Based Speech Synthesis</b>	
<i>T. Dutoit</i> .....	437
21.1 Basics .....	437

21.2	Concatenative Synthesis with a Fixed Inventory .....	438
21.3	Unit-Selection-Based Synthesis .....	447
21.4	Statistical Parametric Synthesis .....	450
21.5	Conclusion .....	453
	<b>References</b> .....	453
<b>22</b>	<b>Linguistic Processing for Speech Synthesis</b>	
	<i>R. Sproat</i> .....	457
22.1	Why Linguistic Processing is Hard .....	457
22.2	Fundamentals: Writing Systems and the Graphical Representation of Language .....	457
22.3	Problems to be Solved and Methods to Solve Them .....	458
22.4	Architectures for Multilingual Linguistic Processing .....	465
22.5	Document-Level Processing .....	465
22.6	Future Prospects .....	466
	<b>References</b> .....	467
<b>23</b>	<b>Prosodic Processing</b>	
	<i>J. van Santen, T. Mishra, E. Klabbers</i> .....	471
23.1	Overview .....	471
23.2	Historical Overview .....	475
23.3	Fundamental Challenges .....	476
23.4	A Survey of Current Approaches .....	477
23.5	Future Approaches .....	484
23.6	Conclusions .....	485
	<b>References</b> .....	485
<b>24</b>	<b>Voice Transformation</b>	
	<i>Y. Stylianou</i> .....	489
24.1	Background .....	489
24.2	Source-Filter Theory and Harmonic Models .....	490
24.3	Definitions .....	492
24.4	Source Modifications .....	494
24.5	Filter Modifications .....	498
24.6	Conversion Functions .....	499
24.7	Voice Conversion .....	500
24.8	Quality Issues in Voice Transformations .....	501
24.9	Summary .....	502
	<b>References</b> .....	502
<b>25</b>	<b>Expressive/Affective Speech Synthesis</b>	
	<i>N. Campbell</i> .....	505
25.1	Overview .....	505
25.2	Characteristics of Affective Speech .....	506
25.3	The Communicative Functionality of Speech .....	508
25.4	Approaches to Synthesizing Expressive Speech .....	510
25.5	Modeling Human Speech .....	512

25.6	Conclusion .....	515
	<b>References</b> .....	515

## Part E Speech Recognition

<b>26</b>	<b>Historical Perspective of the Field of ASR/NLU</b>	
	<i>L. Rabiner, B.-H. Juang</i> .....	521
26.1	ASR Methodologies .....	521
26.2	Important Milestones in Speech Recognition History .....	523
26.3	Generation 1 – The Early History of Speech Recognition .....	524
26.4	Generation 2 – The First Working Systems for Speech Recognition .....	524
26.5	Generation 3 – The Pattern Recognition Approach to Speech Recognition .....	525
26.6	Generation 4 – The Era of the Statistical Model .....	530
26.7	Generation 5 – The Future .....	534
26.8	Summary .....	534
	<b>References</b> .....	535
<b>27</b>	<b>HMMs and Related Speech Recognition Technologies</b>	
	<i>S. Young</i> .....	539
27.1	Basic Framework .....	539
27.2	Architecture of an HMM-Based Recognizer .....	540
27.3	HMM-Based Acoustic Modeling .....	547
27.4	Normalization .....	550
27.5	Adaptation .....	551
27.6	Multipass Recognition Architectures .....	554
27.7	Conclusions .....	554
	<b>References</b> .....	555
<b>28</b>	<b>Speech Recognition with Weighted Finite-State Transducers</b>	
	<i>M. Mohri, F. Pereira, M. Riley</i> .....	559
28.1	Definitions .....	559
28.2	Overview .....	560
28.3	Algorithms .....	567
28.4	Applications to Speech Recognition .....	574
28.5	Conclusion .....	582
	<b>References</b> .....	582
<b>29</b>	<b>A Machine Learning Framework for Spoken-Dialog Classification</b>	
	<i>C. Cortes, P. Haffner, M. Mohri</i> .....	585
29.1	Motivation .....	585
29.2	Introduction to Kernel Methods .....	586
29.3	Rational Kernels .....	587
29.4	Algorithms .....	589
29.5	Experiments .....	591
29.6	Theoretical Results for Rational Kernels .....	593

