Springer Handbook of Speech Processing

Benesty
Sondhi
Huang
Editors

Springer

Springer Handbook of Speech Processing

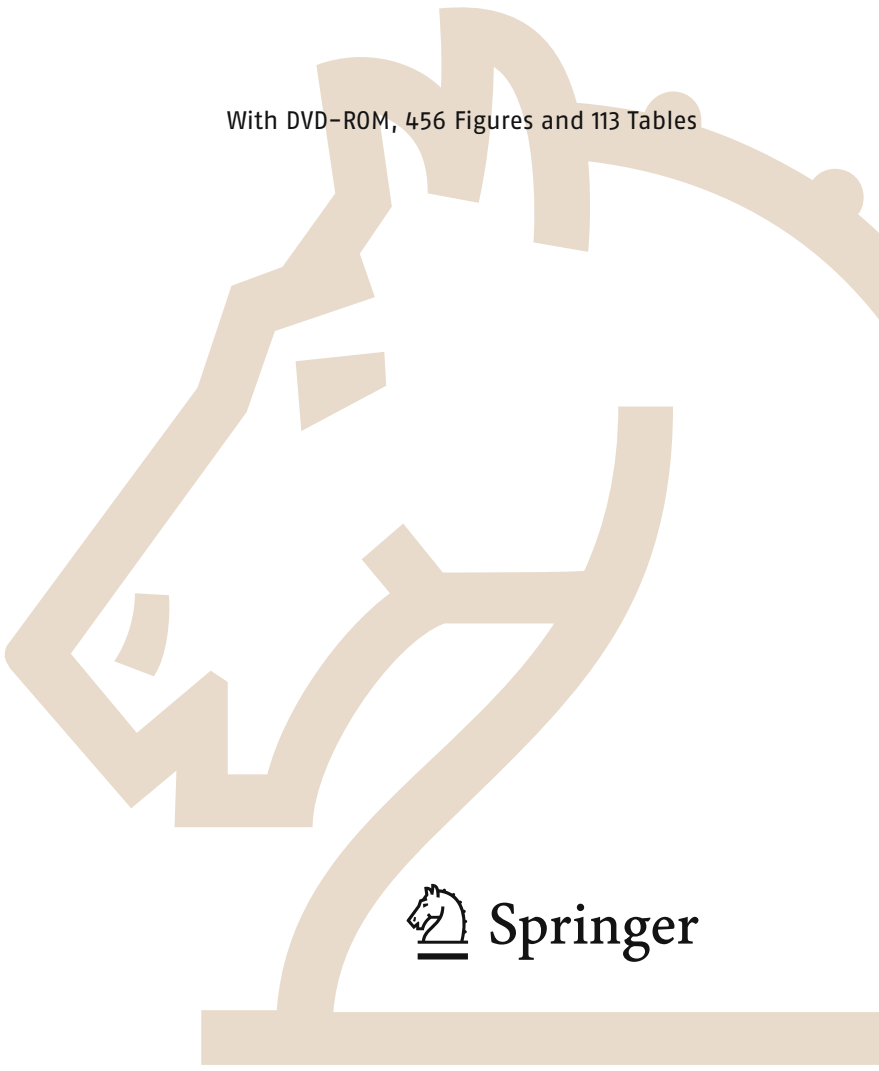
Springer Handbooks provide a concise compilation of approved key information on methods of research, general principles, and functional relationships in physical sciences and engineering. The world's leading experts in the fields of physics and engineering will be assigned by one or several renowned 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@ieee.org

Library of Congress Control Number: 2007931999

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

Springer is a part of Springer Science+Business Media

springer.com

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

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

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

Preface

Jacob Benesty

M. Mohan Sondhi

Yiteng Huang

The achievement of this Springer Handbook is the result of a wonderful journey that started in March 2005 at the 30th International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Two of the editors-in-chief (Benesty and Huang) met in one of the long corridors of the Pennsylvania Convention Center in Philadelphia with Dr Dieter Merkle from Springer. Together we had a very nice discussion about the conference and immediately an idea came up for a handbook. After a short discussion we converged without too much hesitation on a handbook of *speech processing*. It was quite surprising to see that, even after 30 years of ICASSP and more than half a century of research in this fundamental area, there was still no major book summarizing the important aspects of speech processing. We thought that the time was ripe for such a large project. Soon after we got home, a third editor-in-chief (Sondhi) joined the efforts.

We had a very clear objective in our minds: to summarize, in a reasonable number of pages, the most important and useful aspects of speech processing. The content was then organized accordingly. This task was not easy since we had to find a good balance between feasible ideas and new trends. As we all know, practical ideas can be viewed as *old stuff* while emerging ideas can be criticized for not having passed the test of time; we hope that we have succeeded in finding a good compromise. For this we relied on many authors who are well established and are recognized as experts in their field, from all over the world, and from academia as well as from industry.

From *simple* consumer products such as cell phones and MP3 players to more-sophisticated projects such as human–machine interfaces and robots that can obey orders, speech technologies are now everywhere. We believe that it is just a matter of time before more applications of the science of speech become impossible to miss in our daily life. So we believe that this Springer Handbook will play a fundamental role in the sustainable progress of speech research and development.

This handbook is targeted at three categories of readers: graduate students of speech processing, professors and researchers in academia and research labs who are active in this field, and engineers in industry who need to understand or implement specific algorithms for their speech-related products. The handbook could also be used as a text for one or more graduate courses on signal processing for speech and various aspects of speech processing and applications.

For the completion of such an ambitious project we have many people to thank. First, we would like to thank the many authors who did a terrific job in delivering very high-quality chapters. Second, we are very grateful to the members of the editorial board who helped us so much in organizing the content and structure of this book, taking part in all phases of this project from conception to completion. Third, we would like to thank all the reviewers, who helped us to improve the quality of the material. Last, but not least, we would like to thank the Springer team for their availability and very professional work. In particular, we appreciated the help of Dieter Merkle, Christoph Baumann, Werner Skolaut, Petra Jantzen, and Claudia Rau.

We hope this Springer Handbook will inspire many great minds to find new research ideas or to implement algorithms in products.

Montreal, Basking Ridge, Murray Hill
October 2007

Jacob Benesty
M. Mohan Sondhi
Yiteng Huang

List of Editors

Editors-in-Chief

Jacob Benesty, Montreal
M. Mohan Sondhi, Basking Ridge
Yiteng (Arden) Huang, Murray Hill

Part Editors

Part A: Production, Perception, and Modeling of Speech

M. M. Sondhi, Basking Ridge

Part B: Signal Processing for Speech

Y. Huang, Murray Hill; J. Benesty, Montreal

Part C: Speech Coding

W. B. Kleijn, Stockholm

Part D: Text-to-Speech Synthesis

S. Narayanan, Los Angeles

Part E: Speech Recognition

L. Rabiner, Piscataway; B.-H. Juang, Atlanta

Part F: Speaker Recognition

S. Parthasarathy, Sunnyvale

Part G: Language Recognition

C.-H. Lee, Atlanta

Part H: Speech Enhancement

J. Chen, Murray Hill; S. Gannot, Ramat-Gan; J. Benesty, Montreal

Part I: Multichannel Speech Processing

J. Benesty, Montreal; I. Cohen, Haifa; Y. Huang, Murray Hill

List of Authors

Alex Acero
Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA
e-mail: *alexac@microsoft.com*

Jont B. Allen
University of Illinois
ECE
Urbana, IL 61801, USA
e-mail: *JontAllen@ieee.org*

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

Frédéric Bimbot
IRISA (CNRS & INRIA) - METISS
Pièce C 320 - Campus Universitaire de Beaulieu
35042 Rennes, France
e-mail: *bimbot@irisa.fr*

Thomas Brand
Carl von Ossietzky Universität Oldenburg
Sektion Medizinphysik
Haus des Hörens, Marie-Curie-Str. 2
26121 Oldenburg, Germany
e-mail: *thomas.brand@uni-oldenburg.de*

Nick Campbell
Knowledge Creating Communication Research Centre
Acoustics & Speech Research Project, Spoken Language Communication Group
2-2-2 Hikaridai
619-0288 Keihanna Science City, Japan
e-mail: *nick@nict.go.jp*

William M. Campbell
MIT Lincoln Laboratory
Information Systems Technology Group
244 Wood Street
Lexington, MA 02420-9108, USA
e-mail: *wcampbell@ll.mit.edu*

Rolf Carlson
Royal Institute of Technology (KTH)
Department of Speech, Music and Hearing
Lindstedtsvägen 24
10044 Stockholm, Sweden
e-mail: *rolf@speech.kth.se*

Jingdong Chen
Bell Laboratories
Alcatel-Lucent
600 Mountain Ave
Murray Hill, NJ 07974, USA
e-mail: *jingdong@research.bell-labs.com*

Juin-Hwey Chen
Broadcom Corp.
5300 California Avenue
Irvine, CA 92617, USA
e-mail: *rchen@broadcom.com*

Israel Cohen
Technion–Israel Institute of Technology
Department of Electrical Engineering
Technion City
Haifa 32000, Israel
e-mail: *icohen@ee.technion.ac.il*

Jordan Cohen
SRI International
300 Ravenswood Drive
Menlo Park, CA 94019, USA
e-mail: *jrc@speech.sri.com*

Corinna Cortes
Google, Inc.
Google Research
76 9th Avenue, 4th Floor
New York, NY 10011, USA
e-mail: *corinna@google.com*

Eric J. Diethorn
Avaya Labs Research
Multimedia Technologies Research Department
233 Mt. Airy Road
Basking Ridge, NJ 07920, USA
e-mail: *ejd@avaya.com*

Simon Doclo
Katholieke Universiteit Leuven
Department of Electrical Engineering (ESAT-SCD)
Kasteelpark Arenberg 10 bus 2446
3001 Leuven, Belgium
e-mail: *simon.doclo@esat.kuleuven.be*

Jasha Droppo
Microsoft Research
Speech Technology Group
One Microsoft Way
Redmond, WA 98052, USA
e-mail: *jdroppo@microsoft.com*

Thierry Dutoit
Faculté Polytechnique de Mons FPMs
TCTS Laboratory
Bvd Dolez, 31
7000 Mons, Belgium
e-mail: *thierry.dutoit@fpms.ac.be*

Gary W. Elko
mh acoustics LLC
25A Summit Ave
Summit, NJ 07901, USA
e-mail: *gwe@mhacoustics.com*

Sadaoki Furui
Tokyo Institute of Technology Street
Department of Computer Science
2-12-1 Ookayama, Meguro-ku
152-8552 Tokyo, Japan
e-mail: *furui@cs.titech.ac.jp*

Sharon Gannot
Bar-Ilan University
School of Electrical Engineering
Ramat-Gan 52900, Israel
e-mail: *gannot@eng.biu.ac.il*

Mazin E. Gilbert
AT&T Labs, Inc., Research
180 Park Ave.
Florham Park, NJ 07932, USA
e-mail: *mazin@research.att.com*

Michael M. Goodwin
Creative Advanced Technology Center
Audio Research
1500 Green Hills Road
Scotts Valley, CA 95066, USA
e-mail: *mgoodwin@atc.creative.com*

Volodya Grancharov
Multimedia Technologies
Ericsson Research, Ericsson AB
Torshamnsgatan 23, Kista, KI/EAB/TVA/A
16480 Stockholm, Sweden
e-mail: *volodya.grancharov@ericsson.com*

Björn Granström
Royal Institute of Technology (KTH)
Department for Speech, Music and Hearing
Lindstedsvägen 24
10044 Stockholm, Sweden
e-mail: *bjorn@speech.kth.se*

Patrick Haffner
AT&T Labs-Research
IP and Voice Services
200 S Laurel Ave.
Middletown, NJ 07748, USA
e-mail: *haffner@research.att.com*

Roar Hagen
Global IP Solutions
Magnus Ladulsgatan 63B
118 27 Stockholm, Sweden
e-mail: *roar.hagen@gipscorp.com*

Mary P. Harper
University of Maryland
Center for Advanced Study of Language
7005 52nd Avenue
College Park, MD 20742, USA
e-mail: *mharper@casl.umd.edu*

Jürgen Herre
Fraunhofer Institute for Integrated Circuits
(Fraunhofer IIS)
Audio and Multimedia
Am Wolfsmantel 33
91058 Erlangen, Germany
e-mail: *hrr@iis.fraunhofer.de*

Wolfgang J. Hess
University of Bonn
Institute for Communication Sciences, Dept. of Communication, Language, and Speech
Poppelsdorfer Allee 47
53115 Bonn, Germany
e-mail: *wgh@ifk.uni-bonn.de*

Kiyoshi Honda
Université de la Sorbonne Nouvelle-Paris III
Laboratoire de Phonétique et de Phonologie, ATR Cognitive Information Laboratories
UMR-7018-CNRS, 46, rue Barrault
75634 Paris, France
e-mail: *honda@atr.jp*

Yiteng (Arden) Huang
Bell Laboratories
Alcatel-Lucent
600 Mountain Avenue
Murray Hill, NJ 07974, USA
e-mail: *arden_huang@ieee.org*

Matthieu Hébert
Network ASR Core Technology
Nuance Communications
1500 Université
Montréal, Québec H3A-3S7, Canada
e-mail: *hebert@nuance.com*

Biing-Hwang Juang
Georgia Institute of Technology
School of Electrical & Computer Engineering
777 Atlantic Dr. NW
Atlanta, GA 30332-0250, USA
e-mail: *juang@ece.gatech.edu*

Tatsuya Kawahara
Kyoto University
Academic Center for Computing and Media Studies
Sakyo-ku
606-8501 Kyoto, Japan
e-mail: *kawahara@i.kyoto-u.ac.jp*

Ulrik Kjems
Oticon A/S
9 Kongebakken
2765 Smørum, Denmark
e-mail: *uk@oticon.dk*

Esther Klabbers
Oregon Health & Science University
Center for Spoken Language Understanding, OGI School of Science and Engineering
20000 NW Walker Rd
Beaverton, OR 97006, USA
e-mail: *klabbers@cslu.ogi.edu*

W. Bastiaan Kleijn
Royal Institute of Technology (KTH)
School of Electrical Engineering, Sound and Image Processing Lab
Osquldas väg 10
10044 Stockholm, Sweden
e-mail: *bastiaan.kleijn@ee.kth.se*

Birger Kollmeier
Universität Oldenburg
Medizinische Physik
26111 Oldenburg, Germany
e-mail: *birger.kollmeier@uni-oldenburg.de*

Ermin Kozica
Royal Institute of Technology (KTH)
School of Electrical Engineering, Sound and Image Processing Laboratory
Osquldas väg 10
10044 Stockholm, Sweden
e-mail: *ermin.kozica@ee.kth.se*

Sen M. Kuo
Northern Illinois University
Department of Electrical Engineering
DeKalb, IL 60115, USA
e-mail: *kuo@ceet.niu.edu*

Jan Larsen
Technical University of Denmark
Informatics and Mathematical Modelling
Richard Petersens Plads
2800 Kongens Lyngby, Denmark
e-mail: *jl@imm.dtu.dk*

Chin-Hui Lee
Georgia Institute of Technology
School of Electrical and Computer Engineering
777 Atlantic Drive NW
Atlanta, GA 30332-0250, USA
e-mail: *chl@ece.gatech.edu*

Haizhou Li
Institute for Infocomm Research
Department of Human Language Technology
21 Heng Mui Keng Terrace
Singapore, 119613
e-mail: *hli@i2r.a-star.edu.sg*

Jan Linden
Global IP Solutions
301 Brannan Street
San Francisco, CA 94107, USA
e-mail: *jan.linden@gipscorp.com*

Manfred Lutzky
Fraunhofer Integrated Circuits (IIS)
Multimedia Realtime Systems
Am Wolfsmantel 33
91058 Erlangen, Germany
e-mail: *manfred.lutzky@iis.fraunhofer.de*

Bin Ma
Human Language Technology
Institute for Infocomm Research
21 Heng Mui Keng Terrace
Singapore, 119613
e-mail: *mabin@i2r.a-star.edu.sg*

Michael Maxwell
University of Maryland
Center for Advanced Study of Language
Box 25
College Park, MD 20742, USA
e-mail: *mmaxwell@casl.umd.edu*

Alan V. McCree
MIT Lincoln Laboratory
Department of Information Systems Technology
244 Wood Street
Lexington, MA 02420-9185, USA
e-mail: *mccree@ll.mit.edu*

Bernd Meyer
Carl von Ossietzky Universität Oldenburg
Medical Physics Section, Haus des Hörens
Marie-Curie-Str. 2
26121 Oldenburg, Germany
e-mail: *bernd.meyer@uni-oldenburg.de*

Jens Meyer
mh acoustcis
25A Summit Ave.
Summit, NJ 07901, USA
e-mail: *jmm@mhacoustics.com*

Taniya Mishra
Oregon Health and Science University
Center for Spoken Language Understanding,
Computer Science and Electrical Engineering, OGI
School of Science and Engineering
20000 NW Walker Road
Beaverton, OR 97006, USA
e-mail: *mishra@cslu.ogi.edu*

Mehryar Mohri
Courant Institute of Mathematical Sciences
251 Mercer Street
New York, NY 10012, USA
e-mail: *mohri@cs.nyu.edu*

Marc Moonen
Katholieke Universiteit Leuven
Electrical Engineering Department ESAT/SISTA
Arenberg 10
3001 Leuven, Belgium
e-mail: *marc.moonen@esat.kuleuven.be*

Dennis R. Morgan
Bell Laboratories, Alcatel-Lucent
700 Mountain Avenue 2D-537
Murray Hill, NJ 07974-0636, USA
e-mail: *drrm@bell-labs.com*

David Nahamoo
IBM Thomas J. Watson Research Center
PO BOX 218
Yorktown Heights, NY 10598, USA
e-mail: *nahamoo@us.ibm.com*

Douglas O'Shaughnessy
Université du Québec
INRS Énergie, Matériaux et Télécommunications
(INRS-EMT)
800, de la Gauchetiere Ouest
Montréal, Québec H5A 1K6, Canada
e-mail: *dougo@emt.inrs.ca*

Lucas C. Parra
Steinman Hall, The City College of New York
Department of Biomedical Engineering
140th and Convent Ave
New York, NY 10031, USA
e-mail: *parra@ccny.cuny.edu*

Sarangarajan Parthasarathy
Yahoo!, Applied Research
1MC 743, 701 First Avenue
Sunnyvale, CA 94089-0703, USA
e-mail: *parthas@yahoo-inc.com*

Michael Syskind Pedersen
Oticon A/S
Kongebakken 9
2765 Smørum, Denmark
e-mail: *msp@oticon.dk*

Fernando Pereira
University of Pennsylvania
Department of Computer and Information Science
305 Levine Hall, 3330 Walnut Street
Philadelphia, PA 19104, USA
e-mail: *pereira@cis.upenn.edu*

Michael Picheny
IBM Thomas J. Watson Research Center
Yorktown Heights, NY 10598, USA
e-mail: *Picheny@us.ibm.com*

Rudolf Rabenstein
University Erlangen-Nuremberg
Electrical Engineering, Electronics, and
Information Technology
Cauerstrasse 7/LMS
91058 Erlangen, Germany
e-mail: *rabe@LNT.de*

Lawrence Rabiner
Rutgers University
Department of Electrical and Computer
Engineering
96 Frelinghuysen Road
Piscataway, NJ 08854, USA
e-mail: *lrr@caip.rutgers.edu*

Douglas A. Reynolds
Massachusetts Institute of Technology
Lincoln Laboratory, Information Systems
Technology Group
244 Wood Street
Lexington, MA 02420-9108, USA
e-mail: *dar@ll.mit.edu*

Michael Riley
Google, Inc., Research
111 Eighth AV
New York, NY 10011, USA
e-mail: *riley@google.com*

Aaron E. Rosenberg
Rutgers University
Center for Advanced Information Processing
96 Frelinghuysen Road
Piscataway, NJ 08854-8088, USA
e-mail: *aer@caip.rutgers.edu*

Salim Roukos
IBM T. J. Watson Research Center
Multilingual NLP Technologies
Yorktown Heights, NY 10598, USA
e-mail: *roukos@us.ibm.com*

Jan van Santen
Oregon Health And Science University
OGI School of Science and Engineering,
Department of Computer Science and Electrical
Engineering
20000 NW Walker Rd
Beaverton, OR 97006-8921, USA
e-mail: *vansanten@cslu.ogi.edu*

Ronald W. Schafer
Hewlett-Packard Laboratories
1501 Page Mill Road
Palo Alto, CA 94304, USA
e-mail: *ron.schafer@hp.com*

Juergen Schroeter
AT&T Labs - Research
Department of Speech Algorithms and Engines
180 Park Ave
Florham Park, NJ 07932, USA
e-mail: *schroeter@att.com*

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*

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*

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*

Jan Skoglund
Global IP Solutions
301 Brannan Street
San Francisco, CA 94107, USA
e-mail: *jan.skoglund@gipscorp.com*

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

Sascha Spors
Deutsche Telekom AG, Laboratories
Ernst-Reuter-Platz 7
10587 Berlin, Germany
e-mail: *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*

Richard Sproat
University of Illinois at Urbana-Champaign
Department of Linguistics
Urbana, IL 61801, USA
e-mail: *rws@uiuc.edu*

Yannis Stylianou
Institute of Computer Science
Heraklion, Crete 700 13, Greece
e-mail: *yannis@csd.uoc.gr*

Jes Thyssen
Broadcom Corporation
5300 California Avenue
Irvine, CA 92617, USA
e-mail: *jthyssen@broadcom.com*

Jay Wilpon
Research AT&T Labs
Voice and IP Services
Florham Park, NJ 07932, USA
e-mail: *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*

Arie Yeredor
Tel-Aviv University
Electrical Engineering - Systems
Tel-Aviv 69978, Israel
e-mail: *arie@eng.tau.ac.il*

Steve Young
Cambridge University Engineering Dept
Cambridge, CB21PZ, UK
e-mail: *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*

Contents

List of Abbreviations XXXI

1 **Introduction to Speech Processing**
J. Benesty, M. M. Sondhi, Y. Huang 1
1.1 A Brief History of Speech Processing 1
1.2 Applications of Speech Processing 2
1.3 Organization of the Handbook 4
References 4

Part A Production, Perception, and Modeling of Speech

2 **Physiological Processes of Speech Production**
K. Honda 7
2.1 Overview of Speech Apparatus 7
2.2 Voice Production Mechanisms 8
2.3 Articulatory Mechanisms 14
2.4 Summary 24
References 25

3 **Nonlinear Cochlear Signal Processing and Masking in Speech Perception**
J. B. Allen 27
3.1 Basics 27
3.2 The Nonlinear Cochlea 35
3.3 Neural Masking 45
3.4 Discussion and Summary 55
References 56

4 **Perception of Speech and Sound**
B. Kollmeier, T. Brand, B. Meyer 61
4.1 Basic Psychoacoustic Quantities 62
4.2 Acoustical Information Required for Speech Perception 70
4.3 Speech Feature Perception 74
References 81

5 **Speech Quality Assessment**
V. Grancharov, W. B. Kleijn 83
5.1 Degradation Factors Affecting Speech Quality 84
5.2 Subjective Tests 85
5.3 Objective Measures 90
5.4 Conclusions 95
References 96

Part B Signal Processing for Speech

6 **Wiener and Adaptive Filters**
J. Benesty, Y. Huang, J. Chen 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
References 120

7 **Linear Prediction**
J. Benesty, J. Chen, Y. Huang 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
References 133

8 **The Kalman Filter**
S. Gannot, A. Yeredor 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
References 157

9 **Homomorphic Systems and Cepstrum Analysis of Speech**
R. W. Schafer 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
References 180

10 **Pitch and Voicing Determination of Speech with an Extension Toward Music Signals**
W. J. Hess 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
References 208

11 **Formant Estimation and Tracking**
D. O'Shaughnessy 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
References 226

12 **The STFT, Sinusoidal Models, and Speech Modification**
M. M. Goodwin 229
12.1 The Short-Time Fourier Transform 230
12.2 Sinusoidal Models 242
12.3 Speech Modification 253
References 256

13 **Adaptive Blind Multichannel Identification**
Y. Huang, J. Benesty, J. Chen 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

Part C Speech Coding

14 Principles of Speech Coding
W. B. Kleijn 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

15 Voice over IP: Speech Transmission over Packet Networks
J. Skoglund, E. Kozica, J. Linden, R. Hagen, W. B. Kleijn 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

16 Low-Bit-Rate Speech Coding
A. V. McCree 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

17 Analysis-by-Synthesis Speech Coding
J.-H. Chen, J. Thyssen 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

18 **Perceptual Audio Coding of Speech Signals**
J. Herre, M. Lutzky 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

Part D Text-to-Speech Synthesis

19 **Basic Principles of Speech Synthesis**
J. Schroeter 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

20 **Rule-Based Speech Synthesis**
R. Carlson, B. Granström 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

21 **Corpus-Based Speech Synthesis**
T. Dutoit 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
References 453

22 **Linguistic Processing for Speech Synthesis**
R. Sproat 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
References 467

23 **Prosodic Processing**
J. van Santen, T. Mishra, E. Klabbers 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
References 485

24 **Voice Transformation**
Y. Stylianou 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
References 502

25 **Expressive/Affective Speech Synthesis**
N. Campbell 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
References 515

Part E Speech Recognition

26 **Historical Perspective of the Field of ASR/NLU**
L. Rabiner, B.-H. Juang 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
References 535

27 **HMMs and Related Speech Recognition Technologies**
S. Young 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
References 555

28 **Speech Recognition with Weighted Finite-State Transducers**
M. Mohri, F. Pereira, M. Riley 559
28.1 Definitions 559
28.2 Overview 560
28.3 Algorithms 567
28.4 Applications to Speech Recognition 574
28.5 Conclusion 582
References 582

29 **A Machine Learning Framework for Spoken-Dialog Classification**
C. Cortes, P. Haffner, M. Mohri 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
References 595

30 **Towards Superhuman Speech Recognition**
M. Picheny, D. Nahamoo 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
References 614

31 **Natural Language Understanding**
S. Roukos 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
References 626

32 **Transcription and Distillation of Spontaneous Speech**
S. Furui, T. Kawahara 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
References 647

33 **Environmental Robustness**
J. Droppo, A. Acero 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
References 677

34 **The Business of Speech Technologies**
J. Wilpon, M. E. Gilbert, J. Cohen 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
References 702

35 **Spoken Dialogue Systems**
V. Zue, S. Seneff 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
References 718

Part F Speaker Recognition

36 **Overview of Speaker Recognition**
A. E. Rosenberg, F. Bimbot, S. Parthasarathy 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
References 739

37 **Text-Dependent Speaker Recognition**
M. Hébert 743
37.1 Brief Overview 743
37.2 Text-Dependent Challenges 747
37.3 Selected Results 750
37.4 Concluding Remarks 760
References 760

38 **Text-Independent Speaker Recognition**
D. A. Reynolds, W. M. Campbell 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
References 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 Cancelation for Voice Signals

M. M. Sondhi 903
45.1 Network Echoes 904
45.2 Single-Channel Acoustic Echo Cancelation 915
45.3 Multichannel Acoustic Echo Cancelation 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 **Convolutive 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

List of Abbreviations

2TS	two-tone suppression

A

ACELP	algebraic code excited linear prediction
ACF	autocorrelation function
ACR	absolute category rating
ACS	autocorrelation coefficient sequences
ACeS	Asia Cellular Satellite
ADC	analog-to-digital converter
ADPCM	adaptive differential pulse code modulation
AEC	acoustic echo cancelation
AFE	advanced front-end
AGC	automatic gain control
AGN	automatic gain normalization
AL	averaged acoustic frame likelihood
AMR-WB+	extended wide-band adaptive multirate coder
AMR-WB	wide-band AMR speech coder
AMSC-TMI	American Mobile Satellite Corporation Telesat Mobile Incorporated
AN	auditory nerve
ANC	active noise cancelation
ANN	artificial neural networks
ANOVA	analysis of variance
APA	affine projection algorithm
APC	adaptive predictive coding
APCO	Association of Public-Safety Communications Officials
APP	adjusted test-set perplexity
AR	autoregressive
ARISE	Automatic Railway Information Systems for Europe
ARMA	autoregressive moving-average
ARPA	Advanced Research Projects Agency
ARQ	automatic repeat request
ART	advanced recognition technology
ASAT	automatic speech attribute transcription
ASM	acoustic segment model
ASR	automatic speech recognition
ATF	acoustical transfer function
ATIS	airline travel information system
ATN	augmented transition networks
ATR	advanced telecommunications research
AW	acoustic word

B

BBN	Bolt, Beranek and Newman
BIC	Bayesian information criterion
BILD	binaural intelligibility level difference
BM	blocking matrix
BN	broadcast news
BSD	bark spectral distortion
BSS	blind source separation

C

C	consonants
CA	cochlear amplifier
CAF	continuous adaptation feedback
CART	classification and regression tree
CASA	computational auditory scene analysis
CAT	cluster adaptive training
CCR	comparison category rating
CDCN	codeword-dependent cepstral normalization
CDF	cumulative distribution function
CDMA	code division multiple access
CE	categorical estimation
CELP	code-excited linear prediction
CF	characteristic frequency
CF	coherence function
CH	call home
CHN	cepstral histogram normalization
CIS	caller identification system
CMLLR	constrained MLLR
CMOS	comparison mean opinion score
CMR	co-modulation masking release
CMS	cepstral mean subtraction
CMU	Carnegie Mellon University
CMVN	cepstral mean and variance normalization
CNG	comfort noise generation
COC	context-oriented clustering
CR	cross-relation
CRF	conditional random fields
CRLB	Cramèr–Rao lower bound
CS-ACELP	conjugate structure ACELP
CS-CELP	conjugate structure CELP
CSJ	corpus of spontaneous Japanese
CSR	continuous speech recognition
CTS	conversational telephone speech
CVC	consonant–vowel–consonant
CVN	cepstral variance normalization
CZT	chirp z-transform

D

DAC	digital-to-analog
DAG	directed acyclic graph
DAM	diagnostic acceptability measure

DARPA	Defense Advanced Research Projects Agency
DBN	dynamic Bayesian network
DCF	detection cost function
DCR	degradation category rating
DCT	discrete cosine transform
DET	detection error tradeoff
DF	disfluency
DFA	deterministic finite automata
DFT	discrete Fourier transform
DFW	dynamic frequency warping
DM	dialog management
DMOS	degradation mean opinion score
DP	dynamic programming
DPCM	differential PCM
DPMC	data-driven parallel model combination
DRT	diagnostic rhyme test
DSP	digital signal processing
DT	discriminative training
DTFT	discrete-time Fourier transform
DTW	dynamic time warping
DoD	Department of Defense

E

EC	equalization and cancelation
ECOC	error-correcting output coding
EER	equal error rate
EGG	electroglottography
EKF	extended Kalman filter
ELER	early-to-late energy ratio
EM	estimate–maximize
EM	expectation maximization
EMG	electromyographic
EMLLT	extended maximum likelihood linear transform
ER AAC-LD	error resilient low-delay advanced audio
ERB	equivalent rectangular bandwidth
ERL	echo return loss
ERLE	echo return loss enhancement
EVRC	enhanced variable rate coder
eX-CELP	extended CELP

F

FA	false accept
FAP	fast affine projection
FB-LPC	forward backward linear predictive coding
FBF	fixed beamformer
FBS	filter bank summation
FC	functional contour
FCDT	frame-count-dependent thresholding
FEC	frame erasure concealment
FFT	fast Fourier transform
FIFO	first-in first-out
FIR	finite impulse response
FM	forward masking
FMLLR	maximum-likelihood feature-space regression
FR	filler rate
FRC	functional residual capacity
FRLS	fast recursive least-squares
FSM	finite state machine
FSN	finite state network
FSS	frequency selective switch
FST	finite state transducer
FT	Fourier transform
FTF	fast transversal filter
FVQ	fuzzy vector quantization
FXLMS	filtered-X LMS

G

GCC	generalized cross-correlation
GCI	glottal closure instant
GEVD	generalized eigenvalue decomposition
GLDS	generalized linear discriminant sequence
GLR	generalized likelihood ratio
GMM	Gaussian mixture model
GPD	generalized probabilistic descent
GSC	generalized sidelobe canceller
GSM	Groupe Spéciale Mobile
GSV	GMM supervector

H

HLDA	heteroscedastic LDA
HLT	human language technologies
HMIHY	How May I Help You
HMM	hidden Markov models
HMP	hidden Markov processes
HNM	harmonic-plus-noise model
HOS	higher-order statistics
HPF	high-pass filter
HRTF	head-related transfer function
HSD	honestly significant difference
HSR	human speech recognition
HTK	hidden Markov model toolkit

I

IAI	International Association for Identification
ICA	independent component analysis
IDA	Institute for Defense Analyses
IDFT	inverse DFT
IDTFT	inverse discrete-time Fourier transform
IETF	Internet Engineering Task Force
IFSS	inverse frequency selective switch
IHC	inner hair cells
II	information index

IIR	infinite impulse response
iLBC	internet low-bit-rate codec
IMCRA	improved minima-controlled recursive averaging
IMDCT	inverse MDCT
IMM	interacting multiple model
IO	input–output
IP	internet protocol
IP	interruption (disfluent) point
IPA	International Phonetic Alphabet
IPNLMS	improved PNLMS
IQMF	QMF synthesis filterbank
IR	information retrieval
IS	Itakura–Saito
ISU	information state update
ITU	International Telecommunication Union
IVR	interactive voice response

J

JADE	joint approximate diagonalization of eigenmatrices
JND	just-noticeable difference

K

KL	Kullbach–Leibler
KLT	Karhunen–Loève transform

L

LAN	local-area network
LAR	log-area-ratio
LCMV	linearly constrained minimum-variance
LD-CELP	low-delay CELP
LDA	linear discriminant analysis
LDC	Linguistic Data Consortium
LDF	linear discriminant function
LID	language identification
LL	log-likelihood
LLAMA	learning library for large-margin classification
LLR	(log) likelihood ratio
LMFB	log mel-frequency filterbank
LMR	linear multivariate regression
LMS	least mean square
LNRE	large number of rare events
LP	linear prediction
LPC	linear prediction coefficients
LPC	linear predictive coding
LPCC	linear predictive cepstral coefficient
LRE	language recognition evaluation
LRT	likelihood-ratio test
LSA	latent semantic analysis
LSA	log-spectral amplitude
LSF	line spectral frequency
LSI	latent semantic indexing
LSR	low sampling rates
LTI	linear time invariant
LTIC	linear time-invariant causal
LTP	long term prediction
LVCSR	large vocabulary continuous speech recognition

M

M-step	maximization stage
MA	moving average
MAP	maximum a posteriori
MBE	multiband excited
MBR	minimum Bayes-risk
MBROLA	multiband resynthesis overlap-add
MC	multicategory
MCE	minimum classification error
MCN	multichannel Newton
MDC	multiple description coding
MDCT	modified discrete cosine transform
MDF	multidelay filter
MDP	Markov decision process
MELP	mixed excitation linear prediction
MFCC	mel-filter cepstral coefficient
MFoM	maximal figure-of-merit
MIMO	multiple-input multiple-output
MIPS	million instructions per second
ML	maximum-likelihood
MLLR	maximum-likelihood linear regression
MLP	multilayer perceptron
MMI	maximum mutual information
MMSE-LSA	MMSE of the log-spectral amplitude
MMSE-SA	MMSE of the spectral amplitude
MMSE	minimum mean-square error
MNB	measuring normalizing blocks
MNRU	modulated noise reference unit
MOPS	million operations per second
MOS	mean opinion score
MP	matching pursuit
MPE	minimum phone error
MPEG	Moving Pictures Expert Ggroup
MPI	minimal pairs intelligibility
MPLPC	multipulse linear predictive coding
MRI	magnetic resonance imaging
MRT	modified rhyme test
MS	minimum statistics
MSA	modern standard Arabic
MSD	minimum significant difference
MSE	mean-square error
MSG	maximum stable gain
MSNR	maximum signal-to-noise ratio
MSVQ	multistage VQ
MUI	multimodal user interface
MUSHRA	multi stimulus test with hidden reference and anchor

MVE	minimum verification error
MVIMP	my voice is my password
MW	maximum wins
MuSIC	multiple signal classification

N

NAB	North American Business News
NASA	National Aeronautics and Space Administration
NATO	North Atlantic Treaty Organization
NFA	nondeterministic finite automata
NFC	noise feedback coding
NIST	National Institute of Standards and Technology
NLG	natural language generation
NLMS	normalized least-mean-square
NLP	natural language processing
NLU	natural language understanding
NN	neural network
NSA	National Security Agency
NTT	Nippon Telephone & Telegraph
NUU	nonuniform units

O

O	object
OAE	otoacoustic emissions
OHC	outer-hair cells
OLA	overlap-and-add
OOV	out-of-vocabulary
OQ	open quotient
OR	out-of-vocabulary rate
OSI	open systems interconnection reference

P

P-PRLM	parallel PRLM
PARCOR	partial correlation coefficients
PBFD	partitioned-block frequency-domain
PCA	principal component analysis
PCBV	phonetic-class-based verification
PCFG	probabilistic context-free grammar
PCM	pulse-code modulation
PDA	pitch determination algorithms
PDC	personal digital cellular
PDF	probability density function
PDP	parallel distributed processing
PDS	positive-definite symmetric
PEAQ	perceptual quality assessment for digital audio
PESQ	perceptual evaluation of speech quality
PGG	photoglottography
PICOLA	pointer interval controlled overlap and add
PIV	particle image velocity
PLC	packet loss concealment
PLP	perceptual linear prediction
PMC	parallel model combination
PNLMS	proportionate NLMS
POS	part-of-speech
PP	word perplexity
PPRLM	parallel PRLM
PR	phone recognizer
PRA	partial-rank algorithm
PRLM	phoneme recognition followed by language modeling
PSD	power spectral density
PSI-CELP	pitch synchronous innovation CELP
PSI	pitch synchronous innovation
PSQM	perceptual speech quality measure
PSRELP	pitch-synchronous residual excited linear prediction
PSTN	public switched telephone network
PVT	parallel voice tokenization

Q

QA	question answering
QMF	quadrature mirror filter
QP	quadratic program
QoS	quality-of-service

R

RD	rate-distortion
RASTA	relative spectra
RATZ	multivariate Gaussian-based cepstral normalization
RCELP	relaxed CELP
REW	rapidly evolving waveform
RF	radio frequency
RIM	repair interval model
RIR	room impulse response
RL	reticular lamina
RLS	recursive least-squares
RM	resource management
RMS	root mean square
RMSE	root-mean-square error
ROC	receiver operating characteristic
RPA	raw phone accuracy
RPD	point of the reparandum
RPE-LTP	regular-pulse excitation with long-term prediction
RR	reprompt rates
RS	Reed–Solomon
RSVP	resource reservation protocol
RT	rich transcription

RTM	resonant tectorial membrane
RTP	real-time transport protocol

S

S2S	Speech-to-speech
S	subject
SAT	speaker-adaptive trained
SB	switchboard
SCTE	Society of Cable Telecommunications Engineers
SD-CFG	stochastic dependency context-free grammar
SD	spectral distortion
SDC	shifted delta cepstral
SDR-GSC	speech distortion regularized generalized sidelobe canceller
SDW-MWF	speech distortion-weighted multichannel Wiener filter
SEW	slowly evolving waveform
SGML	standard generalized markup language
SI	speech intelligibility
SII	speech intelligibility index
SIMO	single-input multiple-output
SISO	single-input single-output
SL	sensation level
SLM	statistical language model
SLS	spoken language system
SM	sinusoidal models
SMS	speaker model synthesis
SMT	statistical machine translation
SMV	selectable mode vocoder
SNR	signal-to-noise ratio
SOLA	synchronized overlap add
SOS	second-order statistics
SPAM	subspace-constrained precision and means
SPIN	saturated Poisson internal noise
SPINE	speech in noisy environment
SPL	sound pressure level
SPLICE	stereo piecewise linear compensation for environment
SQ	speed quotient
SR	speaking rate
SRT	speech reception threshold
SSML	speech synthesis markup language
STC	sinusoidal transform coder
STFT	short-time Fourier transform
SU	sentence unit
SUI	speech user interface
SUNDIAL	speech understanding and dialog
SUR	Speech Understanding Research
SVD	singular value decomposition
SVM	support vector machines
SWB	switchboard
SegSNR	segmental SNR

T

T–F	time–frequency
TBRR	transient beam-to-reference ratio
TC	text categorization
TCP	transmission control protocol
TCX	transform coded excitation
TD-PSOLA	time-domain pitch-synchronous overlap-add
TD	time domain
TDAC	time-domain aliasing cancelation
TDBWE	time-domain bandwidth extension
TDMA	time-division multiple-access
TDOA	time difference of arrival
TDT	topic detection and tracking
TF-GSC	transfer-function generalized sidelobe canceller
TFIDF	term frequency inverse document frequency
TFLLR	term frequency log-likelihood ratio
TFLOG	term frequency logarithmic
TI	transinformation index
TIA	Telecommunications Industry Association
TITO	two-input-two-output
TM	tectorial membrane
TM	tympanic membrane
TMJ	temporomandibular joint
TNS	temporal noise shaping
TPC	transform predictive coder
TSNFC	two-stage noise feedback coding
TTS	text-to-speech
ToBI	tone and break indices

U

UBM	universal background model
UCD	unit concatenative distortion
UDP	user datagram protocol
UE	user experience
UKF	unscented Kalman filter
ULD	ultra-low delay
USD	unit segmental distortion
USM	upward spread of masking
UT	unscented transform
UVT	universal voice tokenization

V

V	verb
V	vowels
VAD	voice activity detector
VBAP	vector based amplitude panning
VCV	vowel–consonant–vowel
VLSI	very large-scale integration

VMR-WB	variable-rate multimode wide-band
VOT	voice onset time
VQ	vector quantization
VRCP	voice recognition call processing
VRS	variable rate smoothing
VSC	vector space characterization
VSELP	vector sum excited linear prediction
VT	voice tokenization
VTLN	vocal-tract-length normalization
VTR	vocal tract resonance
VTS	vector Taylor-series
VoIP	voice over IP

W

WDRC	wide dynamic-range multiband compression
WER	word error rate
WFSA	weighted finite-state acceptors
WFST	weighted finite-state transducer
WGN	white Gaussian noise
WI	waveform interpolation
WLAN	wireless LAN
WLS	weighted least-squares
WMOPS	weighted MOPS
WSJ	Wall Street Journal
WSOLA	waveform similarity OLA
WiFi	wireless fidelity

X

XML	extensible mark-up languages
XOR	exclusive-or

Z

ZC	zero crossing
ZIR	zero-input response
ZSR	zero-state response

1. Introduction to Speech Processing

J. Benesty, M. M. Sondhi, Y. Huang

In this brief introduction we outline some major highlights in the history of speech processing. We briefly describe some of the important applications of speech processing. Finally, we introduce the reader to the various parts of this handbook.

1.1 A Brief History of Speech Processing 1
1.2 Applications of Speech Processing 2
1.3 Organization of the Handbook 4
References 4

1.1 A Brief History of Speech Processing

Human beings have long been motivated to create machines that can talk. Early attempts at understanding speech production consisted of building mechanical models to mimic the human vocal apparatus. Two such examples date back to the 13th century, when the German philosopher Albertus Magnus and the English scientist Roger Bacon are reputed to have constructed metal talking heads. However, no documentation of these devices is known to exist. The first documented attempts at making speaking machines came some five hundred years later. In 1769 Kratzenstein constructed resonant cavities which, when he excited them by a vibrating reed, produced the sounds of the five vowels a, e, i, o, and u. Around the same time, and independently of this work, Wolfgang von Kempelen constructed a mechanical speech synthesizer that could generate recognizable consonants, vowels, and some connected utterances. His book on his research, published in 1791, may be regarded as marking the beginnings of speech processing. Some 40 years later, Charles Wheatstone constructed a machine based essentially on von Kempelen's specifications [1.1–3].

Interest in mechanical analogs of the human vocal apparatus continued well into the 20th century. Mimics of the type of von Kempelen's machine were constructed by several people besides Wheatstone, e.g., Joseph Faber, Richard Paget, R. R. Riesz, et al.

It is known that as a young man Alexander Graham Bell had the opportunity to see Wheatstone's implementation. He too made a speaking machine of that general nature. However, it was his other invention – the telephone – that provided a major impetus to modern speech processing. Nobody could have guessed at that time the impact the telephone would have, not only on the way people communicate with each other but also on research in speech processing as a science in its own right. The availability of the speech waveform as an electrical signal shifted interest from mechanical to electrical machines for synthesizing and processing speech.

Some attempts were made in the 1920s and 1930s to synthesize speech electrically. However it is Homer Dudley's work in the 1930s that ushered in the modern era of speech processing. His most important contribution was the clear understanding of the *carrier* nature of speech [1.4]. He developed the analogy between speech signals and modulated-carrier radio signals that are used, for instance, for the transmission or broadcast of audio signals. In the case of the radio broadcast, the message to be transmitted is the audio signal which has frequencies in the range of 0–20 kHz. Analogously, the message to be transmitted in the case of speech is carried mainly by the time-varying shape of the vocal tract, which in turn is a representation of the *thoughts* the speaker wishes to convey to the listener. The movements of the vocal tract are at syllabic rates, i. e., at frequencies between 0 and 20 Hz. In each case – electromagnetic and acoustic – the message is in a frequency range unsuitable for transmission. The solution in each case is to imprint the message on a carrier. In the electromagnetic case the carrier is usually a high-frequency sinusoidal wave. In the acoustic case the carrier can be one of several signals. It is the quasi periodic signal provided by the vocal cords for voiced speech, and a noise-like signal provided by turbulence at a constriction for fricative and aspirated sounds. Or it can be a combination of these for voiced fricative sounds. Indeed, the selection of the carrier as well as the changes in intensity and fundamental frequency of the

vocal cords may be conveniently regarded as additional parts of the message.

Being an electrical engineer himself, Dudley proceeded to exploit this insight to construct an *electrical* speech synthesizer which dispensed with all the mechanical devices of von Kempelen's machine. Electrical circuits were used to generate the carriers. And the message (i. e., the characteristics of the vocal tract) was imprinted on the carrier by passing it through a time-varying filter whose frequency response was adjusted to simulate the transfer characteristics of the vocal tract.

With the collaboration of Riesz and Watkins, Dudley implemented two highly acclaimed devices based on this principle – the Voder and the Vocoder. The Voder was the first versatile talking machine able to produce arbitrary sentences. It was a system in which an operator manipulated a keyboard to control the sound source and the filter bank. This system was displayed with great success at the New York World Fair of 1939. It could produce speech of much better quality than had been possible with the mechanical devices, but remained essentially a curiosity. The Vocoder, on the other hand had a much more serious purpose. It was the first attempt at *compressing* speech. Dudley estimated that since the message in a speech signal is carried by the slowly time-varying filters, it should be possible to send adequate information for the receiver to be able to reconstruct a telephone speech signal using a bandwidth of only about 150 Hz, which is about 1/20 the bandwidth required to send the speech signal. Since bandwidth was very expensive in those days, this possibility was extremely attractive from a commercial point of view.