29.7 Conclusion .....	594
<b>References</b> .....	595
<b>30 Towards Superhuman Speech Recognition</b>	
<i>M. Picheny, D. Nahamoo</i> .....	597
30.1 Current Status .....	597
30.2 A Multidomain Conversational Test Set .....	598
30.3 Listening Experiments .....	599
30.4 Recognition Experiments .....	601
30.5 Speculation .....	607
<b>References</b> .....	614
<b>31 Natural Language Understanding</b>	
<i>S. Roukos</i> .....	617
31.1 Overview of NLU Applications .....	618
31.2 Natural Language Parsing .....	620
31.3 Practical Implementation .....	623
31.4 Speech Mining .....	623
31.5 Conclusion .....	625
<b>References</b> .....	626
<b>32 Transcription and Distillation of Spontaneous Speech</b>	
<i>S. Furui, T. Kawahara</i> .....	627
32.1 Background .....	627
32.2 Overview of Research Activities on Spontaneous Speech .....	628
32.3 Analysis for Spontaneous Speech Recognition .....	632
32.4 Approaches to Spontaneous Speech Recognition .....	635
32.5 Metadata and Structure Extraction of Spontaneous Speech .....	640
32.6 Speech Summarization .....	644
32.7 Conclusions .....	647
<b>References</b> .....	647
<b>33 Environmental Robustness</b>	
<i>J. Droppo, A. Acero</i> .....	653
33.1 Noise Robust Speech Recognition .....	653
33.2 Model Retraining and Adaptation .....	656
33.3 Feature Transformation and Normalization .....	657
33.4 A Model of the Environment .....	664
33.5 Structured Model Adaptation .....	667
33.6 Structured Feature Enhancement .....	671
33.7 Unifying Model and Feature Techniques .....	675
33.8 Conclusion .....	677
<b>References</b> .....	677
<b>34 The Business of Speech Technologies</b>	
<i>J. Wilpon, M. E. Gilbert, J. Cohen</i> .....	681
34.1 Introduction .....	682

34.2 Network-Based Speech Services .....	686
34.3 Device-Based Speech Applications .....	692
34.4 Vision/Predications of Future Services – Fueling the Trends .....	697
34.5 Conclusion .....	701
<b>References</b> .....	702
<b>35 Spoken Dialogue Systems</b>	
<i>V. Zue, S. Seneff</i> .....	705
35.1 Technology Components and System Development .....	707
35.2 Development Issues .....	712
35.3 Historical Perspectives .....	714
35.4 New Directions .....	715
35.5 Concluding Remarks .....	718
<b>References</b> .....	718
<b>Part F Speaker Recognition</b>	
<b>36 Overview of Speaker Recognition</b>	
<i>A. E. Rosenberg, F. Bimbot, S. Parthasarathy</i> .....	725
36.1 Speaker Recognition .....	725
36.2 Measuring Speaker Features .....	729
36.3 Constructing Speaker Models .....	731
36.4 Adaptation .....	735
36.5 Decision and Performance .....	735
36.6 Selected Applications for Automatic Speaker Recognition .....	737
36.7 Summary .....	739
<b>References</b> .....	739
<b>37 Text-Dependent Speaker Recognition</b>	
<i>M. Hébert</i> .....	743
37.1 Brief Overview .....	743
37.2 Text-Dependent Challenges .....	747
37.3 Selected Results .....	750
37.4 Concluding Remarks .....	760
<b>References</b> .....	760
<b>38 Text-Independent Speaker Recognition</b>	
<i>D. A. Reynolds, W. M. Campbell</i> .....	763
38.1 Introduction .....	763
38.2 Likelihood Ratio Detector .....	764
38.3 Features .....	766
38.4 Classifiers .....	767
38.5 Performance Assessment .....	776
38.6 Summary .....	778
<b>References</b> .....	779

**Part G Language Recognition**

**39 Principles of Spoken Language Recognition**  
*C.-H. Lee* ..... 785  
 39.1 Spoken Language ..... 785  
 39.2 Language Recognition Principles ..... 786  
 39.3 Phone Recognition Followed by Language Modeling (PRLM) ..... 788  
 39.4 Vector-Space Characterization (VSC) ..... 789  
 39.5 Spoken Language Verification ..... 790  
 39.6 Discriminative Classifier Design ..... 791  
 39.7 Summary ..... 793  
**References** ..... 793

**40 Spoken Language Characterization**  
*M. P. Harper, M. Maxwell* ..... 797  
 40.1 Language versus Dialect ..... 798  
 40.2 Spoken Language Collections ..... 800  
 40.3 Spoken Language Characteristics ..... 800  
 40.4 Human Language Identification ..... 804  
 40.5 Text as a Source of Information on Spoken Languages ..... 806  
 40.6 Summary ..... 807  
**References** ..... 807