We have devoted so much space here to Dudley's work because his ideas were the basis of practically all the work on speech signal processing that followed. The description of speech in terms of a carrier (or excitation function) and its modulation (or the time-varying spectral envelope) is still – 70 years later – the basic representation. The parameters used to quantify these components, of course, have evolved in various ways. Besides the channel Vocoder (the modern name for Dudley's Vocoder) many other types of Vocoders have been invented, e.g., formant Vocoder, voice-excited Vocoder. Besides speech compression, Dudley's description was also considered for other applications such as secure voice systems, and the sound spectrograph and its use for communication with the deaf.

Unfortunately, the quality achieved by analog implementations of Vocoders never reached a level acceptable for commercial telephony. Nevertheless they found useful applications for military purposes where poor speech quality was tolerated. The Vocoder representation was also the basis of a speech secrecy system that found extensive use during World War II.

Another example of an analog implementation of Dudley's representation is the sound spectrograph. This is a device that displays the distribution of energy in a speech signal as a function of frequency, and the evolution of this distribution in time. This tool has been extremely useful for investigating properties of speech signals. A real time version of the spectrograph was intended for use as a device for communication with the deaf. That, however, was not very successful. A few people were able to identify about 300 words after 100 hours of training. However, it turned out to be too difficult a task to be practical.

During more than three decades following Dudley's pioneering work, a great amount of research was done on various aspects and properties of speech – properties of the speech production mechanisms, the auditory system, psychophysics, etc. However, except for the three applications mentioned above, little progress was made in speech signal processing and its applications. Exploitation of this research for practical applications had to wait for the general availability of digital hardware starting in the 1970s. Since then much progress has been made in speech coding for efficient transmission, speech synthesis, speech and speaker recognition, and hearing aids [1.5–7]. In the next section we discuss some of these developments.

Today, the area of speech processing is very vast and rich as can be seen from the contents of this Handbook. While we have made great progress since the invention of the telephone, research in the area of speech processing is still very active, and many challenging problems remain unsolved.

1.2 Applications of Speech Processing

As mentioned above, one of the earliest goals of speech processing was that of coding speech for efficient transmission. This was taken to be synonymous with reduction of the bandwidth required for transmitting speech. Several advances were needed before the modern success in speech coding was achieved. First, the

notions of information theory introduced during the late 1940s and 1950s brought the realization that the proper goal was the reduction of information rate rather than bandwidth. Second, hardware became available to utilize the sampling theorem to convert a continuous band-limited signal to a sequence of discrete samples. And quantization of the samples allowed digitization of a band-limited speech signal, thus making it usable for digital processing. Finally, the description of a speech signal in terms of linear prediction coefficients (LPC) provided a very convenient representation [1.8–11]. (The theory of predictive coding was in fact developed in 1955. However, its application to speech signals was not made until the late 1970s.)

A telephone speech signal, limited in frequency from 0 to 3.4 kHz, requires 64 kbps (kilobits per second) to be transmitted without further loss of quality. With modern speech compression techniques, the bit rate can be reduced to 13 kbps with little further degradation. For commercial telephony a remaining challenge is to reduce the required bit rate further but without sacrificing quality. Today, the rate can be lowered down to 2.4 kbps while maintaining very high intelligibility, but with a significant loss in quality. Some attempts have been made to reduce the bit rate down to 300 bps, e.g., for radio communication with a submarine. However the quality and intelligibility at these low bit rates are very poor.

Another highly successful application of speech processing is automatic speech recognition (ASR). Early attempts at ASR consisted of making deterministic models of whole words in a small vocabulary (say 100 words) and recognizing a given speech utterance as the word whose model comes closest to it. The introduction of hidden Markov models (HMMs) in the early 1980s provided a much more powerful tool for speech recognition [1.12–14]. Today many products have been developed that successfully utilize ASR for communication between humans and machines. And the recognition can be done for continuous speech using a large vocabulary, and in a speaker-independent manner. Performance of these devices, however, deteriorates in the presence of reverberation and even low levels of ambient noise. Robustness to noise, reverberation, and characteristics of the transducer, is still an unsolved problem.

The goal of ASR is to recognize speech accurately regardless of who the speaker is. The complementary problem is that of recognizing a speaker from his/her voice, regardless of what words he/she is speaking. At present this problem appears to be solvable only if the speaker is one of a small set of N known speakers. A variant of the problem is speaker *verification*, in which the aim is to automatically verify the claimed identity of a speaker. While speaker recognition requires the selection of one out of N possible outcomes, speaker verification requires just a yes/no answer. This problem can be solved with a high degree of accuracy for much larger populations. Speaker verification has application wherever access to data or facilities has to be controlled. Forensics is another area of application. The problem of reduced performance in the presence of noise, as mentioned above for ASR, applies also to speaker recognition and speaker verification.

A third application of speech processing is that of synthesizing speech corresponding to a given text. When used together with ASR, speech synthesis allows a complete two-way spoken interaction between humans and machines. Speech synthesis is also a way to communicate for persons unable to speak. Its use for this purpose by the famous physicist Stephen Hawking is well known.

Early attempts at speech synthesis consisted of deriving the time-varying spectrum for the sequence of phonemes of a given text sentence. From this the corresponding time variation of the vocal tract was estimated, and the speech was synthesized by exciting the time-varying vocal tract with periodic or noise-like excitation as appropriate. The quality of the synthesis was significantly improved by concatenating pre-stored units (i. e., short segments such as diphones, triphones) after modifying them to fit the context. Today the highest-quality speech is synthesized by the unit *selection* method in which the units are selected from a large amount of stored speech and concatenated with little or no modification.

Finally we might mention the application of speech processing to aids for the handicapped. Hearing aid technology has made considerable progress in the last two decades. Part of this progress is due to a slow but steady improvement in our knowledge of the human hearing mechanism. A large part is due to the availability of high-speed digital hardware. At present performance of hearing aids is still poor under noisy and reverberant conditions.

A potentially useful application of speech processing to aid the handicapped is to display the shape of one's vocal tract as one speaks. By trying to match one's vocal tract shape to a displayed shape, a deaf person can learn correct pronunciation. Some attempts to implement this idea have been made, but have still been only in the realm of research.

Another useful application is a reading aid for the blind. The idea is to have a device to scan printed text from a book, and synthesize speech from the

scanned text. Coupled with a device to change speaking rate, this forms a useful aid for the blind. Several products offering this application are available on the market.

Many other application examples are described in the various parts of this handbook. We invite the reader to browse this volume on speech processing to find topics relevant to his/her specific interests.

1.3 Organization of the Handbook

This handbook on speech processing is a comprehensive source of knowledge in speech technology and its applications. It is organized as follows. This volume is divided into nine parts. For each part we invited at least one associate editor (AE) to handle it. All the AEs are very well-known researchers in their respective area of research. Part A (AE: M. M. Sondhi) contains four chapters on production, perception, and modeling of speech signals. Part B (AEs: Y. Huang and J. Benesty) concerns signal processing tools for speech, in eight chapters. Part C (AE: B. Kleijn) covers five chapters on speech coding. In part D (AE: S. Narayanan), the areas of text-to-speech synthesis are presented in seven chapters. Part E (AEs: L. Rabiner and B.-H. Juang), with 10 chapters, is a comprehensive overview on speech recognition. Part F (AE: S. Parthasarathy) contains three chapters on speaker recognition. Part G (AE: C.-H. Lee) is about language identification and contains four chapters. In part H (AEs: J. Chen, S. Gannot, and J. Benesty), various aspects of speech enhancement are developed in seven chapters. Finally the last section, part I (AEs: J. Benesty, I. Cohen, and Y. Huang), presents the important aspects of multichannel speech processing in four chapters.

References

1.1 H. Dudley, T.H. Tarnoczy: The speaking machine of Wolfgang von Kempelen, J. Acoust. Soc. Am. **22**, 151–166 (1950)

1.2 G. Fant: *Acoustic Theory of Speech Production* (Mouton, 's-Gravenhage 1960)

1.3 J.L. Flanagan: *Speech Analysis, Synthesis and Perception* (Springer, New York 1972)

1.4 H. Dudley: The carrier nature of speech, Bell Syst. Tech. J. **19**(4), 495–515 (1940)

1.5 L.R. Rabiner, R.W. Schafer: *Digital Processing of Speech Signals* (Prentice Hall, Englewood Cliffs 1978)

1.6 S. Furui, M.M. Sondhi (Eds.): *Advances in Speech Signal Processing* (Marcel Dekker, New York 1992)

1.7 B. Gold, N. Morgan: *Speech and Audio Signal Processing* (Wiley, New York 2000)

1.8 P. Elias: Predictive coding I, IRE Trans. Inform. Theory **1**(1), 16–24 (1955)

1.9 P. Elias: Predictive coding II, IRE Trans. Inform. Theory **1**(1), 24–33 (1955)

1.10 B.S. Atal, M.R. Schroeder: Adaptive predictive coding of speech, Bell Syst. Tech. J. **49**(8), 1973–1986 (1970)

1.11 B.S. Atal: The history of linear prediction, IEEE Signal Proc. Mag. **23**(2), 154–161 (2006)

1.12 L.R. Rabiner: A tutorial on hidden Markov models and selected applications in speech recognition, Proc. IEEE **77**(2), 257–286 (1989)

1.13 L.R. Rabiner, B.-H. Juang: *Fundamentals of Speech Recognition* (Prentice-Hall, Englewood Cliff 1993)

1.14 F. Jelinek: *Statistical Methods for Speech Recognition* (MIT, Boston 1998)

Part A Production

Part A Production, Perception, and Modeling of Speech

Ed. by M. M. Sondhi

2 Physiological Processes of Speech Production
K. Honda, Paris, France

3 Nonlinear Cochlear Signal Processing and Masking in Speech Perception
J. B. Allen, Urbana, USA

4 Perception of Speech and Sound
B. Kollmeier, Oldenburg, Germany
T. Brand, Oldenburg, Germany
B. Meyer, Oldenburg, Germany

5 Speech Quality Assessment
V. Grancharov, Stockholm, Sweden
W. B. Kleijn, Stockholm, Sweden

2. Physiological Processes of Speech Production

K. Honda

Speech sound is a wave of air that originates from complex actions of the human body, supported by three functional units: generation of air pressure, regulation of vibration, and control of resonators. The lung air pressure for speech results from functions of the respiratory system during a prolonged phase of expiration after a short inhalation. Vibrations of air for voiced sounds are introduced by the vocal folds in the larynx; they are controlled by a set of laryngeal muscles and airflow from the lungs. The oscillation of the vocal folds converts the expiratory air into intermittent airflow pulses that result in a buzzing sound. The narrow constrictions of the airway along the tract above the larynx also generate transient source sounds; their pressure gives rise to an airstream with turbulence or burst sounds. The resonators are formed in the upper respiratory tract by the pharyngeal, oral, and nasal cavities. These cavities act as resonance chambers to transform the laryngeal buzz or turbulence sounds into the sounds with special linguistic functions. The main articulators are the tongue, lower jaw, lips, and velum. They generate patterned movements to alter the resonance characteristics of the supra-laryngeal airway. In this chapter, contemporary views on phonatory and articulatory mechanisms are summarized to illustrate the physiological processes of speech production, with brief notes on their observation techniques.

2.1 **Overview of Speech Apparatus** 7
2.2 **Voice Production Mechanisms** 8
2.2.1 Regulation of Respiration 8
2.2.2 Structure of the Larynx 9
2.2.3 Vocal Fold and its Oscillation 10
2.2.4 Regulation of Fundamental Frequency (F_0) 12
2.2.5 Methods for Measuring Voice Production 13
2.3 **Articulatory Mechanisms** 14
2.3.1 Articulatory Organs 14
2.3.2 Vocal Tract and Nasal Cavity 18
2.3.3 Aspects of Articulation in Relation to Voicing 19
2.3.4 Articulators' Mobility and Coarticulation 22
2.3.5 Instruments for Observing Articulatory Dynamics 23
2.4 **Summary** 24
References 25

2.1 Overview of Speech Apparatus

The speech production apparatus is a part of the motor system for respiration and alimentation. The form of the system can be characterized, when compared with those of other primates, by several unique features, such as small red lips, flat face, compact teeth, short oral cavity with a round tongue, and long pharynx with a low larynx position. The functions of the system are also uniquely advanced by the developed brain with the language areas, direct neural connections from the cortex to motor nuclei, and dense neural supply to each muscle. Independent control over phonation and articulation is a human-specific ability. These morphological and neural changes along human evolution reorganized the original functions of each component into an integrated motor system for speech communication.

The speech apparatus is divided into the organs of phonation (voice production) and articulation (settings of the speech organs). The phonatory organs (lungs and larynx) create voice source sounds by setting the driving air pressure in the lungs and parameters for vocal fold vibration at the larynx. The two organs together adjust the pitch, loudness, and quality of the voice, and further generate prosodic patterns of speech. The articulatory organs give resonances or modulations to the voice source and generate additional sounds for some consonants. They consist of the lower jaw, tongue,

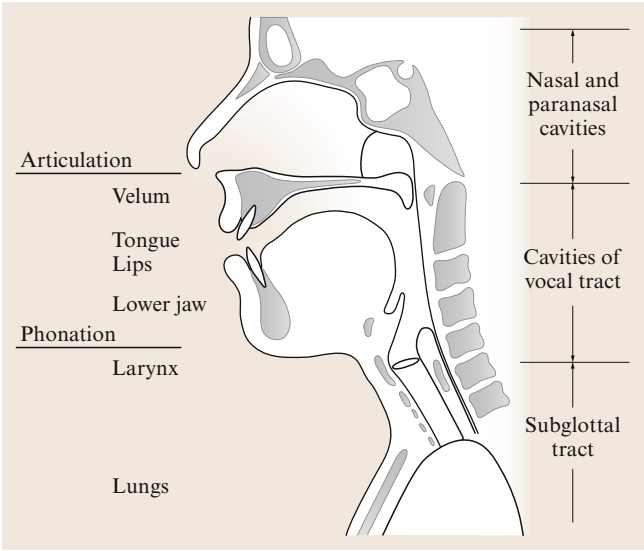


Fig. 2.1 Sketch of a speech production system. Physiological processes of speech production are realized by combined sequential actions of the speech organs for phonation and articulation. These activities result in sound propagation phenomena at the three levels: subglottal cavities, cavities of the vocal tract, and nasal and paranasal cavities

lips, and the velum. The larynx also takes a part in the articulation of voiced/voiceless distinctions. The tongue and lower lip attach to the lower jaw, while the velum is loosely combined with other articulators. The constrictor muscles of the pharynx and larynx also participate in articulation as well as in voice quality control. The phonatory and articulatory systems influence each other mutually, while changing the vocal tract shape for producing vowels and consonants. Figure 2.1 shows a schematic drawing of the speech production system.

2.2 Voice Production Mechanisms

Generation of voice source requires adequate configuration of the airflow from the lungs and vocal fold parameters for oscillation. The sources for voiced sounds are the airflow pulses generated at the larynx, while those for some consonants (i. e., stops and fricatives) are airflow noises made at a narrow constriction in the vocal tract. The expiratory and inspiratory muscles together regulate relatively constant pressure during speech. The laryngeal muscles adjust the onset/offset, amplitude, and frequency of vocal fold vibration.

2.2.1 Regulation of Respiration

The respiratory system is divided into two segments: the conduction airways for ventilation between the atmosphere and the lungs, and the respiratory tissue of the lungs for gas exchange. Ventilation (i. e., expiration and inhalation) is carried out by movements of the thorax, diaphragm, and abdomen. These movements involve actions of respiratory muscles and elastic recoil forces of the system. During quiet breathing, the lungs expand to inhale air by the actions of inspiratory muscles (diaphragm, external intercostal, etc.), and expel air by the elastic recoil force of the lung tissue, diaphragm, and cavities of the thorax and abdomen. In effort expiration, the expiratory muscles (internal intercostals, abdominal muscles, etc.) come into action. The inspiratory and expiratory muscles work alternately, making the thorax expand and contract during deep breathing.

During speech production, the respiratory pattern changes to a longer expiratory phase with a shorter inspiratory phase during quiet breathing. Figure 2.2 shows a conventional view of the respiratory pattern during

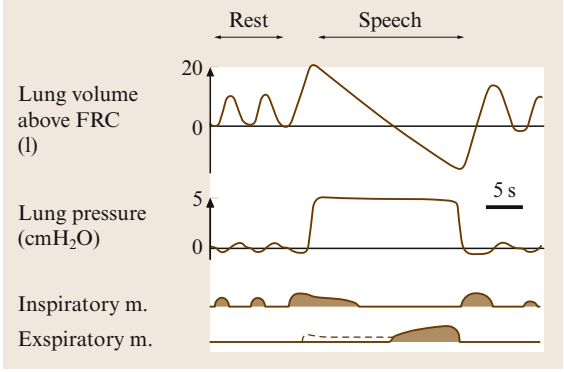


Fig. 2.2 Respiratory pattern during speech. Top two curves show the changes in the volume and pressure in the lungs. The bottom two curves show schematic activity patterns of the inspiratory and expiratory muscles (after [2.1]). The dashed line for the expiratory muscles indicates their predicted activity for expiration

speech [2.1]. The thorax is expanded by inspiration prior to initiation of speech, and then compressed by elastic recoil force by the tissues of the respiratory system to the level of the functional residual capacity (FRC). The lung pressure during speech is kept nearly constant except for the tendency of utterance initial rise and final lowering. In natural speech, stress and emphasis add local pressure increases. The constant lung pressure is due to the actions of the inspiratory muscles to prevent excessive airflow and maintain the long expiratory phase. As speech continues, the lung volume decreases gradually below the level of FRC, and the lung pressure is then maintained by the actions of the expiratory muscles that actively expel air from the lung. It has been argued whether the initiation of speech involves only the elastic recoil forces of the thorax to generate expiratory airflow. Indeed, a few studies have suggested that not only the thoracic system but also the abdominal system assists the regulation of expiration during speech [2.2, 3], as shown by the dashed line in Fig. 2.2. Thus, the contemporary view of speech respiration emphasizes that expiration of air during speech is not a passive process but a controlled one with co-activation of the inspiratory and expiratory muscles.

2.2.2 Structure of the Larynx

The larynx is a small cervical organ located at the top of the trachea making a junction to the pharyngeal cavity: it primarily functions to prevent foreign material from entering the lungs. The larynx contains several rigid structures such as the cricoid, thyroid, arytenoid, epiglottic, and other smaller cartilages. Figure 2.3a shows the arrangement of the major cartilages and the hyoid bone. The cricoid cartilage is ring-shaped and supports the lumen of the laryngeal cavity. It offers two bilateral articulations to the thyroid and arytenoid cartilages at the cricothyroid and cricoarytenoid joints, respectively. The thyroid cartilage is a shield-like structure that offers attachments to the vocal folds and the vestibular folds. The arytenoid cartilages are bilateral tetrahedral cartilages that change in location and orientation between phonation and respiration. The whole larynx is mechanically suspended from the hyoid bone by muscles and ligaments.

The gap between the free edges of the vocal folds is called the *glottis*. The space is divided into two portions by the vocal processes of the arytenoid cartilages: the membranous portion in front (essential for vibration) and cartilaginous portion in back (essential for respiration). The glottis changes its form in various ways during speech: it narrows by adduction and widens by abduction of the vocal folds. Figure 2.3b shows that this movement is carried out by the actions of the intrinsic laryngeal muscles that attach to the arytenoid cartilages. These muscles are functionally divided into the adductor and abductor muscles. The adductor muscles include the thyroarytenoid muscles, lateral cricoarytenoid, and arytenoid muscles, and the abductor muscle is the posterior cricoarytenoid muscle. The glottis also changes in length according to the length of the vocal folds, which takes place mainly at the membranous portion. The length of the glottis shows a large developmental sexual variation. The membranous length on average is 10 mm in adult females and 16 mm in adult males, while the cartilaginous length is about 3 mm for both [2.4].

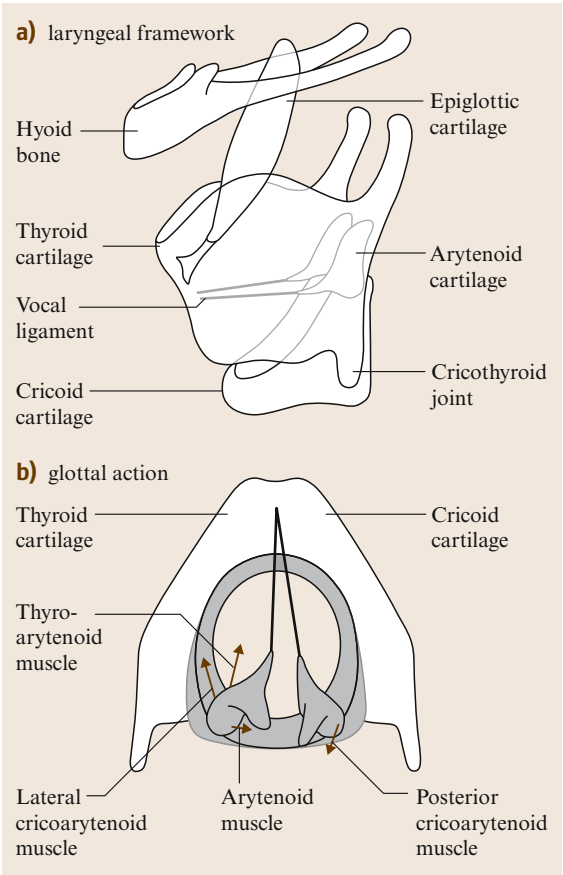


Fig. 2.3a,b Laryngeal framework and internal structures. **(a)** Oblique view of the laryngeal framework, which includes the hyoid bone and four major cartilages. **(b)** Adduction (*left*) and abduction (*right*) of the glottis and the effects of the intrinsic laryngeal muscles

2.2.3 Vocal Fold and its Oscillation

The larynx includes several structures such as the subglottic dome, vocal folds, ventricles, vestibular folds, epiglottis, and aryepiglottic folds, as shown in Fig. 2.4a. The vocal folds run anteroposteriorly from the vocal processes of the arytenoid cartilages to the internal surface of the thyroid cartilage. The vocal fold tissue consists of the thyroarytenoid muscle, vocal ligament, lamina propria, and mucous membrane. They form a special layer structure that yields to aerodynamic forces to oscillate, which is often described as the *body-cover* structure [2.5].

During voiced speech sounds, the vocal folds are set into vibration by pressurized air passing through the membranous portion of the narrowed glottis. The glottal airflow thus generated induces wave-like motion of the vocal fold membrane, which appears to propagate from the bottom to the top of the vocal fold edges. When this oscillatory motion builds up, the vocal fold membranes on either side come into contact with each other, resulting in repetitive closing and opening of the glottis. Figure 2.4b shows that vocal fold vibration repeats four phases within a cycle: the closed phase, opening phase, open phase, and closing phase. The conditions that determine vocal fold vibration are the stiffness and mass of the vocal folds, the width of the glottis, and the pressure difference across the glottis.

The aerodynamic parameters that regulate vocal fold vibration are the transglottal pressure difference and glottal airflow. The former coincides with the measure of subglottal pressure during mid and low vowels, which is about 5–10 cm H_2O in comfortable loudness and pitch (1 cm H_2O = 0.98 hPa). The latter also coincides with the average measure of oral

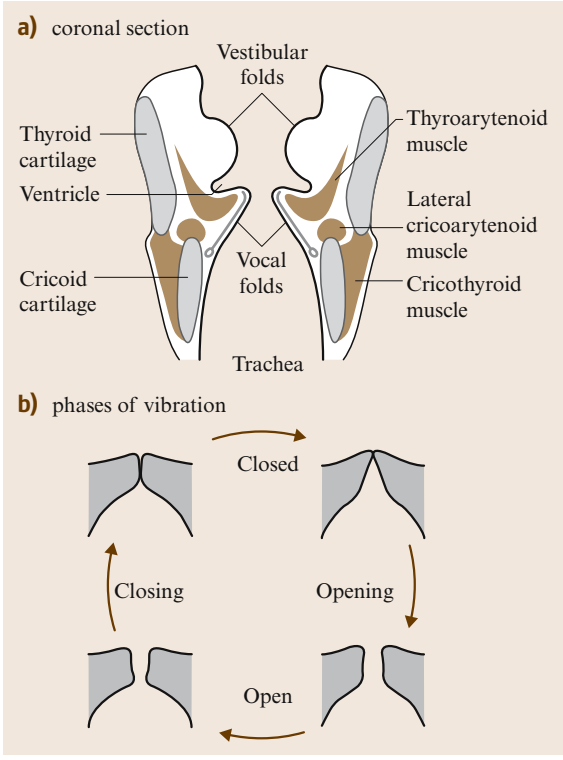


Fig. 2.4a,b Vocal folds and their vibration pattern. **(a)** Coronal section of the larynx, showing the tissues of the vocal and vestibular (false) folds. The cavity of the larynx includes supraglottic and subglottic regions. **(b)** Vocal-fold vibration pattern and glottal shapes in open phases. As the vocal-fold edge deforms in a glottal cycle, the glottis follows four phases: closed, opening, open and closing

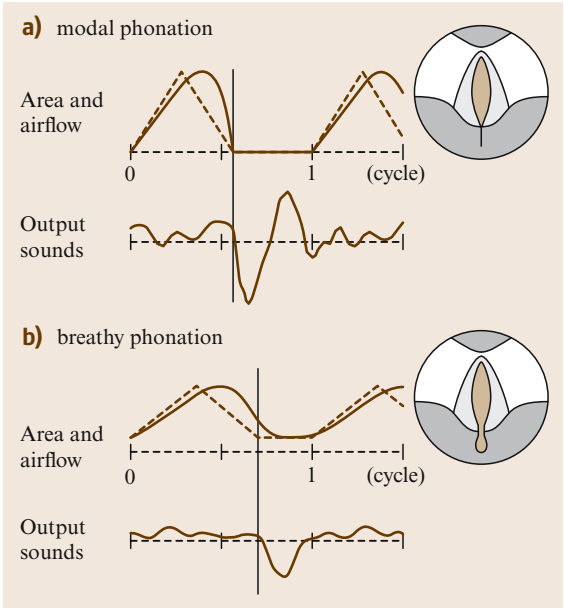


Fig. 2.5a,b Changes in glottal area and airflow in relation to output sounds during 1.5 glottal cycles from glottal opening, with glottal shapes at peak opening (in the circles). **(a)** In modal phonation with complete glottal closure in the closed phase, glottal closure causes abrupt shut-off of glottal airflow and strong excitation of the air in the vocal tract during the closed phase. **(b)** In breathy phonation, the glottal closure is incomplete, and the airflow wave includes a DC component, which results in weak excitation of the tract

airflow during vowel production, which is roughly 0.1–0.2 l/s. These values show a large individual variation: the pressure range is 4.2–9.6 cm H_2O in males and 4.4–7.6 cm H_2O in females, while the airflow rate ranges between 0.1–0.3 l/s in males and 0.09–0.2 l/s in females [2.6].

Figure 2.5 shows schematically the relationship between the glottal cycle and volumic airflow change in normal and breathy phonation. The airflow varies within each glottal cycle, reflecting the cyclic variation of the glottal area and subglottal pressure. The glottal area curve roughly shows a triangular pattern, while the airflow curve shows a skew of the peak to the right due to the inertia of the air mass within the glottis [2.7]. The closure of the glottis causes a discontinuous decrease of the glottal airflow to zero, which contributes the main source of vocal tract excitation, as shown in Fig. 2.5a. When the glottal closure is more abrupt, the output sounds are more intense with richer harmonic components [2.8]. When the glottal closure is incomplete in soft and breathy voices or the cartilaginous portion of the glottis is open to show the *glottal chink*, the airflow includes a direct-current (DC) component and exhibits a gradual decrease of airflow, which results in a more sinusoidal waveform and a lower intensity of the output sounds, as shown in Fig. 2.5b.

Laryngeal control of the oscillatory patterns of the vocal folds is one of the major factors in voice quality control. In sharp voice, the open phase of the glottal cycle becomes shorter, while in soft voice, the open phase becomes longer. The ratio of the open phase within a glottal cycle is called the *open quotient* (OQ), and the ratio of the closing slope to the opening slope in the glottal cycle is called the *speed quotient* (SQ). These two parameters determine the slope of the spectral envelope. When the open phase is longer (high OQ) with a longer closing phase (low SQ), the glottal airflow becomes more sinusoidal, with weak harmonic components. Contrarily, when the open phase is shorter (low OQ), glottal airflow builds up to pulsating waves with rich harmonics. In modal voice, all the vocal fold layers are involved in vibration, and the membranous glottis is completely closed during the closed phase of each cycle. In falsetto, only the edges of the vocal folds vibrate, glottal closure becomes incomplete, and harmonic components reduce remarkably.

The oscillation of the vocal folds during natural speech is quasiperiodic, and cycle-to-cycle variation are observed in speech waveforms as two types of measures: *jitter* (frequency perturbation) and *shimmer* (amplitude perturbation). These irregularities appear to arise from combinations of biomechanical (vocal fold asymmetry), neurogenic (involuntary activities of laryngeal muscles), and aerodynamic (fluctuations of airflow and subglottal pressure) factors. In sustained phonation of normal voice, the jitter is about 1% in frequency, and the shimmer is about 6% in amplitude.

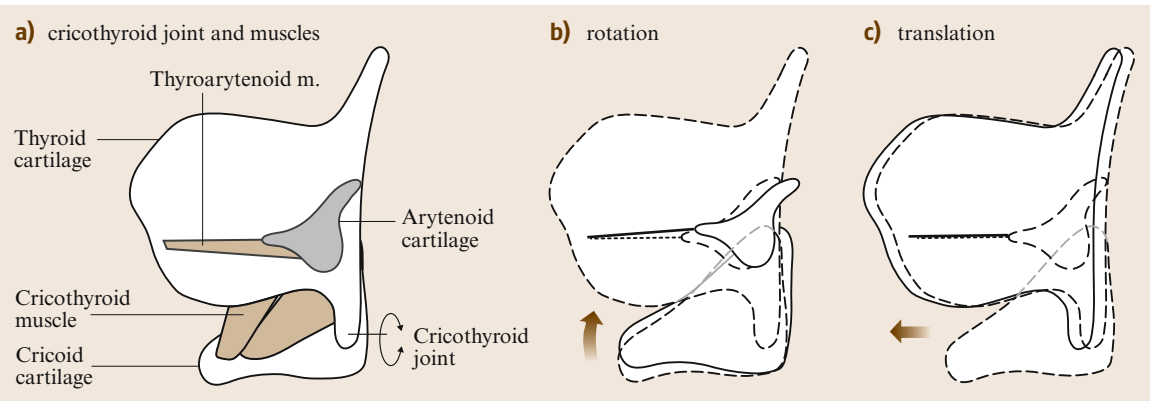


Fig. 2.6a–c Cricothyroid joint and F_0 regulation mechanism. **(a)** The cricothyroid joint is locally controlled by the thyroarytenoid and two parts of the cricothyroid muscles: Pars recta (anterior) and pars obliqua (posterior). As F_0 rises, the thyroid cartilage advances and cricoid cartilage rotates to the direction to stretch the vocal folds, which leads to the increases in the stiffness of vocal fold tissue and in the natural resonance frequency of the vocal folds. **(b)** Rotation of the cricothyroid joint is caused mainly by the action of the pars recta to raise the cricoid arch. **(c)** Translation of the joint is produced mainly by the pars obliqua

2.2.4 Regulation of Fundamental Frequency (F_0)

The fundamental frequency (F_0) of voice is the lowest harmonic component in voiced sounds, which conforms to the natural frequency of vocal fold vibration. F_0 changes depending on two factors: regulation of the length of the vocal folds and adjustment of aerodynamic factors that satisfy the conditions necessary for vocal fold vibration. In high F_0, the vocal folds become thinner and longer; while in low F_0, the vocal folds become shorter and thicker. As the vocal folds are stretched by separating their two attachments (the anterior commissure and vocal processes), the mass per unit length of the vocal fold tissue is reduced while the stiffness of the tissue layer involved in vibration increases. Thus, the mass is smaller and the stiffness is greater for higher F_0 than lower F_0, and it follows that the characteristic frequency of vibrating tissue increases for higher F_0. The length of the vocal folds is adjusted by relative movement of the cricoid and thyroid cartilages. Its natural length is a determinant factor of individual difference in F_0. The possible range of F_0 in adult speakers is about 80–400 Hz in males, and about 120–800 Hz in females.

The thyroid and cricoid cartilages are articulated at the cricothyroid joint. Any external forces applied to this joint cause rotation and translation (sliding) of the joint, which alters the length of the vocal folds. It is well known that the two joint actions are brought about by the contraction of the cricothyroid muscle to approximate the two cartilages at their front edges. Figure 2.6 shows two possible actions of the cricothyroid muscle on the joint: rotation by the pars recta and translation of the pars obliqua [2.9]. Questions still remain as to whether each part of the cricothyroid conducts pure actions of rotation or translation, and as to which part is more responsible for determining F_0.

The extrinsic laryngeal muscles can also apply external forces to this joint as a supplementary mechanism for regulating F_0 [2.10]. The most well known among the activities of the extrinsic muscles in this regulation is the transient action of the sternohyoid muscle observed as F_0 falls. Since this muscle pulls down the hyoid bone to lower the entire larynx, larynx lowering has long been thought to play a certain role in F_0 lowering. Figure 2.7 shows a possible mechanism of F_0 lowering by vertical larynx movement revealed by magnetic resonance imaging (MRI). As the cricoid cartilage descends along the anterior surface of the cervical spine, the cartilage rotates in a direction that shortens the vocal folds because the cervical spine shows anterior convexity at the level of the cricoid cartilage [2.11].

Aerodynamic conditions are an additional factor that alters F_0, as seen in the local rises of the subglottal pressure during speech at stress or emphasis. The increase of the subglottal air pressure results in a larger airflow rate and a wider opening of the glottis, which causes greater deformation of the vocal folds with larger average tissue stiffness. The rate of F_0 increase due to the subglottal pressure is reported to be about 2–5 Hz/cmH_2O when the chest cavity is compressed externally, and is observed to be 5–15 Hz/cmH_2O, when

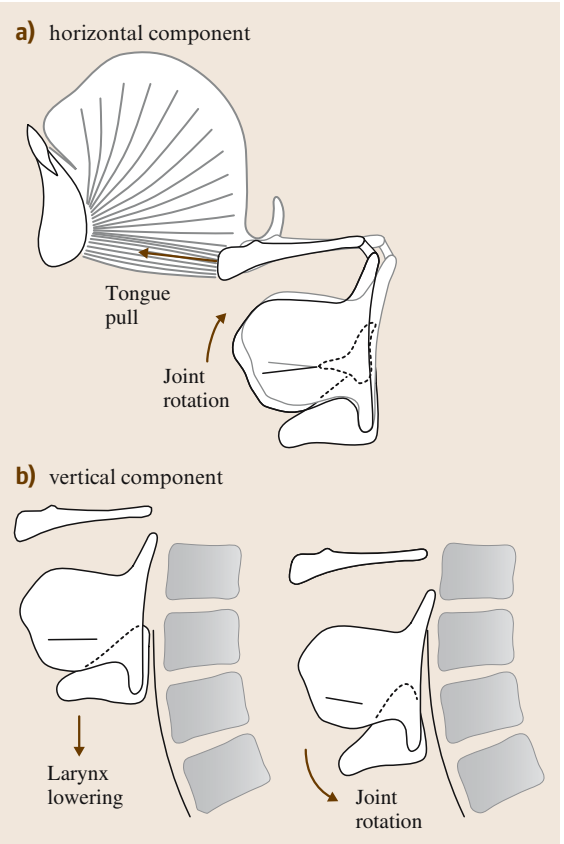


Fig. 2.7a,b Extrinsic control of F_0. Actions of the cricothyroid joint are determined not only by the cricothyroid muscle but also by other laryngeal muscles. Any external forces applied to the joint can activate the actions of the joint. **(a)** In F_0 raising, advancement of the hyoid bone possibly apply a force to rotate the thyroid cartilage. **(b)** In F_0 lowering, the cricoid cartilage rotates as its posterior plate descends along the anterior convexity of the cervical spine

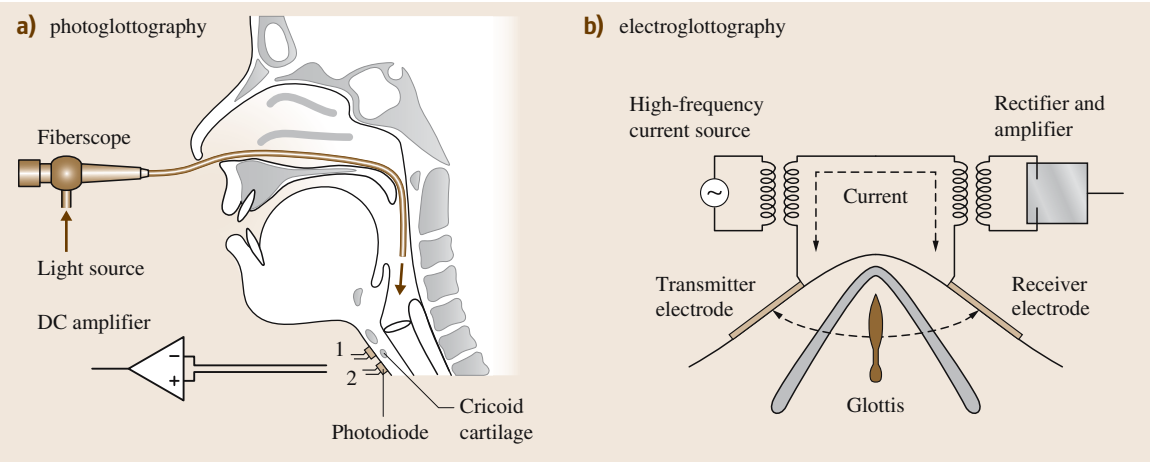


Fig. 2.8a,b Glottographic methods. **(a)** PGG with fiberscopy uses a photodetector attached near the cricothyroid cartilage in two locations: one attachment for measuring vibrations, and two attachment for glottal gestures. **(b)** EGG uses a pair of electrodes on the skin above the thyroid lamina to form a induction circuit to record electrical currents passed through the vocal-fold edges

it is measured between the beginning and end of speech utterances.

2.2.5 Methods for Measuring Voice Production

Speech production mechanisms arise from the functions of the internal organs of the human body that are mostly invisible. Therefore, better understanding of speech production processes relies on the development of observation techniques. The lung functions in speech can be assessed by the tools for aerodynamic measurements, while examination of the larynx functions during speech requires special techniques for imaging and signal recording.

Monitoring Respiratory Functions

Respiratory functions during speech are examined by recording aerodynamic measurements of lung volume, airflow, and pressure. Changes in lung volume are monitored with several types of plethysmography (e.g., whole-body, induction, and magnetic). The airflow from the mouth is measured with pneumotachography using a mask with pressure probes (differential-pressure anemometry) or thermal probes (hot-wire anemometry). Measurements of the subglottal pressure require a tracheal puncture of a needle with a pressure sensor or a thin catheter-type pressure transducer inserted from the nostril to the trachea via the cartilaginous part of the glottis.

Laryngeal Endoscopy

Imaging of the vocal folds during speech has been conducted with a combination of an endoscope and video camera. A solid-type endoscope is capable of observing vocal fold vibration with stroboscopic or real-time digital imaging techniques during sustained phonation. The flexible endoscope is beneficial for video recording of glottal movements during speech with a fiber optic bundle inserted into the pharynx through the nostril via the velopharyngeal port. Recently, an electronic type of flexible endoscope with a built-in image sensor has become available.

Glottography

Glottography is a technique to monitor vocal fold vibration as a waveform. Figure 2.8 shows two types of glottographic techniques. Photoglottography (PGG) detects light intensity modulated by the glottis using an optical sensor. The sensor is placed on the neck and a flexible endoscope is used as a light source. The signal from the sensor corresponds to the glottal aperture size, reflecting vocal fold vibration and glottal adduction–abduction movement. Electroglottography (EGG) records the contact of the left and right vocal fold edges during vibration. High-frequency current is applied to a pair of surface electrodes placed on the skin above the thyroid lamina, which detect a varying induction current that corresponds to the change in vocal fold contact area.

2.3 Articulatory Mechanisms

Speech articulation is the most complex motor activity in humans, producing concatenations of phonemes into syllables and syllables into words using movements of the speech organs. These articulatory processes are conducted within a phrase of a single expiratory phase with continuous changes of vocal fold vibration, which is one of the human-specific characteristics of sound production mechanisms.

2.3.1 Articulatory Organs

Articulatory organs are composed of the rigid organ of the lower jaw and soft-tissue organs of the tongue, lips, and velum, as illustrated in Fig. 2.9. These organs together alter the resonance of the vocal tract in various ways and generate sound sources for consonants in the vocal tract. The tongue is the most important articulatory organ, and changes the gross configuration of the vocal tract. Deformation of the whole tongue determines vowel quality and produces palatal, velar, and pharyngeal consonants. Movements of the tongue apex and blade contribute to the differentiation of dental and alveolar consonants and the realization of retroflex consonants. The lips deform the open end of the vocal tract by various types of gestures, assisting the production of vowels and labial consonants. Actions of these soft-tissue organs are essentially based on contractions of the muscles within these organs, and their mechanism is often compared with the *muscular hydrostat*. Since the tongue and lips have attachments to the lower jaw, they are interlocked with the jaw to open the mouth. The velum controls opening and closing of the velopharyngeal port, and allows distinction between nasal and oral sounds. Additionally, the constrictor muscles of the pharynx adjust the lateral width of the pharyngeal cavity, and their actions also assist articulation for vowels and back consonants.

Upper Jaw

The upper jaw, or the maxilla with the upper teeth, is the structure fixed to the skull, forming the palatal dome on the arch of the alveolar process with the teeth. It forms a fixed wall of the vocal tract and does not belong to the articulatory organs: yet it is a critical structure for speech articulation because it provides the frame of reference for many articulatory gestures. The structures of the upper jaw offer the location for contact or approximation by many parts of the tongue such as the apex, blade, and dorsum. The phonetics literature describes the place of articulation as classified according to the locations of lingual approximation along the upper jaw for dental, alveolar, and palatal consonants. The hard palate is covered by the thick mucoperiosteum, which has several transverse lines of mucosal folds called the *palatine rugae*.

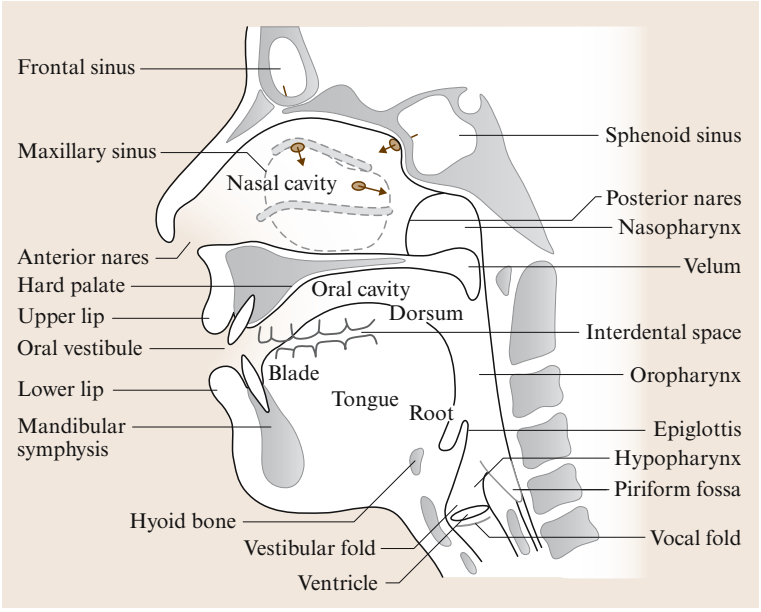


Fig. 2.9 Illustration of the articulatory system with names of articulators and cavities

Lower Jaw

The lower jaw, or the mandible with the lower teeth, is the largest rigid motor organ among the speech production apparatus. Its volume is about 100 cm^3. As well as playing the major role in opening and closing the mouth, it provides attachments for many speech muscles and supports the tongue, lips, and hyoid bone.

Figure 2.10 shows the action of the jaw and the muscles used in speech articulation. The mandible articulates with the temporal bone at the temporomandibular joint (TMJ) and brings about jaw opening–closing actions by rotation and translation. The muscles that control jaw movements are generally called the masticatory muscles. The jaw opening muscles are the digastric and lateral pterygoid muscles. The strap muscles, such as the geniohyoid and sternohyoid, also assist jaw opening. The jaw closing muscles include the masseter, temporalis, and medial pterygoid muscles. While the larger muscles play major roles in biting and chewing, comparatively small muscles are used for speech articulation. The medial pterygoid is mainly used for jaw closing in articulation, and the elastic recoil force of the connective tissues surrounding the mandible is another factor for closing the jaw from its open position.

Tongue

The tongue is an organ of complex musculature [2.12]. It consists of a round body occupying its main mass and a short blade with an apex. Its volume is approximately 100 cm^3, including the muscles in the tongue floor. The tongue body moves in the oral cavity by variously deforming its voluminous mass, while the tongue blade alters its shape and changes the angle of the tongue apex. Deformation of the tongue tissue is caused by contractions of the extrinsic and intrinsic tongue muscles, which are illustrated schematically in Fig. 2.11.

The extrinsic tongue muscles are those that arise outside of the tongue and end within the tongue tissue. This group includes four muscles, the genioglossus, hyoglossus, styloglossus, and palatoglossus muscles, although the former three muscles are thought to be involved in the articulation of the tongue. The palatoglossus muscle participates in the lowering of the velum as discussed later.

The genioglossus is the largest and strongest muscle in the tongue. It begins from the posterior aspect of the mandibular symphysis and runs along the midline of the tongue. Morphologically, it belongs to the triangular muscle, and its contraction effects differ across portions of the muscle. Therefore, the genioglossus is divided functionally into the anterior, middle, and posterior bundles. The anterior and middle bundles run midsagittally, and their contraction makes the midline groove of the tongue for the production of front vowels. The anterior bundle often makes a shallow notch on the tongue surface called the *lingual fossa* and assists elevation of the tongue apex. The middle bundle runs obliquely, and advances the tongue body for front vowels. The posterior bundle of the genioglossus runs midsagittally and

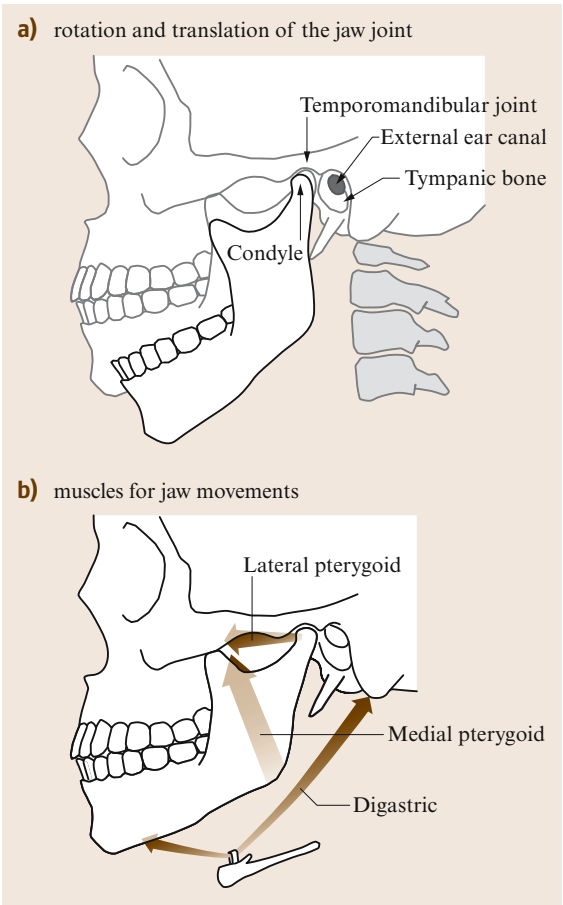


Fig. 2.10a,b Actions of the temporomandibular joint and muscles for jaw opening and closing. **(a)** The lower jaw opens by rotation and translation of the mandible at the temporomandibular joint. Jaw translation is needed for wide opening of the jaw because jaw rotation is limited by the narrow space between the condyle and tympanic bone. **(b)** Jaw opening in speech depends on the actions of the digastric and medial pterygoid muscles with support of the strap muscles. Jaw closing is carried out by the contraction of the lateral pterygoid muscle and elastic recoil forces of the tissues surrounding the jaw

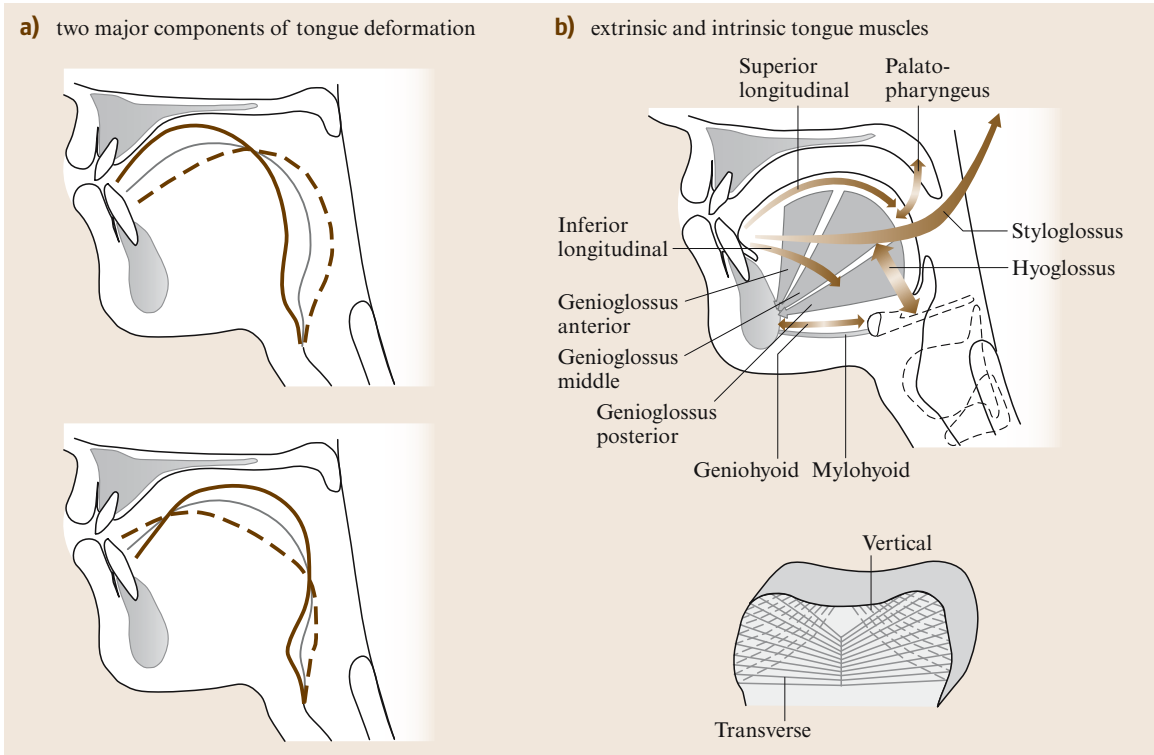


Fig. 2.11a,b Actions of the tongue and its musculature. **(a)** Major components of tongue deformation are high-front vs.low-back (*top*) and high back versus low front (*bottom*) motions, (after [2.14]). **(b)** Lateral view (*top*) shows the extrinsic and intrinsic muscles of the tongue with two tongue floor muscles. Coronal section (*bottom*) shows additional intrinsic muscles

also spreads laterally, reaching a wide area of the tongue root. This bundle draws the tongue root forward and elevates the upper surface of the tongue for high vowels and anterior types of oral consonants. The hyoglossus is a bilateral thin-sheet muscle, which arises from the hyoid bone, runs upward along the sides of the tongue, and ends in the tongue tissue, intermingling with the styloglossus. Its contraction lowers the tongue dorsum and pushes the tongue root backward for the production of low vowels. The styloglossus is a bilateral long muscle originating from the styloid process on the skull base, running obliquely to enter the back sides of the tongue. Within the tongue, it runs forward to reach the apex of the tongue, while branching downward to the hyoid bone and medially toward the midline. Although the extra-lingual bundle of the styloglossus runs obliquely, it pulls the tongue body straight back at the insertion point because the bundle is surrounded by fatty and muscular tissues. The shortening of the intra-lingual bundle draws the tongue apex backward and causes an upward bunching of the tongue body [2.13]. Each of the extrinsic tongue muscles has two functions: drawing of the relevant attachment point toward the origin, and deforming the tongue tissue in the orthogonal orientation. The resulting deformation of the tongue can be explained by two antagonistic pairs of extrinsic muscles: posterior genioglossus versus styloglossus, and anterior genioglossus versus hyoglossus. The muscle arrangement appears to be suitable for tongue body movements in the vertical and horizontal dimensions.

The intrinsic tongue muscle is a group of muscles that have both their origin and termination within the tongue tissue. They include four bilateral muscles: the superior longitudinal, inferior longitudinal, transverse, and vertical muscles. The superior and inferior longitudinal muscles operate on the tongue blade to produce vertical and horizontal movements of the tongue tip. The transverse and vertical muscles together compress the tongue tissue medially to change the cross-sectional shape of the tongue.

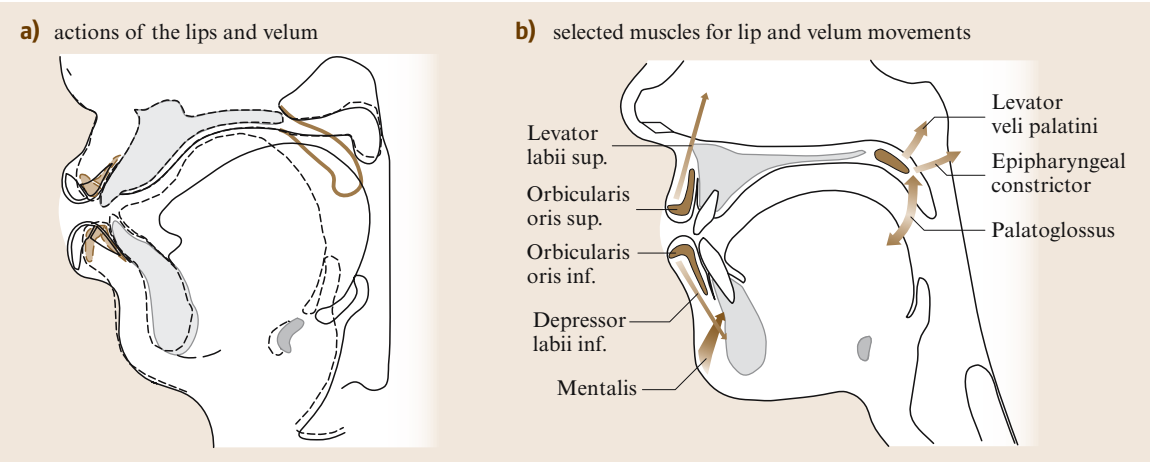


Fig. 2.12a,b Actions of the lips and velum, and their muscles. **(a)** Trace of MRI data in the production of /i/ and /u/ with lip protrusion show that two parts of the orbicularis oris, marginal (*front*) and peripheral (*back*) bundles demonstrate their geometrical changes within the vermillion tissue. The shapes of the velum also vary greatly between the rest position (thick gray line) and vowel articulation. **(b)** Five labial muscles are shown selectively from among many facial muscles. The velum shape is determined by the elevator, constrictor, and depressor (palatopharyngeus)

There are two muscles that support the tongue floor: the geniohyoid and mylohyoid muscles. The geniohyoid runs from the genial process of the mandibular symphysis to the body of the hyoid bone. This muscle has two functions: opening the jaw for open vowels and advancing the hyoid bone to help raise F_0. The mylohyoid is a sheet-like muscle beneath the tongue body that stretches between the mandible and the hyoid bone to support the entire tongue floor. This muscle supports the tongue floor to assist articulation of high front vowels and oral consonants.

Lips and Velum

The lips are a pair of soft-tissue organs consisting of many muscles. Their functions resemble those of the tongue because they partly adhere to the mandible and partly run within the soft tissue of the lips. The vermillion, or the part of red skin, is the unique feature of the human lips, which transmits phonetic signals visually. The deformation of the lips in speech can be divided into three components. The first is opening/closing of the lip aperture, which is augmented by jaw movement. The second is rounding/spreading of the lip tissue, produced by the changes in their left–right dimension. The third is protrusion/retraction of the lip gesture, generated by three-dimensional deformation of the entire lip tissue.

The muscles that cause deformation of the lips are numerous. Figure 2.12 shows only a few representative muscles of the lips. The orbicularis oris is the muscle that surrounds the lips, consisting of two portions; the marginal and peripheral bundles. Contraction of the marginal bundles near the vermillion borders is thought to produce lip rounding without protrusion. Contraction of the peripheral bundles that run in the region around the marginal bundles compresses the lip tissue circumferentially to advance the vermillion in lip protrusion [2.15]. The mentalis arises from the mental part of the mandible to the lip surface, and its contraction elevates the lower lip by pulling the skin at the mental region. The levator labii superior elevates the upper lip, and the depressor labii inferior depresses the lower lip relative to the jaw. The superior and inferior angli oris muscles move the lip corners up and down, respectively, which makes facial expressions rather than speech articulation.

The exact mechanism of lip protrusion is still in question. Tissue bunching by muscle shortening as a general rule for the organs of muscle does not fully apply to the phenomenon of lip protrusion. This is because, as the vermilion thickens in lip protrusion, it does not compress on the teeth; its dental surface often detaches from the teeth (Figure 2.12a). A certain three-dimensional stress distribution within the entire labial tissue must be considered to account for the causal factors of lip protrusion.

The velum, or the soft palate, works as a valve behind the hard palate to control the velopharyngeal port, as shown in Fig. 2.12a. Elevation of the velum is carried

out during the production of oral sounds, while lowering takes place during the production of nasal sounds. The action of the velum to close the velopharyngeal port is not a pure hinge motion but is accompanied by the deformation of the velum tissue with narrowing of the nasopharyngeal wall. In velopharyngeal closure, the levator veli palatine contracts to elevate the velum, and the superior pharyngeal constrictor muscle produces concentric narrowing of the port. In velopharyngeal opening, the palatoglossus muscle assists active lowering of the velum.

2.3.2 Vocal Tract and Nasal Cavity

The vocal tract is an acoustic space where source sounds for speech propagate. Vowels and consonants rely on strengthening or weakening of the spectral components of the source sound by resonance of the air column in the vocal tract. In the broad definition, the vocal tract includes all the air spaces where acoustic pressure variation takes place in speech production. In this sense, the vocal tract divides into three regions: the subglottal tract, the tract from the glottis to the lips, and the nasal cavities.

The subglottal tract is the lower respiratory tract below the glottis down to the lungs via the trachea and bronchial tubes. The length of the trachea from the glottis to the carina is 10–15 cm in adults, including the length of the subglottic laryngeal cavity (about 2 cm). Vocal source sounds propagate from the glottis to the trachea, causing the subglottal resonance in speech spectra. The resonance frequencies of the subglottal airway are estimated to be 640, 1400, and 2100 Hz [2.16]. The second subglottal resonance is often observed below the second formant of high vowels.

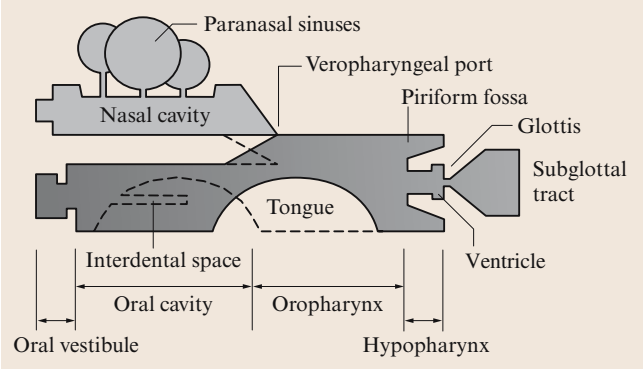


Fig. 2.13 Acoustic design of the vocal tract. Passages from the subglottal tract to two output ends at the lips and nares are shown with the effects of tongue and velar movements. The resonance of the vocal tract above the supraglottic laryngeal cavity determines major the vowel formants (F_1, F_2, and F_3). The resonance of the subglottal tract and interdental space interacts with the vowel formants, while the hypopharyngeal cavities and other small cavities cause local resonances and antiresonances in the higher-frequency region

The vocal tract, according to the conventional definition, is the passage of vocal sounds from the glottis to the lips, where source sounds propagate and give rise to the major resonances. The representative values for the length of the main vocal tract from the glottis to the lips are 15 cm in adult females and 17.5 cm in adult males. According to the measurement data based on the younger population, vocal tract lengths are 14 cm in females and 16.5 cm in males [2.17, 18], which are shorter than the above values. Considering the elongation of the vocal tract during a course of life, the above representative values appear reasonable. While the oral cavity length is maintained by the rigid structures of the skull and jaw, the pharyngeal cavity length increases due to larynx lowering before and after puberty. Thus, elongation of the pharyngeal cavity is the major factor in the developmental variation in vocal tract length.

The vocal tract anatomically divides into four segments: the hypopharyngeal cavities, the mesopharynx, the oral cavity, and the oral vestibule (lip tube). The hypopharyngeal part of the vocal tract consists of the supraglottic laryngeal cavity (2 cm long) and the bilateral conical cavities of the piriform fossa (2 cm long). The mesopharynx extends from the aryepiglottic fold to the anterior palatal arch. The oral cavity is the segment from the anterior palatal arch to the incisors. The oral vestibule extends from the incisors to the lip opening. The latter shows an anterior convexity, which often makes it difficult to measure the exact location of lip opening.

The vocal tract is not a simple uniaxial tube but has a complex three-dimensional construction. The immobile wall of the vocal tract includes the dental arch and the palatal dome. The posterior pharyngeal wall is almost rigid, but it allows subtle changes in convexity and orientation. The soft walls include the entire tongue surface, the velum with the uvula, the lateral pharyngeal wall, and the lip tube. The shape of the vocal tract varies individually due to a few factors. First, the lateral width of the upper and lower jaws relative to the pharyngeal cavity width affects tongue articulation and results in a large individual variation of vocal tract shape observed midsagittally. Second, the mobility of the jaw

depending on the location of the mandibular symphysis relative to the skull can vary the openness of vowels. Third, the size of the tongue relative to the oral and pharyngeal cavities varies individually; the larger the tongue size, the smaller the articulatory space for vowels.

Figure 2.13 shows a schematic drawing of the vocal tract and nasal cavity. The vocal tract has nearly constant branches such as the piriform fossa (entrance to the esophagus) and the vallecula (between the tongue root and epiglottis). The vocal tract also has controlled branches to the nasal cavity at the velopharyngeal port and to the *interdental space* (the space bounded by the upper and lower teeth and the lateral cheek wall). The latter forms a pair of side-branches when the tongue is in a higher position as in /i/ or /e/, while it is unified with the oral cavity when the tongue is in a lower position as in /a/.

The nasal cavity is an accessory channel to the main vocal tract. Its horizontal dimension from the anterior nares to the posterior wall of the epipharynx is approximately 10–11 cm. The nasal cavity can be divided into the single-tube segment (the velopharyngeal region and epipharynx) and the dual-tube segment (the nasal cavity proper and nasal vestibule). Each of the bilateral channels of the nasal cavity proper has a complex shape of walls with the three turbinates with thick mucous membrane, which makes a narrower cross section compared with the epipharyngeal area [2.19]. The nasal cavity has its own side-branches of the paranasal sinuses; the maxillary, sphenoid, ethomoid, and frontal sinuses.

The nasal cavity builds nasal resonance to accomplish phonetic features of nasal sounds and nasalized vowels. The paranasal sinuses also contribute to acoustic characteristics of the nasal sounds. The nasal murmur results from these characteristics: a Helmholtz resonance of the entire nasopharyngeal tract from the glottis to the anterior nares and regional Helmholtz resonances caused by the paranasal sinuses, together characterized by a resonance peak at 200–300 Hz and spectral flattening up to 2 kHz [2.20, 21]. The nasal resonance could takes place even in oral vowels with a complete closure of the velopharyngeal port: the soft tissue of the velum transmits the pressure variation in the oral cavity to the nasal cavity, which would enhances sound radiation for close vowels and voiced stops.

2.3.3 Aspects of Articulation in Relation to Voicing

Here we consider a few phonetic evidences that can be considered as joint products of articulation and phonation. Vowel production is the typical example for this topic, in view of its interaction with the larynx. Regulation of voice quality, which has been thought to be a laryngeal phenomenon, is largely affected by the lower part of the vocal tract. The voiced versus voiceless distinction is a pertinent issue of phonetics that involves both phonatory and articulatory mechanisms.

Production of Vowels

The production of vowels is the result of the joint action of phonatory and articulatory mechanisms. In this process, the larynx functions as a source generator, and the vocal tract plays the role of an acoustic filter to modulate the source sounds and radiate from the lip opening, as described by the *source-filter* theory [2.22, 23]. The quality of oral vowels is determined by a few peak frequencies of vocal tract resonance (formants). In vowel production, the vocal tract forms a *closed tube* with the closed end at the glottis and the open end at the lip opening. Multiple reflections of sound wave between the two ends of the vocal tract give rise to vowel formants (F_1, F_2, F_3). The source-filter theory has been supported by many studies as the fundamental concept explaining the acoustic process of speech production, which is further discussed in the next section.

Vowel articulation is the setup for the articulatory organs to determine vocal tract shape for each vowel. When the jaw is in a high position and the tongue is in a high front position, the vocal tract assumes the shape for /i/. Contrarily, when the jaw is in a low position and the tongue is in a low back position, the vocal tract takes the shape for /a/. The articulatory organ that greatly influences vocal tract shape for vowels is the tongue. When the vocal tract is modeled as a tube with two segments (front and back cavities), the movements of the tongue body between its low back and high front positions creates contrasting diverging and converging shapes of the main vocal tract. Jaw movement enhances these changes in the front cavity volume, while pharyngeal constriction assists in the back cavity volume. When the vocal tract is modeled as a tube with three segments, the movements of the tongue body between its high back and low front positions determine the constriction or widening of the vocal tract in its middle portion. The velum also contributes to the articulation of open vowels by decreasing the area of the vocal tract at the velum or making a narrow branch to the nasal cavity. The lip tube is another factor for vowel articulation that determines the vocal tract area near the open end.

Although muscular control for vowel articulation is complex, a simplified view can be drawn based

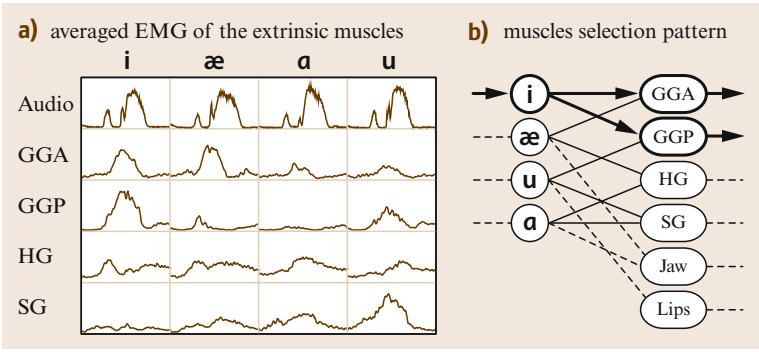


Fig. 2.14a,b Tongue EMG data during VCV utterances and muscle selection pattern in vowel articulation. **(a)** Averaged EMG data for four English corner vowels are shown for the major muscles of the tongue: the anterior genioglossus (GGA), posterior genioglossus (GGP), hyoglossus (HG), and styloglossus (SG). **(b)** The systematic variation observed in the muscle–vowel matrix suggests a muscle selection pattern

on electromyographic (EMG) data obtained from the tongue muscles [2.24]. Figure 2.14a shows a systematic pattern of muscle activities for CVC (consonant-vowel-consonant) utterances with /p/ and four English corner vowels. The anterior and posterior genioglossus are active for front vowels, while the styloglossus and hyoglossus are active for back vowels. These muscles also show a variation depending on vowel height. These observations are shown schematically in Fig. 2.14b: the basic control pattern for vowel articulation is the selection of two muscles among the four extrinsic muscles of the tongue [2.25].

As the tongue or jaw moves for vowel articulation, they apply forces to the surrounding organs and cause secondary effects on vowel sounds. For example, articulation of high vowels such as /i/ and /u/ is mainly produced by contraction of the posterior genioglossus, which is accompanied by forward movement of the hyoid bone. This action applies a force to rotate the thyroid cartilage in a direction that stretches the vocal folds. In evidence, higher vowels tend to have a higher F_0, known as the *intrinsic vowel* F_0 [2.26,27]. When the jaw opens to produce open vowels, jaw rotation compresses the tissue behind the mandibular symphysis, which applies a force to rotate the thyroid cartilage in the opposite direction, thereby shortening the vocal folds. Thus, the jaw opening has the secondary effect of lowering the intrinsic F_0 for lower vowels.

Supra-Laryngeal Control of Voice Quality

The laryngeal mechanisms controlling voice quality were described in an earlier section. In this section, the supra-laryngeal factors are discussed. Recent studies have shown evidence that the resonances of the hypopharyngeal cavities determine the spectral envelope in the higher frequencies above 2.5 kHz by causing an extra resonance and antiresonances [2.28–31]. The hypopharyngeal cavities include a pair of vocal-tract

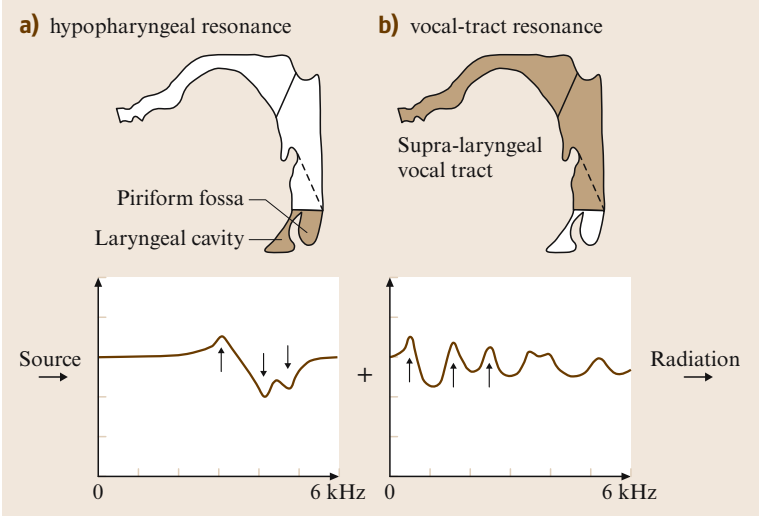


Fig. 2.15a,b Vocal-tract resonance with hypopharyngeal cavity coupling in vowel production. **(a)** The supraglottal laryngeal cavity contributes a resonance peak at 3–3.5 kHz, and the bilateral cavities of the piriform fossa cause antiresonances at 4–5 kHz. **(b)** The main vocal tract above the laryngeal cavity determines the major vowel formants

side-branches formed by the piriform fossa. Each fossa maintains a relatively constant cavity during speech, which is collapsed only in deep inhalation by the wide abduction of the arytenoid cartilage. The piriform fossa causes one or two obvious antiresonances in the higher frequencies above 4 kHz [2.29] and affects the surrounding formants. The laryngeal cavity above the vocal folds also contributes to shaping the higher frequencies [2.28, 32]. The supraglottic laryngeal cavity, from the ventricles to the aryepiglottic folds via the ventricular folds, forms a type of Helmholtz resonator and gives rise to a resonance at higher frequencies of 3–3.5 kHz. This resonance can be counted as the fourth formant (F4) but it is actually an *extra formant* to the resonance of the vocal tract above the laryngeal cavity [2.30]. When the glottis opens in the open phase of vocal fold vibration, the supraglottic laryngeal cavity no longer constitutes a typical Helmholtz resonator, and demonstrates a strong damping of the resonance, which is observed as the disappearance of the affiliated extra formant. Therefore, the laryngeal cavity resonance shows a cyclic nature during vocal fold vibration, and it is possibly absent in breathy phonation or pathological conditions with insufficient glottal closure [2.31]. Figure 2.15 shows an acoustic model of the vocal tract to illustrate this coupling of the hypopharyngeal cavities.

The hypopharyngeal cavities are not an entirely fixed structure but vary due to physiological efforts to control F_0 and voice quality. A typical case of the hypopharyngeal adjustment of voice quality is found in the *singing formant* [2.28]. When high notes are produced by opera singers, the entire larynx is pulled forward due to the advanced position of the tongue, which widens the piriform fossa to deepens the fossa's antiresonances, resulting in a decrease of the frequency of the adjacent lower formant (F_5). When the supraglottic laryngeal cavity is constricted, its resonance (F_4) comes down towards the lower formant (F_3). Consequently, the third to fifth formants come closer to each other and generate a high resonance peak observed near 3 kHz.

Regulation of Voiced and Voiceless Sounds

Voiced and voiceless sounds are often attributed to the glottal state with and without vocal fold vibration, while their phonetic characteristics result from phonatory and articulatory controls over the speech production system. In voiced vowels, the vocal tract forms a closed tube with no significant constrictions except for the narrow laryngeal cavity. On the other hand, in whispered vowels, the membranous glottis is closed, and the supraglottic laryngeal cavity forms an extremely narrow channel continued from the open cartilaginous glottis, with a moderate constriction of the lower pharynx. Devoiced vowels exhibit a wide open glottis and a reduction of tongue articulation. Phonetic distinctions of voiced and voiceless consonants further involve fine temporal control over the larynx

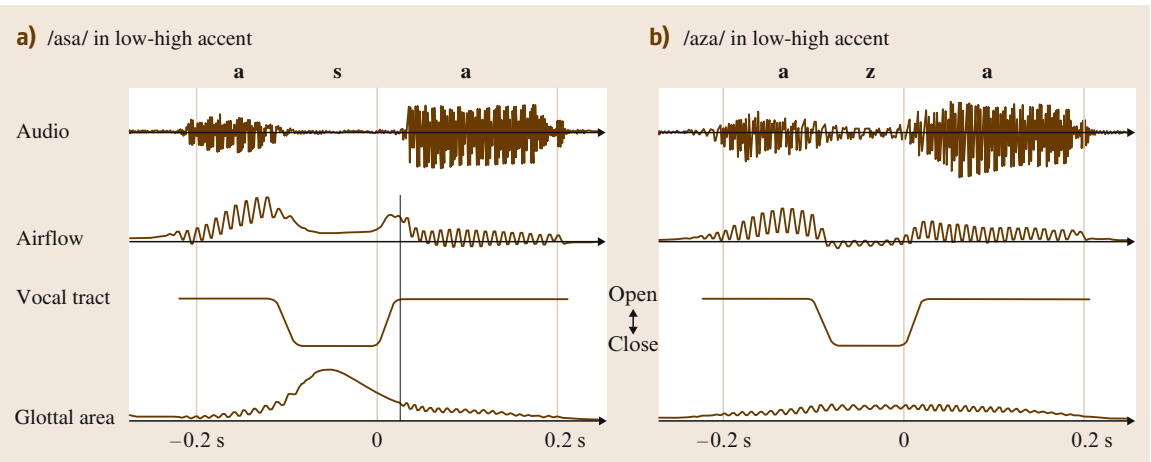


Fig. 2.16a,b Laryngeal articulatory patterns in producing VCV utterances with voiceless and voiced fricatives as in /asa/ and /aza/. From the top to bottom, speech signals, oral airflow, schematic patterns of vocal tract constriction, and glottal area variations are shown schematically. This figure is based on the author's recent experiment with anemometry with an open-type airflow transducer and photoglotto-graphy with an external lighting technique, conducted by *Dr. Shinji Maeda* (ENST) and the author

and supra-laryngeal articulators in language-specific ways.

In the production of voiced consonants, vocal fold vibration typically continues during the voiced segments. In voiced stops and fricatives, the closure or narrowing of the vocal tract results in decrease in glottal airflow and transglottal pressure difference. The glottal airflow during the stop closure is maintained during the closure due to the increases in vocal tract volume: the expansion of the oral cavity (jaw lowering and cheek wall expansion) and the expansion of the pharyngeal cavity (lateral wall expansion and larynx lowering). During the closure period, air pressure variations are radiated not only from the vocal tract wall but also from the anterior nares due to transvelar propagation of the intra-oral sound pressure into the nasal cavities.

In the production of voiceless consonants, vocal fold vibration is suppressed due to a rapid reduction of the transglottal pressure difference and abduction of the vocal folds. During stop closures, the intra-oral pressure builds up to reach the subglottal pressure, which enhances the rapid airflow after the release of the closure. Then, vocal fold vibration restarts with a delay to the release, which is observed as a long voice onset time (VOT) for voiceless stops. The process of suppressing vocal fold vibration is not merely a passive aerodynamic process on the vocal folds, but is assisted by a physiological process to control vocal fold stiffness. The cricothyroid muscle has been observed to increase its activity in producing voiceless consonants. This activity results in a high–falling F_0 pattern during the following vowel, contributing a phonetic attribute to voiceless consonants [2.33]. In glottal stops, vocal fold vibration stops due to forced adduction of the vocal folds with an effort closure of the supraglottic laryngeal cavity.

Figure 2.16 illustrates the time course of the processes during vowel-consonant-vowel (VCV) utterances with a voiceless fricative in comparison to the case with a voiced fricative. The voiceless segment initiates with glottal abduction and alveolar constriction, and vocal fold vibration gradually fades out during the phase of glottal opening. After reaching the maximum glottal abduction, the glottis enters the adduction phase, followed by the release of the alveolar constriction. Then, the glottis becomes narrower and vocal fold vibration restarts. There is the time lag between the release of the constriction and full adduction of the glottis, which results in the peak flow seen in Fig. 2.16a, presumably accompanied by aspiration sound at the glottis.

2.3.4 Articulators' Mobility and Coarticulation

The mobility of speech articulators varies across organs and contributes certain phonetic characteristics to speech sounds. Rapid movements are essential to a sequence from one distinctive feature to another, as observed in the syllable /sa/ from a narrow constriction to the vocalic opening, while gradual movements are found to produce nasals and certain labial sounds. These variations are principally due to the nature of articulators with respect to their mobility. The articulatory mechanism involves a complex system that is built up by organs with different motor characteristics. Their variation in temporal mobility may be explained by a few biological factors. The first is the phylogenetic origin of the organs: the tongue muscles share their origin with the fast motor systems such as the eyeball or finger, while other muscles such as in the lips or velum originate from the slow motor system similar to the musculature of the alimentary tract. The second is the innervation density to each muscle: the faster organs are innervated by thicker nerve bundles, and vise versa, which derives from an adaptation of the biological system to required functions. In fact, the human hypoglossal nerve that supplies the tongue muscles is much thicker than that of other members of the primate family. The third is the composition of muscle fiber types in the musculature, which varies from organ to organ. The muscles in the larynx have a high concentration of the ultrafast fibers (type 2B), while the muscle to elevate the velum predominantly contains the slow fibers (type 1). In accordance with these biological views, the rate of the articulators movement indexed by the maximum number of syllables per second follows the order of the tongue apex, body, and lips: the tongue moves at a maximum rate of 8.2 syllables per second at the apex, and 7.1 syllables per second with the back of the tongue, while the lips and facial structures move at a maximum rate of 2.5–3 syllables per second [2.34]. More recent measurements indicate that the lips are slower than the tongue apex but faster than the tongue dorsum. The velocities during speech tasks reach 166 mm/sec for the lower lip, 196 mm/sec for the tongue tip, and 129 mm/sec for the tongue dorsum [2.35]. The discrepancy between these two reports regarding the mobility of the lips may be explained by the types of movements measured: opening–closure movement by the jaw–lower lip complex is faster than the movement of the lips themselves, such as protrusion and spreading.

It is often noted that speech is characterized by asynchrony among articulatory movements, and the degree

of asynchrony varies with the feature to be realized. Each articulator does not necessarily strictly keep pace with other articulators in a syllable sequence. The physiological basis of this asynchrony may be explained by the mobility of the articulatory organs and motor precision required for the target of articulation. The slower articulators such as the lips and velum tend to exhibit marked coarticulation in production of labial and nasal sounds. In stop–vowel–nasal sequences (such as /tan/), the velopharyngeal port is tightly closed at the stop onset and the velum begins to lower before the nasal consonant. Thus, the vowel before the nasal consonant is partly nasalized. When the vowel /u/ is preceded by /s/, the lips start to protrude during the consonant prior to the rounded vowel.

The articulators' mobility also contributes some variability to speech movements. The faster articulators such as parts of the tongue show various patterns from target undershooting to overshooting. In articulation of close–open–close vowel sequences such as /iai/, tongue movements naturally show undershooting for the open vowel. In contrast, when the alveolar voiceless stop /t/ is placed in the open vowel context as in /ata/, the tongue blade sometimes shows an extreme overshoot with a wide contact on the hard palate because such articulatory variations do not significantly affect the output sounds. On the contrary, in alveolar and postalveolar fricatives such as /s/ and /sh/, tongue movements also show a dependence on articulatory precision because the position of the tongue blade must be controlled precisely to realize the narrow passage for generating frication sounds. The lateral /l/ is similar to the stops with respect to the palatal contact, while the rhotic /r/ with no contact to the palate can show a greater extent of articulatory variations from retroflex to bunched types depending on the preceding sounds.

2.3.5 Instruments for Observing Articulatory Dynamics

X-ray and palatography have been used as common tools for articulatory observation. Custom instruments are also developed to monitor articulatory movements, such as the X-ray microbeam system and magnetic sensor system. The various types of newer medical imaging techniques are being used to visualize the movements of articulatory system using sonography and nuclear magnetic resonance. These instruments are generally large scale, although relatively compact instruments are becoming available (e.g., magnetic probing system or portable ultrasound scanner).

Palatography

The palatograph is a compact device to record temporal changes in the contact pattern of the tongue on the palate. There are traditional static and modern dynamic types. The dynamic type is called electropalatography, or dynamic palatography, which employs an individually customized palatal plate to be placed on the upper jaw. As shown in Fig. 2.17a, this system employs a palatal plate with many surface electrodes to monitor electrical contacts on the tongue's surface.

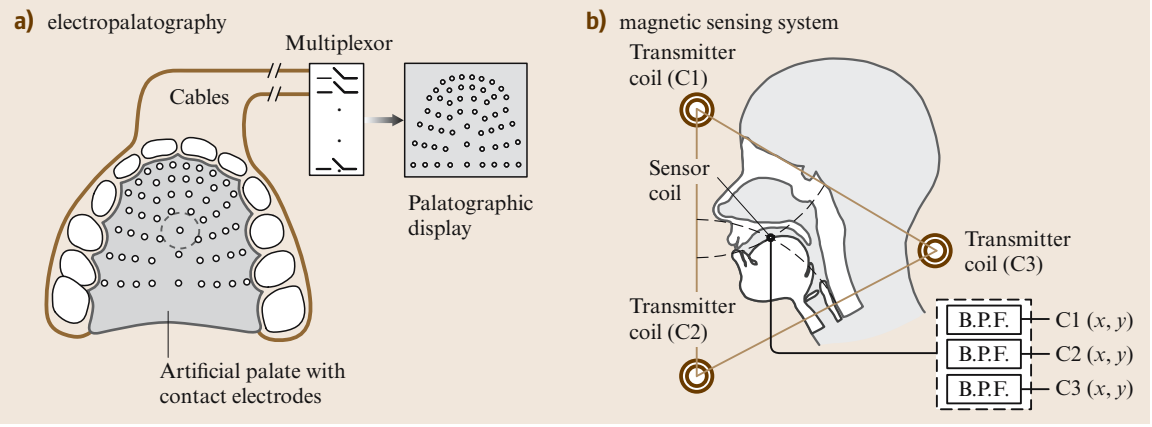


Fig. 2.17a,b Electropalatography and magnetic sensing system. **(a)** Electropalatography displays tongue–palate contact patterns by detecting weak electrical current caused by the contact between the electrodes on the artificial palate and the tongue tissue. **(b)** Magnetic sensing system is based on detection of alternate magnetic fields with different frequencies using miniature sensor coils

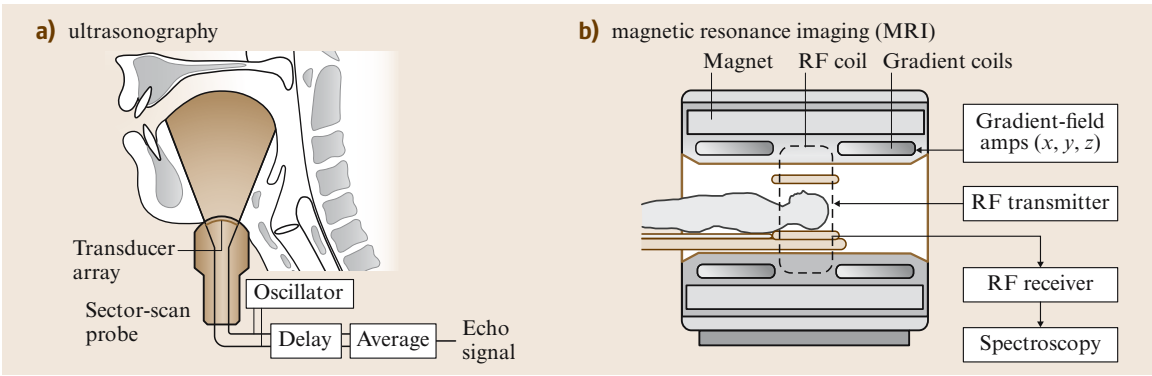


Fig. 2.18a,b Medical imaging techniques. **(a)** Ultrasound scanner uses an array of transmitters and receivers to detect echo signals from regions where the ultrasound signals reflect strongly such as at the tissue-air boundaries on the tongue surface. **(b)** Magnetic resonance imaging (MRI) generates strong static magnetic field, controlled gradient fields in the three directions, and radio-frequency (RF) pulses. Hydrogen atoms respond to the RF pulses to generate echo signals, which are detected with a receiver coil for spectral analysis

Marker Tracking System

A few custom devices have been developed to record movements of markers attached on the articulatory organs. X-ray microbeam and magnetic sensing systems belong to this category. Both can measure 10 markers simultaneously. The X-ray microbeam system uses a computer-controlled narrow beam of high-energy X-rays to track small metal pellets attached on the articulatory organs. This system allows automatic accurate measurements of pellets with a minimum X-ray dosage.

The magnetic sensing system (magnetometer, or magnetic articulograph) is designed to perform the same function as the microbeam system without X-rays. The system uses a set of transmitter coils that generate alternate magnetic fields and miniature sensor coils attached to the articulatory organs, as shown in Fig. 2.17b. The positions of the receiver coils are computed from the filtered signals from the coils.

Medical Imaging Techniques

X-ray cinematography and X-ray video fluorography have been used for re-cording articulatory movements in two-dimensional projection images. The X-ray images show clear outlines of rigid structures, while they provide less-obvious boundaries for soft tissue. The outline of the tongue is enhanced by the application of liquid contrast media on the surface. Metal markers are often used to track the movements of flesh points on the soft-tissue articulators.

Ultrasonography is a diagnostic technique to obtain cross-sectional images of soft-tissues in real time. Ultrasound scanners consist of a sound probe (phased-array piezo transducer and receiver) and image processor, as illustrated in Fig. 2.18a. The probe is attached to the skin below the tongue to image the tongue surface in the sagittal or coronal plane.

Magnetic resonance imaging (MRI), shown in Fig. 2.18b, is a developing medical technique that excels at soft-tissue imaging of the living body. Its principle relies on excitation and relaxation of the hydrogen nuclei in water in a strong homogeneous magnetic field in response to radio-frequency (RF) pulses applied with variable gradient magnetic fields that determine the slice position. MRI is essentially a method for recording static images, while motion imaging setups with stroboscopic or real-time techniques have been applied to the visualization of articulatory movements or vocal tract deformation three-dimensionally [2.36].

2.4 Summary

This chapter described the structures of the human speech organs and physiological mechanisms for producing speech sounds. Physiological processes during speech are multidimensional in nature as described in this chapter. Discoveries of their component mechanisms have been dependent on technical developments

for visualizing the human body and analyses of biological signals, and this is still true today. For example, the hypopharyngeal cavities have long been known to exist, but their acoustic role was underestimated until recent MRI observations. The topics in this chapter were chosen with the author's hope to provide a guideline for the sophistication of speech technologies by reflecting the real and detailed processes of human speech production. Expectations from these lines of studies include speech analysis by recovering control parameters of articulatory models from speech sounds, speech synthesis with full handling of voice quality and individual vocal characteristics, and true speech recognition through biologic, acoustic, and phonetic characterizations of input sounds.

References

2.1 M.H. Draper, P. Ladefoged, D. Whittenridge: Respiratory muscles in speech, J. Speech Hearing Res. **2**, 16–27 (1959)

2.2 T.J. Hixon, M. Goldman, J. Mead: Kinematics of the chest wall during speech production: volume displacements of the rib cage, abdomen, and lung, J. Speech Hearing Res. **16**, 78–115 (1973)

2.3 G. Weismer: Speech production. In: *Handbook of Speech-Language Pathology and Audiology*, ed. by N.J. Lass, L.V. McReynolds, D.E. Yoder (Decker, Toronto 1988) pp. 215–252

2.4 J. Kahane: A morphological study of the human prepubertal and pubertal larynx, Am. J. Anat. **151**, 11–20 (1979)

2.5 M. Hirano, Y. Kakita: Cover-body theory of vocal cord vibration. In: *Speech Science*, ed. by R. Daniloff (College Hill, San Diego 1985) pp. 1–46

2.6 E.B. Holmberg: Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice, J. Acoust. Soc. Am. **84**, 511–529 (1988)

2.7 M.R. Rothenberg: Acoustic interaction between the glottal source and the vocal tract. In: *Vocal Fold Physiology*, ed. by K.N. Stevens, M. Hirano (Univ. Tokyo Press, Tokyo 1981) pp. 305–328

2.8 G. Fant, J. Liljencrants, Q. Lin: A four-parameter model of glottal flow, Speech Transmission Laboratory – Quarterly Progress and Status Report (STL-QPSR) **4**, 1–13 (1985)

2.9 B.R. Fink, R.J. Demarest: *Laryngeal Biomechanics* (Harvard Univ. Press, Cambridge 1978)

2.10 J.E. Atkinson: Correlation analysis of the physiological features controlling fundamental frequency, J. Acoust. Soc. Am. **63**, 211–222 (1978)

2.11 K. Honda, H. Hirai, S. Masaki, Y. Shimada: Role of vertical larynx movement and cervical lordosis in F0 control, Language Speech **42**, 401–411 (1999)

2.12 H. Takemoto: Morphological analysis of the human tongue musculature for three-dimensional modeling, J. Speech Lang. Hearing Res. **44**, 95–107 (2001)

2.13 S. Takano, K. Honda: An MRI analysis of the extrinsic tongue muscles during vowel production, Speech Commun. **49**, 49–58 (2007)

2.14 S. Maeda: Compensatory articulation during speech: evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In: *Speech Production and Speech Modeling*, ed. by W.J. Hardcastle, A. Marchal (Kluwer Academic, Dartrecht 1990) pp. 131–149

2.15 K. Honda, T. Kurita, Y. Kakita, S. Maeda: Physiology of the lips and modeling of lip gestures, J. Phonetics **23**, 243–254 (1995)

2.16 K. Ishizaka, M. Matsudaira, T. Kaneko: Input acoustic-impedance measurement of the subglottal system, J. Acoust. Soc. Am. **60**, 190–197 (1976)

2.17 U.G. Goldstein: An articulatory model for the vocal tracts of growing children. Ph.D. Thesis (Massachusetts Institute of Technology, Cambridge 1980)

2.18 W.T. Fitch, J. Giedd: Morphology and development of the human vocal tract: A study using magnetic resonance imaging, J. Acoust. Soc. Am. **106**, 1511–1522 (1999)

2.19 J. Dang, K. Honda, H. Suzuki: Morphological and acoustic analysis of the nasal and paranasal cavities, J. Acoust. Soc. Am. **96**, 2088–2100 (1994)

2.20 O. Fujimura, J. Lindqvist: Sweep-tone measurements of the vocal tract characteristics, J. Acoust. Soc. Am. **49**, 541–557 (1971)

2.21 S. Maeda: The role of the sinus cavities in the production of nasal vowels, Proc. IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP'82), Vol. 2 (1982) pp. 911–914, Paris

2.22 T. Chiba, M. Kajiyama: *The Vowel – Its Nature and Structure* (Tokyo-Kaiseikan, Tokyo 1942)

2.23 G. Fant: *Acoustic Theory of Speech Production* (Mouton, The Hague 1960)

2.24 T. Baer, P. Alfonso, K. Honda: Electromyography of the tongue muscle during vowels in /*pVp/ environment, Ann. Bull. RILP **22**, 7–20 (1988)

2.25 K. Honda: Organization of tongue articulation for vowels, J. Phonetics **24**, 39–52 (1996)

2.26 I. Lehiste, G.E. Peterson: Some basic considerations in the analysis of intonation, J. Acoust. Soc. Am. **33**, 419–425 (1961)

2.27 K. Honda: Relationship between pitch control and vowel articulation. In: *Vocal Fold Physiology*, ed. by D.M. Bless, J.H. Abbs (College-Hill, San Diego 1983) pp. 286–297

2.28 J. Sundberg: Articulatory interpretation of the singing formant, J. Acoust. Soc. Am. **55**, 838–844 (1974)

2.29 J. Dang, K. Honda: Acoustic characteristics of the piriform fossa in models and humans, J. Acoust. Soc. Am. **101**, 456–465 (1996)
2.30 H. Takemoto, S. Adachi, T. Kitamura, P. Mokhtari, K. Honda: Acoustic roles of the laryngeal cavity in vocal tract resonance, J. Acoust. Soc. Am. **120**, 2228–2238 (2006)
2.31 T. Kitamura, H. Takemoto, S. Adachi, P. Mokhtari, K. Honda: Cyclicity of laryngeal cavity resonance due to vocal fold vibration, J. Acoust. Soc. Am. **120**, 2239–2249 (2006)
2.32 I.R. Titze, B.H. Story: Acoustic interactions of the voice source with the lower vocal tract, J. Acoust. Soc. Am. **101**, 2234–2243 (1997)
2.33 A. Lofqvist, N.S. McGarr, K. Honda: Laryngeal muscles and articulatory control, J. Acoust. Soc. Am. **76**, 951–954 (1984)
2.34 R.G. Daniloff: Normal articulation processes. In: *Normal Aspect of Speech, Hearing, and Language*, ed. by F.D. Minifie, T.J. Hixon, F. Williams (Prentice-Hall, Englewood Cliffs 1983) pp. 169–209
2.35 D.P. Kuehn, K.L. Moll: A cineradiographic study of VC and CV articulatory velocities, J. Phonetics **4**, 303–320 (1976)
2.36 K. Honda, H. Takemoto, T. Kitamura, S. Fujita, S. Takano: Exploring human speech production mechanisms by MRI, IEICE Info. Syst. **E87-D**, 1050–1058 (2004)

3. Nonlinear Cochlear Signal Processing and Masking in Speech Perception

J. B. Allen

There are many classes of masking, but two major classes are easily defined: *neural masking* and *dynamic masking*. Neural masking characterizes the internal noise associated with the neural representation of the auditory signal, a form of loudness noise. Dynamic masking is strictly cochlear, and is associated with cochlear outer-hair-cell processing. This form is responsible for dynamic nonlinear cochlear gain changes associated with sensorineural hearing loss, the upward spread of masking, two-tone suppression and forward masking. The impact of these various forms of masking are critical to our understanding of speech and music processing. In this review, the details of what we know about nonlinear cochlear and basilar membrane signal processing is reviewed, and the implications of neural masking is modeled, with a comprehensive historical review of the masking literature. This review is appropriate for a series of graduate lectures on nonlinear cochlear speech and music processing, from an auditory point of view.