**41 Automatic Language Recognition Via Spectral and Token Based Approaches**  
*D. A. Reynolds, W. M. Campbell, W. Shen, E. Singer* ..... 811  
 41.1 Automatic Language Recognition ..... 811  
 41.2 Spectral Based Methods ..... 812  
 41.3 Token-Based Methods ..... 815  
 41.4 System Fusion ..... 818  
 41.5 Performance Assessment ..... 820  
 41.6 Summary ..... 823  
**References** ..... 823

**42 Vector-Based Spoken Language Classification**  
*H. Li, B. Ma, C.-H. Lee* ..... 825  
 42.1 Vector Space Characterization ..... 826  
 42.2 Unit Selection and Modeling ..... 827  
 42.3 Front-End: Voice Tokenization and Spoken Document Vectorization ..... 830  
 42.4 Back-End: Vector-Based Classifier Design ..... 831  
 42.5 Language Classification Experiments and Discussion ..... 835  
 42.6 Summary ..... 838  
**References** ..... 839

**Part H Speech Enhancement**

**43 Fundamentals of Noise Reduction**  
*J. Chen, J. Benesty, Y. Huang, E. J. Diethorn* ..... 843  
 43.1 Noise ..... 843  
 43.2 Signal Model and Problem Formulation ..... 845  
 43.3 Evaluation of Noise Reduction ..... 846  
 43.4 Noise Reduction via Filtering Techniques ..... 847  
 43.5 Noise Reduction via Spectral Restoration ..... 857  
 43.6 Speech-Model-Based Noise Reduction ..... 863  
 43.7 Summary ..... 868  
**References** ..... 869

**44 Spectral Enhancement Methods**  
*I. Cohen, S. Gannot* ..... 873  
 44.1 Spectral Enhancement ..... 874  
 44.2 Problem Formulation ..... 875  
 44.3 Statistical Models ..... 876  
 44.4 Signal Estimation ..... 879  
 44.5 Signal Presence Probability Estimation ..... 881  
 44.6 A Priori SNR Estimation ..... 882  
 44.7 Noise Spectrum Estimation ..... 888  
 44.8 Summary of a Spectral Enhancement Algorithm ..... 891  
 44.9 Selection of Spectral Enhancement Algorithms ..... 896  
 44.10 Conclusions ..... 898  
**References** ..... 899

**45 Adaptive Echo Cancellation for Voice Signals**  
*M. M. Sondhi* ..... 903  
 45.1 Network Echoes ..... 904  
 45.2 Single-Channel Acoustic Echo Cancellation ..... 915  
 45.3 Multichannel Acoustic Echo Cancellation ..... 921  
 45.4 Summary ..... 925  
**References** ..... 926

**46 Dereverberation**  
*Y. Huang, J. Benesty, J. Chen* ..... 929  
 46.1 Background and Overview ..... 929  
 46.2 Signal Model and Problem Formulation ..... 931  
 46.3 Source Model-Based Speech Dereverberation ..... 932  
 46.4 Separation of Speech and Reverberation via Homomorphic Transformation ..... 936  
 46.5 Channel Inversion and Equalization ..... 937  
 46.6 Summary ..... 941  
**References** ..... 942

**47 Adaptive Beamforming and Postfiltering**  
*S. Gannot, I. Cohen*..... 945

47.1 Problem Formulation..... 947

47.2 Adaptive Beamforming..... 948

47.3 Fixed Beamformer and Blocking Matrix..... 953

47.4 Identification of the Acoustical Transfer Function..... 955

47.5 Robustness and Distortion Weighting..... 960

47.6 Multichannel Postfiltering..... 962

47.7 Performance Analysis..... 967

47.8 Experimental Results..... 972

47.9 Summary..... 972

47.A Appendix: Derivation of the Expected Noise Reduction  
 for a Coherent Noise Field..... 973