3.1 **Basics** 27
3.1.1 Function of the Inner Ear 28
3.1.2 History of Cochlear Modeling 31

3.2 **The Nonlinear Cochlea** 35
3.2.1 Cochlear Modeling 35
3.2.2 Outer-Hair-Cell Transduction 41
3.2.3 Micromechanics 42

3.3 **Neural Masking** 45
3.3.1 Basic Definitions 47
3.3.2 Empirical Models 51
3.3.3 Models of the JND 51
3.3.4 A Direct Estimate of the Loudness JND 52
3.3.5 Determination of the Loudness SNR 54
3.3.6 Weber–Fraction Formula 54

3.4 **Discussion and Summary** 55
3.4.1 Model Validation 55
3.4.2 The Noise Model 55

References 56

3.1 Basics

Auditory masking is critical to our understanding of speech and music processing. There are many classes of masking, but two major classes are easily defined. These two types of masking and their relation to nonlinear (NL) speech processing and coding are the focus of this chapter.

The *first* class of masking, denoted *neural masking*, is due to internal *neural noise*, characterized in terms of the intensity *just noticeable difference*, denoted $\Delta I(I, f, T)$ (abbreviated $\mathrm{JND_I}$) and defined as the *just discriminable change in intensity*. The $\mathrm{JND_I}$ is a function of intensity I, frequency f and stimulus type T (e.g., noise, tones, speech, music, etc.). As an *internal noise*, the $\mathrm{JND_I}$ may be modeled in terms of a loudness (i. e., perceptual intensity) noise density along the length of the cochlea ($0 \le X \le L$), described in terms of a *partial loudness JND* ($\Delta\mathcal{L}(X, T)$, a.k.a. $\mathrm{JND}_{\mathcal{L}}$). The cochlea or inner ear is the organ that converts signals from acoustical to neural signals. The loudness JND is a function of the *partial loudness* $\mathcal{L}(X)$, defined as the loudness contribution coming from each cochlear *critical band*, or more generally, along some *tonotopic central auditory representation*. The critical band is a measure of cochlear bandwidth at a given cochlear *place* X. The loudness JND plays a major role in speech and music coding since coding quantization noise may be masked by this internal quantization (i. e., *loudness noise*).

The *second* masking class, denoted here as *dynamic masking*, comes from the NL mechanical action of cochlear *outer-hair-cell* (OHC) signal processing. It can have two forms, simultaneous and nonsimultaneous, also known as *forward masking*, or *post-masking*. Dynamic-masking (i. e., nonlinear OHC signal processing) is well known (i. e., there is a historical literature

on this topic) to be intimately related to questions of cochlear frequency selectivity, sensitivity, dynamic range compression and *loudness recruitment* (the loss of loudness dynamic range). Dynamic masking includes the *upward spread of masking* (USM) effect, or in neural processing parlance, *two-tone suppression* (2TS). It may be underappreciated that NL OHC processing (i. e., dynamic masking) is largely responsible for *forward masking* (FM, or post-stimulus masking), which shows large effects over long time scales. For example OHC effects (FM/USM/2TS) can be as large as 50 dB, with an FM *latency* (return to base line) of up to 200 ms. *Forward masking* (FM) and NL OHC *signal onset enhancement* are important to the detection and identification of perceptual features of a speech signal. Some research has concluded that forward masking is not related to OHC processing [3.1, 2], so the topic remains controversial. Understanding and modeling NL OHC processing is key to many speech processing applications. As a result, a vibrant research effort driven by the National Institute of Health on OHC biophysics has ensued.

This OHC research effort is paying off at the highest level. Three key examples are notable. *First* is the development of wide dynamic-range multiband compression (WDRC) hearing aids. In the last 10–15 years WDRC signal processing (first proposed in 1937 by researchers at Bell Labs [3.3]), revolutionized the hearing-aid industry. With the introduction of compression signal processing, hearing aids now address the recruitment problem, thereby providing speech audibility over a much larger dynamic range, at least in quiet. The problems of the impaired ear given speech in noise is poorly understood today, but this problem is likely related to the effects of NL OHC processing. This powerful circuit (WDRC) is not the only reason hearing aids of today are better. Improved electronics and transducers have made significant strides as well. In the last few years the digital barrier has finally been broken, with digital signal processing hearing aids now becoming common.

A *second* example is the development of otoacoustic emissions (OAE) as a hearing diagnostic tool. Pioneered by David Kemp and Duck Kim, and then developed by many others, this tool allows for cochlear evaluation of neonates. The identification of cochlear hearing loss in the first month has dramatically improves the lives of these children (and their parents). While it is tragic to be born deaf, it is much more tragic for the deafness to go unrecognized until the child is three years old, when they fail to learn to talk. If you cannot hear you do not learn to talk. With proper and early cochlear implant intervention, these kids can lead nearly normal-hearing lives and even talk on the phone. However they cannot understand speech in noise. It is at least possible that this loss is due to the lack of NL OHC processing.

A *third* example of the application of NL OHC processing to speech processing is still an underdeveloped application area. The key open problem here is: *How does the auditory system, including the NL cochlea, followed by the auditory cortex, processes human speech*? There are many aspects of this problem including speech coding, speech recognition in noise, hearing aids and language learning and reading disorders in children. If we can solve the *robust phone decoding problem*, we will fundamentally change the effectiveness of human-machine interactions. For example, the ultimate hearing aid is the hearing aid with built in robust speech feature detection and phone recognition. While we have no idea when this will come to be, and it is undoubtedly many years off, when it happens there will be a technology revolution that will change human communications.

In this chapter several topics will be reviewed. First is the history of cochlear models including extensions that have taken place in recent years. These models include both macromechanics and micromechanics of the tectorial membrane and hair cells. This leads to comparisons of the basilar membrane, hair cell, and neural frequency tuning. Hearing loss, loudness recruitment, as well as other key topics of modern hearing health care, are discussed. The role of NL mechanics and dynamic range are reviewed to help the reader understand the importance of modern wideband dynamic range compression hearing aids as well as the overall impact of NL OHC processing.

Any reader desiring further knowledge about cochlear anatomy and function or a basic description of hearing, they may consult *Pickles* [3.4], *Dallos* [3.5], *Yost* [3.6].

3.1.1 Function of the Inner Ear

The goal of cochlear modeling is to refine our understanding of how auditory signals are processed. The two main roles of the cochlea are to separate the input acoustic signal into overlapping frequency bands, and to compress the large acoustic intensity range into the much smaller mechanical and electrical dynamic range of the inner hair cell. This is a basic question of information processing by the ear. The eye plays a similar role as a peripheral organ. It breaks the light image into rod- and cone-sized pixels, as it compresses the dynamic range of the visual signal. Based on the intensity JND, the corresponding visual dynamic range is about nine to

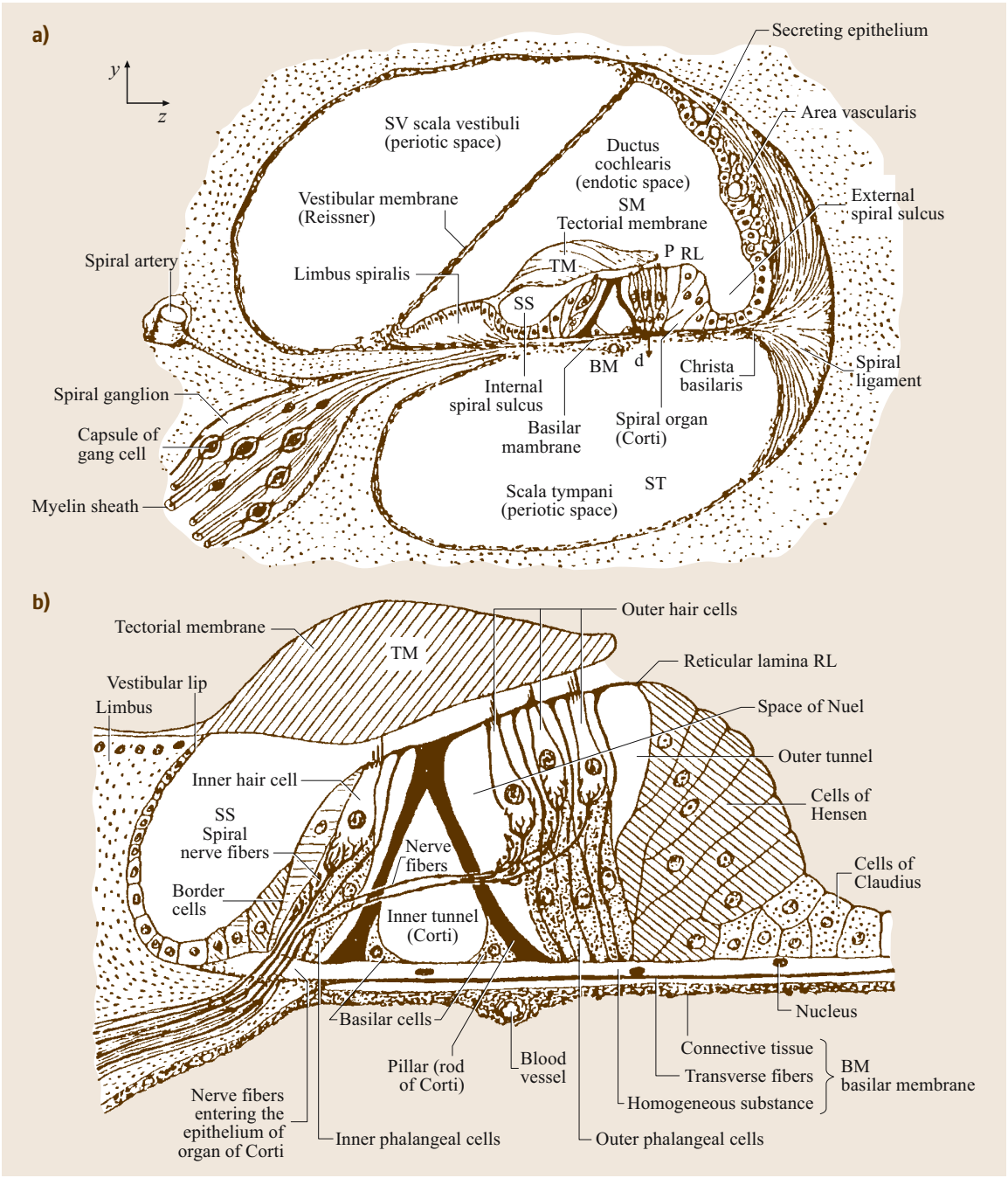


Fig. 3.1a,b On the left we see all the major structures of the cochlea **(a)**. The three chambers are filled with fluid. Reissner's membrane is an electrical barrier and is not believed to play a mechanical role. The right panel **(b)** shows the inner and outer hair cells, pillar cells and other supporting structures, the basilar membrane (BM), and the tectorial membrane (TM)

ten orders of magnitude of intensity [3.7, 8], while the ear has about 11 to 12. The stimulus has a relatively high information rate. Neurons are low-bandwidth channels. The eye and the ear must cope with this problem by reducing the stimulus to a large number of low bandwidth signals. It is then the job of the cortex to piece these pixel signals back together, to reconstruct the world as we see and hear it.

The acoustic information coding starts in the cochlea (Fig. 3.1a) which is composed of three major chambers formed by Reissner's membrane and the basilar membrane (BM). Mechanically speaking, there are only two chambers, as Reissner's membrane is only for electrical isolation of the scala media (SM) [3.4, 5]. Figure 3.1b shows a blown-up view of the organ of Corti where the inner hair cells (IHC) and outer hair cells (OHC) sit between the BM and the tectorial membrane (TM). As the BM moves up and down, the TM shears against the reticular lamina (RL), causing the cilia of the inner and outer hair cells to bend. The afferent auditory nerve fibers that are connected to the inner hair cells carry the signal information into the auditory system. Many fewer efferent fibers bring signals from the auditory system to the base of the outer hair cells. The exact purpose of these efferent fibers remains unknown.

Inner Hair Cells

In very general terms, the role of the cochlea is to convert sound at the eardrum into neural pulse patterns along approximately 30 000 neurons of the human auditory (VIIIth) nerve. After being filtered by the cochlea, a low-level pure tone has a narrow spread of excitation which excites the cilia of about 40 contiguous inner hair cells [3.5, 9, 10]. The IHC excitation signal has a narrow bandwidth and a center frequency that depends on the inner hair cell's location along the basilar membrane. Each hair cell is about 10 μm in diameter while the human basilar membrane is about 35 mm in length (35 000 μm). Thus the neurons of the auditory nerve encode the responses of about 3500 inner hair cells which form a single row of cells along the length of the BM. Each inner-hair-cell voltage is a low-pass-filtered representation of the detected inner-hair-cell cilia displacement [3.11]. Each hair cell is connected to many neurons, having a wide range of spontaneous firing rates and thresholds [3.12]. In the cat, for example, approximately 15–20 neurons encode each of these narrow band inner hair cells with a neural timing code. It is commonly accepted that all mammalian cochleae are similar in function except the frequency range of operation differs between species (e.g., human $\approx 0.1-20$ kHz and cat $\approx 0.3-50$ kHz). It is widely believed that the neuron information channel between the hair cell and the *cochlear nucleus* is a combination of the mean firing rate and the relative timing between neural pulses (spikes). The mean firing rate is reflected in the loudness coding, while the relative timing carries more subtle cues, including for example pitch information such as speech voicing distinctions.

Outer Hair Cells

As shown in Fig. 3.1b there are typically three (occasionally four) outer hair cells (OHCs) for each inner hair cell (IHCs), leading to approximately 12 000 OHCs in the human cochlea. Outer hair cells are used for intensity dynamic-range control. This is a form of NL signal processing, not dissimilar to Dolby sound processing. This form of processing was inspired by cochlear function, and was in use long before it was patented by Dolby, in movie sound systems developed by Bell Labs in the 1930s and 1940s. Telephone speech is similarly compressed [3.13] via μ-law coding. It is well known (as was first proposed by *Lorente de Nó* [3.14] and *Steinberg* [3.3]) that noise damage of *nerve cells* (i. e., OHCs) leads to a reduction of dynamic range, a disorder clinically named *loudness recruitment*. The word *recruitment*, which describes the abnormal growth of loudness in the impaired ear, is a seriously misleading term, since nothing is being recruited [3.15].

We may describe cochlear processing two ways: first in terms of the signal representation at various points in the system; and second, in terms of models which are our most succinct means of conveying the conclusions of years of detailed and difficult experimental work on cochlear function. The body of experimental knowledge has been very efficiently represented (to the extent that it is understood) in the form of these mathematical models. When no model exists (e.g., because we do not understand the function), a more basic description via the experimental data is necessary. Several good books and review papers that make excellent supplemental reading are available [3.4, 8, 16, 17].

For pedagogical purposes this chapter has been divided into four parts. Besides this introduction, we include sections on the NL cochlea, neural masking, and finally a brief discussion. Section 3.2 discusses dynamic masking due to NL aspects of the cochlear outer hair cells. This includes the practical aspects, and theory, of the upward spread of masking (USM) and two-tone suppression. Section 3.3 discusses neural masking, the JND, loudness recruitment, the loudness signal-to-noise ratio

(SNR), and the Weber fraction. Section 3.4 provides a brief summary.

3.1.2 History of Cochlear Modeling

Typically the cochlea is treated as an uncoiled long thin box, as shown in Fig. 3.2a. This represents the starting point for the macromechanical models.

Macromechanics

In his book *On the Sensations of Tone Helmholtz* [3.18] likened the cochlea to a bank of highly tuned resonators selective to different frequencies, much like a piano or a harp [3.19, p. 22–58], with each string representing a different place X on the basilar membrane. This model as proposed was quite limited since it leaves out key features, the most important of which is the cochlear fluid coupling between the mechanical resonators. But given the early publication date, the great master of physics and psychophysics Helmholtz shows deep insight and his studies provided many very important contributions.

The next major contribution by *Wegel* and *Lane* [3.20] stands in a class of its own even today, as a double-barreled paper having both deep psychophysical and modeling insight. Fletcher published much of the Wegel and Lane data one year earlier [3.21]. It is not clear to me why Wegel and Lane are always quoted for these results rather than Fletcher. In Fletcher's 1930 modeling paper, he mentioned that he was the subject in the Wegel and Lane study. It seems to me that Fletcher deserves some of the credit. The paper was the first to quantitatively describe the details of how a high level low frequency tone affects the audibility of a second low-level higher-frequency tone (i. e., the *upward spread of masking*). It was also the first publication to propose a *modern* model of the cochlea, as shown in Fig. 3.2b. If Wegel and Lane had been able to solve the model equations implied by their circuit (of course they had no computer to do this), they would have predicted cochlear traveling waves. It was their mistake, in my opinion, to make this a single paper. The modeling portion of their paper has been totally overshadowed by their experimental results. Transmission line theory had been widely exploited by Campbell, the first mathematical research at AT&T research (ca. 1898) with the invention of the wave filter [3.22, 23], which had been used for speech articulation studies [3.24–26], and Fletcher and Wegel were fully utilizing Campbell's important discoveries.

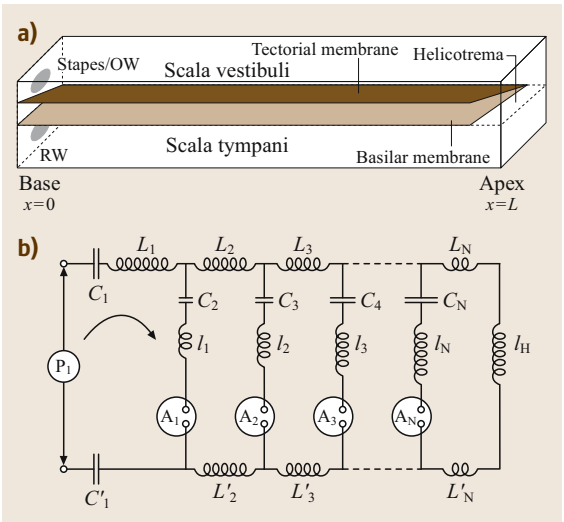


Fig. 3.2a,b On the left **(a)** see the basic 2-D box model of the cochlea. The *Base* ($x = 0$) is the high-frequency end of the cochlea while the *Apex* ($x = L$) carries the low frequencies. On the right **(b)** the 1924 Wegel and Lane electrical equivalent circuit. The model is built from a cascade of electrical sections

It was the experimental observations of G. von Békésy starting in 1928 on human cadaver cochleae which unveiled the physical nature of the basilar membrane traveling wave. What von Békésy found (consistent with the 1924 Wegel and Lane model) was that the cochlea is analogous to a *dispersive* transmission line where the different frequency components which make up the input signal travel at different speeds along the basilar membrane, thereby isolating each frequency component at a different place X along the basilar membrane. He properly identified this dispersive wave a *traveling wave,* just as Wegel and Lane had predicted in their 1924 model of the cochlea.

Over the intervening years these experiments have been greatly improved, but von Békésy's fundamental observation of the traveling wave still stands. His original experimental results, however, are *not* characteristic of the responses seen in more-recent experiments, in many important ways. These differences are believed to be due to the fact that Békésy's cochleae were dead, and because of the high sound levels his experiments required. He observed the traveling wave using stroboscopic light, in dead human cochleae, at sound levels well above 140 dB – SPL.

Today we find that for a pure tone input the traveling wave has a more sharply defined location on the basilar membrane than that observed by von Békésy. In fact, according to measurements made over the last 20 years, the response of the basilar membrane to a pure tone

can change in amplitude by more than five orders of magnitude per millimeter of distance along the basilar membrane (e.g., 300 dB/oct is equivalent to 100 dB/mm in the cat cochlea).

The One-Dimensional Model of the Cochlea

To describe this response it is helpful to call upon the macromechanical *transmission line models* of *Wegel* [3.20] (Fig. 3.2b) and *Fletcher* [3.27], first quantitatively analyzed by *Zwislocki* [3.28, 29], *Ranke* [3.30], *Peterson* and *Bogert* [3.31], *Fletcher* [3.32, 33]. This popular transmission line model is now denoted the *one-dimensional* (1-D), or *long-wave* model.

Zwislocki [3.28] was first to quantitatively analyze Wegel and Lane's macromechanical cochlear model, explaining Békésy's traveling wave observations. The stapes input pressure P_1 is at the left, with the input velocity V_1, as shown by the arrow, corresponding to the stapes velocity. This model represents the mass of the fluids of the cochlea as electrical inductors and the BM stiffness as capacitors. Electrical circuit networks are useful when describing mechanical systems. This is possible because of an electrical to mechanical analog that relates the two systems of equations. Electrical circuit elements comprise a de facto standard for describing such equations. It is possible to write down the equations that describe the system from the circuit of Fig. 3.2b, by those trained in the art. Engineers and scientists frequently find it easier to *read* and think in terms of these pictorial circuit diagrams, than to interpret the corresponding equations.

BM Impedance. During the following discussion it is necessary to introduce the concept of a *one-port* (two-wire) impedance. *Ohm's law* defines the impedance as

$$\text{Impedance} = \frac{\text{effort}}{\text{flow}} \,. \tag{3.1}$$

In an electrical system the impedance is the ratio of a voltage (effort) over a current (flow). In a mechanical system it is the force (effort) over the velocity (flow).

For *linear time-invariant causal* (LTIC) systems (i. e., an impedance), *phasor* notation is very useful, where the tone is represented as the real part (Re) of the complex exponential

$$\mathrm{e}^{\mathrm{i}2\pi ft+\mathrm{i}\phi} \equiv \cos(2\pi ft+\phi)+\mathrm{i}\sin(2\pi ft+\phi) \,. \tag{3.2}$$

The symbol $\equiv$ denotes *equivalence*. It means that the quantity to the left of $\equiv$ is defined by the quantity on the right. More specifically, impedance is typically defined in the frequency domain using *Laplace transform* notation, in terms of a damped tone

$$A\,\mathrm{e}^{\sigma t}\cos(2\pi ft+\phi) \equiv A\,\mathrm{Re}\,\mathrm{e}^{st+\mathrm{i}\phi} \tag{3.3}$$

excitation, characterized by the tone's amplitude A, phase ϕ and *complex Laplace frequency* $s \equiv \sigma + \mathrm{i}2\pi f$. When a function such as $Z(s)$ is shown as a function of the complex frequency s, this means that its inverse Laplace transform $z(t) \leftrightarrow Z(s)$ must be *causal*. In the time domain, the voltage may be found from the current via a convolution with $z(t)$. Three classic examples of such impedances are presented next.

Example 3.1: The impedance of the tympanic membrane (TM, or eardrum) is defined in terms of a pure tone pressure in the ear canal divided by the resulting TM volume velocity (the velocity times the area of TM motion) [3.34, 35]. The pressure (effort) and volume velocity (flow) referred to here are conventionally described using complex numbers, to account for the phase relationship between the two.

Example 3.2: The impedance of a spring is given by the ratio of the force $F(f)$ to velocity $V(f) = sX(f)$ with displacement X

$$Z(s) \equiv \frac{F}{V} = \frac{K}{s} = \frac{1}{sC} \,, \tag{3.4}$$

where the spring constant K is the stiffness, C the compliance, and s is the complex radian frequency. The stiffness is represented electrically as a capacitor (as parallel lines in Fig. 3.2b). Having $s = \sigma + \mathrm{i}2\pi f$ in the denominator indicates that the impedance of a spring has a phase of $-\pi/2$ (e.g., $-90°$). Such a phase means that when the velocity is $\cos(2\pi ft)$, the force is $\sin(2\pi ft)$. This follows from Hooke's law

$$F = KX = \frac{K}{s}sX = \frac{K}{s}V \,. \tag{3.5}$$

Example 3.3: From *Newton's law* $F = Ma$ where F is the force, M is the mass, and acceleration $a(s) = sV(s)$ (i. e., the acceleration in the time domain is $\mathrm{d}v(t)/\mathrm{d}t$). The electrical element corresponding to a mass is an *inductor,* indicated in Fig. 3.2b by a coil. Thus for a mass $Z(s) = sM$.

From these relations the magnitude of the impedance of a spring decreases as $1/f$, while the impedance magnitude of a mass is proportional to f. The stiffness with

its −90° phase is called a *lagging* phase, while the mass with its +90° phase is called a *leading* phase.

Different points along the basilar membrane are represented by the cascaded sections of the lumped transmission line model of Fig. 3.2b. The position X along the model is called the *place* variable and corresponds to the longitudinal position along the cochlea. The series (horizontal) inductors (coils) denoted by L_k represent the fluid mass (inertia) along the length of the cochlea, while the shunt elements represent the mechanical (acoustical) impedance of the corresponding partition (organ of Corti) impedance, defined as the pressure drop across the partition divided by its volume velocity per unit length

$$Z_{\mathrm{p}}(s, X) = \frac{K_{\mathrm{p}}(X)}{s} + R_{\mathrm{p}}(X) + sM_{\mathrm{p}} \,, \tag{3.6}$$

where $K(X)$ is the partition stiffness, and R_{p} is the partition resistance. Each inductor going to ground (l_i in Fig. 3.2b) represents the partition plus fluid mass per unit length M_{p} of the section. Note that sM, R_{p} and K/s are impedances, but the mass M and stiffness K are not. The partition stiffness decreases exponentially along the length of the cochlea, while the mass is frequently approximated as being independent of place.

As shown in Fig. 3.3a, for a given input frequency the BM impedance magnitude has a local minimum at the shunt resonant frequency, where the membrane that can move in a relatively unrestricted manner. The shunt *resonance* has special significance because at this resonance frequency $F_{\mathrm{cf}}(X)$ the inductor and the capacitor reactance cancel each other, creating an acoustic *hole*, where the only impedance element that contributes to the flow resistance is R_{p}. Solving for $F_{\mathrm{cf}}(X)$

$$\frac{K_{\mathrm{p}}(X)}{2\pi \mathrm{i} F_{\mathrm{cf}}} + 2\pi \mathrm{i} F_{\mathrm{cf}} M_{\mathrm{p}} = 0 \tag{3.7}$$

defines the *cochlear map function*, which is a key concept in cochlear modeling:

$$F_{\mathrm{cf}}(X) \equiv \frac{1}{2\pi} \sqrt{\frac{K_{\mathrm{p}}(X)}{M_{\mathrm{p}}}} \,. \tag{3.8}$$

The inverse of this function specifies the location of the *hole* $X_{\mathrm{cf}}(f)$ as shown in Fig. 3.3a. In the example of Fig. 3.3a two frequencies are show, at 1 and 8 kHz, with corresponding resonant points shown by $X_{\mathrm{cf}}(1)$ and $X_{\mathrm{cf}}(8)$.

Basal to $X_{\mathrm{cf}}(f)$ in Fig. 3.3a, the basilar membrane is increasingly stiff, and apically (to the right of the resonant point), the impedance is mass dominated. The above description is dependent on the input frequency f since the location of the hole is frequency dependent. In this apical region the impedance has little influence since almost no fluid flows past the low-impedance hole. This description is key to our understanding of why the various frequency components of a signal are splayed out along the basilar membrane.

If one puts a pulse of current in at the stapes, the highest frequencies that make up the pulse would be shunted close to the stapes since at high frequencies the hole is near the stapes, while the lower frequencies would continue down the line. As the low-pass pulse travels down the basilar membrane, the higher frequencies are progressively removed, until almost nothing is left when the pulse reaches the end of the model (the helicotrema end, the apex of the cochlea).

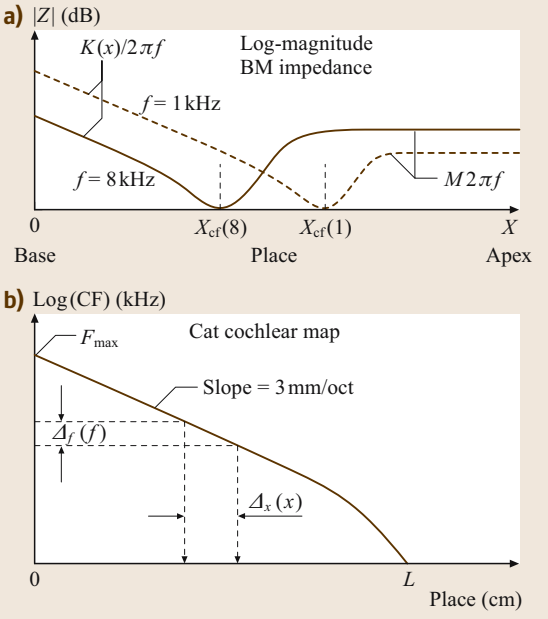


Fig. 3.3 **(a)** Plot of the log-magnitude of the impedance as a function of place for two different frequencies of 1 and 8 kHz showing the impedance; the region labeled $K(X)$ is the region dominated by the stiffness and has impedance $K(X)/s$. The region labeled M is dominated by the mass and has impedance sM. The characteristic places for 1 and 8 kHz are shown as X_{cf}. **(b)** Cochlear map of the cat following Liberman and Dodds. The resonance frequency depends on place according to the *cochlear map function* **(b)**. A *critical bandwidth* $\Delta_f(f)$ and a *critical spread* $\Delta_x(X)$ area related through the cochlear map

When a single tone is played, the response in the base increases in proportion to the BM compliance (inversely with the stiffness) until there is is a local maximum just before the traveling wave reaches the resonant hole, at which point the response plummets, since the fluid flow is shorted by the hole. For a fixed stimulus frequency f there is a maximum along the place axis called the *characteristic place*, denoted by $X_{\mathrm{cf}}^{(p)}(f)$. Likewise at a given place X as a function of frequency there is a local maximum called the *characteristic frequency*, denoted by $F_{\mathrm{cf}}^{(p)}(X)$. The relation between the peak in place as a function of frequency or of the peak in frequency as a function of place is also called the *cochlear map*. There is serious confusion with conventional terminology here. The resonant frequency of the BM impedance mathematically defines F_{cf} and specifies the frequency on the base of the high-frequency steep portion of the tuning slope, *not* the peak. However the peak is used as the visual cue, *not* the base of the high-frequency slope. These two definitions differ by a small factor (that is ignored) that depends directly on the high-frequency slope of the response. Over most of the frequency range this slope is huge, resulting in a very small factor, justifying its being ignored. However at very low frequencies the slope is shallow and the factor can then be large. The *droop* in the cochlear map seen in Fig. 3.3b at the apex ($x = L$) may be a result of these conflicting definitions. The cochlear map function $F_{\mathrm{cf}}(X)$ plays a key role in cochlear mechanics, has a long history, and is known by many names [3.27, 36–40], the most common today being *Greenwood's function*. In the speech literature it is called the *Mel scale*.

The spread of the response around the peak for a fixed frequency is denoted the *critical spread* $\Delta_x(f)$, while the frequency spread at a given place is called the *critical band* denoted $\Delta_f(X)$. As early as 1933 it was clear that the critical band must exist, as extensively discussed by *Fletcher* and *Munson* [3.41]. At any point along the BM the critical band is proportional to the *critical ratio* $\kappa(X)$, defined as the ratio of pure tone detection intensity at threshold in a background of white noise, to the spectral level of the noise [3.42], namely

$$\Delta_f(X) \propto \kappa(X) \,. \tag{3.9}$$

In the next section we shall show how the the relations between these various quantities are related via the cochlear map.

Derivation of the Cochlear Map Function. The derivation of the cochlear map is based on *counting* critical bands as shown by *Fletcher* [3.10] and popularized by *Greenwood* [3.43]. The *number of critical bands* N_{cb} may be found by integrating the critical band density over both frequency and place, and equating these two integrals, resulting in the cochlear map $F_{\mathrm{cf}}(X)$:

$$N_{\mathrm{cb}} \equiv \int_0^{X_{\mathrm{cf}}} \frac{\mathrm{d}X}{\Delta_x(X)} = \int_0^{F_{\mathrm{cf}}} \frac{\mathrm{d}f}{\Delta_f(f)} \,. \tag{3.10}$$

There are approximately 20 pure-tone frequency JNDs per critical band [3.37], [3.42, p. 171], and *Fletcher* showed that the *critical ratio* expressed in dB $\kappa_{\mathrm{dB}}(X)$ is of the form $aX + b$, where a and b are constants [3.10]. As verified by *Greenwood* [3.43, p. 1350, (1)] the critical bandwidth in Hz is therefore

$$\Delta_f(X) \propto 10^{\kappa_{\mathrm{dB}}(X)/10} \,. \tag{3.11}$$

The critical spread $\Delta_x(X)$ is the effective width of the energy spread on the basilar membrane for a pure tone. Based on a suggestion by Fletcher, *Allen* showed that for the cat, $\Delta_x(X)$ corresponds to about 2.75 times the basilar membrane width with $W_{\mathrm{bm}}(X) \propto \mathrm{e}^X$ [3.10]. It is reasonable to assume that the same relation would hold in the human case.

The direct observation of the cochlear map in the cat was made by *Liberman* [3.44] and *Liberman* and *Dodds* [3.45], and they showed the following empirical formula fit the data

$$F_{\mathrm{cf}}(X) = 456(10^{2.1(1-X/L)} - 0.8) \,, \tag{3.12}$$

where the length of the cat cochlea is $L = 21$ mm, and X is measured from the stapes [3.44]. The same formula may be used for the human cochlea if $L = 35$ mm is used, the 456 is replaced by 165.4, and 0.8 by 0.88. Based on (3.12), and as defined in Fig. 3.3b, the *slope* of the cochlear map is 3 mm/oct for the cat and 5 mm/oct for the human, as may be determined from the formula $L \log_{10}(2)/2.1$ with $L = 21$ or 35 for cat and human, respectively.

For a discussion of work after 1960 on the critical band see *Allen* [3.10] and *Hartmann* [3.17].

3.2 The Nonlinear Cochlea

3.2.1 Cochlear Modeling

In cochlear modeling there are two fundamental intertwined complex problems, *cochlear frequency selectivity* and *cochlear/OHC nonlinearity*. Wegel and Lane's 1924 transmission line wave theory was a most important development, since it was published 26 years prior to the experimental results of von Békésy, and it was based on a simple set of physical principles, conservation of fluid mass, and a spatially variable basilar membrane stiffness. It gives insight into both the NL cochlea, as well has two-dimensional (2-D) model frequency-selective wave-transmission effects (mass loading of the BM).

Over a 15 year period starting in 1971, there was a paradigm shift. Three discoveries rocked the field:

1. nonlinear compressive basilar membrane and inner-hair-cell measures of neural-like cochlear frequency selectivity [3.47, 48],
2. otoacoustic (ear canal) nonlinear emissions [3.49], and
3. motile outer hair cells [3.50].

Today we know that these observations are related, and all involve outer hair cells. A theory (e.g., a computational model) is needed to tie these results together. Many groups are presently working out such theories.

On the modeling side during the same period (the 1970's) all the variants of Wegel and Lane 1-D linear theory were becoming dated because:

1. numerical model results became available, which showed that 2- and three-dimensional (3-D) models were more frequency selective than the 1-D model,
2. experimental basilar membrane observations showed that the basilar membrane motion had a nonlinear compressive response growth, and
3. improved experimental basilar membrane observations became available which showed increased nonlinear cochlear frequency selectivity.

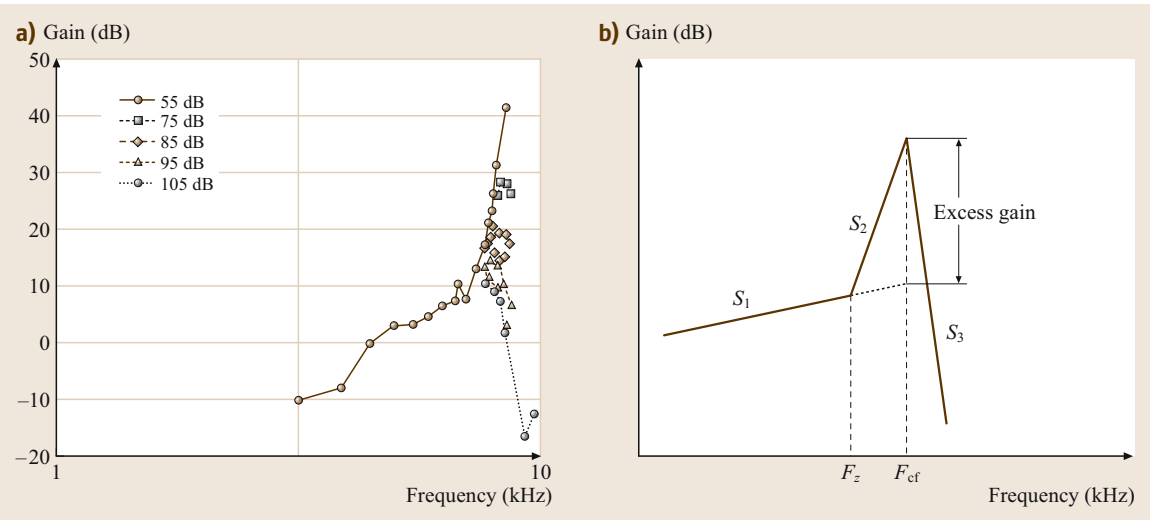


Fig. 3.4a,b There are six numbers that characterize every curve, three slopes (S_1, S_2, S_3), in dB/oct, two frequencies (F_z, F_{cf}), and the *excess gain* characterizes the amount of gain at F_{cf} relative to the gain defined by S_1. The excess gain depends on the input level for the case of a nonlinear response like the cochlea. Rhode found up to ≈ 35 dB of excess gain at 7.4 kHz and 55 dB − SPL, relative to the gain at 105 dB − SPL. From of the 55 dB − SPL curve of **(a)** (the most sensitive case), and his Table I, $S_1 = 9$, $S_2 = 86$, and $S_3 = -288$ (dB/oct), $F_z = 5$ kHz, $F_{cf} = 7.4$ kHz, and an *excess gain* of 27 dB. Rhode reported $S_1 = 6$ dB/oct, but 9 seems to be a better fit to the data, so 9 dB/oct is the value we have used for our comparisons. **(a)** Response of the basilar membrane for his most sensitive animal. The graduations along the abscissa are at 0.1, 1.0 and 10.0 kHz (after [3.46, Fig. 9a]) **(b)** Basic definition of the 6 parameters for characterizing a tuning curve: slopes S_1, S_2, S_3, frequencies F_z and F_{cf}, and the excess gain

Because these models and measures are still under development today [the problem has not yet (ca. 2007) been solved], it is necessary to describe the data rather than the models. Data that drives these nonlinear cochlear measures include:

- The upward spread of masking (USM), first described quantitatively by Wegel and Lane in 1924,
- Distortion components generated by the cochlea and described by *Wegel* and *Lane* [3.20], *Goldstein* and *Kiang* [3.52], *Smoorenburg* [3.53], *Kemp* [3.54], *Kim* et al. [3.55], *Fahey* and *Allen* [3.56] and many others,
- Normal loudness growth and recruitment in the impaired ear [3.3, 41],
- The frequency dependent neural two-tone suppression observed by *Sachs* and *Kiang* [3.57], *Arthur* et al. [3.58], *Kiang* and *Moxon* [3.59], *Abbas* and *Sachs* [3.60], *Fahey* and *Allen* [3.56], *Pang* and *Guinan* [3.61], and others,
- The frequency-dependent basilar membrane response-level compression first described by *Rhode* [3.46, 47],
- The frequency-dependent inner-hair-cell receptor potential level compression, first described by *Sellick* and *Russell* [3.48], *Russell* and *Sellick* [3.62].
- Forward masking data that shows a linear return to baseline after up to 0.2 s [3.63]. There may be compelling evidence that OHCs are the source of forward masking.

We shall discuss each of these, but two related measures are the most important for understanding these NL masking effects, the upward spread of masking (USM) and two-tone suppression (2TS).

Basilar Membrane Nonlinearity. The most basic early and informative of these nonlinear effects was the NL basilar membrane measurements made by *Rhode* [3.46, 47], as shown in Fig. 3.4a, showing that the basilar membrane displacement to be a highly NL function of level. For every four dB of pressure level increase on the input, the output displacement (or velocity) only changed one dB. This compressive nonlinearity depends on frequency, and only occurs near the most sensitive region (e.g., the tip of the tuning curve). For other frequencies the system was either linear, namely, one dB of input change gave one dB of output change for frequencies away from the best frequency, or very close to linear. This NL effect was highly dependent on the health of the animal, and would decrease or would not be present at all, when the animal was not in its physiologically pristine state.

An important and useful measure of cochlear linear and nonlinear response first proposed by *Rhode* [3.46, Fig. 8], is shown in Fig. 3.4b which describes cochlear tuning curves by straight lines on log–log coordinates. Such straight line approximations are called *Bode plots* in the engineering literature. The *slopes* and *break points*, defined as the locations where the straight lines cross, characterize the response.

Otoacoustic Emissions. A few years after Rhode's demonstration of cochlear nonlinearity, David Kemp observed otoacoustic emissions (tonal sound emanating from the cochlea and NL *echos* to clicks and tone bursts) [3.49, 54, 64–66]. Kemp's findings were like a jolt to the field, which led to a cottage industry of objective testing of the auditory system, including both cochlear and middle ear tests.

Motile OHCs. Subsequently, *Brownell* et al. [3.50] discovered that isolated OHCs change their length when placed in an electric field, thus that the outer hair cell is motile. This then led to the intuitive and widespread proposal that outer hair cells act as voltage-controlled motors that directly drive the basilar membrane on a cycle by cycle basis. It seems quite clear, from a great deal of data, that the OHC onset response time is on the order of one cycle or so of the BM impulse response, because the first peak is linear [3.67]. The release time must be determined by the OHC membrane properties, which is slow relative to the attack. Thus OHC NL processing is the basis for both the frequency asymmetry of simultaneous (upward versus downward spread) and temporal (forward versus backward) masking.

As summarized in Fig. 3.5, OHCs provide feedback to the BM via the OHC receptor potential, which in turn is modulated by both the position of the basilar

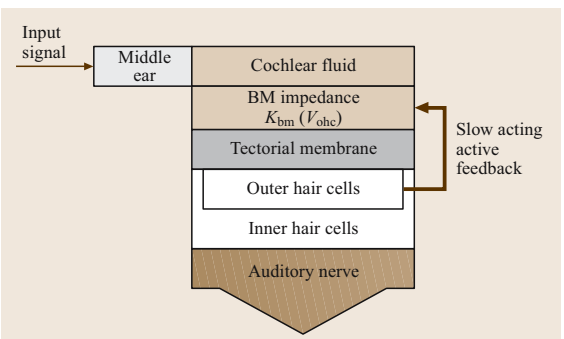


Fig. 3.5 Block flow diagram of the inner ear (after *Allen* [3.51])

membrane (forming a fast feedback loop), and alternatively by the efferent neurons that are connected to the outer hair cells (forming a slow feedback loop). The details of all this are the topic of a great deal of present research.

OHCs are the one common element that link all the NL data previously observed, and a missing piece of the puzzle that most needs to be understood before any model can hope to succeed in predicting basilar membrane, hair cell, and neural tuning, or NL compression. Understanding the outer hair cell's two-way mechanical transduction is viewed as the key to solving the problem of the cochlea's dynamic range.

Historically the implication that hair cells might play an important role in cochlear mechanics go back at least to 1936 when loudness recruitment was first reported by *Fowler* [3.68] in a comment by *R. Lorente de Nó* [3.14] stating that cochlear hair cells are likely to be involved in loudness recruitment.

The same year *Steinberg* and *Gardner* [3.3] were explicit about the action of recruitment when they concluded:

> *When someone shouts, such a deafened person suffers practically as much discomfort as a normal hearing person would under the same circumstances. Furthermore for such a case, the effective gain in loudness afforded by amplification depends on the amount of variable type loss present. Owing to the expanding action of this type of loss it would be necessary to introduce a corresponding compression in the amplifier in order to produce the same amplification at all levels.*

Therefore as early as 1937 there was a clear sense that cochlear hair cells were related to dynamic range compression.

More recently, theoretical attempts to explain the difference in tuning between normal and damaged cochleae led to the suggestion that OHCs could influence BM mechanics. In 1983 *Neely* and *Kim* [3.69] concluded:

> *We suggest that the negative damping components in the model may represent the physical action of outer hair cells, functioning in the electrochemical environment of the normal cochlea and serving to boost the sensitivity of the cochlea at low levels of excitation.*

In 1999 yet another (a fourth) important discovery was made, that the outer-hair-cell mechanical stiffness depends on the voltage across its membrane [3.70, 71]. This change in stiffness, coupled with the naturally occurring internal static pressure, may well account for the voltage dependent accompanying length changes (the cell's voltage dependent motility). This view follows from the block diagram feedback model of the organ of Corti shown in Fig. 3.5 where the excitation to the OHC changes the cell voltage V_{ohc}, which in turn changes the basilar stiffness [3.51]. This is one of several possible theories that have been put forth.

This experimental period set the stage for explaining the two most dramatic NL measures of cochlear response, the upward spread of masking and its related neural correlate, two-tone suppression, and may well turn out to be the explanation of the nonlinear forward-masking effect as well [3.63].

Simultaneous Dynamic-Masking

The psychophysically measured *upward spread of masking* (USM) and the neurally measured *two-tone suppression* (2TS) are closely related dynamic-masking phenomena. Historically these two measures have been treated independently in the literature. As will be shown, it is now clear that they are alternative objective measures of the same OHC compressive nonlinearity. Both involve the dynamic suppression of a basal (high-frequency) probe due to the simultaneous presentation of an apical (low-frequency) suppressor. These two views (USM versus 2TS) nicely complement each other, providing a symbiotic view of cochlear nonlinearity.

Upward Spread of Masking (USM). In a classic paper, *Mayer* [3.72] was the first to describe the asymmetric nature of masking [3.63,73]. Mayer made his qualitative observations with the use of clocks, organ pipes and tuning forks, and found that that the spread of masking is a strong function of the probe-to-masker frequency ratio (f_p/f_m) [3.63].

In 1923, Fletcher published the first quantitative results of tonal masking. In 1924, Wegel and Lane extended Fletcher's experiments (Fletcher was the subject [3.27, p. 325]) using a wider range of tones. Wegel and Lane then discuss the results in terms of their 1-D model described above. As shown in Fig. 3.6a, Wegel and Lane's experiments involved presenting listeners with a masker tone at frequency $f_m = 400\,\text{Hz}$ and intensity I_m (the abscissa), along with a probe tone at frequency f_p (the parameter used in the figure). At each masker intensity and probe frequency, the threshold probe intensity $I_p^*(I_m)$ is determined, and displayed relative to its threshold *sensation level* (SL) (the ordinate

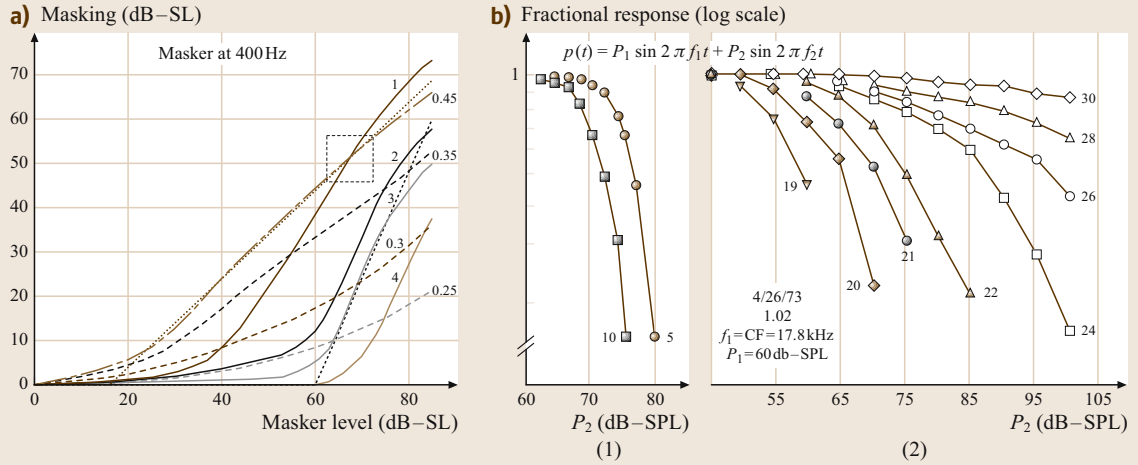


Fig. 3.6a,b On the left **(a)** we see the psychoacoustic measure of 2TS, called the upward spread of masking. On the right **(b)** are related measures taken in the auditory nerve by a procedure called two-tone suppression (2TS). Low- and high-side masking or suppression have very different thresholds and slopes. These suppression slopes and thresholds are very similar between 2TS and the USM. **(a)** Upward spread of masking as characterized by Wegel and Lane in 1924. The *solid lines* correspond to the probe being higher than the 400 Hz masker, while the *dashed lines* correspond to the 400 Hz probe lower than the masker. On the *left* we see upward spread of masking functions from Wegel and Lane for a 400 Hz low-frequency masker. The abscissa is the masker intensity I_m in dB − SL while the ordinate is the threshold probe intensity $I_p^*(I_m)$ in dB − SL. The frequency of the probe f_p, expressed in kHz, is the parameter indicated on each curve. The *dashed box* shows that the masking due to a 1 kHz tone becomes more than that at 450 Hz, for a 400 Hz probe. This is the first observation of *excitation pattern migration* with input intensity. **(b)** Two-tone suppression (2TS) input–output (IO) functions from *Abbas* and *Sachs* [3.60, Fig. 8]. On the *left* (1) is low-side suppression and on the *right* (2) we see high-side suppression. In 2TS the suppressor plays the role of the masker and the probe the role of the maskee. Note that the threshold of suppression for low-side suppressor (masker) is close to 70 dB − SPL, which is similar to human low-side suppressors, the case of the Wegel and Lane USM (1) (60–70 dB − SPL). The onset of suppression for high-side suppressors is close to the neuron's CF threshold of 50 dB, as elaborated further in Fig. 3.7a

is the probe level at threshold [dB − SL]). The asterisk indicates a threshold measure.

In Fig. 3.6a $f_m = 400$ Hz, I_m is the abscissa, f_p is the parameter on each curve, in kHz, and the threshold probe intensity $I_p^*(I_m)$ is the ordinate. The dotted line superimposed on the 3 kHz curve $(I_m/10^{60/10})^{2.4}$ represents the suppression threshold at 60 dB − SL which has a slope of 2.4 dB/dB. The dotted line superimposed on the 0.45 kHz curve has a slope of 1 and a threshold of 16 dB − SL.

Three regions are clearly evident: the *downward spread* of masking ($f_p < f_m$, dashed curves), *critical band* masking ($f_p \approx f_m$, dashed curve marked 0.45), and the *upward spread* of masking ($f_p > f_m$, solid curves) [3.74].

Critical band masking has a slope close to 1 dB/dB (the superimposed dotted line has a slope of 1). Four years later *Riesz* [3.75] shows critical band masking obeys the *near miss to Weber's law*, as described in Sect. 3.3.2. The downward spread of masking (the dashed lines in Fig. 3.6a) has a low threshold intensity and a variable slope that is less than one dB/dB, and approaches 1 at high masker intensities. The upward spread of masking (USM), shown by the solid curves, has a threshold near 50 dB re sensation level (e.g., 65 dB − SPL), and a growth just less than 2.5 dB/dB. The dotted line superimposed on the $f_p = 3$ kHz curve has a slope of 2.4 dB/dB and a threshold of 60 dB − SL.

The dashed box shows that the upward spread of masking of a probe at 1 kHz can be greater than the masking within a critical band (i. e., $f_p = 450$ Hz $> f_m = 400$ Hz). As the masker frequency is increased, this *crossover effect* occurs in a small frequency region (i. e., 1/2 octave) above the masker frequency. The crossover is a result of a well-documented NL *response migration*, of the excitation pattern with

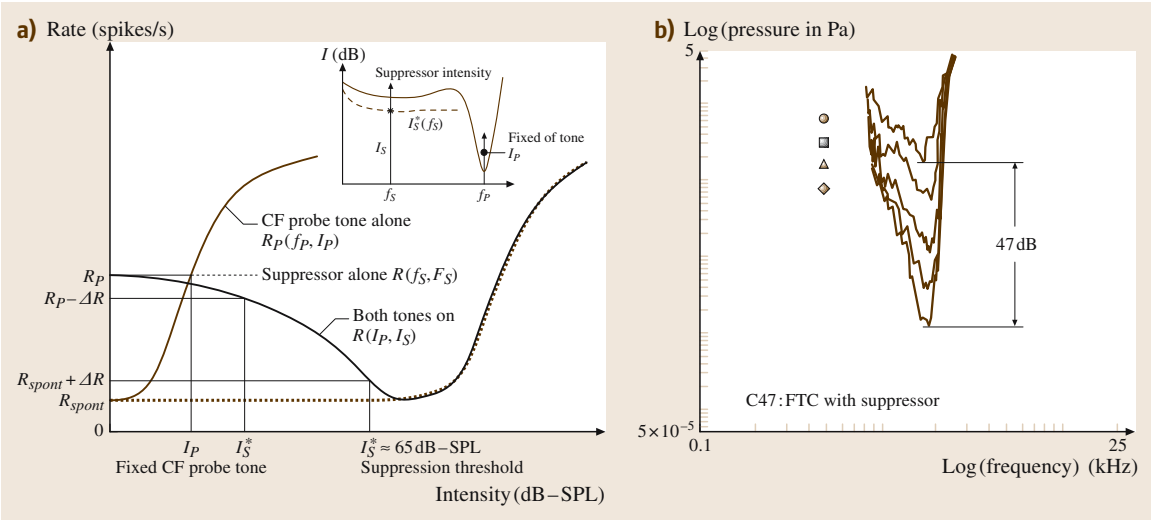


Fig. 3.7 **(a)** Definitions of 2TS low-side masking procedure (see (3.13) and (3.14)). **(b)** Example of 2TS (low-side masking) in the cat auditory nerve (AN). A cat neural tuning curve taken with various *low-side* suppressors present (suppressor below the best frequency), as indicated by the symbols. The tuning curve with the lowest threshold is for no suppressor. When the suppressor changes by 20 dB, the F_{cf} threshold changes by 36 dB. Thus for a 2 kHz neuron, the slope is 36/20, or 1.8. These numbers are similar to those measure by *Delgutte* [3.80]. One Pa = 94 dB – SPL

stimulus intensity, described in a wonderful paper by *McFadden* [3.76]. Response migration was also observed by *Munson* and *Gardner* in a classic paper on forward masking [3.77]. This important migration effect is beyond the scope of the present discussion, but is reviewed in [3.74, 78, 79] discussed in the caption of Fig. 3.10.

The upward spread of masking is important because it is easily measured psychophysically in normal hearing people, is robust, well documented, and nicely characterizes normal outer-hair-cell nonlinearities. The psychophysically measured USM has correlates in basilar membrane and hair cell signals, and is known as two-tone suppression (2TS) in the auditory nerve literature, as discussed in the caption of Fig. 3.6b.

Two-Tone Suppression. The neural correlate of the psychophysically measured USM is called *two-tone suppression* (2TS). As shown in the insert of Fig. 3.7a, first a neural tuning curve is measured. A pure tone probe at intensity $I_p(f_p)$, and frequency f_p, is placed a few dB (e.g., 6 to 10) above threshold at the characteristic (best) frequency of the neuron F_{cf} (i. e., $f_p = F_{cf}$). In 2TS a suppressor tone plays the role of the masker. There are two possible *thresholds*. The intensity of the suppressor tone $I_s(f_s)$ at frequency f_s is increased until either

1. the rate response to either the probe alone $R(I_p, I_s = 0)$ decreases by a small increment Δ_R, or
2. drops to the small increment Δ_R, just above the undriven spontaneous rate $R(0, 0)$.

These two criteria are defined in Fig. 3.6b and may be written

$$R_p\left(I_p, I_s^*\right) \equiv R(I_p, 0) - \Delta_R \tag{3.13}$$

and

$$R_{spont}\left(I_p, I_s^*\right) \equiv R(0, 0) + \Delta_R\,; \tag{3.14}$$

Δ_R indicates a fixed small but statistically significant constant change in the rate (e.g., $\Delta_R = 20$ spikes/s is a typical value). The threshold suppressor intensity is defined as $I_s^*(f_s)$, and as before the $*$ indicates the threshold suppressor intensity. The two threshold definitions (3.13) and (3.14) are very different, and both are useful. The difference in intensity between the two thresholds is quite large, and the more common measure used by *Abbas* and *Sachs* [3.60] is (3.13). The second measure (3.14) is consistent with neural tuning curve suppression, and is therefore the more interesting of the two. It corresponds to suppression of the probe to threshold.

Neural data of *Abbas* and *Sachs* [3.60, Fig. 8] are reproduced in Fig. 3.6b. For this example (see entry in lower-right just below 105), F_{cf} is 17.8 kHz, and the

$f_p = F_{cf}$ probe intensity $20 \log 10(|P_1|)$ is 60 dB. The label on the curves is the frequency f_1. The threshold intensity of the associated neural tuning curve is has a low spontaneous rate and a 50–55 dB threshold. The left panel of Fig. 3.6b is for apical suppressors that are lower in frequency than the characteristic frequency (CF) probe ($f_s < f_p$). In this case the threshold is just above 65 dB − SPL. The suppression effect is relatively strong and almost independent of frequency. In this example the threshold of the effect is less than 4 dB apart (the maximum shift of the two curves) at suppressor frequencies f_s of 10 and 5 kHz (a one octave separation).

The right panel shows the case $f_s > f_p$. The suppression threshold is close to the neuron's threshold (e.g., 50 dB − SPL) for probes at 19 kHz, but increases rapidly with frequency. The strength of the suppression is weak in comparison to the case of the left panel ($f_s < f_p$), as indicated by the slopes of the family of curves.

The Importance of the Criterion. The data of Fig. 3.6b uses the first suppression threshold definition (3.13) R_p (a small drop from the probe driven rate). In this case the F_{cf} probe is well above its detection threshold at the suppression threshold, since according to definition (3.13), the probe is just detectably reduced, and thus audible. With the second suppression threshold definition (3.14), the suppression threshold corresponds to the detection threshold of the probe. Thus (3.14), *suppression to the spontaneous rate*, is appropriate for Wegel and Lane's masking data where the probe is at its detection threshold $I_p^*(I_m)$. Suppression threshold definition (3.14) was used when taking the 2TS data of Fig. 3.7b, where the suppression threshold was estimated as a function of suppressor frequency.

To be consistent with a detection threshold criterion, such as the detection criterion used by Wegel and Lane in psychophysical masking, (3.14) must be used. To have a tuning curve pass through the F_{cf} probe intensity of a 2TS experiment (i. e., be at threshold levels), it is necessary to use the suppression to rate criterion given by (3.14). This is shown in Fig. 3.7b where a family of tuning curves is taken with different suppressors present. As described by *Fahey* and *Allen* [3.56, Fig. 13], when a probe is placed on a specific tuning curve of Fig. 3.7b, corresponding to one of the suppressor level symbols of Fig. 3.7b, and a suppression threshold is measured, that suppression curve will fall on the corresponding suppression symbol of Fig. 3.7b. There is a symmetry between the tuning curve measured in the presents of a suppressor, and a suppression threshold obtained with a given probe. This symmetry only holds for criterion (3.14), the detection threshold criterion, which is appropriate for Wegel and Lane's data. If one uses (3.13) as in [3.60] they will not see this symmetry as cleary.

Suppression Threshold. Using the criterion (3.14), *Fahey* and *Allen* [3.56, Fig. 13] showed that the suppression threshold $I_s^*(I_p)$ in the tails is near 65 dB − SPL (0.04 Pa). This is true for suppressors between 0.6 and 4 kHz. A small amount of data are consistent with the threshold being constant to much higher frequencies, but the Fahey and Allen data are insufficient on that point.

Suppression Slope. *Delgutte* has written several insightful papers on masking and suppression [3.80–82]. He estimated how the intensity growth slope (in dB/dB) of 2TS varies with suppressor frequency for several probe frequencies [3.80]. As may be seen in his figure, the suppression growth slope for the case of a low frequency apical suppressor on a high frequency basal neuron (the case of the left panel of Fig. 3.6b), is ≈ 2.4 dB/dB. This is the same slope as for Wegel and Lane's 400 Hz masker, 3 kHz probe USM data shown in Fig. 3.6a. For suppressor frequencies greater than the probe's ($f_s > f_p$), Delgutte reports a slope that is significantly less than 1 dB/dB. Likewise Wegel and Lane's data has slopes much less than 1 for the downward spread of masking.

One may conclude that USM and 2TS data show systematic and quantitative correlations between the threshold levels and slopes. The significance of these correlations has special importance because

1. they come from very different measurement methods, and
2. Wegel and Lane's USM are from human, while the 2TS data are from cat, yet they show similar responses. This implies that the cat and human cochleae may be quite similar in their NL responses.

The USM and 2TS threshold and growth slope (e.g., 50 dB − SL and 2.4 dB/dB) are important features that must be fully understood and modeled before we can claim to understand cochlear function. While there have been several models of 2TS [3.83–85] as discussed in some detail by *Delgutte* [3.80], none are in quantitative agreement with the data. The two-tone suppression model of *Hall* [3.84] is an interesting contribution to this problem because it qualitatively explores many of the key issues. Finally forward-masking data also show related nonlinear properties that we speculate may turn out to be related to NL OHC function as well [3.78, 86, 87].

3.2.2 Outer-Hair-Cell Transduction

The purpose of this section is to address two intimately intertwined problems *cochlear frequency selectivity* and *cochlear nonlinearity*. The fundamental question in cochlear research today is: *What is the role of the outer hair cell (OHC) in cochlear mechanics?* The OHC is the source of the NL effect, and the end product is dynamic masking, including the USM, 2TS and forward masking, all of which include dramatic amounts of gain and tuning variation. The issues are the nature of the NL transformations of the BM, OHC cilia motion, and OHC soma motility, at a given location along the basilar membrane.

The prevailing and popular *cochlear-amplifier* view is that the OHC provides *cochlear sensitivity* and *frequency selectivity* [3.5, 88–94]. The alternative view, argued here, is that the OHC compresses the excitation to the inner hair cell, thereby providing dynamic range expansion.

There is an important difference between these two views. The *first* view deemphasizes the role of the OHC in providing dynamic range control (the OHC's role is to improve sensitivity and selectivity), and assumes that the NL effects result from OHC saturation.

The *second* view places the dynamic range problem as the top priority. It assumes that the sole purpose of the OHC nonlinearity is to provide dynamic range compression, and that the OHC plays no role in either sensitivity or selectivity, which are treated as important but independent issues. Of course other views besides these two are possible.

The Dynamic-Range Problem

The question of how the large (up to 120 dB) dynamic range of the auditory system is attained has been a long standing problem which remains fundamentally incomplete. For example, *recruitment*, the most common symptom of neurosensory hearing loss, is best characterized as the loss of dynamic range [3.3, 10, 15, 95]. Recruitment results from outer-hair-cell damage [3.96]. To successfully design hearing aids that deal with the problem of recruitment, we need models that improve our understanding of *how* the cochlea achieves its dynamic range.

Based on a simple analysis of the IHC voltage, one may prove that the dynamic range of the IHC must be less than 65 dB [3.97]. In fact it is widely accepted that IHC dynamic range is less than 50 dB.

The IHC's transmembrane voltage is limited at the high end by the cell's open circuit (unloaded) membrane voltage, and at the low end by thermal noise. There are two obvious sources of thermal noise, cilia Brownian motion, and Johnson (shot) noise across the cell membrane (Fig. 3.8).

The obvious question arises: *How can the basic cochlear detectors (the IHCs) have a dynamic range of less than* 50 dB *(a factor of* 0.3×10^2*), and yet the auditory system has a dynamic range of up to* 120 dB

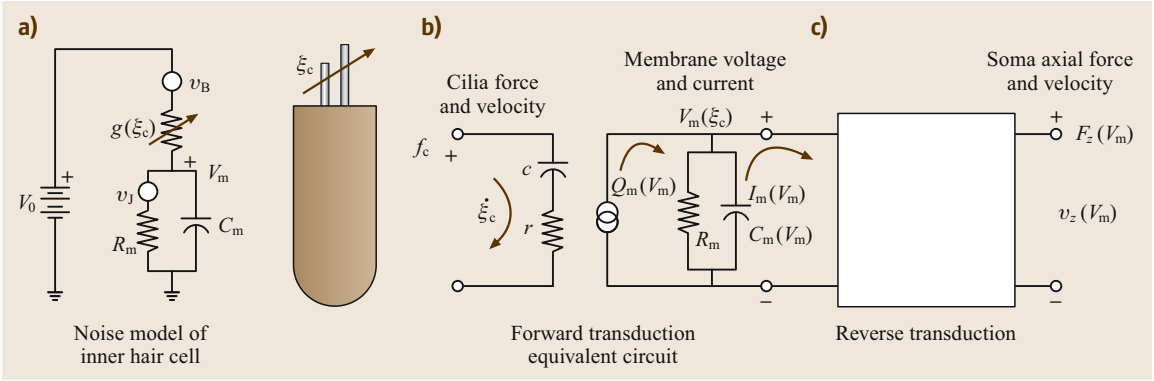


Fig. 3.8a–c On the far left **(a)** is the electrical equivalent circuit model of an IHC with thermal noise sources due to the cell leakage resistance Johnson and shot noise v_J and the Brownian motion of the cilia, represented by the voltage noise source v_B. The cilia force f_c and velocity $\dot{\xi}_c$ are the stimulus (input) variables to the forward transduction **(b)**, and are loaded by the mechanical impedance of the cilia viscous drag r and compliance c. **(c)** For OHCs, when the cilia move, current flows into the cell charging the membrane capacitance, thus changing the membrane voltage V_m. This membrane capacitance $C_m(V_m)$ is voltage dependent (i. e., it is NL). The membrane voltage has also been shown to control the cell's soma axial stiffness. It follows that the axial force $F_z(V_m)$ the cell can deliver, and the axial velocity $\mathcal{V}_z(V_m)$ of the cell, must also depend on the membrane voltage. The precise details of how all this works is unknown

(a factor of 10^6*)?* The huge amount of indirect evidence has shown that this increased dynamic range results from mechanical NL signal compression provided by outer hair cells. This dynamic-range compression shows up in auditory psychophysics and in cochlear physiology in many ways.

This thus forms the basic dynamic-range dilemma.

Outer-Hair-Cell Motility Model

A most significant finding in 1985 was of OHC *motility*, namely that the OHC changes its length by up to 5% in response to the cell's membrane voltage [3.50, 99, 100]. This less than 5% change in length must account for a 40 dB (100 times) change in cochlear sensitivity. This observation led to a significant increases in research on the OHC cell's motor properties.

In 1999 it was shown that the cell's longitudinal soma stiffness changes by at least a factor of 2 ($> 100\%$), again as a function of cell membrane voltage [3.70, 71]. A displacement of the cilia in the direction of the tallest cilia, which is called a *depolarizing* stimulus, decreases the magnitude of the membrane voltage $|V_m|$, *decreases* the longitudinal soma stiffness, and *decreases* the cell soma length. A hyperpolarizing stimulus increases the stiffness and extends the longitudinal soma length.

Given this much larger relative change in stiffness (a factor of 2) compared to the relative change in length (a factor of 1.05), for a maximum voltage change, it seems possible, or even likely, that the observed length changes (the motility) are simply a result of the voltage dependent stiffness. For example, imagine a spring stretched by applying a constant force (say a weight), and then suppose that the spring's stiffness decreases. It follows from Hooke's law (3.5) that the spring's length will *increase* when the stiffness decreases.

Each cell is stretched by its internal static pressure $\mathcal{P}$ [3.101], and its stiffness is voltage controlled [3.70, 71]. The voltage dependent relative stiffness change is much greater than the relative motility change. Thus we have the necessary conditions for stiffness-induced motility.

3.2.3 Micromechanics

Unlike the case of macromechanical models, the physics of every micromechanical model differs significantly. This is in part due to the lack of direct experimental evidence of physical parameters of the cochlea. This is an important and very active area of research (e.g., [3.102]).

To organize our discussion of cochlear micromechanics, we represent each radial cross-section through the cochlear partition (Fig. 3.1b) as a linear two-port network. A general formalization in transmission matrix form of the relation between the basilar membrane *input* pressure $P(x, s)$ and velocity $V(x, s)$ and the OHC *output* cilia bundle shear force $f(x, s)$ and shear velocity

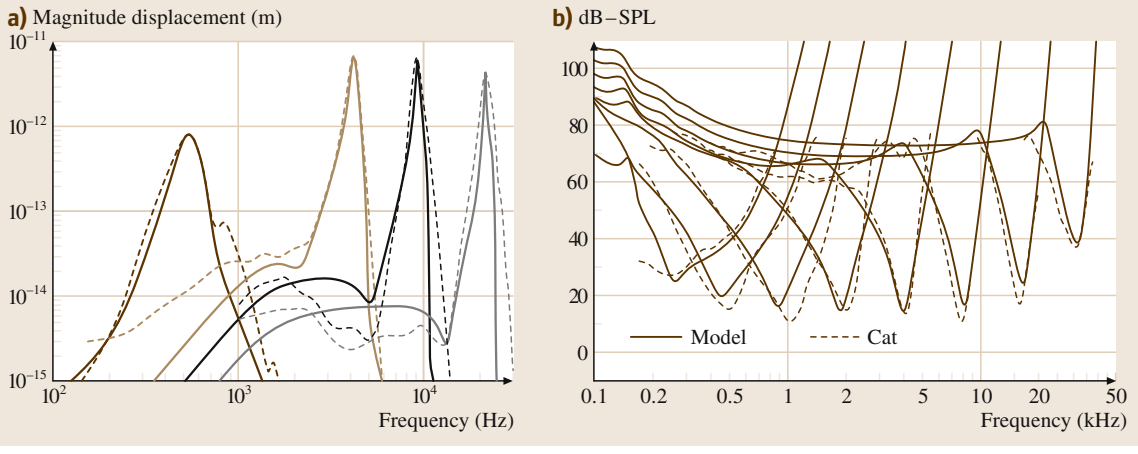


Fig. 3.9a,b The tuning curves shown by the dashed lines are the average of single nerve fiber responses from six cats obtained by M. C. Liberman and B. Delgutte. **(a)** Comparison between neural data and the computed model excitation patterns from Allen's passive RTM model (transfer function format). This CA model assumes an IHC cilia bundle displacement of about 50 pm at the neural rate threshold. **(b)** Comparison between neural data computed tuning curves from *Neely*'s active model [3.98]. This CA model assumes an IHC cilia bundle displacement of 300 pm (0.3 nm) at the neural rate threshold

$v(x, s)$

$$\begin{pmatrix} P \\ V \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} f \\ v \end{pmatrix} , \qquad (3.15)$$

where A, B, C, and D are complex functions of place X and radian frequency s.

Passive BM Models

The most successful *passive* model of cochlear tuning is the resonant tectorial membrane (RTM) model [3.9, 104]. The RTM model starts from the assumption that the slope S_2 of BM tuning is insufficient to account for the slope S_2 of neural tuning, as seen in Fig. 3.4b. This sharpening is accounted for by a reflection in the tectorial membrane, introducing an antiresonance (*spectral zero*) at frequency F_z (Fig. 3.4b), which is about half an octave below the resonant frequency F_{cf} of the basilar membrane. As described by *Allen* and *Neely* [3.9], the detailed A, B, C, D elements of (3.15) are given by *Allen* [3.104], *Allen* and *Neely* [3.9].

As described in *Allen* [3.105], the response ratio of IHC cilia bundle displacement to basilar membrane displacement is defined as $H_{ihc}(x, s)$. The parameters of the RTM model may be chosen such that model results fit the experimental neural threshold tuning curves closely, as shown in Fig. 3.9a.

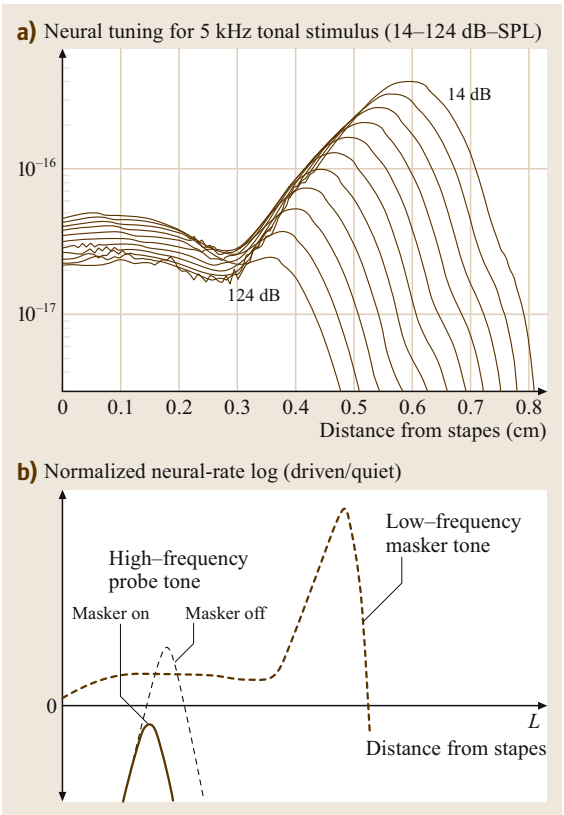


Fig. 3.10a,b In **(a)** results of model calculations by *Sen* and *Allen* [3.103] are shown of a NL BM stiffness model. On the right shows a cartoon of what might happen to the excitation pattern of a low-level probe when a suppressor is turned on given such a nonlinearity. The presence of the suppressor causes the probe to be suppressed and shifted slightly toward the base when the stiffness is decreased with increased level. It may be inferred from Fig. 3.3a that, if the BM stiffness is reduced, the location of the maximum will shift to the base, as is seen in real data. **(a)** Compression in the NL RTM model. Note how the response at the peak is reduced as the BM stiffness changes, causing the peak to shift to the base. As this happens the response in the tail region between $0 \le X \le 0.3$ cm becomes more sensitive, and thus shows an expansive NL response. All of these effects have been seen in real BM data. **(b)** Cartoon showing the effect of a low-side masker on a high-frequency tone as a function of position along the basilar membrane. When the suppressor is turned on, the CF of the high-frequency probe becomes less sensitive and shifts to higher frequencies. We model this effect in the panel on the left as BM stiffness that depends on level (i. e., $K_p(I_s)$)

The Nonlinear RTM Model. The resonant tectorial membrane (RTM) model is made NL by control of the BM stiffness via OHC's stiffness is based on Fig. 3.1b. The OHC soma stiffness has been shown to be voltage dependent by *Dallos* et al. [3.106] and dependent on prestin in the membrane wall [3.107]. If an elastic connection is assumed where the TM attaches to the Limbus, and if this elasticity is similar to that of the cilia of the OHC, then the resulting transfer function between the BM and IHC cilia is strongly filtered at low frequencies [3.51,103,108,109]. Such models are actively under consideration [3.102].

It is postulated that the decrease in OHC stiffness accompanying cilia stimulation results in a decrease of the net BM partition stiffness $K_p(x)$ (i. e., increasing compliance) of (3.6). As shown in Fig. 3.3, this decrease in the local BM stiffness would result in the partition excitation pattern shifting basally towards the stapes. Such shifts in the BM response patterns are commonly seen. Another way to view this is shown in Fig. 3.10. This migration of the excitation pattern, combined with the assumption that the TM has a high-pass characteristic, means that the cilia excitation gain at CF is nonlinearly compressed

as the intensity increases. This compression effect is shown in a cartoon format in Fig. 3.10b, while Fig. 3.10a shows the actual calculated model results. Note how the bandwidth $\Delta_f(X)$) remains approximately constant as a function of input intensity.

Sewell [3.110] has nicely demonstrated that as the voltage driving the hair cells changes, the neural gain in dB at CF changes proportionally. It is not yet known why the dB gain is proportional to the voltage (1 dB/mv), however this would explain why forward masking decays linearly in dB value with time, after a strong excitation, since the membrane voltage $V_m(t)$ is proportional to e^{-t/τ_m}, due to the OHC membrane's $\tau_m = RC$ time constant. In my view, explaining the proportionality between the neural threshold in dB and the linear membrane voltage, is key.

Discussion. Two important advantages of the NL RTM model include its physically based assumptions (described above), and its simplicity. Given these physical assumptions, we show next that the NL RTM model can explain:

1. the basal-ward half-octave traveling-wave migration as a function of increasing intensity [3.76],
2. the upward spread of masking (USM) [3.20, 21], two-tone suppression (2TS) (see Sect. 3.2.1),
3. distortion product generation [3.49,55,56,111–113],
4. normal and recruiting loudness growth, and
5. hypersensitive tails [3.45].

From the steep 2.5 dB/dB slope of the USM and 2TS (Fig. 3.6a) it seems necessary that the low-frequency suppressor is turning down the high-frequency probe even though the growth of the masker at the high frequency's place is linear with masker level, as shown in Fig. 3.10b.

Active BM Models

One obvious question about active cochlear models is *Are they really necessary?* At least three attempts to answer this question based on detailed comparisons of basilar membrane responses have concluded that the measured responses *cannot* be accounted for by a passive cochlear model [3.93, 114–117].

The CA Hypothesis. The most popular active micromechanical theory is called the *cochlear amplifier* (CA) hypothesis. The concept of the *cochlear amplifier*, originated by Gold, Kemp, Kim and Neely, and named by H. Davis, refers to a hypothetical mechanism within the cochlear partition which increases the *sensitivity* of basilar membrane vibrations to low-level sounds and, at the same time, increases the *frequency selectivity* of these vibrations [3.94]. The CA adds mechanical energy to the cochlear partition at acoustic frequencies by drawing upon the electrical and mechanical energy available from the outer hair cells. In response to a tone, the CA adds mechanical energy to the cochlear traveling wave in the region defined by S_2 as it approaches the place of maximum response. This energy is reabsorbed at other places along the cochlear partition. The resulting improvement in sensitivity of the ear due to the CA is thought to be 40 dB, or more under certain conditions; however, the details of how this amplification might be accomplished are still unknown [3.118, 119]. A general discussion of this model is presented in *Geisler* [3.90], and in *Allen* and *Fahey* [3.91].

It is presumed that this OHC action amplifies the BM signal energy on a cycle-by-cycle basis, increasing the sensitivity [3.69, 92]. In some of the models it is assumed that this cycle-by-cycle pressure (force) due to the OHCs causes the sharp BM tuning tip. In most of these models, the CA is equivalent to introducing a frequency-dependent negative damping (resistance) into the BM impedance [3.120]. Nonlinear compression is introduced by assuming that the resistance is signal level dependent. This NL resistance model was first described by *Hall* [3.84] for the case of $R > 0$. Thus the CA model is an extension of Hall's model to the case of $R < 0$. In several models NL negative damping is obtained with a nonlinear stiffness and a small delay. The addition of a small delay introduces a negative real part into the impedance. In mathematical physics, NL damping resonators are described by *van der Pol equations*, while NL stiffness resonators are described by *Duffing equations* [3.121].

Allen and *Fahey* [3.91] developed a method for directly measuring the cochlear amplifier (CA) gain. All of the studies to date using this method have found no gain. However many researchers continue to believe that the CA has gain. Given that the gain is order 40–50 dB this is difficult to understand. A nice summary of this situation has been recently published in *Shera* and *Guinan* [3.120]. The reasons for the failure to directly measure any CA gain are complex and multifaceted, and many important questions remain open. One possibility that remains open is that the many observed large NL OHC BM effects we see are not due to cycle-by-cycle power amplification of the BM traveling wave.

Discussion and Summary

Discussion. Both active and passive BM models are reasonably successful at simulating the neural thresh-

old response tuning curves. Thus we may need to look elsewhere to contrast the difference between these two approaches, such as 2TS/USM. While the passive RTM model is easily made NL with the introduction of $K_{ohc}(V_m)$, differences between *nonlinear* RTM and CA models have not yet been investigated. The CA and RTM models differ in their interpretation of damaged cochlear responses. In CA models, the loss of sensitivity of the cochlea with damage is interpreted as a loss of CA gain while in passive models, the loss of sensitivity has been interpreted as a 2:1 change in the BM stiffness [3.122].

The discovery of OHC motility demonstrates the existence of a potential source of mechanical energy within the cochlear partition which is suitably positioned to influence vibrations of the basilar membrane. It is still an open question whether this source of energy is sufficient to power a CA at high frequencies.

One possible advantage of the CA is that of improving the signal-to-noise ratio in front of the IHC detector. A weakness of the CA models has been their lack of specificity about the physical realization of the active elements. Until we have a detailed physical representation for the CA, RTM models have the advantage of being simpler and more explicit.

The discovery by He and Dallos that the OHC soma stiffness is voltage dependent is an exciting development for the NL passive RTM model, as it greatly simplifies the implementation of the physical model. The RTM model has been in disfavor because many feel it does not account for basilar membrane tuning. This criticism is largely due to the experimental results of physiologists who have measured the BM–ear canal transfer function, and found the tuning of BM velocity to be similar to neural threshold response data. Much of the experimental BM data, however, are not convincing on this point, with the BM slope S_2 (Fig. 3.4b) generally being much smaller than that of neural responses [3.97]. The question of whether an active model is required to simulate measured BM responses is still being debated.

Better estimates of the amplitude of cilia bundle displacement at a given sound pressure level directly address the sensitivity questions. If the estimate of *Russell* of 30 mV/degree is correct [3.123], then the cochlear sensitivity question may be resolved by having very sensitive detectors. Also, better estimates are needed of the ratio of the BM frequency response to the IHC frequency response, both at high and low frequencies. Rhode's approach of using the slopes of Fig. 3.4b rather than traditional ad hoc bandwidth measures, is a useful tool in this regard. The bandwidth 10 dB down relative to the peak has been popular, but arbitrary and thus poor, criterion in cochlear research. A second, somewhat better, bandwidth measure is Fletcher's *equivalent rectangular bandwidth* discussed in *Allen* [3.10].

Summary. This section has reviewed what we know about the cochlea. The *Basics* section reviews the nature of modeling and briefly describes the anatomy of the inner ear, and the function of inner and outer hair cells. In Sect. 3.1.2 we reviewed the history of cochlear modeling. The Wegel and Lane paper was a key paper that introduced the first detailed view of masking, and in the same paper introduced the first modern cochlear model Fig. 3.2b. We presented the basic tools of cochlear modeling, *impedance*, and introduced the *transmission matrix* method (two-port analysis). We describe how these models work in intuitive terms, including how the basilar membrane may be treated as having a frequency dependent acoustic hole. The location of the hole, as a function of frequency, is called the cochlear map. This hole keeps fluid from flowing beyond a certain point, producing the cochlear traveling wave.

We reviewed and summarized the NL measures of cochlear response. Since these data are not fully understood, and have not been adequately modeled, this is the most difficult section. However it is worth the effort to understand these extensive data and to appreciate the various relations between them, such as the close parallel between two-tone suppression and the upward spread of masking, and between loudness recruitment and outer hair cell damage.

We review several models of the hair cell, including forward and reverse transduction. Some of this material is recently published, and the view of these models could easily change over the next few years as we better understand reverse transduction.

Finally in Sect. 3.2.3 we reviewed the basics of micromechanics. We have presented the two basic types of models, *passive* and *active* models, with a critical review of each.

3.3 Neural Masking

When modeling human psychophysics one must carefully distinguish the external *physical* variables, which we call Φ variables, from the internal *psychophysical* variables, or Ψ variables. It may be helpful to note that

Φ and Ψ sound similar to the initial syllable of the words *phy*sical and *psy*chological, respectively [3.124]. Psychophysical modeling seeks a transformation from the Φ domain to the Ψ domain. The Φ intensity of a sound is easily quantified by direct measurement. The Ψ intensity is the loudness. The idea that loudness could be quantified was first suggested by *Fechner* [3.125] in 1860, who raised the question of the quantitative transformation between the physical and psychophysical intensity. For a recent review of this problem, and a brief summary of its long history, see *Schlauch* et al. [3.126]. This section is based on an earlier report by *Allen* [3.79], and *Allen* and *Neely* [3.127].

An increment in the intensity of a sound that results in a *just noticeable difference* is called an intensity JND. Fechner suggested quantifying the intensity-loudness growth transformation by counting the number of the *loudness JNDs* between two intensity values. However, after many years of work, the details of the relationship between loudness and the intensity JNDs remain unclear [3.128–130].

The contribution of *Allen* and *Neely* [3.127] and *Allen* [3.79] is that it takes a new view of the problem of the intensity JND and loudness by merging the 1953 Fletcher neural excitation pattern model of loudness [3.10, 131] with auditory signal detection theory [3.132].

It is generally accepted that the intensity JND is the physical correlate of the psychological-domain uncertainty corresponding to the psychological intensity representation of a signal. Along these lines, for long duration pure tones and wide-band noise, we assume that the Ψ-domain intensity is the loudness, and that the loudness JND results from loudness *noise* due to its stochastic representation.

To model the intensity JND we must define a *decision variable* associated with loudness and its random fluctuations. We call this loudness random decision variable the *single-trial loudness*. Accordingly we define the loudness and the loudness JND in terms of the first and second moments of the single-trial loudness, that is the mean and variance of the distribution of the single-trial loudness decision variable. We also define the ratio of the mean loudness to the loudness standard deviation as the *loudness signal-to-noise ratio*, $\mathrm{SNR_L}$.

Our ultimate goal in this work is to use signal detection theory to unify masking and the JND, following the 1947 outline of this problem by *Miller* [3.133]. Tonal data follows the *near miss to Weber's law* (thus does not obey Weber's law), while the wide-band noise data does obey Weber's law. We will show that the transformation of the Φ-domain (intensity) JND data (both tone and noise) into the Ψ domain (loudness) unifies these two types of JND data, since $\mathrm{SNR_L}(L)$ is the same for both the tone and noise cases. To help understand these results, we introduce the concept of a near miss to Stevens' law, which we show cancels the near–miss to Weber's law, giving the invariance in $\mathrm{SNR_L}$ for the tone case [3.127]. This work has applications in speech and audio coding.

For the case of tones, we have chosen to illustrate our theoretical work using the classical intensity modulation measurements of *Riesz* [3.75] who measured the intensity JND using small, low-frequency (3 Hz), sinusoidal modulation of tones. *Modern* methods generally use *pulsed* tones which are turned on and off somewhat abruptly, to make them suitable for a two-alternative, forced-choice (2AFC) paradigm. This transient could trigger cochlear forward masking. Riesz's modulation method has a distinct advantage for characterizing the internal signal detection process, because it maintains a nearly steady-state small-signal condition within the auditory system, minimizing any cochlear forward masking component. The interpretation of intensity JNDs is therefore simplified since underlying stochastic processes are stationary.

An outline of this neural masking section is as follows. After some basic definitions in Sect. 3.3.1 and a review of historical models (e.g., Weber and Fechner), in Sect. 3.3.2, we explore issues surrounding the relation between the intensity JND and loudness, for the special cases of tones in quiet and for wide-band noise. First, we look at formulae for counting the number of intensity and loudness JNDs and we use these formulae, together with decision-theoretic principles, to relate loudness to the intensity JND. We then review the loudness-JND theory developed by *Hellman* and *Hellman* [3.134], which provided the inspiration for the present work. Next, we empirically estimate the loudness SNR, defined as the mean loudness over the loudness variance, and proportional to $L/\Delta L$, as a function of both intensity and loudness, using the tonal JND data of *Riesz* [3.75] and the loudness growth function of *Fletcher* and *Munson* [3.41]. We then repeat this calculation for Miller's wide-band noise JND and loudness data. Finally we propose a model of loudness that may be used to compute the JND. This model merges Fletcher's neural excitation pattern model of loudness with signal detection theory.

3.3.1 Basic Definitions

We need a flexible yet clear notation that accounts for important time fluctuations and modulations that are present in the signals, such as beats and gated signals. We include a definition of *masked threshold* because we view the intensity JND as a special case of the masked threshold [3.133]. We include a definition of *beats* so that we can discuss their influence on Riesz's method for the measurement of intensity JNDs.

Intensity. In the time domain, it is common to define the Φ *intensity* in terms of the time-integrated squared signal pressure $s(t)$, namely,

$$I_s(t) \equiv \frac{1}{\varrho c T} \int_{t-T}^{t} s^2(t)\,\mathrm{d}t \,, \tag{3.16}$$

where T is the integration time and ϱc is the specific acoustic impedance of air. The *intensity level* is defined as I_s/I_{ref}, and the *sound pressure level* as $|s|/s_{\mathrm{ref}}$, where the reference intensity is I_{ref} or $10^{-10}\,\mu\mathrm{W/cm^2}$ and the reference pressure $s_{\mathrm{ref}} = 20\,\mu\mathrm{Pa}$. These two reference levels are equivalent at only one temperature, but both seem to be in use. Equivalence of the pressure and intensity references requires that $\varrho c = 40\,\mathrm{cgs}$ Rayls. At standard atmospheric pressure, this is only true when the temperature is about 39 °C. Such levels are typically expressed in dB units.

Intensity of Masker plus Probe. The JND is sometimes called *self-masking*, to reflect the view that it is determined by the internal noise of the auditory system. To model the JND it is useful to define a more-general measure called the *masked threshold*, which is defined in the Φ domain in terms of a nonnegative pressure scale factor α applied to the probe signal $p(t)$ that is then added to the masking pressure signal $m(t)$. The relative intensity of the probe and masker is varied by changing α. Setting $s(t) = m(t) + \alpha p(t)$, we denote the combined intensity as

$$I_{\mathrm{m+p}}(t,\alpha) \equiv \frac{1}{\varrho c T} \int_{t-T}^{t} [m(t) + \alpha p(t)]^2\,\mathrm{d}t \,. \tag{3.17}$$

The unscaled probe signal $p(t)$ is chosen to have the same long-term average intensity as the masker $m(t)$, defined as I. Let $I_{\mathrm{m}}(t)$ be the intensity of the masker with no probe ($\alpha = 0$), and $I_{\mathrm{p}}(t,\alpha) = \alpha^2 I$ be the intensity of the scaled probe signal with no masker. Thus

$$I \equiv I_{\mathrm{m+p}}(t,0) = I_{\mathrm{m}}(t) = I_{\mathrm{p}}(t,1) \,.$$

Because of small fluctuations in I_{m} and I_{p} due to the finite integration time T, this equality cannot be exactly true. We are specifically ignoring these small rapid fluctuations – when these rapid fluctuations are important, our conclusions and model results must be reformulated.