47.B Appendix: Equivalence Between Maximum SNR  
 and LCMV Beamformers..... 974

**References**..... 975

**48 Feedback Control in Hearing Aids**  
*A. Spriet, S. Doclo, M. Moonen, J. Wouters*..... 979

48.1 Problem Statement..... 980

48.2 Standard Adaptive Feedback Canceller..... 982

48.3 Feedback Cancellation Based on Prior Knowledge  
 of the Acoustic Feedback Path..... 986

48.4 Feedback Cancellation Based on Closed-Loop System Identification..... 990

48.5 Comparison..... 995

48.6 Conclusions..... 997

**References**..... 997

**49 Active Noise Control**  
*S. M. Kuo, D. R. Morgan*..... 1001

49.1 Broadband Feedforward Active Noise Control..... 1002

49.2 Narrowband Feedforward Active Noise Control..... 1006

49.3 Feedback Active Noise Control..... 1010

49.4 Multichannel ANC..... 1011

49.5 Summary..... 1015

**References**..... 1015

**Part I Multichannel Speech Processing**

**50 Microphone Arrays**  
*G. W. Elko, J. Meyer*..... 1021

50.1 Microphone Array Beamforming..... 1021

50.2 Constant-Beamwidth Microphone Array System..... 1029

50.3 Constrained Optimization of the Directional Gain..... 1030

50.4 Differential Microphone Arrays..... 1031

50.5 Eigenbeamforming Arrays..... 1034

50.6 Adaptive Array Systems..... 1037

50.7 Conclusions..... 1040

**References**..... 1040

**51 Time Delay Estimation and Source Localization**  
*Y. Huang, J. Benesty, J. Chen*..... 1043

51.1 Technology Taxonomy..... 1043

51.2 Time Delay Estimation..... 1044

51.3 Source Localization..... 1054

51.4 Summary..... 1061

**References**..... 1062

**52 Convolutional Blind Source Separation Methods**  
*M. S. Pedersen, J. Larsen, U. Kjems, L. C. Parra*..... 1065

52.1 The Mixing Model..... 1066

52.2 The Separation Model..... 1068

52.3 Identification..... 1071

52.4 Separation Principle..... 1071

52.5 Time Versus Frequency Domain..... 1076

52.6 The Permutation Ambiguity..... 1078

52.7 Results..... 1084

52.8 Conclusion..... 1084

**References**..... 1084

**53 Sound Field Reproduction**  
*R. Rabenstein, S. Spors*..... 1095

53.1 Sound Field Synthesis..... 1095

53.2 Mathematical Representation of Sound Fields..... 1096

53.3 Stereophony..... 1100

53.4 Vector-Based Amplitude Panning..... 1103

53.5 Ambisonics..... 1104

53.6 Wave Field Synthesis..... 1109

**References**..... 1113

**Acknowledgements**..... 1115

**About the Authors**..... 1117

**Detailed Contents**..... 1133

**Subject Index**..... 1161

## 50. Microphone Arrays

G. W. Elko, J. Meyer

This chapter introduces various types of microphone array beamforming systems and discusses some of the fundamental theory of their operation, design, implementation, and limitations. It is shown that microphone arrays have the ability to offer directional gains that can significantly improve the quality of signal pickup in reverberant and noisy environments.

Hands-free audio communication is now a major feature in mobile communication systems as well as audio and video conferencing systems. One problem that becomes evident to users of these systems is the decrease in communication quality due to the pickup of room reverberation and background noise. In the past, this problem was dealt with by using microphones placed close to the desired talker or source. Although this simple solution has proven to be quite effective, it also has its drawbacks. First, it is not always possible or desirable to place the microphone very close to the talker's mouth. Second, by placing the microphone close to the talker's mouth, one has to deal with rapid level variation as the talker moves his or her mouth relative to the microphone. Third is the negative impact of speech plosives (air-flow transients generated by plosive sounds) and forth, microphone structure-borne handling noise has a detrimental effect. Finally, for directional microphone elements, there is also a nearfield

50.1	<b>Microphone Array Beamforming</b> .....	1021
50.1.1	Delay-and-Sum Beamforming.....	1023
50.1.2	Filter-and-Sum Beamforming.....	1028
50.1.3	Arrays with Directional Elements.....	1028
50.2	<b>Constant-Beamwidth Microphone Array System</b> .....	1029
50.3	<b>Constrained Optimization of the Directional Gain</b> .....	1030
50.4	<b>Differential Microphone Arrays</b> .....	1031
50.5	<b>Eigenbeamforming Arrays</b> .....	1034
50.5.1	Spherical Array.....	1034
50.5.2	Eigenbeamformer.....	1035
50.5.3	Modal Beamformer.....	1037
50.6	<b>Adaptive Array Systems</b> .....	1037
50.6.1	Constrained Broadband Arrays.....	1038
50.7	<b>Conclusions</b> .....	1040
	<b>References</b> .....	1040

proximity effect (where the frequency response of the microphone is modulated by the relative position of the microphone to the mouth). With these issues in mind, it is of interest to investigate other potential solutions. One solution is to use beamforming microphone arrays, which can offer significant directional gain so as to result in similar audio performance to that of closely placed microphones.