Beats. Rapid fluctuations having frequency components outside the bandwidth of the period T_{second} rectangular integration window are very small and will be ignored (T is assumed to be large). Accordingly we drop the time dependence in terms I_{m} and I_{p}. The beats between $m(t)$ and $p(t)$ of these signals are within a common critical band. Slowly varying correlations, between the probe and masker having frequency components within the bandwidth of the integration window, may *not* be ignored, as with beats between two tones separated in frequency by a few Hz. Accordingly we keep the time dependence in the term $I_{\mathrm{m+p}}(t,\alpha)$ and other slow–beating time dependent terms. In the Φ domain these beats are accounted for as a probe–masker correlation function $\rho_{\mathrm{mt}}(t)$ [3.132, p. 213].

Intensity Increment $\delta I(t,\alpha)$. Expanding (3.17) and solving for the *intensity increment* δI we find

$$\delta I(t,\alpha) \equiv I_{\mathrm{m+p}}(t,\alpha) - I = \left[2\alpha\rho_{\mathrm{mp}}(t) + \alpha^2\right] I \,, \tag{3.18}$$

where

$$\rho_{\mathrm{mp}}(t) = \frac{1}{\varrho c T I} \int_{t-T}^{t} m(t)\,p(t)\,\mathrm{d}t \tag{3.19}$$

defines a normalized cross-correlation function between the masker and the probe. The correlation function must lie between -1 and 1.

Detection Threshold. As the probe-to-masker ratio α is increased from zero, the probe can eventually be detected. We specify the probe *detection threshold* as α_*, where the asterisk indicates the threshold value of α where a subject can discriminate intensity $I_{\mathrm{m+p}}(t,\alpha_*)$ from intensity $I_{\mathrm{m+p}}(t,0)$ 50% of the time, corrected for chance (i. e., obtain a 75% correct score in a direct comparison of the two signals [3.132, p. 129]). The quantity $\alpha_*(t,I)$ is the probe to masker root-mean-square (RMS) pressure ratio at the detection threshold. It is a function of the masker intensity I and, depending on the experimental setup, time. α_* summarizes the experimental measurements.

Masked Threshold Intensity. When $\rho_{mp} = 0$, the *masked threshold intensity* is defined in terms of α_* as

$$I_p^*(I) \equiv I_p(\alpha_*) = \alpha_*^2 I \,,$$

which is the threshold intensity of the probe in the presence of the masker.

The masked threshold intensity is a function of the stimulus modulation parameters. For example, tone maskers and narrow-band noise maskers of equal intensity, and therefore approximately equal loudness, give masked thresholds that are about 20 dB different [3.135]. As a second example, when using the method of beats [3.75], the just–detectable modulation depends on the beat frequency. With *modern* 2AFC methods, the signals are usually gated on and off (100% modulation) [3.136]. According to *Stevens* and *Davis* [3.137, p. 142]

> *A gradual transition, such as the sinusoidal variation used by Riesz, is less easy to detect than an abrupt transition; but, as already suggested, an abrupt transition may involve the production of unwanted transients.*

One must conclude that the *relative masked threshold* [i. e., $\alpha_*(t, I)$] is a function of the modulation conditions, and depends on ρ_{mp}, and therefore T.

Ψ-Domain Temporal Resolution. When modeling time-varying psychological decision variables, the relevant integration time T is not the duration defined by the Φ intensity (3.16), rather the integration time is determined in the Ψ domain. This important Ψ-domain model parameter is called *loudness temporal integration* [3.138]. It was first explicitly modeled by *Munson* in 1947 [3.139].

The Φ-domain temporal resolution (T) is critical to the definition of the JND in Riesz's experiment because it determines the measured intensity of the beats. The Ψ-domain temporal resolution plays a different role. Beats cannot be heard if they are faster than, and therefore *filtered* out by, the Ψ domain response. The Ψ-domain temporal resolution also impacts results for gated stimuli, such as in the 2AFC experiment, though its role is poorly understood in this case. To model the JND as measured by Riesz's method of just-detectable beats, one must know the Ψ-domain resolution duration to calculate the probe–masker effective correlation $\rho_{mp}(t)$ in the Ψ domain. It may be more practical to estimate the Ψ domain resolution from experiments that estimate the degree of correlation, as determined by the beat modulation detection threshold as a function of the beat frequency f_b.

In summary, even though Riesz's modulation detection experiment is technically a masking task, we treat it, following *Riesz* [3.75], *Miller* [3.133], and *Littler* [3.16], as characterizing the intensity JND. It follows that the Ψ-domain temporal resolution plays a key role in intensity JND and masking models.

The Intensity JND ΔI. The intensity *just-noticeable difference* (JND) is

$$\Delta I(I) \equiv \delta(t, \alpha_*) \,, \tag{3.20}$$

the intensity increment at the masked threshold, for the special case where the probe signal is equal to the masking signal ($p(t) = m(t)$). From (3.18) with α set to threshold α_* and $\rho_{mp}(t) = 1$

$$\Delta I(I) = (2\alpha_* + \alpha_*^2) I \,. \tag{3.21}$$

It is traditional to define the intensity JND to be a function of I, rather than a function of $\alpha(I)$, as we have done here. We shall treat both notations as equivalent [i. e., $\Delta I(I)$ or $\Delta I(\alpha)$].

An important alternative definition for the special case of the *pure-tone JND* is to let the masker be a pure tone, and let the probe be a pure tone of a slightly different frequency (e.g., a beat frequency difference of $f_b = 3$ Hz). This was the definition used by *Riesz* [3.75]. Beats are heard at $f_b = 3$ Hz, and assuming the period of 3 Hz is within the passband of the Ψ temporal resolution window, $\rho_{mp}(t) = \sin(2\pi f_b t)$. Thus

$$\Delta I(t, I) = \left[2\alpha_* \sin(2\pi f_b t) + \alpha_*^2\right] I \,. \tag{3.22}$$

If the beat period is less than the Ψ temporal resolution window, the beats are *filtered* out by the auditory brain (the effective ρ_{mn} is small) and we do not hear the beats. In this case $\Delta I(I) = \alpha_*^2 I$. This model needs to be tested [3.139].

Internal Noise. It is widely accepted that the pure-tone intensity JND is determined by the *internal noise* of the auditory system [3.140, 141], and that ΔI is proportional to the standard deviation of the Ψ-domain decision variable that is being discriminated in the intensity detection task, reflected back into the Φ domain. The usual assumption, from signal detection theory, is that $\Delta I = d'\sigma_I$, where d' is defined as the proportionality between the change in intensity and the variance $d' \equiv \Delta I/\sigma_I$. Threshold is typically when $d' = 1$ but can

depend on the the experimental design; σ_I is the intensity standard deviation of the Φ-domain intensity due to Ψ-domain auditory noise [3.15, 17, 127].

Hearing Threshold. The *hearing threshold* (or unmasked threshold) *intensity* may be defined as the intensity corresponding to the first (lowest intensity) JND. The hearing threshold is represented as $I_p^*(0)$ to indicate the probe intensity when the masker intensity is small (i. e., $I \to 0$). It is believed that internal noise is responsible for the hearing threshold.

Loudness L. The *loudness* L of a sound is the Ψ intensity. The *loudness growth function* $L(I)$ depends on the stimulus conditions. For example $L(I)$ for a tone and for wide-band noise are not the same functions. Likewise the loudness growth function for a 100 ms tone and a 1 s tone differ. When defining a *loudness scale* it is traditional to specify the intensity, frequency, and duration of a tone such that the loudness growth function is one [$L(I_{ref}, f_{ref}, T_{ref}) = 1$ defines a loudness scale]. For the sone scale, the reference signal is a $I_{ref} = 40\,\text{dB} - \text{SPL}$ tone at $f_{ref} = 1\,\text{kHz}$ with duration $T_{ref} = 1\,\text{s}$. For Fletcher's LU scale the reference intensity is the hearing threshold, which means that 1 sone = 975 LU [3.42] for a normal hearing person. Fletcher's LU loudness scale seems a more-natural scale than the sone scale used in the American National Standards Institute (ANSI) and International Organization for Standardization (ISO) standards.

The Single-Trial Loudness. A fundamental postulate of psychophysics is that all decision variables (i. e., Ψ variables) are random variables, drawn from some probability space [3.132, Chap. 5]. For early discussions of this point see *Montgomery* [3.142] and p. 144 of *Stevens* and *Davis* [3.137]. To clearly indicate the distinction between random and nonrandom variables, a tilde ($\sim$) is used to indicate a random variable. As a mnemonic, we can think of the $\sim$ as a *wiggle* associated with randomness.

We define the loudness decision variable as the *single-trial loudness* $\widetilde{L}$, which is the sample loudness heard on each stimulus presentation. The loudness L is then the expected value of the single-trial loudness $\widetilde{L}$

$$L(I) \equiv \mathcal{E}\widetilde{L}(I)\,. \tag{3.23}$$

The second moment of the single-trial loudness

$$\sigma_L^2 \equiv \mathcal{E}(\widetilde{L} - L)^2 \tag{3.24}$$

defines the loudness *variance* σ_L^2 and *standard deviation* σ_L.

Derived Definitions

The definitions given above cover the basic variables. However many alternative forms (various normalizations) of these variables are used in the literature. These derived variables were frequently formed with the hope of finding an invariance in the data. This could be viewed as a form of modeling exercise that has largely failed (e.g., the near miss to Weber's law), and the shear number of combinations has led to serious confusions [3.138, p. 152]. Each normalized variable is usually expressed in dB, adding an additional unnecessary layer of confusion to the picture. For example, *masking* is defined as the masked threshold normalized by the unmasked (quiet) threshold, namely

$$M \equiv \frac{I_p^*(I_m)}{I_p^*(0)}\,.$$

It is typically quoted in dB re sensation level (dB − SL). The intensity JND is frequently expressed as a *relative JND* called the *Weber fraction* defined by

$$J(I) \equiv \frac{\Delta I(I)}{I}\,. \tag{3.25}$$

From the signal detection theory premise that $\Delta I = d'\sigma_I$ [3.17], J is just the reciprocal of an effective signal-to-noise ratio defined as

$$\text{SNR}_I(I) \equiv \frac{I}{\sigma_I(I)} \tag{3.26}$$

since

$$J = d'\frac{\sigma_I}{I} = \frac{d'}{\text{SNR}_I}\,. \tag{3.27}$$

One conceptual difficulty with the Weber fraction J is that it is an *effective* signal-to-noise ratio, expressed in the Φ (physical) domain, but determined by a Ψ (psychophysical) domain mechanism (internal noise), as may be seen from Fig. 3.11.

Loudness JND ΔL. Any suprathreshold Ψ-domain increments may be quantified by corresponding Φ domain increments. The *loudness* JND $\Delta L(I)$ is defined as the change in loudness $L(I)$ corresponding to the intensity JND $\Delta I(I)$. While it is not possible to measure ΔL directly, we assume that we may expand the loudness function in a Taylor series (Fig. 3.11), giving

$$L(I + \Delta I) = L(I) + \Delta I \left.\frac{dL}{dI}\right|_I + \text{HOT}\,,$$

where HOT represents *higher-order terms*, which we shall ignore. If we solve for

$$\Delta L \equiv L(I + \Delta I) - L(I) \tag{3.28}$$

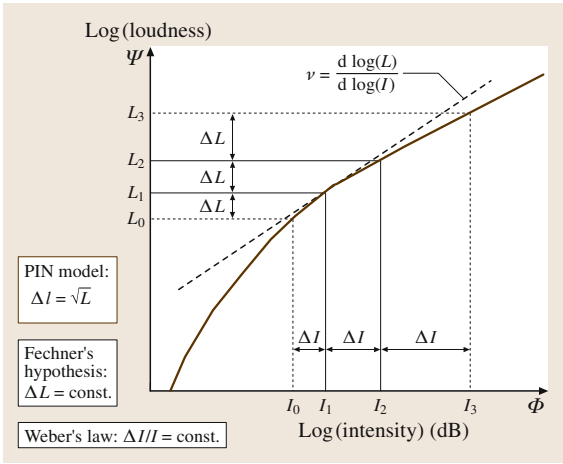


Fig. 3.11 Summary of all historical ideas about psychophysics and the relations between the Φ and Ψ variables. Along the abscissa we have the physical variable, intensity, and along the ordinate, the psychological variable loudness. The *curve* represents the loudness, on a log-intensity log-loudness set of scales. A JND in loudness is shown as ΔL and it depends on loudness, as described by the *Poisson internal noise* (PIN) model shown in the box on the left. Fechner assumed that ΔL was constant, which we now know to be incorrect. The loudness JND is reflected back into the physical domain as an intensity JND ΔI, which also depends on level. Weber's law, is therefore not true in general (but is approximately true for wide-band noise). Our analysis shows that the loudness SNR and the intensity SNR must be related by the slope of the loudness growth function, as given by (3.32). These relations are verified in Fig. 3.12, as discussed in detail in *Allen* and *Neely* [3.127]

we find

$$\Delta L = \Delta I \left.\frac{\mathrm{d}L}{\mathrm{d}I}\right|_{\mathrm{I}} . \tag{3.29}$$

We call this expression the *small*-JND approximation. The above shows that the loudness JND $\Delta L(I)$ is related to the intensity JND $\Delta I(I)$ by the slope of the loudness function, evaluated at intensity I. According to the signal detection model, the standard deviation of the single-trial loudness is proportional to the loudness JND, namely

$$\Delta L = d' \sigma_{\mathrm{L}} . \tag{3.30}$$

A more explicit way of expressing this assumption is

$$\frac{\Delta L}{\Delta I} = \frac{\sigma_{\mathrm{L}}}{\sigma_{\mathrm{I}}} , \tag{3.31}$$

where d' in both the Φ and Ψ domains is the same and thus cancels.

Loudness SNR. In a manner analogous to the Φ-domain $\mathrm{SNR_I}$, we define the Ψ-domain loudness SNR as $\mathrm{SNR_L}(L) \equiv L/\sigma_{\mathrm{L}}(L)$. Given (3.30), it follows that

$$\mathrm{SNR_I} = \nu \mathrm{SNR_L} , \tag{3.32}$$

where ν is the slope of the log-loudness function with respect to log-intensity. If we express the loudness as a power law

$$L(I) = I^{\nu}$$

and let $x = \log(I)$ and $y = \log(L)$, then $y = \nu x$. If the change of ν with respect to dB – SPL is small, then $\mathrm{d}y/\mathrm{d}x \approx \Delta y/\Delta x \approx \nu$. Since $\mathrm{d}\log(y) = \mathrm{d}y/y$ we get

$$\frac{\Delta L}{L} = \nu \frac{\Delta I}{I} . \tag{3.33}$$

Equation (3.32) is important because

1. it tells us how to relate the SNRs between the Φ and Ψ domains,
2. every term is dimensionless,
3. the equation is simple, since $\nu \approx 1/3$ is approximately constant above 40 dB – SL (i. e., Stevens' law), and because
4. we are used to seeing and thinking of loudness, intensity, and the SNR, on log scales, and ν as the slope on log–log scales.

Counting JNDs. While the concept of counting JNDs has been frequently discussed in the literature, starting with Fechner, unfortunately the actual counting formula (i. e., the equation) is rarely provided. As a result of a literature search, we found the formula in *Nutting* [3.143], *Fletcher* [3.21], *Wegel* and *Lane* [3.20], *Riesz* [3.75], *Fletcher* [3.144], and *Miller* [3.133].

To derive the JND counting formula, (3.29) is rewritten as

$$\frac{\mathrm{d}I}{\Delta I} = \frac{\mathrm{d}L}{\Delta L} . \tag{3.34}$$

Integrating over an interval gives the total number of intensity JNDs

$$N_{12} \equiv \int_{I_1}^{I_2} \frac{\mathrm{d}I}{\Delta I} = \int_{L_1}^{L_2} \frac{\mathrm{d}L}{\Delta L} , \tag{3.35}$$

where $L_1 = L(I_1)$ and $L_2 = L(I_2)$. Each integral counts the total number of JNDs in a different way between I_1 and I_2 [3.75,144]. The number of JNDs must be the same regardless of the domain (i. e., the abscissa variable), Φ or Ψ.

3.3.2 Empirical Models

This section reviews some earlier empirical models of the JND and its relation to loudness relevant to our development.

Weber's Law

In 1846 it was suggested by Weber that $J(I)$ is independent of I. According to (3.21) and (3.25)

$$J(I) = 2\alpha_* + \alpha_*^2 \,.$$

If J is constant, then α_* must be constant, which we denote by $\alpha_*(\not I)$ (we strike out I to indicate that α_* is not a function of intensity). This expectation, which is called Weber's law [3.145], has been successfully applied to many human perceptions. We refer the reader to the helpful and detailed review of these questions by *Viemeister* [3.129], *Johnson* et al. [3.146], and *Moore* [3.147].

Somewhat frustrating is the empirical observation that $J(I)$ is not constant for the most elementary case of a pure tone [3.75, 136]. This observation is referred to as *the near miss to Weber's law* [3.148].

Weber's law does make one simple prediction that is potentially important. From (3.35) along with Weber's law $J_0 \equiv J(\not I)$ we see that the formula for the number of JNDs is

$$N_{12} = \int_{I_1}^{I_2} \frac{\mathrm{d}I}{J_0 I} = \frac{1}{J_0} \ln\left(\frac{I_2}{I_1}\right) . \tag{3.36}$$

It remains unexplained why Weber's law holds as well as it does [3.149, 150, p. 721] (it holds approximately for the case of wide band noise), or even why it holds at all. Given the complex and NL nature of the transformation between the Φ and Ψ domains, coupled with the belief that the noise source is in the Ψ domain, it seems unreasonable that a law as simple as Weber's law could hold in any general way. A transformation of the JND from the Φ domain to the Ψ domain greatly clarifies the situation.

Fechner's Postulate

In 1860 Fechner postulated that the loudness JND $\Delta L(I)$ is a constant [3.125, 130, 151, 152]. We are only considering the auditory case of Fechner's more general theory. We shall indicate such a constancy with respect to I as $\Delta L(\not I)$ (as before, we strike out the I to indicate that ΔL is *not* a function of intensity). As first reported by *Stevens* [3.153], we shall show that Fechner's postulate is not generally true.

The Weber–Fechner Law

It is frequently stated [3.152] that Fechner's postulate ($\Delta L(\not I)$) and Weber's law ($J_0 \equiv J(\not I)$) lead to the conclusion that the difference in loudness between any two intensities I_1 and I_2 is proportional to the logarithm of the ratio of the two intensities, namely

$$\frac{L(I_2) - L(I_1)}{\Delta L} = \frac{1}{J_0} \log\left(\frac{I_2}{I_1}\right) . \tag{3.37}$$

This is easily seen by eliminating N_{12} from (3.36) and by assuming Weber's law and Fechner's hypothesis. This result is called *Fechner's law* (also called the *Weber–Fechner law*). It is not true because of the faulty assumptions, Weber's law and Fechner's postulate.

3.3.3 Models of the JND

Starting in 1923, *Fletcher* and *Steinberg* studied loudness coding of pure tones, noise, and speech [3.21, 154–156], and proposed that loudness was related to neural spike count [3.41], and even provided detailed estimates of the relation between the number of spikes and the loudness in sones [3.42, p. 271]. In 1943 *De Vries* first introduced a photon-counting Poisson process model as a theoretical basis for the threshold of vision [3.157]. *Siebert* [3.140] proposed that Poisson point-process noise, resulting from the neural rate code, acts as the internal noise that limits the frequency JND [3.136, 150]. A few years later [3.158], and independently [3.159] *McGill* and *Goldberg* [3.160] proposed that the Poisson internal noise (PIN) model might account for the intensity JND, but they did not find this to produce a reasonable loudness growth function. *Hellman* and *Hellman* [3.134] further refined the argument that Poisson noise may be used to relate the loudness growth to the intensity JND, and they found good agreement between the JND and realistic loudness functions.

Given Poisson noise, the variance is equal to the mean, thus

$$\Delta L(L) \propto \sqrt{L} \,. \tag{3.38}$$

This may also be rewritten as $\sigma_L^2 \propto L$. We would expect this to hold if the assumptions of *McGill* [3.148] (i. e., the PIN model) are valid.

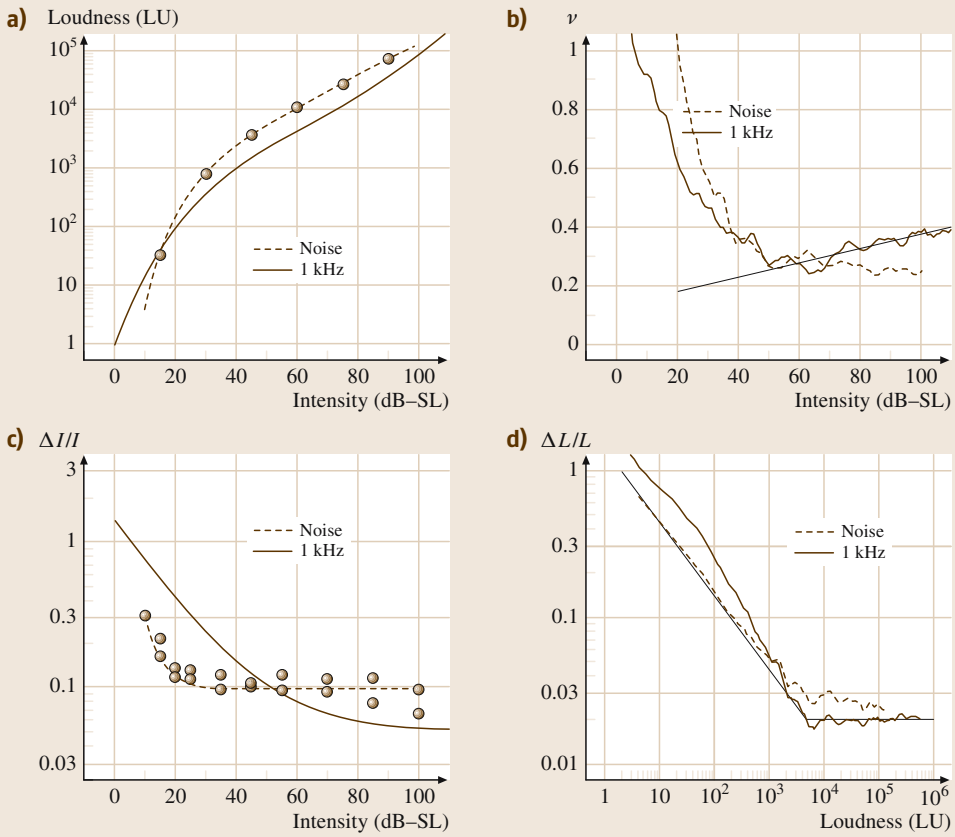


Fig. 3.12a–d In 1947 Miller measured the JND_I and the loudness level for two subjects using wide-band modulated noise (0.15–7 kHz) for levels between 3 and 100 dB − SL. The noise (*dashed line*) and pure tone (*solid line*) loudness are shown in **(a)**. The similarity between $\Delta L/L$ derived from the loudness curves for pure tones and for noise provide an almost perfect fit to the SPIN model which results from assuming the noise is neural point-process noise. See the text for a summary of these results. The direct derivation of ΔL based on pure tone JND and loudness data from *Miller* [3.133], *Riesz* [3.75], *Fletcher* and *Munson* [3.41].

In the following we directly compare the loudness–growth function of Fletcher and Munson to the number of JNDs N_{12} from Riesz [3.75, 127] to estimate $\Delta L/L$.

3.3.4 A Direct Estimate of the Loudness JND

Given its importance, it is important to estimate ΔL directly from its definition (3.28), using Riesz's $\Delta I(I)$ and Fletcher and Munson's 1933 estimate of $L(I)$.

Miller's 1947 famous JND paper includes wide-band-noise loudness-level results. We transformed these JND data to loudness using *Fletcher* and *Munson* [3.41] reference curve (i. e., Fig. 3.12a).

Loudness Growth, Recruitment, and the OHC

In 1924 *Fletcher* and *Steinberg* published an important paper on the measurement of the loudness of speech signals [3.155]. In this paper, when describing the growth of loudness, the authors state

> *the use of the above formula involved a summation of the cube root of the energy rather than the energy.*

This cube–root dependence had first been described by *Fletcher* the year before [3.21].

In 1930 *Fletcher* [3.27] postulated that there was a monotonic relationship between central nerve firings rates and loudness. Given a tonal stimulus at the ear

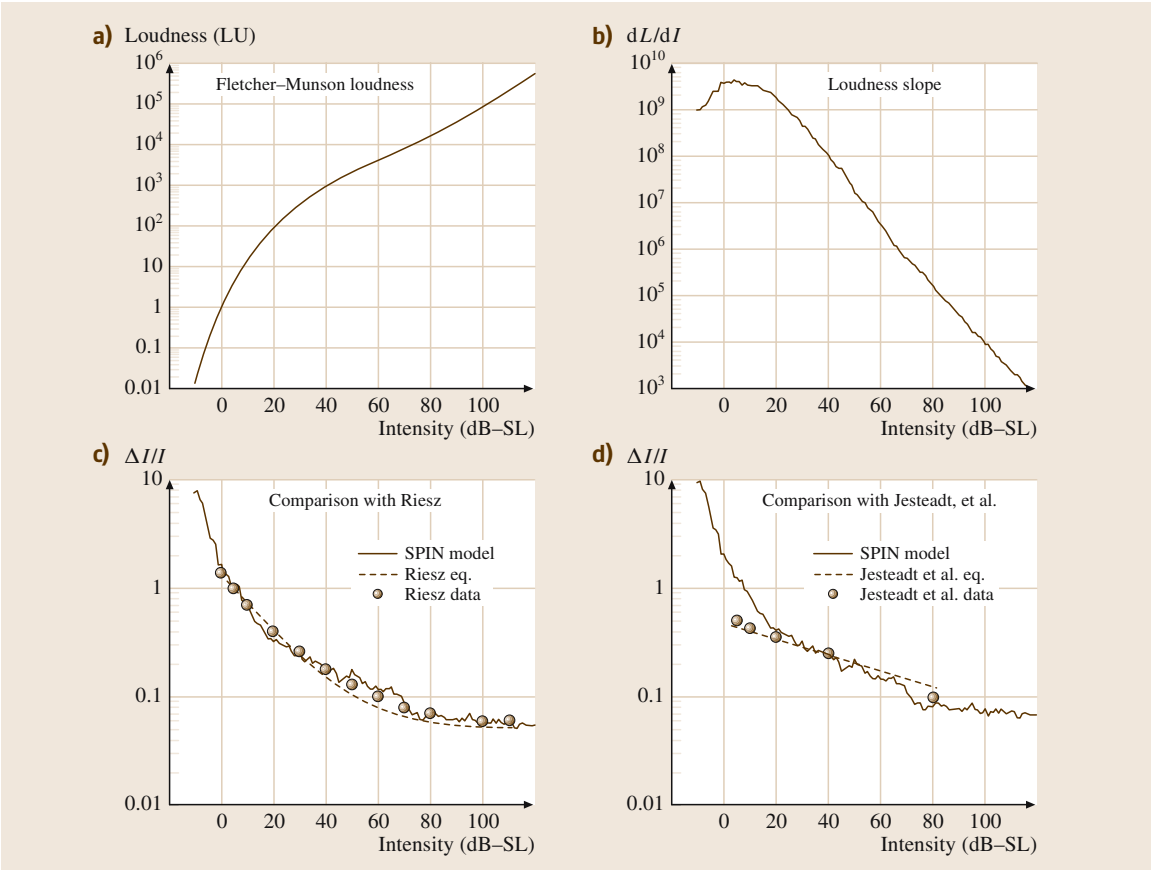


Fig. 3.13a–d Test of the SPIN model against the classic results of *Riesz* [3.75], *Jesteadt* et al. [3.136]. Test of the model derived on the *left* based on a comparison between loudness data and intensity JND data at 1 kHz, using the SPIN model

drum, Stevens' law says that the loudness is given by

$$L \equiv L(f, x, I) \propto I^{\nu} , \qquad (3.39)$$

where (f, x, I) are the frequency, place, and intensity of the tone, respectively. The exponent ν has been experimentally established to be in the range between 1/4 and 1/3 for long duration pure tones at 1 kHz. *Fletcher* and *Munson* [3.41] found $\nu \approx 1/4$ at high intensities and approximately 1 near threshold. Although apparently it has not been adequately documented, ν seems to be close to 1 for the *recruiting ear* [3.15].

Recruitment. What is the source of Fletcher's cube-root loudness growth (i. e., Stevens' law)? Today we know that cochlear outer hair cells are the source of the cube-root loudness growth observed by Fletcher.

From noise trauma experiments on animals and humans, we may conclude that recruitment (abnormal loudness growth) occurs in the cochlea [3.3, 96]. *Steinberg* and *Gardner* described such a loss as a *variable loss* (i. e., sensory neural loss) and partial recruitment as a mixed loss (i. e., having a conductive component) [3.3, 161]. They and Fowler verified the conductive component by estimating the air–bone gap. In a comment to Fowler's original presentation on loudness recruitment in 1937, the famous anatomist *Lorente de Nó* theorized that recruitment is due to hair cell damage [3.14]. *Steinberg* and *Gardner* clearly understood recruitment, as is indicated in the following quote [3.3, p. 20]

> *Owing to the expanding action of this type of loss it would be necessary to introduce a corresponding compression in the amplifier in order to produce the same amplification at all levels.*

This compression/loss model of hearing and hearing loss, along with the loudness models of *Fletcher* and *Munson* [3.41], are basic to an eventual quantitative understanding of NL cochlear signal processing and the cochlea's role in detection, masking and loudness in normal and impaired ears. The work by *Fletcher* [3.162] and *Steinberg* and *Gardner* [3.3], and work on modeling hearing loss and recruitment [3.122] support this view.

In summary, many studies conclude that the cube-root loudness growth starts with the NL compression of basilar membrane motion due to stimulus-dependent voltage changes within the OHC.

3.3.5 Determination of the Loudness SNR

In Fig. 3.12 we show a summary of $L(I)$, $\nu(I)$, $J(I)$ and $\Delta L/L = d'/\mathrm{SNR_L}$ for the tone and noise data.

The pure-tone and wide-band noise JND results may be summarized in terms of the loudness $\mathrm{SNR_L}(L)$ data shown in Fig. 3.12d where we show $\Delta L/L = d'/\mathrm{SNR_L}$, as a function of loudness.

For noise below $55\,\mathrm{dB} - \mathrm{SL}$ ($L < 5000\,\mathrm{LU}$) the loudness signal-to-noise ratio $\mathrm{SNR_L} \equiv L/\sigma_\mathrm{L}$ decreases as the square root of the loudness. For a loudness greater than $5000\,\mathrm{LU}$ ($N \approx 5$ sones), $\Delta L/L \approx 0.025$ fn both tones and noise (Fig. 3.12d)

In the lower-right panel (Fig. 3.12d) we provide a functional summary of $\Delta L/L$ for both tones and noise with the light solid line described by

$$\frac{\Delta L(L)}{L} = h\,[\min(L, L_0)]^{-1/2}\ , \tag{3.40}$$

where $h = \sqrt{2}$ and $L_0 = 5000\,\mathrm{LU}$ (≈ 5 sone). We call this relation the *saturated Poisson internal noise* (SPIN) model. With these parameter values, (3.40) appears to be a lower bound on the relative loudness $\mathrm{JND_L}$ for both tones and noise. From (3.33) $\Delta L/L = \nu(I)J(I)$. Note how the product of $\nu(I)$ and $J(I)$ is close to a constant for tones above 5000 LU.

In Fig. 3.12b the second top panel shows the exponent $\nu(I)$ for both Fletcher and Munson's and Miller's loudness growth function. In the lower-left panel (Fig. 3.12c) we see $\Delta I/I$ versus I for Miller's subjects, Miller's equation, and Riesz's JND equation.

Near miss to Stevens' Law

For tones the intensity exponent $\nu(I)$ varies systematically between 0.3 and 0.4 above $50\,\mathrm{dB} - \mathrm{SL}$, as shown by the solid line in the upper-right panel of Fig. 3.12b. We have highlighted this change in the power law with intensity for a 1 kHz tone in the upper-right panel with a light solid straight line. It is logical to call this effect the *near miss to Stevens' law,* since it cancels the near miss to Weber's law, giving a constant relative loudness JND $\Delta L/L$ for tones.

Figure 3.13a shows the Fletcher–Munson loudness data from Table III in [3.41]. The upper-right panel (Fig. 3.13b) is the slope of the loudness with respect to intensity (LU cm^2/W).In the lower–right (Fig. 3.13d) we compare the SPIN model relative JND (3.43) (with $h = 3.0$), and the relative JND computed from the *Jesteadt* et al. [3.136] formula (dashed line) and data from their Table B-I (circles). They measured the JND using pulsed tones for levels between 5 and 80 dB. The Jesteadt et al. data were taken with gated stimuli (100% modulation) and 2AFC methods. It is expected that the experimental method would lead to a different value of h than the valued required for Riesz's data set. The discrepancy between 0 and 20 dB may be due to the 100% modulation for these stimuli. The fit from 20 to $80\,\mathrm{dB} - \mathrm{SL}$ is less than a 5% maximum error, and much less in terms of RMS error. Note the similarity in slope between the model and the data.

3.3.6 Weber–Fraction Formula

In this section we derive the relation between the Weber fraction $J(I)$ given the loudness $L(I)$ starting from the *small-JND approximation*

$$\Delta L = \Delta I L'(I)\,, \tag{3.41}$$

where $L'(I) \equiv \mathrm{d}L/\mathrm{d}I$. If we solve this equation for ΔI and divide by I we find

$$J(I) \equiv \frac{\Delta I}{I} = \frac{\Delta L}{IL'(I)}\ . \tag{3.42}$$

Finally we substitute the SPIN model (3.40)

$$J(I) = \frac{hL(I)}{IL'(I)}\,[\min(L(I), L_0)]^{-1/2}\ . \tag{3.43}$$

This formula is the same as that derived by *Hellman* and *Hellman* [3.134], when $L \le L_0$. In Fig. 3.13c we plot (3.43) labeled *SPIN-model* with $h = 2.4$ and $L_0 = 10\,000\,\mathrm{LU}$. For levels between 0 and $100\,\mathrm{dB} - \mathrm{SL}$, the SPIN model (solid curve) fit to Riesz's data and Riesz's formula is excellent. Over this 100 dB range the curve defined by the loudness function fits as well as the curve defined by Riesz's formula [3.127]. The excellent fit gives us further confidence in the basic assumptions of the model.

3.4 Discussion and Summary

Inspired by the Poisson internal noise (PIN)-based theory of *Hellman* and *Hellman* [3.134], we have developed a theoretical framework that can be used to explore the relationship between the pure-tone loudness and the intensity JND. The basic idea is to combine Fletcher's neural excitation response pattern model of loudness with signal detection theory. We defined a random decision variable called the single-trial loudness. The *mean* of this random variable is the loudness, while its *standard deviation* is proportional to the loudness JND. We define the loudness signal-to-noise ratio $\mathrm{SNR_L}$ as the ratio of loudness (the signal) to standard deviation (a measure of the noise).

3.4.1 Model Validation

To evaluate the model we have compared the loudness data of *Fletcher* and *Munson* [3.41] with the intensity JND data of *Riesz* [3.75], for tones. A similar comparison was made for noise using loudness and intensity JND data from *Miller* [3.133]. We were able to unify the tone and noise data by two equivalent methods in Fig. 3.12d. Since the loudness SNR is proportional to the ratio of the loudness to the JND $L/\Delta L$, the SNR is also a piecewise power-law function which we call the SPIN model. All the data are in excellent agreement with the SPIN model, providing support for the validity of this theory.

The above discussion has

- drawn out the fundamental nature of the JND,
- shown that the PIN loudness model holds below 5 sone (5000 LU) (the solid line in the lower right panel of Fig. 3.11 below 5000 LU obeys the PIN model, and the data for both tones and wide band noise fall close to this line below 5000 LU) (one sone is 975 LU [3.127, p. 3631], thus 5000 LU = 5.13 LU. From the loudness scale this corresponds to a 1 kHz pure tone at 60 dB − SL),
- shown that above 5 sone the PIN model fails and the loudness SNR remains constant.

3.4.2 The Noise Model

The SPIN Model

Equation (3.40) summarizes our results on the relative loudness JND for both tones and noise. Using this formula along with (3.32), the JND may be estimated for tones and noise once the loudness has been determined, by measurement, or by model. Fechner's postulate, that the loudness JND is constant, is not supported by our analysis, in agreement with *Stevens* [3.153].

The PIN Model

The success of the PIN model is consistent with the idea that the pure-tone loudness code is based on neural discharge rate. This theory should apply between threshold and moderate intensities (e.g., < 60 dB) for *frozen stimuli* where the JND is limited by internal noise.

CNS Noise

Above 60 dB − SL we find that the loudness signal-to-noise ratio saturated (Fig. 3.12d) with a constant loudness SNR between 30 and 50 for both the tone and noise conditions, as summarized by Ekman's law [3.163]. We conclude that the Hellman and Hellman theory must be modified to work above 5 sones.

Weber's Law

It is significant that, while both $J(I)$ and $\nu(I)$ vary with intensity, the product is constant above 60 dB − SL. Given that $J = d'/\nu\mathrm{SNR_L}$, the saturation in $\mathrm{SNR_L}$ explains Weber's law for wideband signals (since ν and $\mathrm{SNR_L}$ for that case are constant) as well as the near miss to Weber's law for tones, where ν is not constant (the near miss to Stevens' law, Fig. 3.12a).

Generalization to Other Data

If $\sigma_L(L, I)$ depends on L, and is independent of I, then the $\mathrm{SNR_L}(L)$ should not depend on the nature of the function $L(I)$ (i. e., it should be true for any $L(I)$). This prediction is supported by our analysis summarized by (3.40). It will be interesting to see how $\mathrm{SNR_L}$ depends on L and I for subjects having a hearing-loss-induced recruitment, and how well this theory explains other data in the literature, such as loudness and JNDs with masking-induced recruitment [3.126].

Conditions for Model Validity

To further test the SPIN model, several conditions must be met. First the loudness and the JND must have been measured under the same stimulus conditions. Second, the internal noise must be the dominant factor in determining the JND. This means that the stimuli must be frozen (or have significant duration and bandwidth), and the subjects well trained in the task. As the signal uncertainty begins to dominate the internal noise, as it does in

the cases of roving the stimulus, the intensity JND will become independent of the loudness.

As discussed by *Stevens* and *Davis* [3.164, pp. 141-143], JND data are quite sensitive to the modulation conditions. The *Riesz* [3.75] and *Munson* [3.165] data make an interesting comparison because they are taken under steady–state conditions and are long duration tonal signals. Both sets of experimental data (i. e., Riesz and Munson) were taken in the same laboratory within a few years of each other. In 1928 Wegel, Riesz, and Munson were all members of Fletcher's department. *Riesz* [3.75] states that he used the same methods as *Wegel* and *Lane* [3.20], and it is likely that *Munson* [3.165] did as well.

Differences in the signal conditions are the most likely explanation for the differences observed in the intensity JND measurements of Riesz and Jesteadt shown in Fig. 3.13d. One difference between the data of *Riesz* [3.75] and *Jesteadt* et al. [3.136] is that Riesz varied the amplitude of the tones in a sinusoidal manner with a small (i. e., just detectable) modulation index, while Jesteadt et al. alternated between two intervals of different amplitude, requiring that the tones be gated on and off (i. e., a 100% modulation index).

The neural response to transient portions of a stimulus is typically larger than the steady-state response (e.g., neural overshoot) and, therefore, may dominate the perception of stimuli with large, abrupt changes in amplitude. The fact that the intensity JND is sensitive to the time interval between two tones of different amplitude [3.164] is another indication that neural overshoot may play a role.

It would be interesting to check the SPIN model on loudness and JND data taken using gated signals, given the observed sensitivity to the modulation. While these JND data are available [3.136], one would need loudness data taken with identical (or at least similar) modulations. We are not aware of such data.