### 50.1 Microphone Array Beamforming

The propagation of acoustic waves in space is a function of both space and time coordinates. In general, the mathematical representation of a propagating acoustic wave is a four-dimensional function: three spatial dimensions and one time variable. Acoustic wave propagation can be modeled by a linearized scalar acoustic wave equation that describes the relationships between the physical quantities of acoustic pressure and density variation of the medium [50.1],

$$\nabla^2 p = \frac{1}{c^2} \frac{\partial^2 p}{\partial t^2}, \quad (50.1)$$

where  $p$  is the instantaneous acoustic pressure fluctuation of the sound and  $c$  is the propagation speed of sound in the medium. There are a few major underlying assumptions in the derivation of (50.1), but this simplified linear model is adequate for laying the foundation for array beamforming of acoustic signals. The dimen-

sionality of the wave equation can easily be seen from (50.1), where  $p$  is a function of three space variables and one time variable. Therefore the scalar acoustic pressure field in space can be represented as  $p(\mathbf{r}, t)$ , where  $\mathbf{r}$  is the measurement position in the acoustic field and  $t$  is the time dependence. The scalar acoustic pressure field must satisfy (50.1) at all points in space. By using the spatial Fourier transform, the acoustic field can also equivalently be represented in the wavevector–frequency domain. This representation has been widely used in the fields of geophysics, structural, underwater and aeroacoustics, and can be quite useful in the analysis and design of microphone array beamformers.

Applying the four-dimensional Fourier transform to the time–space scalar acoustic field yields,

$$P(\mathbf{k}, \omega) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(\mathbf{r}, t) e^{-i(\omega t - \mathbf{k}^T \mathbf{r})} d\mathbf{r} dt, \quad (50.2)$$

where the acoustic pressure  $P$  is capitalized to indicate a transform to the frequency domain,  $\mathbf{k}$  is the wavevector and the superscript ‘T’ represents the transpose operator. It may seem that this transformation has unnecessarily complicated the description of the scalar acoustic pressure field, but as will be seen later, it has transformed the field description into a form that makes the analysis of acoustic field space–time functions analogous to the field of multidimensional signal processing. Since the Fourier transform is a linear transformation, one can write an inverse relationship,

$$p(\mathbf{r}, t) = \frac{1}{(2\pi)^4} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P(\mathbf{k}, \omega) e^{i(\omega t - \mathbf{k}^T \mathbf{r})} d\mathbf{k} d\omega, \quad (50.3)$$

Equation (50.3) directly shows that any acoustic field  $p(\mathbf{r}, t)$  can be represented as an infinite number of propagating plane waves with appropriate complex weighting. This interpretation can easily be seen from the equation for a single propagating plane wave with frequency  $\omega_0$  and wavevector  $\mathbf{k}_0$ ,

$$p(\mathbf{r}, t) = A_0 e^{-i(\omega_0 t - \mathbf{k}_0^T \mathbf{r})}, \quad (50.4)$$

where  $A_0$  is the plane-wave amplitude. Note that for generality, one can allow  $A_0$  to be complex, but for this discussion it is assumed that  $A_0$  is real. The wavenumber–frequency spectrum of this single propagating plane wave is simply,

$$P(\mathbf{k}, \omega) = A_0 \delta(\mathbf{k} - \mathbf{k}_0) \delta(\omega - \omega_0). \quad (50.5)$$

Equation (50.5) shows the Fourier mapping of infinite continuous functions in both time and space to be a point in the wavevector–frequency space.

Now consider a general weighting function of the space–time pressure distribution with a four-dimensional linear shift-invariant filter having an impulse response  $h(\mathbf{r}, t)$ . Then, the output signal is the convolution of the space–time acoustic pressure signal and the spatial weighting function such that,

$$y(\mathbf{r}, t) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(\mathbf{r} - \hat{\mathbf{r}}, t - \tau) p(\hat{\mathbf{r}}, \tau) d\hat{\mathbf{r}} d\tau. \quad (50.6)$$

Equation (50.6) is a filtering operation that can equivalently be represented in the wavevector–frequency domain as,

$$Y(\mathbf{k}, \omega) = H(\mathbf{k}, \omega) P(\mathbf{k}, \omega). \quad (50.7)$$

Equation (50.6) and (50.7) show the direct analogy between the space–time wavevector–frequency representation of spatial filtering and the well-known results from multidimensional linear systems theory. Thus, by using specific spatial weighting functions, one can filter the acoustic field to investigate the propagating directions, amplitude, and phases of plane waves traveling in space. Filtering the spatiotemporal acoustic scalar pressure field is known in the field of array signal processing as *beamforming*. Thus, if one wants to *look* at plane waves propagating from the direction of the unit vector  $\mathbf{k}_0 / \|\mathbf{k}_0\|$ , one would design a filter such that,

$$H(\mathbf{k}, \omega) = \delta(\mathbf{k} - \mathbf{k}_0) G(\omega), \quad (50.8)$$

where  $G(\omega)$  is the desired frequency response to waves propagating in the direction of the wavevector  $\mathbf{k}_0$ .