References

3.1 E. Relkin, C. Turner: A reexamination of forward masking in the auditory nerve, J. Acoust. Soc. Am. **84**(2), 584–591 (1988)

3.2 M. Hewitt, R. Meddis: An evaluation of eight computer models of mammalian inner hair-cell function, J. Acoust. Soc. Am. **90**(2), 904–917 (1991)

3.3 J. Steinberg, M. Gardner: Dependence of hearing impairment on sound intensity, J. Acoust. Soc. Am. **9**, 11–23 (1937)

3.4 J.O. Pickles: *An Introduction to the Physiology of Hearing* (Academic, London 1982)

3.5 P. Dallos: Cochlear neurobiology. In: *The Cochlea*, ed. by P. Dallos, A. Popper, R. Fay (Springer, New York 1996) pp. 186–257

3.6 W.A. Yost: *Fundamentals of Hearing, An Introduction* (Academic, San Diego, London 2006)

3.7 S. Hecht: Vision II, The nature of the photoreceptor process. In: *Handbook of General Experimental Psychology*, ed. by C. Murchison (Clark Univ. Press, Worcester 1934)

3.8 G. Gescheider: *Psychophysics: The Fundamentals*, 3rd edn. (Lawrence Erlbaum, Mahwah 1997)

3.9 J.B. Allen, S. Neely: Micromechanical models of the cochlea, Phys. Today **45**(7), 40–47 (1992)

3.10 J.B. Allen: Harvey Fletcher's role in the creation of communication acoustics, J. Acoust. Soc. Am. **99**(4), 1825–1839 (1996)

3.11 A. Hudspeth, D. Corey: Sensitivity, polarity, and conductance change in the response of vertebrate hair cells to controlled mechanical stimuli, Proc. Nat. Acad. Sci. **74**(6), 2407–2411 (1977)

3.12 M. Liberman: Single-neuron labeling in the cat auditory nerve, Science **216**, 163–176 (1982)

3.13 J. Steinberg: Stereophonic sound-film system-pre- and post-equalization of compandor systems, J. Acoust. Soc. Am. **13**, 107–114 (1941), B-1327

3.14 R.L. de Nó: The diagnosis of diseases of the neural mechanism of hearing by the aid of sounds well above threshold, T. Am. Otol. Soc. **27**, 219–220 (1937looseness1)

3.15 S.T. Neely, J.B. Allen: Relation between the rate of growth of loudness and the intensity DL. In: *Modeling Sensorineural Hearing Loss*, ed. by W. Jesteadt (Lawrence Erlbaum, Mahwah 1997) pp. 213–222

3.16 T.S. Littler: *The Physics of the Ear* (Pergamon, Oxford 1965)

3.17 W.M. Hartmann: *Signals, Sound, and Sensation* (AIP Press, Woodbury 1997)

3.18 H.L.F. Helmholtz: *On the Sensations of Tone* (Dover, New York 1954), 1863

3.19 H.L.F. Helmholtz: *Helmholtz's popular scientific lectures* (Dover, New York 1962), 1857

3.20 R.L. Wegel, C. Lane: The auditory masking of one pure tone by another and its probable relation to the dynamics of the inner ear, Phys. Rev. **23**, 266–285 (1924)

3.21 H. Fletcher: Physical measurements of audition and their bearing on the theory of hearing, J. Franklin Inst. **196**(3), 289–326 (1923)

3.22 G. Campbell: On loaded lines in telephonic transmission, Phil. Mag. **5**, 313–331 (1903)

3.23 G. Campbell: Physical theory of the electric wave filter, Bell System Tech. J. **1**(1), 1–32 (1922)
3.24 G.A. Campbell: Telephonic intelligibility, Phil. Mag. **19**(6), 152–159 (1910)
3.25 H. Fletcher: The nature of speech and its interpretation, J. Franklin Inst. **193**(6), 729–747 (1922)
3.26 H. Fletcher, J. Steinberg: Articulation testing methods, J. Acoust. Soc. Am. **1**(2.2), 17–113 (1930), Intelligibility Lists pp. 65–113)
3.27 H. Fletcher: A space-time pattern theory of hearing, J. Acoust. Soc. Am. **1**(1), 311–343 (1930)
3.28 J. Zwislocki: Theorie der Schneckenmechanik, Acta Otolaryngol. **72**, 1–112 (1948)
3.29 J. Zwislocki: Theory of the acoustical action of the cochlea, J. Acoust. Soc. Am. **22**, 779–784 (1950)
3.30 O. Ranke: Theory operation of the cochlea: A contribution to the hydrodynamics of the cochlea, J. Acoust. Soc. Am. **22**, 772–777 (1950)
3.31 L.C. Peterson, B.P. Bogert: A dynamical theory of the cochlea, J. Acoust. Soc. Am. **22**, 369–381 (1950)
3.32 H. Fletcher: On the dynamics of the cochlea, J. Acoust. Soc. Am. **23**, 637–645 (1951)
3.33 H. Fletcher: Acoustics, Phys. Today **4**, 12–18 (1951)
3.34 S. Puria, J.B. Allen: Measurements and model of the cat middle ear: Evidence for tympanic membrane acoustic delay, J. Acoust. Soc. Am. **104**(6), 3463–3481 (1998)
3.35 J.B. Allen, P.S. Jeng, H. Levitt: Evaluating human middle ear function via an acoustic power assessment, J. Rehabil. Res. Dev. **42**(4), 63–78 (2005)
3.36 H. Fletcher, W. Munson: Relation between loudness and masking, J. Acoust. Soc. Am. **9**, 1–10 (1937)
3.37 H. Fletcher: Loudness, masking and their relation to the hearing process and the problem of noise measurement, J. Acoust. Soc. Am. **9**, 275–293 (1938)
3.38 H. Fletcher: Auditory patterns, Rev. Mod. Phys. **12**(1), 47–65 (1940)
3.39 J. Steinberg: Positions of stimulation in the cochlea by pure tones, J. Acoust. Soc. Am. **8**, 176–180 (1937), cochlear map estimate, Monograph B-973
3.40 D.D. Greenwood: Auditory masking and the critical band, J. Acoust. Soc. Am. **33**(4), 484–502 (1961)
3.41 H. Fletcher, W. Munson: Loudness, its definition, measurement, and calculation, J. Acoust. Soc. Am. **5**, 82–108 (1933)
3.42 H. Fletcher: *Speech and Hearing in Communication* (Krieger, Huntington 1953)
3.43 D.D. Greenwood: Critical bandwidth and the frequency coordinates of the basilar membrane, J. Acoust. Soc. Am. **33**, 1344–1356 (1961)
3.44 M. Liberman: The cochlear frequency map for the cat: Labeling auditory-nerve fibers of known characteristic frequency, J. Acoust. Soc. Am. **72**(5), 1441–1449 (1982)
3.45 M. Liberman, L. Dodds: Single neuron labeling and chronic cochlear pathology III: Stereocilia damage and alterations of threshold tuning curves, Hearing Res. **16**, 55–74 (1984)
3.46 W. Rhode: Some observations on cochlear mechanics, J. Acoust. Soc. Am. **64**, 158–176 (1978)
3.47 W. Rhode: Observations of the vibration of the basilar membrane in squirrel monkeys using the Mössbauer technique, J. Acoust. Soc. Am. **49**, 1218–1231 (1971)
3.48 P. Sellick, I. Russell: Intracellular studies of cochlear hair cells: Filling the gap between basilar membrane mechanics and neural excitation. In: *Evoked Electrical Activity in the Auditory Nervous System*, ed. by F. Naunton, C. Fernandez (Academic, New York 1978) pp. 113–140
3.49 D. Kemp: Stimulated acoustic emissions from within the human auditory system, J. Acoust. Soc. Am. **64**, 1386–1391 (1978)
3.50 W. Brownell, C. Bader, D. Bertran, Y. de Rabaupierre: Evoked mechanical responses of isolated cochlear outer hair cells, Science **227**, 194–196 (1985)
3.51 J.B. Allen: OHCs shift the excitation pattern via BM tension. In: *Diversity in Auditory Mechanics*, ed. by E. Lewis, G. Long, R. Lyon, P. Narins, C. Steele, E. Hecht-Poinar (World Scientific, Singapore 1997) pp. 167–175
3.52 J.L. Goldstein, N. Kiang: Neural correlates of the aural combination tone 2f1-f2, Proc. IEEE **56**, 981–992 (1968)
3.53 G. Smoorenburg: Combination tones and their origin, J. Acoust. Soc. Am. **52**(2), 615–632 (1972)
3.54 D. Kemp: Evidence of mechanical nonlinearity and frequency selective wave amplification in the cochlea, Arch. Oto-Rhino-Laryngol. **224**, 37–45 (1979)
3.55 D. Kim, J. Siegel, C. Molnar: Cochlear nonlinear phenomena in two-tone responses, Scand. Audiol. **9**, 63–82 (1979)
3.56 P.F. Fahey, J.B. Allen: Nonlinear phenomena as observed in the ear canal and at the auditory nerve, J. Acoust. Soc. Am. **77**(2), 599–612 (1985)
3.57 M.B. Sachs, Y.S. Kiang: Two-tone inhibition in auditory-nerve fibers, J. Acoust. Soc. Am. **43**, 1120–1128 (1968)
3.58 R. Arthur, R. Pfeiffer, N. Suga: Properties of "two-tone inhibition" in primary auditory neurons, J. Physiol. (London) **212**, 593–609 (1971)
3.59 N.-S. Kiang, E. Moxon: Tails of tuning curves of auditory-nerve fibers, J. Acoust. Soc. Am. **55**, 620–630 (1974)
3.60 P. Abbas, M. Sachs: Two-tone suppression in auditory–nerve fibers: Extension of a stimulus–response relationship, J. Acoust. Soc. Am. **59**(1), 112–122 (1976)
3.61 X. Pang, J. Guinan: Growth rate of simultaneous masking in cat auditory-nerve fibers: Relationship to the growth of basilar-membrane motion and the origin of two-tone suppression, J. Acoust. Soc. Am. **102**(6), 3564–3574 (1997), beautiful data showing the slope of suppression for low frequency suppressors

3.62 I. Russell, P. Sellick: Intracellular studies of hair cells in the mammalian cochlea, J. Physiol. **284**, 261–290 (1978)

3.63 H. Duifhuis: Consequences of peripheral frequency selectivity for nonsimultaneous masking, J. Acoust. Soc. Am. **54**(6), 1471–1488 (1973)

3.64 D. Kemp: The evoked cochlear mechanical response and the auditory microstructure – evidence for a new element in cochlear mechanics, Scand. Audiol. **9**, 35–47 (1979)

3.65 D. Kemp: Towards a model for the origin of cochlear echoes, Hearing Res. **2**, 533–548 (1980)

3.66 D. Kemp: Otoacoustic emissions, travelling waves and cochlear mechanisms, J. Acoust. Soc. Am. **22**, 95–104 (1986)

3.67 A. Recio, W. Rhode: Basilar membrane responses to broadband stimuli, J. Acoust. Soc. Am. **108**(5), 2281–2298 (2000)

3.68 E. Fowler: A method for the early detection of otosclerosis, Archiv. Otolaryngol. **24**(6), 731–741 (1936)

3.69 S. Neely, D.O. Kim: An active cochlear model showing sharp tuning and high sensitivity, Hearing Res. **9**, 123–130 (1983)

3.70 D. He, P. Dallos: Somatic stiffness of cochlear outer hair cells is voltage-dependent, Proc. Nat. Acad. Sci. **96**(14), 8223–8228 (1999)

3.71 D. He, P. Dallos: Properties of voltage-dependent somatic stiffness of cochlear outer hair cells, J. Assoc. Res. Otolaryngol. **1**(1), 64–81 (2000)

3.72 A.M. Mayer: Research in acoustics, Phil. Mag. **2**, 500–507 (1876), in Benchmark Papers in Acoustics, Vol. 13, edited by E. D. Schubert

3.73 E. Titchener: *Experimental Psychology, A Manual of Laboratory Practice*, Vol. II (Macmillan, London 1923)

3.74 J.B. Allen: A short history of telephone psychophysics, J. Audio Eng. Soc. Reprint **4636**, 1–37 (1997)

3.75 R.R. Riesz: Differential intensity sensitivity of the ear for pure tones, Phys. Rev. **31**(2), 867–875 (1928)

3.76 D. McFadden: The curious half–octave shift: Evidence of a basalward migration of the traveling–wave envelope with increasing intensity. In: *Applied and Basic Aspects of Noise-Induced Hearing Loss*, ed. by R. Salvi, D. Henderson, R. Hamernik, V. Coletti (Plenum, New York 1986) pp. 295–312

3.77 W.A. Munson, M.B. Gardner: Loudness patterns – a new approach, J. Acoust. Soc. Am. **22**(2), 177–190 (1950)

3.78 B. Strope, A. Alwan: A model of dynamic auditory perception and its application to robust word recognition, IEEE Trans. Acoust. Speech Signal Proc. **5**(5), 451–464 (1997)

3.79 J.B. Allen: Psychoacoustics. In: *Wiley Encyclopedia of Electrical and Electronics Engineering*, Vol. 17, ed. by J. Webster (Wiley, New York 1999) pp. 422–437

3.80 B. Delgutte: Two-tone suppression in auditory-nerve fibres: Dependence on suppressor frequency and level, Hearing Res. **49**, 225–246 (1990)

3.81 B. Delgutte: Physiological mechanisms of psychophysical masking: Observations from auditory-nerve fibers, J. Acoust. Soc. Am. **87**, 791–809 (1990)

3.82 B. Delgutte: Physiological models for basic auditory percepts. In: *Auditory Computation*, ed. by H. Hawkins, T. McMullen (Springer, New York 1995)

3.83 L. Kanis, E. de Boer: Two-tone suppression in a locally active nonlinear model of the cochlea, J. Acoust. Soc. Am. **96**(4), 2156–2165 (1994)

3.84 J. Hall: Two-tone distortion products in a nonlinear model of the basilar membrane, J. Acoust. Soc. Am. **56**, 1818–1828 (1974)

3.85 C.D. Geisler, A.L. Nuttall: Two-tone suppression of basilar membrane vibrations in the base of the guinea pig cochlea using "low-side" suppressors, J. Acoust. Soc. Am. **102**(1), 430–440 (1997)

3.86 M. Régnier, J.B. Allen: *The Importance of Across-Frequency Timing Coincidences in the Perception of Some English Consonants in Noise* (ARO, Denver 2007), Abstract

3.87 M. Régnier, J.B. Allen: *Perceptual Cues of Some CV Sounds Studied in Noise* (AAS, Scottsdale 2007), Abstract

3.88 S. Narayan, A. Temchin, A. Recio, M. Ruggero: Frequency tuning of basilar membrane and auditory nerve fibers in the same cochleae, Science **282**, 1882–1884 (1998), shows BM and neural differ by 3.8 dB/oct

3.89 E. deBoer: Mechanics of the cochlea: Modeling efforts. In: *The Cochlea*, ed. by P. Dallos, A. Popper, R. Fay (Springer, New York 1996) pp. 258–317

3.90 D.C. Geisler: *From Sound to Synapse: Physiology of the Mammalian Ear* (Oxford Univ. Press, Oxford 1998)

3.91 J.B. Allen, P.F. Fahey: Using acoustic distortion products to measure the cochlear amplifier gain on the basilar membrane, J. Acoust. Soc. Am. **92**(1), 178–188 (1992)

3.92 S. Neely, D.O. Kim: A model for active elements in cochlear biomechanics, J. Acoust. Soc. Am. **79**, 1472–1480 (1986)

3.93 E. deBoer, A. Nuttall: The mechanical waveform of the basilar membrane, III Intensity effects, J. Acoust. Soc. Am. **107**(3), 1496–1507 (2000)

3.94 D. Kim, S. Neely, C. Molnar, J. Matthews: An active cochlear model with negative damping in the cochlear partition: Comparison with Rhode's ante- and post-mortem results. In: *Psychological, Physiological and Behavioral Studies in Hearing*, ed. by G. van den Brink, F. Bilsen (Delft Univ. Press, Delft 1980) pp. 7–14

3.95 J.B. Allen: DeRecruitment by multiband compression in hearing aids. In: *Psychoacoustics, Speech, and Hearing aids*, ed. by B. Kollmeier (World Scientific, Singapore 1996) pp. 141–152

3.96 W.F. Carver: Loudness balance procedures. In: *Handbook of Clinical Audiology*, 2nd edn., ed. by J. Katz (Williams Wilkins, Baltimore 1978) pp. 164–178, Chap. 15

3.97 J.B. Allen: Nonlinear cochlear signal processing. In: *Physiology of the Ear*, 2nd edn., ed. by A. Jahn, J. Santos-Sacchi (Singular, San Diego 2001) pp. 393–442, Chap. 19

3.98 S. Neely: A model of cochlear mechanics with outer hair cell motility, J. Acoust. Soc. Am. **94**, 137–146 (1992)

3.99 J. Ashmore: A fast motile response in guinea-pig outer hair cells: the molecular basis of the cochlear amplifier, J. Physiol. (London) **388**, 323–347 (1987)

3.100 J. Santos-Sacchi: Reversible inhibition of voltage–dependent outer hair cell motility and capacitance, J. Neurosci. **11**(10), 3096–3110 (1991)

3.101 K. Iwasa, R. Chadwick: Elasticity and active force generation of cochlear outer hair cells, J. Acoust. Soc. Am. **92**(6), 3169–3173 (1992)

3.102 I. Russell, P. Legan, V. Lukashkina, A. Lukashkin, R. Goodyear, G. Richardson: Sharpened cochlear tuning in a mouse with a genetically modified tectorial membrane, Nat. Neurosci. **10**, 215–223 (2007)

3.103 D. Sen, J.B. Allen: Functionality of cochlear micromechanics – as elucidated by the upward spread of masking and two tone suppression, Acoustics Australia **34**(1), 43–51 (2006)

3.104 J.B. Allen: Cochlear micromechanics: A physical model of transduction, J. Acoust. Soc. Am. **68**(6), 1660–1670 (1980)

3.105 J.B. Allen: Cochlear micromechanics – A mechanism for transforming mechanical to neural tuning within the cochlea, J. Acoust. Soc. Am. **62**, 930–939 (1977)

3.106 P. Dallos, D.Z. He, X. Lin, I. Sziklai, S. Mehta, B.N. Evans: Acetylcholine, outer hair cell electromotility, and the cochlear amplifier, J. Neurosci. **17**(6), 2212–2226 (1997)

3.107 P. Dallos: Prestin and the electromechanical responses of outer hair cells, ARO-2002 **25**, 189 (2002)

3.108 J.B. Allen: Derecruitment by multiband compression in hearing aids. In: *The Efferent Auditory System*, ed. by C. Berlin (Singular, San Diego 1999) pp. 73–86, Chap. 4 (includes a CDROM video talk by J. B. Allen in MP3 format)

3.109 J.B. Allen, D. Sen: Is tectorial membrane filtering required to explain two tone suppression and the upward spread of masking?. In: *Recent Developments in Auditory Mechanics*, ed. by H. Wada, T. Takasaka, K. Kieda, K. Ohyama, T. Koike (World Scientific, Singapore 1999) pp. 137–143

3.110 W. Sewell: The effects of furosemide on the endocochlear potential and auditory-nerve fiber tuning curves in cats, Hearing Res. **14**, 305–314 (1984)

3.111 J.B. Allen, P.F. Fahey: Nonlinear behavior at threshold determined in the auditory canal on the auditory nerve. In: *Hearing – Physiological Bases and Psychophysics*, ed. by R. Klinke, R. Hartmann (Springer, Berlin, Heidelberg 1983) pp. 128–134

3.112 J.B. Allen, B.L. Lonsbury-Martin: Otoacoustic emissions, J. Acoust. Soc. Am. **93**(1), 568–569 (1993)

3.113 P.F. Fahey, J.B. Allen: Measurement of distortion product phase in the ear canal of cat, J. Acoust. Soc. Am. **102**(5), 2880–2891 (1997)

3.114 R. Diependaal, E. de Boer, M. Viergever: Cochlear power flux as an indicator of mechanical activity, J. Acoust. Soc. Am. **82**, 917–926 (1987)

3.115 G. Zweig: Finding the impedance of the organ of Corti, J. Acoust. Soc. Am. **89**, 1229–1254 (1991)

3.116 E. deBoer, A. Nuttall: The "inverse problem" solved for a three-dimensional model of the cochlear, III Brushing-up the solution method, J. Acoust. Soc. Am. **105**(6), 3410–3420 (1999)

3.117 E. deBoer, A. Nuttall: The mechanical waveform of the basilar membrane, II From data to models – and back, J. Acoust. Soc. Am. **107**(3), 1487–1496 (2000)

3.118 G. Zweig: Finding the impedance of the organ of Corti, J. Acoust. Soc. Am. **89**(3), 1276–1298 (1991)

3.119 E. deBoer, A. Nuttall: The "inverse problem" solved for a three-dimensional model of the cochlea. II Application to experimental data sets, J. Acoust. Soc. Am. **98**(2), 904–910 (1995)

3.120 C. Shera, J. Guinan: Cochlear traveling-wave amplification, suppression, and beamforming probed using noninvasive caliration of intracochlear distortion sources, J. Acoust. Soc. Am. **121**(2), 1003–1016 (2007)

3.121 L.A. Pipes: *Applied Mathematics for Engineers and Physicists* (McGraw-Hill, New York 1958)

3.122 J.B. Allen: Modeling the noise damaged cochlea. In: *The Mechanics and Biophysics of Hearing*, ed. by P. Dallos, C.D. Geisler, J.W. Matthews, M.A. Ruggero, C.R. Steele (Springer, New York 1991) pp. 324–332

3.123 I.J. Russell, G.P. Richardson, A.R. Cody: Mechanosensitivity of mammalian auditory hair cells in vitro, Nature **321**(29), 517–519 (1986)

3.124 E.G. Boring: *History of Psychophysics* (Appleton–Century, New York 1929)

3.125 G. Fechner: Translation of "Elemente der Psychophysik". In: *Elements of Psychophysics*, Vol. I, ed. by H. Adler (Holt Rinehart Winston, New York 1966)

3.126 R. Schlauch, S. Harvey, N. Lanthier: Intensity resolution and loudness in broadband noise, J. Acoust. Soc. Am. **98**(4), 1895–1902 (1995)

3.127 J.B. Allen, S.T. Neely: Modeling the relation between the intensity JND and loudness for pure tones and wide–band noise, J. Acoust. Soc. Am. **102**(6), 3628–3646 (1997)

3.128 J. Zwislocki, H. Jordan: On the relation of intensity JNDs to loudness and neural noise, J. Acoust. Soc. Am. **79**, 772–780 (1986)

3.129 N.F. Viemeister: Psychophysical aspects of auditory intensity coding. In: *Auditory Function*, ed. by G. Edelman, W. Gall, W. Cowan (Wiley, New York 1988) pp. 213–241, Chap. 7

3.130 C. Plack, R. Carlyon: Loudness perception and intensity coding. In: *Hearing, Handbook of Perception and Cognition*, ed. by B. Moore (Academic, San Diego 1995) pp. 123–160, Chap. 4

3.131 J.B. Allen: Harvey Fletcher 1884–1981. In: *The ASA edition of Speech, Hearing in Communication*, ed. by J.B. Allen (Acoust. Soc. Am., Woodbury 1995) pp. 1–34

3.132 D.M. Green, J.A. Swets: *Signal Detection Theory and Psychophysics* (Wiley, New York 1966)

3.133 G.A. Miller: Sensitivity to changes in the intensity of white noise and its relation to masking and loudness, J. Acoust. Soc. Am. **19**, 609–619 (1947)

3.134 W. Hellman, R. Hellman: Intensity discrimination as the driving force for loudness, Application to pure tones in quiet, J. Acoust. Soc. Am. **87**(3), 1255–1271 (1990)

3.135 J. Egan, H. Hake: On the masking pattern of a simple auditory stimulus, J. Acoust. Soc. Am. **22**, 622–630 (1950)

3.136 W. Jesteadt, C. Wier, D. Green: Intensity discrimination as a function of frequency and sensation level, J. Acoust. Soc. Am. **61**(1), 169–177 (1977)

3.137 S. Stevens, H. Davis: *Hearing, Its Psychology and Physiology* (Acoust. Soc. Am., Woodbury 1983)

3.138 W.A. Yost: *Fundamentals of Hearing, An Introduction* (Academic, San Diego, London 1994)

3.139 W. Munson: The growth of auditory sensation, J. Acoust. Soc. Am. **19**, 584–591 (1947)

3.140 W. Siebert: Some implications of the stochastic behavior of primary auditory neurons, Kybernetik **2**, 205–215 (1965)

3.141 D. Raab, I. Goldberg: Auditory intensity discrimination with bursts of reproducible noise, J. Acoust. Soc. Am. **57**(2), 437–447 (1975)

3.142 H.C. Montgomery: Influence of experimental technique on the measurement of differential intensty sensitivity of the ear, J. Acoust. Soc. Am. **7**, 39–43 (1935)

3.143 P.G. Nutting: The complete form of Fechner's law, Bull. Bureau Standards **3**(1), 59–64 (1907)

3.144 H. Fletcher: *Speech and Hearing* (Van Nostrand, New York 1929)

3.145 E.H. Weber: Der Tastsinn und das Gemainfül. In: *Handwörterbuch der Physiologie*, Vol. 3, ed. by R. Wagner (Vieweg, Braunschweig 1988) pp. 481–588, Chap. 7

3.146 J. Johnson, C. Turner, J. Zwislocki, R. Margolis: Just noticeable differences for intensity and their relation to loudness, J. Acoust. Soc. Am. **93**(2), 983–991 (1993)

3.147 B.C.J. Moore: *An Introduction to the Psychology of Hearing*, 2nd edn. (Academic, London, New York 1982)

3.148 W.J. McGill, J.P. Goldberg: A study of the near–miss involving Weber's law and pure tone intensity discrimination, Percept. Psychophys. **4**, 105–109 (1968)

3.149 D. Green: Audition: Psychophysics and perception. In: *Stevens' Handbook of Experimental Psychology*, ed. by R. Atkinson, R. Herrnstein, G. Lindzey, R. Luce (Wiley, New York 1988) pp. 327–376, Chap. 6

3.150 D. Green: Application of detection theory in psychophysics, Proc. IEEE **58**(5), 713–723 (1970)

3.151 S. Stevens: Mathematics, measurement, and psychophysics. In: *Handbook of Experimental Psychology*, ed. by S. Stevens (Wiley, New York 1951) pp. 1–49, Chap. 1

3.152 R. Luce: *Sound and Hearing* (Lawrence Erlbaum, Hilldale 1993)

3.153 S. Stevens: To honor Fechner and repeal his law, Science **133**(3446), 80–86 (1961)

3.154 H. Fletcher: Physical measurements of audition and their bearing on the theory of hearing, Bell System Tech. J. **ii**(4), 145–180 (1923)

3.155 H. Fletcher, J. Steinberg: The dependence of the loudness of a complex sound upon the energy in the various frequency regions of the sound, Phys. Rev. **24**(3), 306–317 (1924)

3.156 J. Steinberg: The loudness of a sound and its physical stimulus, Phys. Rev. **26**, 507 (1925)

3.157 H. De Vries: The quantum character of light and its bearing upon the threshold of vision, the differential sensitivity and the acuity of the eye, Physica **10**, 553–564 (1943)

3.158 W. Siebert: Stimulus transformations in the peripheral auditory system. In: *Recognizing Patterns*, ed. by P. Kolers, M. Eden (MIT Press, Cambridge 1968) pp. 104–133, Chap. 4

3.159 W. Siebert: *Personal communication* (1989)

3.160 W.J. McGill, J.P. Goldberg: Pure–tone intensity discrimination as energy detection, J. Acoust. Soc. Am. **44**, 576–581 (1968)

3.161 J.C. Steinberg, M.B. Gardner: On the auditory significance of the term hearing loss, J. Acoust. Soc. Am. **11**, 270–277 (1940)

3.162 H. Fletcher: A method of calculating hearing loss for speech from an audiogram, J. Acoust. Soc. Am. **22**, 1–5 (1950)

3.163 G. Ekman: Weber's law and related functions, Psychology **47**, 343–352 (1959)

3.164 S. Stevens, H. Davis: *Hearing, Its Psychology and Physiology* (Acoust. Soc. Am., Woodbury 1938)

3.165 W.A. Munson: An experimental determination of the equivalent loudness of pure tones, J. Acoust. Soc. Am. **4**, 7 (1932), Abstract

4. Perception of Speech and Sound

B. Kollmeier, T. Brand, B. Meyer

The transformation of acoustical signals into auditory sensations can be characterized by psychophysical quantities such as loudness, tonality, or perceived pitch. The resolution limits of the auditory system produce spectral and temporal masking phenomena and impose constraints on the perception of amplitude modulations. Binaural hearing (i. e., utilizing the acoustical difference across both ears) employs interaural time and intensity differences to produce localization and binaural unmasking phenomena such as the binaural intelligibility level difference, i. e., the speech reception threshold difference between listening to speech in noise monaurally versus listening with both ears.

The acoustical information available to the listener for perceiving speech even under adverse conditions can be characterized using the articulation index, the speech transmission index, and the speech intelligibility index. They can objectively predict speech reception thresholds as a function of spectral content, signal-to-noise ratio, and preservation of amplitude modulations in the speech waveform that enter the listener's ear. The articulatory or phonetic information available to and received by the listener can be characterized by speech feature sets. Transinformation analysis allows one to detect the relative transmission error connected with each of these speech features. The comparison across man and machine in speech recognition allows one to test hypotheses and models of human speech perception. Conversely, automatic speech recognition may be improved by introducing human signal-processing principles into machine processing algorithms.

4.1 **Basic Psychoacoustic Quantities** 62
4.1.1 Mapping of Intensity into Loudness 62
4.1.2 Pitch 64
4.1.3 Temporal Analysis and Modulation Perception 65
4.1.4 Binaural Hearing 67
4.1.5 Binaural Noise Suppression 68

4.2 **Acoustical Information Required for Speech Perception** 70
4.2.1 Speech Intelligibility and Speech Reception Threshold (SRT) 70
4.2.2 Measurement Methods 71
4.2.3 Factors Influencing Speech Intelligibility 72
4.2.4 Prediction Methods 72

4.3 **Speech Feature Perception** 74
4.3.1 Formant Features 75
4.3.2 Phonetic and Distinctive Feature Sets 76
4.3.3 Internal Representation Approach and Higher-Order Temporal-Spectral Features 77
4.3.4 Man–Machine Comparison 80

References 81

Acoustically produced speech is a very special sound to our ears and brain. Humans are able to extract the information contained in a spoken message extremely efficiently even if the speech energy is lower than any competing background sound. Hence, humans are able to communicate acoustically even under adverse listening conditions, e.g., in a cafeteria. The process of understanding speech can be subdivided into two stages. First, an auditory pre-processing stage where the speech sound is transformed into its *internal representation* in the brain and special speech features are extracted (such as, e.g., acoustic energy in a certain frequency channel as a function of time, or instantaneous pitch of a speech sound). This process is assumed to be mainly bottom-up with no special preference for speech sounds as compared to other sounds. In other words, the information contained in any of the speech features can be described quite well by the acoustical contents of the input signal to the ears. In a second step, speech pattern recognition takes place under cognitive control where the internally available speech cues are assembled by our brain to convey the underlying message of the speech sound. This

process is assumed to be top-down and cognitive controlled, and is dependent on training, familiarity, and attention. In the context of this handbook we primarily consider the first step while assuming that the second step operates in a nearly perfect way in normal human listeners, thus ignoring the vast field of cognitive psychology, neuropsychology of speech, and psycholinguistics. Instead, we consider the psychoacoustics of transforming speech and other sounds into its *internal representation*. We will concentrate on the acoustical prerequisites of speech perception by measuring and modeling the speech information contained in a sound entering the ear, and finally the speech features that are presumably used by our brain to recognize speech.

4.1 Basic Psychoacoustic Quantities

The ear converts the temporally and spectrally fluctuating acoustic waveform of incoming speech and sound into a stream of auditory percepts. The most important dimensions of auditory perception are:

- the transformation of sound intensity into subjectively perceived loudness,
- the transformation of major frequency components of the sound into subjectively perceived pitch,
- the transformation of different temporal patterns and rhythms into subjectively perceived fluctuations,
- the transformation of the spectro-temporal contents of acoustic signals into subjectively perceived timbre (which is not independent of the dimensions listed above),
- the transformation of interaural disparities (i. e., differences across both ears) and spectro-temporal contents of acoustical signals into the perceived spatial location and spatial extent of an auditory object.

A basic prerequisite for being able to assign these dimensions to a given sound is the ear's ability to internally separate acoustically superimposed sound sources into different auditory streams or objects.

Psychoacoustics is the scientific discipline that measures and models the relation between physical acoustical quantities (e.g., the intensity of a sinusoidal stimulus specified by sound pressure level, frequency, and duration) and their respective subjective impression (e.g., loudness, pitch, and perceived temporal extent).

4.1.1 Mapping of Intensity into Loudness

The absolute threshold in quiet conditions for a continuous sinusoid is highly dependent on the frequency of the pure tone (Fig. 4.1). It shows highest sensitivity in the frequency region around 1 kHz, which also carries most speech information, and increases for low and high frequencies. The normalized threshold in quiet at 1 kHz averaged over a large number of normal hearing subjects is defined as 0 dB sound pressure level (SPL), which corresponds to 20 μPa. As the level of the sinusoid increases, the perceived loudness increases with approximately a doubling of perceived loudness with a level increase by 10 dB. All combinations of sound pressure levels and frequencies of sinusoid that produce the same loudness as a reference 1 kHz sinusoid of a given level (in dB SPL) are denoted as isophones (Fig. 4.1). Hence, the loudness level (in phon) can only be assigned to a sinusoid and not to a multi-frequency mixture of sounds such as speech. This difference is due to the fact that a broadband sound is perceived as being louder than a narrow-band sound at the same

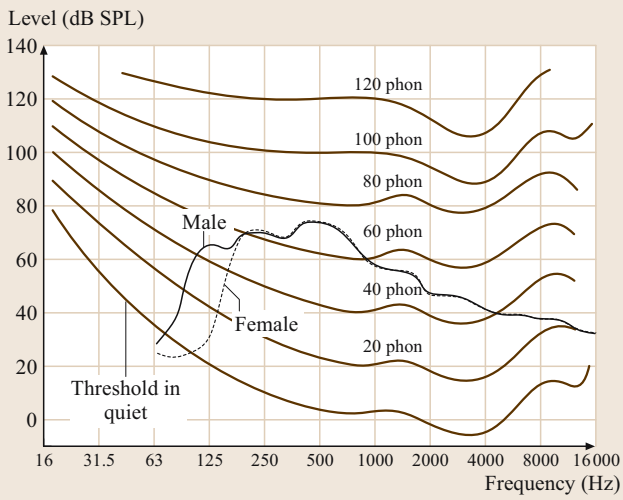


Fig. 4.1 Auditory field described by the threshold in quiet, the isophones, and the uncomfortable listening level of a continuous sinusoid as a function of tone frequency. Also given is an average speech spectrum for male and female speech plotted as a power density

sound pressure level, which puts all of its energy into one *critical band*. In order to express the speech sound pressure level in a way that approximates the human loudness impression, there are several options available:

- Unweighted root-mean-square (RMS) level: the total signal intensity over the audio frequency range is averaged within a certain time window (denoted as *slow*, *fast*, or *impulse*, respectively, for standardized sound level meters, or as the RMS value for a digitized speech signal) and is expressed as dB SPL, i. e., as $10\log(I/I_0)$ where I denotes the signal intensity and I_0 denotes the reference signal intensity at auditory threshold. The usage of the dB scale already takes into account the Weber–Fechner law of psychophysics: roughly, a sound intensity difference of 1 dB can be detected as the just notice-

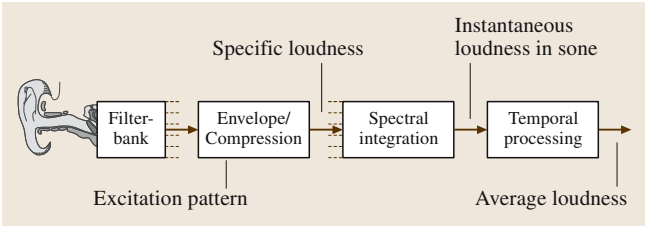


Fig. 4.2 Block diagram of a loudness model

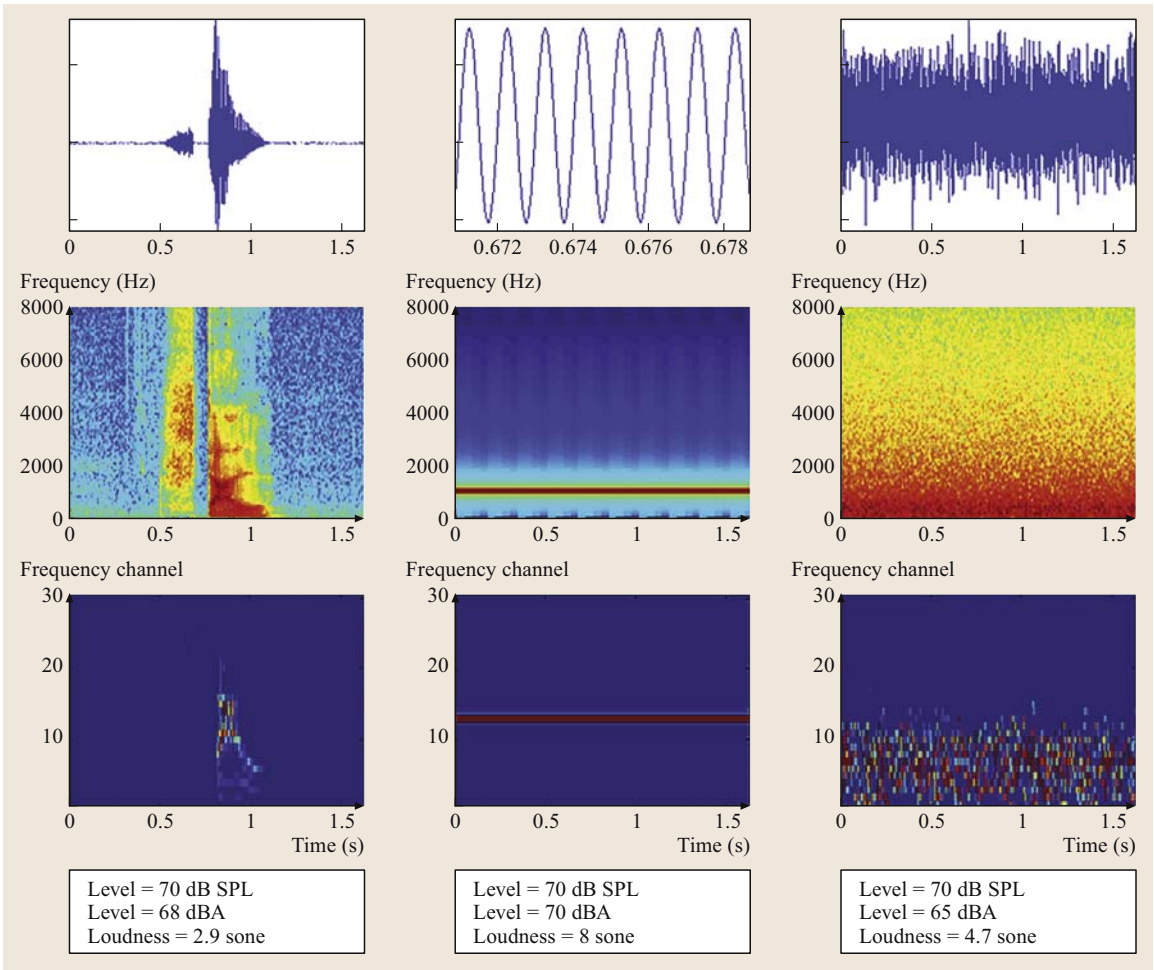


Fig. 4.3 Example plot of a speech sample, 1 kHz sinusoidal tone and a continuous speech-shaped noise sample represented as a waveform, spectrogram, partial loudness pattern and resulting value in dB SPL, dBA and loudness in sone

able intensity difference irrespective of the reference level.

- A-weighted signal level: in order to account for the higher sensitivity of the ear to mid-frequencies at low levels, a spectral weighting function that approximates the isophones between 20 and 30 phon (denoted as A-weighting) is applied to the spectral components of the sound before they are summed up to give the total sound level. The B- and C-weighting curves are available for higher signal levels. Note that the A-, B-, and C-weighted speech levels do not differ too much from the unweighted speech level because the long-term average spectral shape of speech includes most energy in the frequency range close to 1 kHz where the weighting curves coincide. Since none of these definitions include the psychophysical effect of loudness summation across frequency and the temporal integration performed by our ear, a speech sound at a given sound pressure level does not necessarily produce the same perceived loudness as a 1 kHz signal at this level.
- Loudness in sone. A more exact measure of perceived loudness for sounds that differ in frequency contents is given by a loudness calculation scheme based on the sone scale. For a narrow-band sound (like a sinusoid), the loudness in sone is expressed as

$$N[\text{sone}] = (I/I_0)^{\alpha} , \tag{4.1}$$

where I_0 is the reference intensity which is set to 40 dB SPL for a sinusoid at 1 kHz. Since the exponent α amounts to ≈ 0.3 according to *Stevens* and *Zwicker* [4.1], this yields a compression of sound intensity similar to the nonlinear compression in the human ear. For broadband sounds (such as speech), the same compressive power law as given above has to be applied to each *critical* frequency band (see later) before the partial loudness contributions are summed up across frequencies, which results in the total loudness. The detailed loudness calculation scheme (according to ISO 532b) also accounts for the spread of spectral energy of a narrow-band sound into adjacent frequency bands known from cochlear physiology (Sect. 4.1.2). This *leakage* of spectral energy can also be modeled by a bank of appropriately shaped bandpass filters with a limited upper and lower spectral slope (Fig. 4.2).

To account for the time dependency of loudness perception, a temporal integration process is assumed that sums up all intensity belonging to the same auditory object within a period of approximately 200 ms. This roughly models the effect that sounds with a constant sound level are perceived as being louder if their duration increases up to 200 ms while remaining constant in loudness if the duration increases further. For fluctuating sounds with fluctuating instantaneous loudness estimates, the overall loudness impression is dominated by the respective loudness maxima. This can be well represented by considering the 95 percentile loudness (i. e., the loudness value that is exceeded for only 5% of the time) as the *average* loudness value of a sequence of fluctuating sounds [4.1]. For illustration, Fig. 4.3 displays the relation between waveform, spectrogram, partial loudness pattern and resulting value in dB SPL, dBA and loudness in sone for three different sounds.

4.1.2 Pitch

If the frequency of a sinusoid at low frequencies up to 500 Hz is increased, the perceived tone height (or pitch, also denoted as tonality) increases linearly with frequency. At higher frequencies above 1 kHz, however, the perceived pitch increases approximately logarithmically with increasing frequency. The combination of both domains yields the psychophysical mel-scale (Fig. 4.4).

This relation between frequency and subjective frequency perception also represents the mapping of frequencies on the basilar membrane (Sect. 4.2.1 and

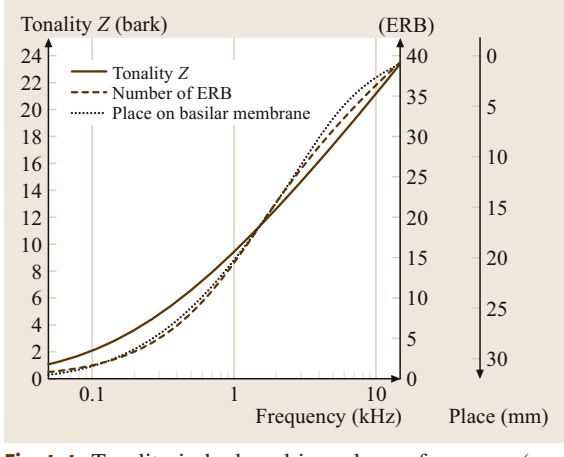


Fig. 4.4 Tonality in bark and in mel over frequency (one bark equals 100 mel). For comparison the (hypothetical) place of maximum excitation on the basilar membrane and the psychoacoustical frequency scale based on equivalent rectangular bandwidth (ERB) is plotted

Fig. 4.4), where frequencies up to approximately 2 kHz occupy half of the basilar membrane and those between 2 kHz and 20 kHz the remaining half. The slope of this function relates to the just noticeable difference (JND) for frequency. The frequency JND is about 3 Hz for frequencies below 500 Hz and about 0.6% for frequencies above 1000 Hz, which is approximately 3 mel. This value amounts to 1/30 of the frequency-dependent bandwidth of the *critical band* which plays a role both in loudness summation (see above) and in the psychophysical effect of spectral masking. All spectral energy that falls into one *critical band* is summed up and *masks* (or disables) the detection of a sinusoidal tone centered within that critical band as long as its level is below this *masked threshold*. According to *Zwicker* et al. [4.2] the auditory critical bandwidth is expressed in bark after the German physicist *Barkhausen* as a function of frequency f_0 (in Hz) as:

$$\begin{aligned} 1\,\text{bark} &= 100\,\text{mel} \\ &\approx 100\,\text{Hz for frequencies below 500 Hz} \\ &\approx 1/5\,f_0 \text{ for frequencies above 500 Hz}\,. \end{aligned} \tag{4.2}$$

The psychophysical frequency scale resulting from integrating the critical bandwidth over frequency is denoted as *bark scale* Z and can be approximated [4.4] by the inverse function of the hyperbolical sinus

$$Z[\text{bark}] = 7 \cdot \text{arcsinh}(f_0/650) \tag{4.3}$$

More-refined measurements of spectral masking performed by *Moore* and *Patterson* [4.5] resulted in the *equivalent rectangular bandwidth* (ERB) as a measure of the psychoacoustical critical bandwidth. It deviates slightly from the bark scale, especially at low frequencies (Fig. 4.4).

It should be noted, however, that the *pitch strength* of a sinusoid decreases steadily as the frequency increases. A much more distinct pitch perception, which is also related to the pitch of musical instruments and the pitch of voiced speech elements, can be perceived for a periodic, broad band sound. For such complex sounds, the term *pitch* should primarily be used to characterize the (perceived) fundamental frequency while *tone height* or *tonality* refers to the psychophysical equivalence of frequency and coincides with pitch only for sinusoids. The perception of pitch results both from temporal cues (i. e., the periodicity information within each critical band) primarily at low frequencies, and from spectral pitch cues, i. e., the regular harmonic structure/line spectrum of a periodic sound primarily dominating at high frequencies. Several theories exist about the perception and relative importance of temporal and spectral pitch cues [4.6]. For the perception of the pitch frequency range of normal speech with fundamental frequencies between approximately 80 Hz and 500 Hz, however, predominantly temporal pitch cues are exploited by our ear. In this range, the pitch JND amounts to approximately 1 Hz, i. e., better resolution occurs for the fundamental frequency of a complex tone than for the audio frequency of a single sinusoid at the fundamental frequency.

4.1.3 Temporal Analysis and Modulation Perception

When perceiving complex sounds such as speech that fluctuate in spectral contents across time, the ear can roughly be modeled as a bank of critical-band wide bandpass filters that transform the speech signal into a number of narrow-band time signals at center frequencies that are equally spaced across the bark scale. Hence, the temporal analysis and resolution within each of these frequency channels is of special importance

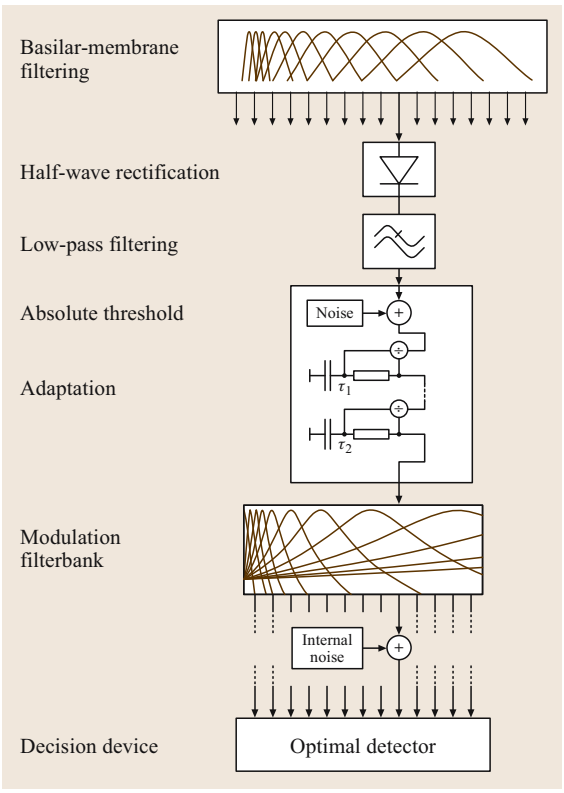


Fig. 4.5 Model of the *effective* signal processing in the auditory system (after [4.3])

for the overall function of our auditory system. For the within-channel analysis, the following phenomena are relevant.

- For center frequencies below ≈ 1000 Hz, the temporal fine structure of the bandpass channel is coded in the auditory nerve and is therefore accessible to the brain. Hence the signal's phase can be exploited during the central processing stages, e.g. for a comparison between different frequency bands to produce a difference in perceived timbre and for a comparison between ears to produce a difference in perceived localization as the phase characteristic is changed. The latter results from extracting the interaural phase difference of a sound signal arriving from a point in space with a certain travel time to either ear (see later).
- For center frequencies above 1 kHz, primarily the envelope of the signal is extracted and analyzed. This makes the ear comparatively phase-deaf above 1 kHz. This envelope extraction is due to the asymmetry between depolarization and hyperpolarization at the synapses between inner hair cells and the auditory nerve as well as due to the temporal integration observed in auditory nerve fibers (Chap. 3). In auditory models this can be modeled to a good approximation by a half-wave rectifier followed by a low-pass filter with a cut-off frequency of ≈ 1 kHz.
- The resulting envelope is subject to a compression and adaptation stage that is required to map the large dynamic range of auditory input signals to the comparatively narrow dynamic range of the nervous system. It is also necessary to set the operation point of the respective further processing stages according to some average value of the current input signal. This compression and adaptation characteristic can either be modeled as a logarithmic compression in combination with the temporal leaky integrator using an *effective* auditory temporal integration window or, alternatively, by a series of nonlinear adaptation loops (Fig. 4.5).

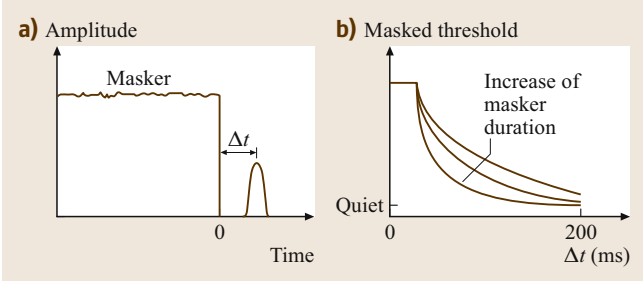


Fig. 4.6a,b Schematic plot of a masked threshold **(a)** of a short probe tone in the presence of (or following) a masking noise burst that extends across a variable amount of time **(b)** (simultaneous and forward masking)

Such an adaptation stage produces the *temporal integration* effect already outlined in Sect. 4.1.1, i. e., all temporal energy belonging to the same acoustical object is summed up within a time window with an effective duration of up to 200 ms. In addition, *temporal masking* is due to this integration or adaptation circuit. A short probe signal (of an intensity higher than the threshold intensity in quiet) will become inaudible in the presence of a *masking* signal if the probe signal is presented either before, during, or after the masker. Hence, the masker extends its masking property both back in time (*backward masking*, extending to approximately 5 ms prior to the onset of the masker), simultaneously with the probe signal (*simultaneous masking*, which becomes less efficient if the masker duration is decreased below 200 ms) and subsequent to the masker (*forward masking*, extending up to 200 ms, Fig. 4.6).

Note that forward masking in speech sounds can prevent detection of soft consonant speech components that are preceded by high-energy vocalic parts of speech.

An important further property of temporal analysis in the auditory system is the perception and analysis of the incoming *temporal envelope fluctuations*. While slow amplitude modulations (modulation frequencies below approximately 4 Hz) are primarily perceived as temporal fluctuations, amplitude modulations between approximately 8 Hz and 16 Hz produce a rolling, R-type roughness percept. Modulations between 16 Hz and approximately 80 Hz are perceived as roughness of a sound. Higher modulation frequencies may be perceived as spectral coloration of the input signal without being resolved in the time domain by the auditory system.

The auditory processing of sounds that differ in their composition of modulation frequencies is best described by the modulation spectrum concept which can be modeled by a modulation filter bank (Fig. 4.5). The separation of different modulation frequencies into separate modulation frequency channels (similar to separating the audio frequencies into different center frequency channels in the inner ear) allows the brain to group together sound elements that are generated from the same sound source even if they interfere with sound elements from a different sound source at the same center frequency. Natural objects are usually characterized by a common

modulation of the emitted frequency components as a function of time. By grouping those sound components that exhibit the same modulation spectrum across different center frequencies, the brain is able to recombine all the sound components of a certain object that are spread out across different audio frequencies. This property of the auditory system is advantageous in performing a figure-background analysis (such as required for the famous *cocktail-party phenomenon*, i. e., a talker can be understood even in the background of a lively party with several interfering voices).