Although it may seem limiting to use plane waves as the underlying basis functions for the representation of a sound-field, it should be noted that other orthogonal representations such as spherical and cylindrical basis functions can themselves be represented as a series of plane waves.

The representation of the wavevector–frequency domain is a natural framework for analysis of the spatial and frequency filtering of beamforming arrays, as will be seen in the following.

In spatial filtering by beamforming, one deals with spatial apertures. An aperture is a region over which energy is received. Where an aperture can either be continuous, as in parabolic dishes, or discretely realized as in microphone arrays with multiple microphones, mixtures of continuous and discrete apertures are possible.

In fact, even discrete microphone arrays are a mixture since no microphone is a perfect point receiver. Although sampled aperture systems must generally perform more signal processing, they offer several advantages over continuous aperture systems. A main advantage in using sampled apertures is that they allow for the possibility of using digital signal processing on the individual array signals. Arrays can be electronically steered, can form multiple simultaneous beams, and can be made adaptive purely by electronic means.

### 50.1.1 Delay-and-Sum Beamforming

Delay-and-sum beamforming, also known as *classical beamforming*, is one of the simplest and oldest techniques for realizing directional array systems. Although not fundamentally limited in bandwidth, early delay-and-sum arrays were used in *narrowband* operation to focus arrays onto a particular point or direction. Since a time delay for narrowband applications can be accomplished with a unique phase shift for each element, narrowband beamforming is typically implemented with phase shifts and is therefore commonly referred to as *phased-array* beamforming. One typical application for microphone arrays is to pick up speech or acoustic signals that are wideband (the desired signals cover many octaves in bandwidth). Therefore most microphone arrays for speech acquisition are implemented as delay-sum beamformers since they are inherently wideband.

Both continuous and discrete beamformers can be seen as implementations of wavevector–frequency filtering as discussed in the previous section. It should be noted that these two beamformers are not mutually exclusive; one can, for instance, have a discrete set of continuous transducers so that the beamformer is a mix of continuous and discrete sampling. Of special interest here is the spatial and frequency response of a finite array of microphones that sample the acoustic field. A classic delay-sum beamformer uses the sum of weighted and delayed samples from an array of  $N$  discretely sampled points of the pressure field. In general, a delay-sum beamformer output  $y$  is formed as,

$$y(t) = \sum_{n=1}^N w_n s(\mathbf{r}_n, t - \tau_n), \quad (50.9)$$

where the weights  $w_n$  are real and the time delay  $\tau_n$  is applied to the measured signals  $s(\mathbf{r}_n, t)$ , at microphone  $n$ . If the incident field is a planewave with wavevector  $\mathbf{k}_0$ , amplitude  $A_0$  and frequency  $\omega_0$ , then,

$$s(\mathbf{r}_n, t) = A_0 e^{i(\omega_0 t - \mathbf{k}_0^T \mathbf{r}_n)}, \quad (50.10)$$

and

$$y(t) = A_0 e^{i\omega_0 t} \sum_{n=1}^N w_n e^{-i(\omega_0 \tau_n + \mathbf{k}_0^T \mathbf{r}_n)}. \quad (50.11)$$

The real weights  $w_n$  scale the signals measured at the positions  $\mathbf{r}_n$ . Thus, from (50.9), the origin of the term *delay-and-sum* can be seen as simply a description of how the classical delay-sum beamformer is realized.

For a causal delay-sum beamformer, the exponent in (50.11) must be less than or equal to zero for all  $n$ . Typically, one element position is selected as the spatial origin. This reference element position therefore defines the vectors  $\mathbf{r}_n$ . Since the array can be steered to any direction in space and one can select any position to define  $\mathbf{r}_n$ , an additional delay may be required in order to maintain causality. (Recall that  $\mathbf{k}_0^T \mathbf{r}_n$  can be either positive or negative depending on the direction of the incident wave relative to the array.) If an end element is selected as the spatial origin, then the causal delay is equal to the maximum time that an acoustic wave takes to transit the array. Smaller causal delays are also possible depending on which microphone position in the array is chosen as the spatial origin reference and the maximum desired steering angle.

If we set the causal time delay equal to  $T_0$ , then the output for an incident plane wave with angular frequency  $\omega_0$  and wavevector  $\mathbf{k}_0$  can be written as,

$$y(t) = A_0 e^{i\omega_0 t} \sum_{n=1}^N w_n e^{-i[\omega_0(T_0 + \tau_n) + \mathbf{k}_0^T \mathbf{r}_n]}. \quad (50.12)$$

The maximum plane-wave response corresponds to the direction where the delays  $\tau_n$  compensate for the propagation delay  $\mathbf{k}_0^T \mathbf{r}_n$  of a plane wave propagating with wavevector  $\mathbf{k}_0$ . This is done by setting the values of  $\tau_n$  such that these delays offset the time lead (or lag) of a desired direction plane wave propagating over the array. Thus, setting the delays to steer the array to the direction of  $\mathbf{k}_0$ , means that the last two terms in the exponential in (50.12) cancel out. The delayed and summed output  $y(t)$  is now such that all microphone signals propagating from the desired direction are added in phase. This direction corresponds to a maximum amplitude output for any selection of  $\tau_n$ . Plane waves propagating from directions other than  $\mathbf{k}_0$  will result in the addition of position-dependent phase variations, and as a result, the output amplitude will be smaller due to destructive interference.

To summarize, classical delay-and-sum beamforming uses delays between each array element that compensate for differences in the propagation delay of

array beam steering [51.1–4] and a good example of the latter is automatic camera pointing for videoconferencing [51.5–8].

The problem of locating radiative point sources using sensor arrays has long been of great research interest given its theoretical as well as practical importance in a great variety of applications, e.g., radar [51.9, 10], underwater sonar [51.11], and seismology [51.12]. In these applications, source localization is more commonly referred to as direction of arrival (DOA) estimation. A class of celebrated approaches is based on high-resolution spectral analysis, including Capon's minimum variance (MV) spectral estimation method [51.13], and eigenanalysis-based techniques, such as the popular multiple signal classification (MUSIC) algorithm [51.14]. These methods perform statistical fit for DOA with respect to a spatio-spectral correlation matrix derived from the signals received at the sensors. By assumption, the source signal needs to be statistically stationary and narrowband, and the source is located in the far field of the sensor array. However, these basic premises are barely met by speech sources. Moreover, the multipath effect is not taken into account in the formulation of DOA estimation. Therefore, the high-resolution DOA estimation algorithms were found to be incompetent for speech source localization, particularly in reverberant acoustic environments.

For speech source localization, microphone array beam scanning and time delay estimation (TDE)-based localization methods are the two most widely used approaches. A beamformer is a spatial filter that operates on the microphone outputs in order to enhance the signal coming from one direction while suppressing noise and interference from other directions. Therefore, steering a microphone array beamformer and scanning across a room for the highest-energy output leads to an estimate of the direction of an active speech source. Using two arrays and by intersecting the two corresponding direction estimates, the source location

is determined [51.15, 16]. While the beam scanning method has the remarkable advantage that it can easily be extended to the case of simultaneously localizing multiple speech sources [51.17], it has some intrinsic drawbacks. Since speech is a typical broadband signal, ideally a broadband beamformer whose beam pattern would be the same across the entire speech spectrum needs to be developed, which is a technically challenging problem. In addition, most beamforming algorithms assume that the speech sound source is in the far field in order to make the design problem analytically tractable. When the speech source is close to the microphone array, performance degradation can be expected. On the contrary, if the speech source is in the far field, the localization resolution may not be satisfactory as the source direction is only selected from a set of discrete beam scanning angles. The number of beam scanning angles is therefore a trade-off between resolution and computational complexity.

Alternatively, a TDE-based source localization algorithm involves a two-step procedure [51.18]. In the first step, a set of relative time differences of arrival (TDOAs) for different microphone pairs are calculated. In the second step, the acoustic source location is estimated from the TDOAs and using the a priori knowledge about the locations of the microphones. TDE-based source localization algorithms have many merits: they cope well with both narrowband and broadband source signals; their localization resolution can be flexibly adjusted by varying the sampling rate and microphone array size; the effect of room reverberation is only of concern in the first-step TDE and several recently developed robust TDE algorithms are promising for future practical use; and in general they are computationally efficient. Therefore, after continuous investigation over the last two decades, TDE-based speech source localization schemes have become the technique of choice, especially in recent digital systems. In this chapter, an overview of the state of the art of this class of source localization methods is presented.

Depending on the surrounding acoustic environment, there are two signal models for the TDE problem: the ideal free-field model and the real reverberant model. The former assumes no room reverberation while the latter uses an acoustic channel impulse response, usually a finite impulse response (FIR) filter, to describe the effect of room reverberation.

## 51.2 Time Delay Estimation

### 51.2.1 Problem Formulation and Signal Models

Suppose that there is one speech sound source and  $N$  microphones. The TDE problem is concerned with the computation of the relative time difference of arrival (TDOA) between different microphone signals.

#### Ideal Free-Field Model

In an anechoic open space, as shown in Fig. 51.1a, the speech source signal  $s(k)$  propagates radiatively and the sound level falls off as a function of distance from the source. Then the signal captured by the  $n$ -th microphone at time  $k$  can be expressed as follows:

$$x_n(k) = \alpha_n s(k - \tau_n) + b_n(k), \quad n = 1, 2, \dots, N, \quad (51.1)$$

where  $\alpha_n$  ( $0 \leq \alpha_n \leq 1$ ) is an attenuation factor due to propagation loss,  $\tau_n$  is the propagation time, and  $b_n(k)$  is additive noise. The time difference of arrival (TDOA) between the  $i$ -th and  $j$ -th microphones is defined as,

$$\tau_{ij} \triangleq \tau_i - \tau_j, \quad i, j = 1, 2, \dots, N. \quad (51.2)$$

The noise signal  $b_n(k)$  is presumed to be a zero-mean, white, Gaussian random process, and is uncorrelated with the source signal as well as the noise signals at other microphones.

#### Real Reverberant Model

While the ideal free-field model has the merit of being simple, it does not take into account room reverberation. Therefore, in a real reverberant environment, the ideal free-field model is inadequate and problematic, and we need a more-comprehensive and more-informative alternative to describe the effect of multipath propagation. The real reverberant model treats the acoustic impulse response with an FIR filter, as illustrated by Fig. 51.1b. In such a single-input multiple-output (SIMO) system, the  $n$ -th microphone signal is given by

$$x_n(k) = h_n * s(k) + b_n(k), \quad n = 1, 2, \dots, N, \quad (51.3)$$

where  $h_n$  is the  $n$ -th channel impulse response, and the symbol  $*$  denotes the linear convolution operator. In vector/matrix form, (51.3) can be rewritten as

$$\mathbf{x}_n(k) = \mathbf{H}_n \cdot \mathbf{s}(k) + \mathbf{b}_n(k), \quad n = 1, 2, \dots, N, \quad (51.4)$$

where

$$\mathbf{x}_n(k) = [x_n(k) \ \dots \ x_n(k-L+1)]^T,$$

$$\mathbf{H}_n = \begin{pmatrix} h_{n,0} & \dots & h_{n,L-1} & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & h_{n,0} & \dots & h_{n,L-1} \end{pmatrix},$$

$$\mathbf{s}(k) = [s(k) \quad s(k-1) \quad \dots \quad s(k-L+1) \\ \dots \quad s(k-2L+2)]^T,$$

$$\mathbf{b}_n(k) = [b_n(k) \ \dots \ b_n(k-L+1)]^T,$$

where  $[\cdot]^T$  denotes the transpose of a vector or a matrix, and  $L$  is the length of the longest channel impulse response in this SIMO system.

In contrast to the ideal free-field model, the time delay  $\tau_n$  in the real reverberant model is an implicit or hidden parameter. Using such a model, TDE can be performed only after the SIMO system is *blindly* identified (since the source signal is unknown), which looks like a more-difficult problem but is fortunately not insurmountable.

### 51.2.2 The Family of the Generalized Cross-Correlation Methods

The generalized cross-correlation (GCC) algorithm proposed by Knapp and Carter [51.19] is so far the most widely used approach to TDE. It employs the ideal free-field model (51.1) and considers only two microphones,

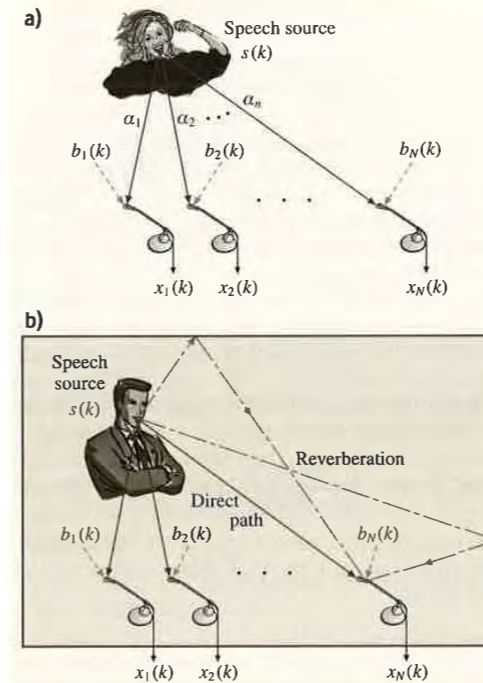


Fig. 51.1a,b Illustration of the two acoustic signal models for time delay estimation: (a) ideal free-field model and (b) real reverberant model