A way of quantitatively measuring the auditory grouping effect is the so-called co-modulation masking release (CMR) depicted in Fig. 4.7.

A probe tone has to be detected against a narrowband, fluctuating noise at the same frequency (*on-frequency masker*). If the adjacent frequency bands are stimulated with uncorrelated noise samples, the threshold of the tone in the modulated noise is comparatively high. However, if the masking noise in the adjacent bands fluctuates with the same amplitude modulations across time (this is usually done by duplicating the on-frequency masker and shifting its center frequency appropriately), the probe tone becomes better audible and a distinct threshold shift occurs. This is called co-modulation masking release. It is mainly due to within-channel cues, i. e., the modulation minima become more distinct with increasing noise bandwidth and hence allow for a better detection of the continuous probe tone at a certain instant of time. It is also due to some across-channel cues and cognitive processing, i. e., the co-modulated components at different frequencies are grouped to form a single auditory object which is distinct from the probe tone. Since speech is usually characterized by a high degree of co-modulation across different frequencies for a single speaker, the co-modulation masking release helps to detect any irregularity which is not co-modulated with the remainder of the speech signal. Such detectable irregularities may reflect, e.g., any speech pathology, a second, faint acoustical object or even speech processing artefacts. The CMR effect is most prominent for amplitude modulation frequencies between approximately 4 Hz and 50 Hz [4.7], which is a region where most of the modulation spectrum energy of speech is located. Hence this effect is very relevant for speech perception.

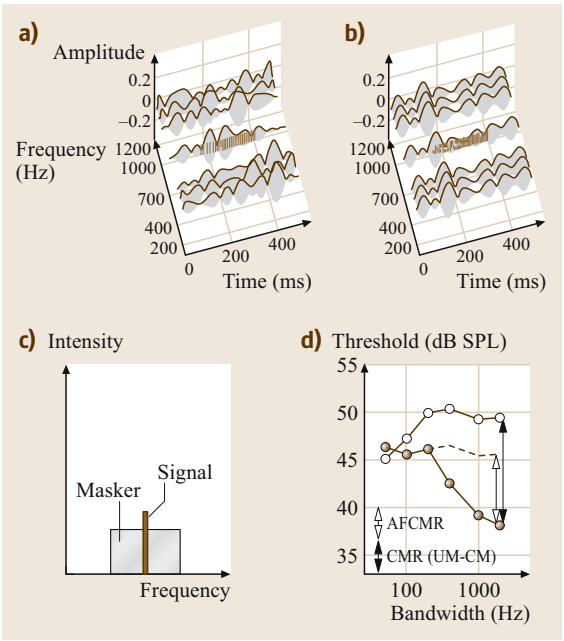


Fig. 4.7a–d Schematic plot of co-modulation masking release (after [4.7]). **(a)** Denotes the temporal-spectral difference in the unmodulated condition (flanking bands are modulated in an uncorrelated way) whereas **(b)** shows the pattern for co-modulated sidebands. In the latter case, the detection of a sinusoid at the on-frequency masking band is facilitated. **(c)**, **(d)** Shows typical psychophysical data for a band-widening experiment with unmodulated (*open symbols*) and co-modulated masker (*filled symbols*). A considerable difference in masked threshold for the sinusoidal signal is observed

4.1.4 Binaural Hearing

Binaural processing, i. e., the central interaction between signal information entering the right and the left ear contributes significantly to

- suppression of subjectively perceived reverberation in closed rooms,
- localization of sound sources in space,
- suppression of *unwanted* sound sources in real acoustical environments.

To perform these tasks, our brain can utilize

- interaural time (or phase) cues, i. e., the central auditory system extracts the travel time difference between the left and right ear,
- interaural intensity difference, i. e., our brain can utilize the head-shadow effect: sound arriving at the ear pointing towards the sound source in space is not attenuated, while the sound arriving at the opposite ear is attenuated,

- spectral changes (coloration) of the sound reaching the inner ear due to interference and scattering effects if the direction of the incoming sound varies (Fig. 4.8).

Normal listeners can localize sound sources with a precision of approximately 1° if sound arrives at the head from the front. This relates to a just noticeable difference (JND) in interaural time difference as small as 10 μs and an interaural level difference JND as low as 1 dB. This remarkable high resolution is due to massive parallel processing at the brain-stem level where the first neural comparison occurs between activation from the right and left ear, respectively.

The binaural performance of our auditory system is extremely challenged in complex acoustical everyday situations characterized by several nonstationary sound sources, reverberation, and a continuous change of the interaural cues due to head movements in space. For perceiving, localizing, and understanding speech in such situations, the following phenomena are relevant:

- Spectral integration of localization cues. Continuous narrow-band signals are hard to localize because their respective interaural time and level difference achieved at the ear level are ambiguous: they can result from any direction within a *cone of confusion*, i. e., a surface that includes all spatial angles centered around the interaural axis that yield the same path difference between right and left ear. For a broadband signal, the comparison across different frequency channels helps to resolve this ambiguity. Also, the onset cues in strongly fluctuating, broadband sounds contain more reliable localization cues than the running cues in the steady-state situation for continuous sounds.
- Precedence effect or the law of the first wavefront. The direct sound (first wavefront) of a sound source hitting the receiver's ears determines the subjective localization percept. Conversely, any subsequent wavefront (that is due to reflections from surrounding structures in a real acoustical environment and hence carries the *wrong* directional information) is not used to create the subjective localization impression. Even though reflections arriving approximately 5–20 ms after the first wavefront are perceivable and their energetic contribution to the total stimulus percept is accessible to the brain, their respective directional information seems to be suppressed. This effect is utilized in some public address loudspeaker systems that deliberately delay the amplified sound in order for the small, unamplified direct sound to reach the listener prior to the amplified sound with the *wrong* directional information.

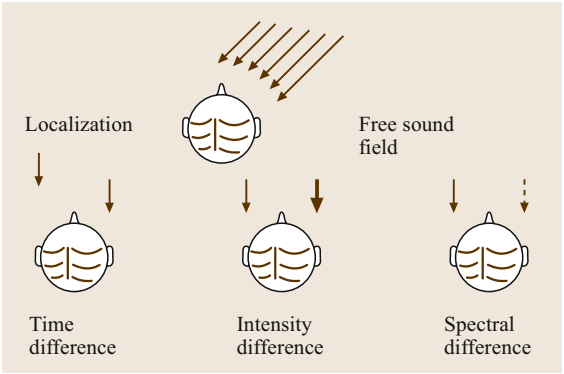


Fig. 4.8 Schematics of interaural cues due to interference and scattering effects that can be utilized by the auditory system

4.1.5 Binaural Noise Suppression

The localization mechanisms described above are not only capable of separating the perceived localization of several simultaneously active acoustical objects. They are also a prerequisite for binaural noise suppression, i. e., an enhancement of the *desired* signal and a suppression of *undesired* parts of the input signals that originate from a different spatial direction. This enhancement is also denoted as binaural release from masking. It can be demonstrated by a tone-in-noise detection experiment where in the reference condition tone and noise are the same at both ears (i. e., exhibit the phase difference 0). The detection threshold can be compared to the threshold using the same noise, but inverting the signal on one side (i. e., a phase difference of π for the signal), which yields a higher detectability. This difference in threshold is denoted as binaural masking level difference and amounts up to 20 dB for short probe tones at frequencies below 1000 Hz.

For speech signals, the binaural unmasking can be measured by comparing the speech reception threshold (i. e., the signal-to-noise ratio required to understand 50% of the presented speech material, see later) for different spatial arrangements of target speech sound and interfering noise: in the reference condition, speech and noise are presented directly in front of the subject, while in the test condition speech comes from the front, but the interfering sound source from the side. The gain in speech reception threshold is called the intelligibil-

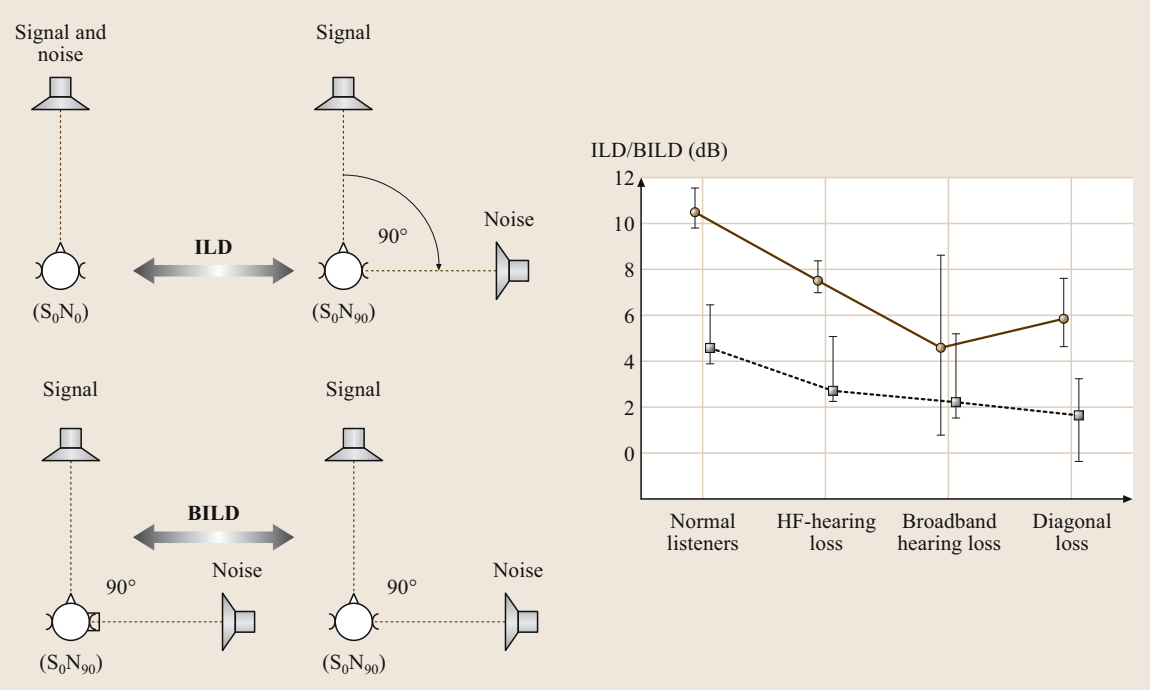


Fig. 4.9 Intelligibility level difference (ILD, *filled circles*) and binaural intelligibility level difference (BILD, *filled squares*) averaged across a group of normal and impaired listeners that differ in the shape of their respective audiogram (high-frequency hearing loss abbreviated as HF-hearing loss). The difference in speech reception threshold across both situations plotted on the left-hand side is plotted as average value and intersubject standard deviation

ity level difference. (The abbreviation ILD is used for this difference, but is also used for interaural level difference.) Intelligibility level difference is due to a monaural effect (i. e., improved signal-to-noise ratio at the ear opposite to the interfering sound source) and a binaural effect. To separate this latter effect, another threshold in the same spatial situation is used where the *worse* ear is plugged and the speech reception threshold is obtained using only the *better* ear, i. e., the ear with the better signal-to-noise ratio. The difference in speech reception threshold (SRT= between the latter two situations (i. e., the difference due to *adding* the *worse* ear) is a purely binaural effect, called the binaural intelligibility level difference (BILD). Figure 4.9 gives an example of the ILD and BILD at an incidence angle for the interfering noise of 90° in an anechoic condition for different groups of listeners that vary in their hearing loss.

A basic model which describes binaural unmasking phenomena quite well for speech signals in complex acoustical environments is a multichannel equalization and cancelation (EC) model such as that depicted in Fig. 4.10 [4.8].

Within each frequency band, an *equalization and cancelation* mechanism [4.9] is used that first delays

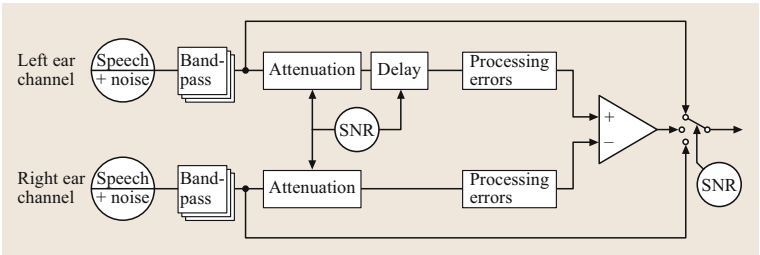


Fig. 4.10 Sketch of a multichannel binaural noise canceling model describing the binaural release from masking for speech in complex environments (after [4.8])

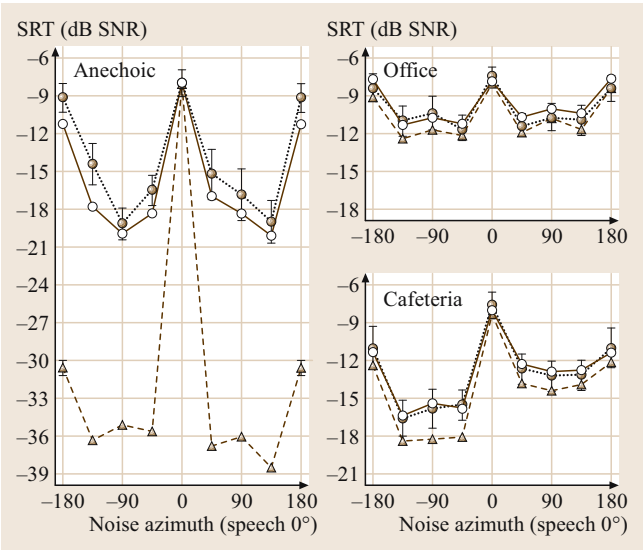


Fig. 4.11 SRT data (*filled symbols*) and predictions for three different acoustical conditions and normal listeners. The *triangles* denote model predictions without introducing appropriate processing errors, whereas the open symbols denote predictions employing internal processing errors, that have been taken from average values in other psychoacoustical tasks (after [4.8])

and amplifies one or both input channels to yield an approximate match (*equalization*) of the composite input signal within each frequency band. In a second, the *cancelation* stage, the signals from both respective sides of the head are (imperfectly) subtracted from each other. Hence, if the masker (after the *equalization* step) is approximately the same in both ears, the cancelation step will eliminate the masker with the exception of some remaining error signal. Conversely, the *desired* signal, which differs in interaural phase and/or intensity relation from the masker, should stay nearly unchanged, yielding an improvement in signal-to-noise ratio. Using an appropriate numerical optimization strategy to fit the respective equalization parameters across frequency, the model depicted in Fig. 4.11 can predict human performance quite well even under acoustically *difficult* situations, such as, e.g., several interfering talkers within a reverberant environment. Note that this model effectively corresponds to an adaptive spatial beam former, i. e., a frequency-dependent optimum linear combination of the two sensor inputs to both ears that yields a directivity optimized to improve the signal-to-noise ratio for a given target direction and interfering background noise. If the model output is used to predict speech intelligibility with an appropriate (monaural) speech intelligibility prediction method [such as, e.g., the speech intelligibility index (SII), see later], the binaural advantage for speech intelligibility in rooms can be predicted quite well (Fig. 4.11 from [4.8]).

Note that in each frequency band only one EC circuit is employed in the model. This reflects the empirical evidence that the brain is only able to cancel out one direction for each frequency band at each instant of time. Hence, the processing strategy adopted will use appropriate compromises for any given real situation.

4.2 Acoustical Information Required for Speech Perception

4.2.1 Speech Intelligibility and Speech Reception Threshold (SRT)

Speech intelligibility (SI) is important for various fields of research, engineering, and diagnostics for quantifying very different phenomena such as the quality of recordings, communication and playback devices, the reverberation of auditoria, characteristics of hearing impairment, benefit using hearing aids, or combinations of these topics. The most useful way to define SI is: *speech intelligibility SI is the proportion of speech items (e.g., syllables, words, or sentences) correctly repeated by (a) listener(s) for a given speech intelligibility test*. This operative definition makes SI directly and quantitatively measurable.

The intelligibility function (Fig. 4.12) describes the listener's speech intelligibility SI as a function of speech level L which may either refer to the sound pressure level (measured in dB) of the speech signal or to the speech-to-noise ratio (SNR) (measured in dB), if the test is performed with interfering noise.

In most cases it is possible to fit the logistic function SI (L) to the empirical data

$$\mathrm{SI}(L) = \frac{1}{A}\left(1 + \mathrm{SI}_{\max}\frac{A-1}{1+\exp\left(-\frac{L-L_{\mathrm{mid}}}{s}\right)}\right), \quad (4.4)$$

with L_{mid}: speech level of the midpoint of the intelligibility function; s: slope parameter, the slope at L_{mid} is

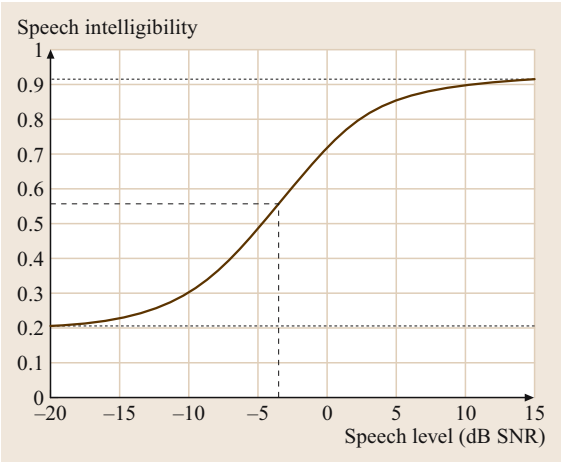


Fig. 4.12 Typical example of SI function (*solid line*) for word intelligibility test (closed response format with five response alternatives). The *dashed line* denotes L_{mid}. The *dotted lines* denote the lower limit $(1/A)$ and the asymptotic maximum $\mathrm{SI}_{\mathrm{asymp}}$ of the SI function. Parameters: $L_{\mathrm{mid}} = -3.5$ dB SNR, $\mathrm{SI}_{\mathrm{max}} = 0.9 (\mathrm{SI}_{\mathrm{asymp}} = 0.92)$, $A = 5$, slope $= 0.05$/dB ($s = 3.6$ dB)

given by $\frac{\mathrm{SI}_{\mathrm{max}}(A-1)}{4As}$; $\mathrm{SI}_{\mathrm{max}}$: parameter for maximum intelligibility which can be smaller than 1 in some cases (e.g., distorted speech signals or listeners with hearing impairment). The asymptotic maximum of SI is given by $\mathrm{SI}_{\mathrm{max}} + (1 - \mathrm{SI}_{\mathrm{max}})/A$. A is the number of response alternatives (e.g., $A = 10$ when the listener should respond in a closed response format for instance using digits between '0' and '9'). In SI tests with *open response format*, like word tests without limiting the number of response alternatives, A is assumed to be infinite, that means

$$\mathrm{SI} = \mathrm{SI}_{\mathrm{max}} \frac{1}{1 + \exp\left(-\frac{L - L_{\mathrm{mid}}}{s}\right)} \quad \text{and}$$

$$\text{slope} = \frac{\mathrm{SI}_{\mathrm{max}}}{4s} \,. \tag{4.5}$$

The primary interest of many applications is the speech reception threshold (SRT) which denotes the speech level (measured in dB), which belongs to a given intelligibility (e.g., SI = 0.5 or 0.7).

The accuracy of SI measurements is given by the binomial distribution. Consequently, the standard error SE(SI) of an SI estimate based on n items (e.g., words) is given by

$$\mathrm{SE}(\mathrm{SI}) = \sqrt{\frac{\mathrm{SI}(1 - \mathrm{SI})}{n}} \,. \tag{4.6}$$

A further increase of this standard error is caused by the fact that SI tests consist of several items (e.g., 50 words) which unavoidably differ in SI. Therefore, SI tests should be constructed in a way that the SI of all items is as homogeneous as possible.

To a first approximation, the standard error of the SRT is equal to $\mathrm{SE}(\mathrm{SI}_{\mathrm{SRT}})$ (the standard error of the SI estimate at the SRT) divided by the slope of the intelligibility function at the SRT. Thus

$$\mathrm{SE}(\mathrm{SRT}) = \frac{\mathrm{SE}(\mathrm{SI}_{\mathrm{SRT}})}{\mathrm{slope}_{\mathrm{SRT}}} \,. \tag{4.7}$$

4.2.2 Measurement Methods

Speech Materials

A speech material (i. e., a set of speech items like words or sentences) is suitable for SI tests when certain requirements are fulfilled: the different speech items have to be homogeneous in SI to yield high measurement accuracy and reproducibility in a limited measuring time, and the distribution of phonemes should be representative of the language being studied. Only speech materials that have been optimized properly by a large number of evaluation measurements can fulfill these requirements.

A large number of SI tests using different materials are available for different languages. An overview of American SI tests can be found in *Penrod* [4.10]. There are different formats, i. e., nonsense syllables, single words, and sentences. Sentences best represent a realistic communication situation. Nonsense syllables and words allow assessing of confusion matrices and analyzing transmission of information. Furthermore, the intelligibility functions of most sentence tests [4.11–16] show slopes between 0.15 and 0.25 per dB, which are considerably steeper than the values obtained with nonsense syllables or single-word tests.

Since the standard deviation of SRT estimates is inversely proportional to the slope of the intelligibility function (see Sect. 4.2.1), these sentence tests are better suited for efficient and reliable SRT measurements than single-word tests.

Presentation Modes

Signals can be presented either via loudspeakers (free field condition) or via headphones. The free field condition is more natural. Drawbacks are a larger experimental effort and difficulties in calibration. Especially in spatial speech/noise situations (see Sect. 4.2.3), small movements of the listener's head may influence the result of the SI measurement.

The advantages of presentation via headphones are: very good reproducibility for each individual listener, smaller experimental effort, and spatial speech/noise conditions can easily be realized using virtual acoustics. Drawbacks are the individual calibration is complicated because headphones may produce different sound pressures in different ears. Measurements with hearing aids are not possible.

Adaptive procedures can be used to concentrate presentation levels near the SRT, which yields highest efficiency for SRT estimates. In sentence tests, each word can be scored independently, which allows one to design adaptive procedures which converge more efficiently than adaptive procedures usually used in psychoacoustics [4.17].

4.2.3 Factors Influencing Speech Intelligibility

Measuring Method

The various speech materials mentioned above generate different results. Therefore, only standardized speech materials or speech materials with well known reference intelligibility functions should be used.

Noise and Room Acoustics

Noise and reverberation reduce SI. Therefore, if SI in silence is to be measured, environmental noise and reverberation have to be minimized (e.g., using sound insulated cabins and damping headphones). On the other hand, SI measurements can be used to investigate the influence of different noises and room acoustics on SI, which is important for the so called *cocktail-party phenomenon* (see later).

Cocktail-Party Phenomenon

The human auditory system has very impressive abilities in understanding a target talker even if maskers, i. e., competitive sound sources like different talkers, are present at the same time. An interesting review of research on this so-called *cocktail-party phenomenon* can be found in *Bronkhorst* [4.18]. The SI in these multi-talker conditions is influenced by many masker properties such as sound pressure level, frequency spectrum, amplitude modulations, spatial direction, and the number of maskers. The spatial configuration of target speaker and masker plays a very important role. Binaural hearing (hearing with both ears) produces a very effective release from masking (improvement of the SRT) of up to 12 dB compared to monaural hearing (hearing with one ear) [4.18].

Hearing Impairment

An introduction to SI in clinical audiology can be found in *Penrod* [4.10]. Hearing impairment can lead to an increase of the SRT, a decrease of the maximum reachable intelligibility SI_{asymp} and a flatter slope of the intelligibility function. The most difficult situations for hearing impaired listeners are noisy environments with many interfering sound sources (*cocktail-party situation*). Therefore, SI tests in noise are important diagnostic tools for assessing the daily-life consequences of a hearing impairment and the benefit of a hearing aid. SI plays a very important role for the research on and the fitting of hearing aids.

4.2.4 Prediction Methods

Articulation Index (AI), Speech Intelligibility Index (SII), and Speech Transmission Index (STI)

The most common methods for the prediction of speech intelligibility are the articulation index (AI) [4.19–21] which was renamed the speech intelligibility index (SII) [4.22], and the speech transmission index (STI) [4.23, 24] [Table 4.1]. The strength of these models is the large amount of empirical knowledge they are based on. All of these models assume that speech is coded by several frequency channels that carry independent information. This can be expressed by

$$\mathrm{AI} = \sum_i \mathrm{AI}_i \,, \tag{4.8}$$

with AI denoting the cumulative articulation index of all channels and AI_i denoting the articulation index of the single channels (including a weighting of the respective channel).

AI and SII are derived from the speech signal by calculating the signal to noise ratio SNR in the different frequency channels:

$$\mathrm{AI} = \sum_i \frac{W_i(\mathrm{SNR}_i + 15)}{30} \,, \tag{4.9}$$

with W_i denoting a frequency channel weighting factor and SNR_i denoting the signal-to-noise ratio in channel i. W_i depends on the speech material used and takes into account that high frequencies are more important for the recognition of consonants than for the recognition of meaningful sentences. The main differences between the different versions of AI and SII are the way they include nonlinearities like distortion, masking, and broadening of frequency bands.

The speech transmission index (STI) uses the modulation transfer function instead of the SNR and is especially successful for predicting SI in auditoria and rooms, because it explicitly takes into account the flattening of the information-carrying speech envelopes due to reverberation.

The transformation of AI, SII, or STI to speech intelligibility requires a nonlinear transformation that has to be fitted to empirical data. The transformation depends on the kind of speech material used and is usually steeper at its steepest point for materials with context (e.g., sentences) compared to single words.

Statistical Methods

The assumption of independent information in different frequency channels does not hold in all situations because synergetic as well as redundant interactions between different channels occur. The speech recognition sensitivity model [4.25, 26] takes these interactions into account using statistical decision theory in order to model the linguistic entropy of speech.

Functional Method

These methods are based on relatively rough parameterizations of speech (i. e., long-term frequency spectrum and sometimes modulation transfer function). The method (Table 4.1) proposed by *Holube* and *Kollmeier* [4.27], however, is based on physiological and psychoacoustical data and is a combination of a functional model of the human auditory system (Sect. 4.1) and a simple automatic speech recognition system (Sect. 4.3). A drawback of this approach is that there is still a large gap between recognition rates of humans and automatic speech recognition systems (for a review see [4.28], Sect. 4.3).

Table 4.1 Examples of methods for the prediction of speech intelligibility (SI)

Method	Signal parameters	Comments
Articulation index, AI [4.19]	Levels and frequency spectra of speech and noise, kind of speech material	Macroscopic model that describes the influence of the frequency content of speech on intelligibility
Articulation index, AI [4.20]	Levels and frequency spectra of speech and noise, kind of speech material	More complex than French and Steinberg version, describes more nonlinear effects, seldom used
Articulation index, AI [4.21]	Levels and frequency spectra of speech and noise	Simplified version based on [4.19], not in use anymore
Speech intelligibility index, SII [4.22]	Levels and frequency spectra of speech and noise, kind of speech material, hearing loss	Revision of ANSI S3.5-1969, includes spread of masking, standard speech spectra, relative importance of frequency bands
Speech transmission index, STI [4.24]	Modulation transfer function	Predicts the change of intelligibility caused by a speech transmission system (e.g., an auditorium) based on the modulation transfer function of the system
Speech recognition sensitivity model, SRS [4.25, 26]	Levels and frequency spectra of speech and noise, number of response alternatives	Alternative to SII, handles frequency band interactions and is better suited for unsteady frequency spectra
Holube and *Kollmeier* [4.27]	Speech and noise signals, hearing loss	Microscopic modeling of signal processing of auditory system combined with simple automatic speech recognition

4.3 Speech Feature Perception

The information-theoretic approach to describing speech perception assumes that human speech recognition is based on the combined, parallel recognition of several acoustical cues that are characteristic for certain speech elements. While a *phoneme* represents the smallest unit of speech information, its acoustic realization (denoted as *phone*) can be quite variable in its acoustical properties. Such a phone is produced in order to deliver a number of acoustical speech cues to the listener who should be able to deduce from it the underlying phoneme. Each speech cue represents one feature value of more- or less-complex speech features like *voicing*, *frication*, or *duration*, that are linked to phonetics and to perception. These speech feature values are decoded by the listener independently of each other and are used for recognizing the underlying speech element (such as, e.g., the represented phoneme). Speech perception can therefore be interpreted as reception of certain values of several speech features in parallel and in discrete time steps.

Each phoneme is characterized by a unique combination of the underlying speech feature values. The articulation of words and sentences produces (in the sense of information theory) a discrete stream of information via a number of simultaneously active channels (Fig. 4.13).

The spoken realization of a given phoneme causes a certain speech feature to assume one out of several different possible values. For example, the speech feature *voicing* can assume the value one (i. e., voiced sound) or the value zero (unvoiced speech sound). Each of these features is transmitted via its own, specific transmission channel to the speech recognition system of the listener. The *channel* consists of the acoustical transmission channel to the listener's ear and the subsequent decoding of the signal in the central auditory system of the receiver (which can be hampered by a hearing impairment or a speech pathology). The listener recognizes the actually assumed values of certain speech features and combines these features to yield the recognized phoneme.

If $p(i)$ gives the probability (or relative frequency) that a specific speech feature assumes the value i and $p'(j)$ gives the probability (or relative frequency, respectively) that the receiver receives the feature value j, and $p(i, j)$ gives the joint probability that the value j is recognized if the value i is transmitted, then the so-called transinformation T is defined as

$$T = -\sum_{i=1}^{N}\sum_{j=1}^{N} p(i, j)\log_2\left(\frac{p(i)p'(j)}{p(i, j)}\right) . \qquad (4.10)$$

The transinformation T assumes its maximal value for perfect transmission of the input values to the output values, i. e., if $p(i, j)$ takes the diagonal form or any permutation thereof. T equals 0 if the distribution of received feature values is independent of the distribution of input feature values, i. e., if $p(i, j) = p(i)p'(j)$. The maximum value of T for perfect transmission (i. e., $p(i, j) = p(i) = p'(j)$) equals the amount of information (in bits) included in the distribution of input feature values H, i. e.,

$$H = \sum_{i=1}^{N} p(i)H(i) = -\sum_{i=1}^{N} p(i)\log_2[p(i)] . \qquad (4.11)$$

In order to normalize T to give values between 0 and 1, the so-called transinformation index (TI) is often

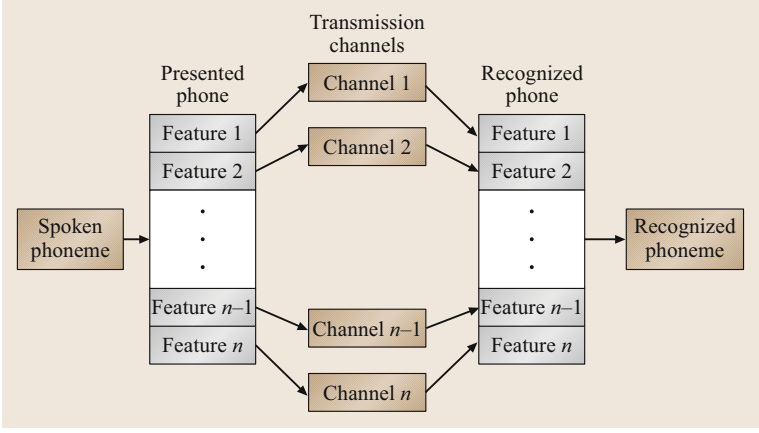


Fig. 4.13 Schematic representation of speech recognition using speech features. Each speech sound is characterized by a combination of speech features that are modeled to be transmitted independently of each other by specialized (noisy) transmission channels

used, i. e.,

$$\mathrm{TI} = T/H \,. \tag{4.12}$$

For speech perception experiments, the distribution $p(i, j)$ can be approximated by a confusion matrix, i. e., a matrix denoting the frequency of a recognized speech element j for all different presented speech elements i. This confusion matrix can be condensed to a confusion matrix of each specific speech feature if for each speech element the corresponding value of the respective feature is assigned. For example, a 20×20 confusion matrix of consonants can be reduced to 2×2 matrix of the feature *voicing* if for each consonant the feature value *voiced* or *unvoiced* is given. The transinformation analysis of this feature-specific confusion matrix therefore allows extracting the transmission of all respective speech features separately and hence can be used to characterize a certain speech information transmission channel. Note however, that such an analysis requires a sufficiently high number of entries in the confusion matrix to appropriately sample $p(i, j)$, which requires a large data set. Also, it is not easy and straightforward to assign appropriate speech features to all of the presented and recognized speech sounds that will allow an adequate analysis of the acoustical deficiencies in the transmission process. From the multitude of different features and feature sets that have been used in the literature to describe both human speech production and speech perception, only a very limited set of the most prominent features can be discussed here ([4.29], for a more-complete coverage).

4.3.1 Formant Features

Vowels are primarily discriminated by their formant structure, i. e., the resonance frequencies of the vocal tract when shaping the vowels. For stationary vowels, the relation between the first and second formant frequencies (F_1 and F_2), respectively, and the perceived vowel is quite well established (Fig. 4.14 and the introduction).

The *classical* theory of vowel perception has the advantage of being linked to the physical process of speech production (i. e., the formant frequencies are closely linked to the position and elevation of the tongue in the vocal tract). However, modern theories of speech perception no longer assume that vowel identification is based solely on the position of the formant frequencies, for the following reasons.

- For short vowels and for vocalic segments of running speech the perceived vowel often differs from the expected spectral shapes, which are based on long, isolated vowels. This indicates that the mapping between formants and perceived vowels shows different boundaries and different regions of overlap depending on the respective speech context.
- The spectral shape of speech in real-life environments varies considerably due to large spectral variations of the room transfer function, and due to the presence of reverberation and background noise. The pure detection and identification of spectral peaks would yield a much less robust perception of vowels than can actually be observed in human listeners.
- Vowel discrimination and speech understanding is even possible under extreme spectral manipulations, such as, e.g., flat spectrum speech [4.30] and slit-filtered speech (i. e., listening to speech through very few spectral slits, [4.31]).

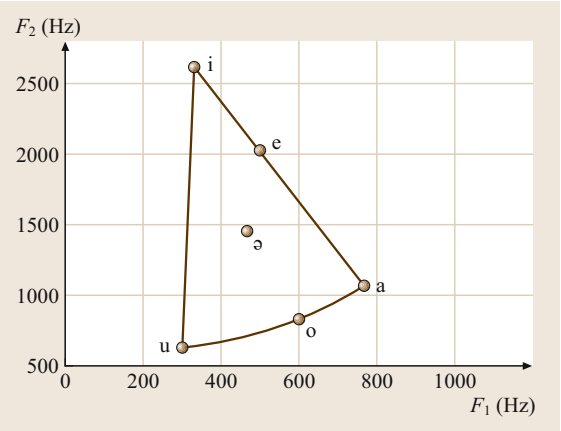


Fig. 4.14 Schematic plot of the perceived vowels as a function of F_1 and F_2 in the vowel triangle (sometimes plotted as a quadrangle) for stationary vowels. Note that the vowel boundaries overlap and that for short vowels and for segments of vowels in real speech these boundaries can vary significantly

These findings indicate that speech perception is at least partially based on temporal cues rather than purely on the detection of spectral peaks or formants. Modern speech perception theories therefore assume that our brain monitors the temporal intensity pattern in each frequency band characterized by a *critical band* filter (Sect. 4.1). By comparing these temporal patterns across a few center frequencies that are not too closely spaced, a reliable estimate of the presented vowel is possible. Such principles are both implemented in state-of-the-art perception models (Sect. 4.1) and in preprocessing/feature extraction strategies for automatic speech recognition systems (Sect. 4.2.3 and the chapters in part E).

4.3.2 Phonetic and Distinctive Feature Sets

The *classical* theory of consonant perception assumes that several phonetically and acoustically defined features are used by the auditory system to decode the underlying, presented consonant. A distinctive feature set combines all binary features that characterize all available consonants in a unique way [4.32]. Both articulatory and (to a lesser degree) acoustical features can be employed to construct such a feature set (see Table 4.2 for an example of a feature set and resulting confusion matrices).

These feature sets where extended and used by *Miller* and *Nicely* [4.33], by *Wang* and *Bilger* [4.34] and subsequently by a large number of researchers to characterize the listeners' ability to discriminate across consonants using, e.g., transinformation analysis (Sect. 4.2.4). Using this approach, the amount of information carried by the specific feature that the receiving side was able to use can be characterized quite well. For example, the confusion matrix listed in Table 4.2 yields a total information transmission index of 0.53 with the features *voicing* assumed to be 0.53 and place 0.46.

However, these phonetic features show only a very weak link to the auditory features actually used by human listeners. From the view point of modern auditory models that assume multichannel temporal energy recording and analysis, most of the phonetic features listed above can be regarded as special prototypes of temporal-spectral patterns that are used by our cognitive system to perform a pattern match between actually presented speech and a stored speech reference database in our brain. Hence, they represent some complex combination of basic auditory perception features that might be defined psychoacoustically or physiologically rather than phonetically.

Table 4.2 Example for a consonant feature set and the construction of confusion matrices for speech-in-noise recognition data with normal-hearing listeners. *Top right panel*: phonetic feature values for eleven consonants. Voicing is a binary feature (with feature values 0 and 1), while manner and place are ternary features. *Middle*: matrix of confusion for consonants, obtained from human listening tests, where noisy speech was presented at an SNR of -10 dB. Matrix element (i, j) denotes how often the consonant in row i was confused with the consonant in column j. *Bottom panels*: confusion matrices for the phonetic features place and voicing, derived from the matrix of confusion for consonants

	Voicing	Manner	Place
p	0	0	0
t	0	0	1
k	0	0	2
b	1	0	0
d	1	0	1
g	1	0	2
s	0	1	1
f	0	1	0
v	1	1	0
n	1	2	2
m	1	2	1

	p	t	k	b	d	g	s	f	v	n	m
p	379	20	131	45	7	31		49	46	4	5
t	3	658	16		33			1	1		
k	42	14	484	10	8	117	1	16	12	6	
b	58	4	51	260	35	88		18	143	16	25
d	5	28	7	21	424	93	1	3	19	49	6
g	11	5	44	43	27	449		9	73	18	7
s		2					702	3			
f	23		3	4			88	556	38		
v	19	7	16	78	22	43	7	51	398	9	30
n	1	5	3	13	51	12		2	23	364	78
m	7	1	4	43	20	25		8	62	95	346

	Anterior	Medail	Posterior
Anterior	2691	335	392
Medail	179	2317	131
Posterior	223	79	1094

	Voiced	Unvoiced
Voiced	3508	375
Unvoiced	367	3191

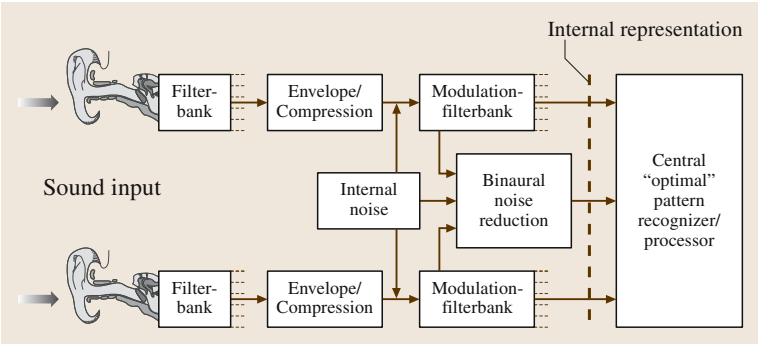


Fig. 4.15 Schematic diagram of a model of the *effective* auditory processing using a front end to transform the incoming speech signal into and internal representation and a subsequent *back end*, ideal recognition stage which is only limited by the internal noise

4.3.3 Internal Representation Approach and Higher-Order Temporal-Spectral Features

The *internal representation* approach of modeling speech reception assumes that the speech signal is transformed by our auditory system with some nonlinear, parallel processing operations into an *internal representation*. This representation is used as the input for a central, cognitive recognition unit which can be assumed to operate as an *ideal observer*, i. e., it performs a pattern match between the incoming internal representation and the multitude of stored internal representations. The accuracy of this recognition process is limited by the *external variability* of the speech items to be recognized, i. e., by their deviation from any of the stored internal templates. It is also limited by the *internal noise* that blurs the received *internal representation* due to neural noise and other physiological and psychological factors. The amount of internal noise can be estimated quite well from psychoacoustical experiments.

Such an internal representation model puts most of the peculiarities and limitations of the speech recognition process into the nonlinear, destructive transformation process from the acoustical speech waveform into its internal representation, assuming that all transformation steps are due to physiological processes that can be characterized completely physiologically or by psychoacoustical means (Fig. 4.15).

Several concepts and models to describe such an internal representation have been developed so far. Some of the basic ideas are as follows.

1. Auditory spectrogram: The basic internal representation assumes that the speech sound is separated into a number of frequency bands (distributed evenly across a psychoacoustically based frequency scale like the bark or ERB scale) and that the compressed frequency-channel-specific intensity is represented

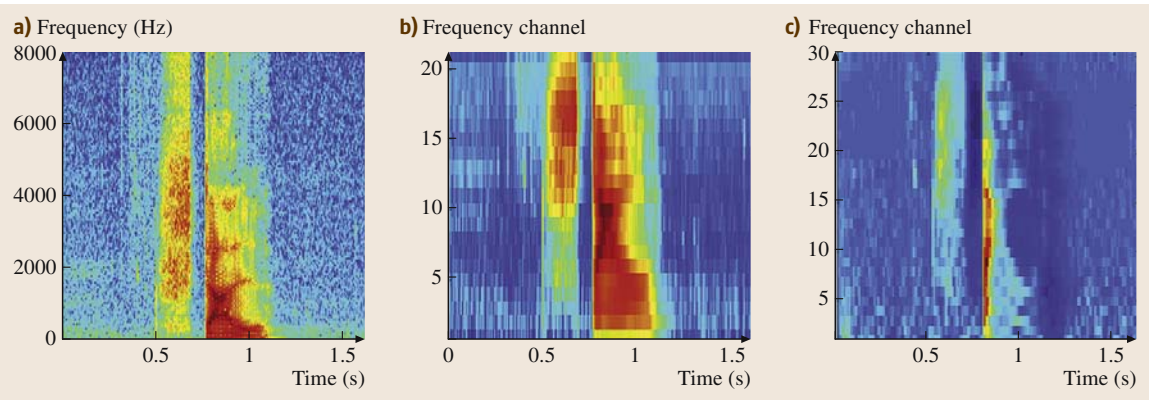


Fig. 4.16a–c Auditory spectrogram representation of the German word *Stall*. It can be represented by a spectrogram **(a)**, a bark spectrogram on a log-loudness scale **(b)** or as a contrast-enhanced version using nonlinear feedback loops (after [4.3, 35]) **(c)**

over time. The compression can either be a logarithmic compression or a loudness-derived power law compression that is also required to represent human intensity resolution and loudness mapping. The temporal representation can also include some temporal contrast enhancement and sluggishness in order to represent forward and backward masking and temporal integration (Sect. 4.1). An example of such a representation is given in Fig. 4.16.

Note that speech intelligibility in noise can be modeled quite well with such an approach [4.27]. In addition, such a transformation into the *internal representation* can be implemented as a robust front end for automatic speech recognition (e.g., [4.35]). Finally, it can be used to predict any perceived deviations of the (coded) speech from the original speech [4.36].

2. Modulation spectrogram: one important property of the internal representation is the temporal analysis within each audio frequency band using the modulation filter bank concept. Temporal envelope fluctuations in each audio frequency channel are spectrally analyzed to yield the modulation spectrum in each frequency band, using either a fixed set of modulation filters (modulation filter bank) or a complete spectral analysis (modulation spectrum). This representation yields the so-called amplitude modulation spectrogram for each instant of time, i. e., a two-dimensional representation of modulation frequency across center audio frequencies (Fig. 4.17).

The physiological motivation for this analysis is the finding of amplitude modulation sensitivity in the auditory brain:adjacent cells are tuned to different modulation frequencies. Their arrangement seems to yield a perpendicular representation of modulation frequencies across center frequencies [4.37]. In addition, psychoacoustical findings of modulation sensitivity can best be described by a set of modulation filter banks [4.3]. The advantage of the modulation spectrogram is that the additional dimension of modulation frequency allows separation of acoustical objects that occupy the same center frequency channel, but are modulated at different rates (considering either the syllabic rate at low modulation frequencies or temporal pitch at higher modulation frequencies). Such a more-refined model of internal representation has been used to predict psychoacoustical effects [4.3] and was also used in automatic speech recognition [4.38].

3. Temporal/spectral ripple or Gabor feature approach: A generalization of the modulation frequency feature detectors in the temporal domain outlined above also considers the spectral analysis of ripples in

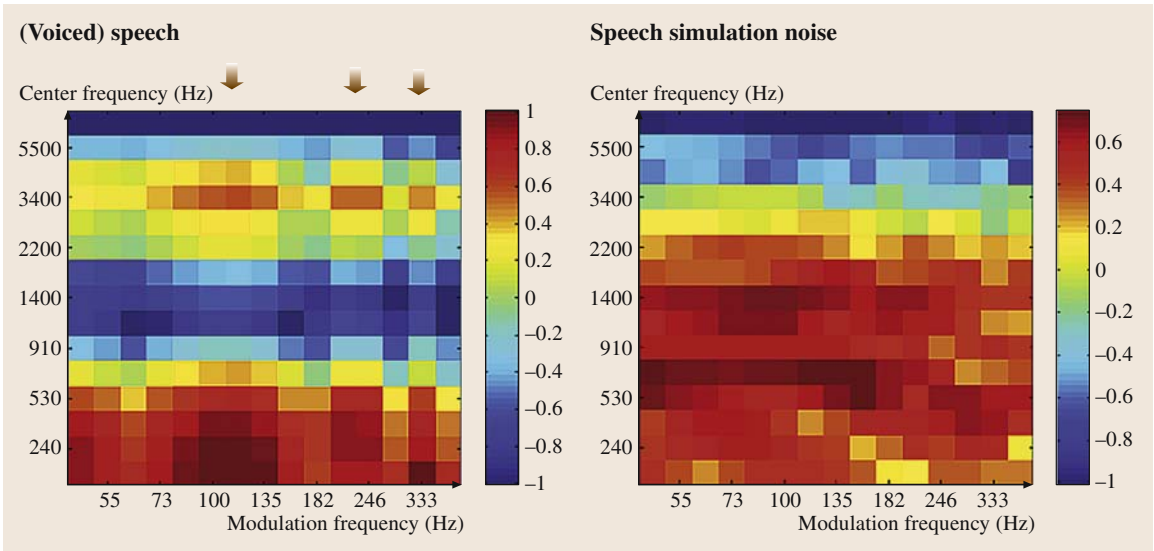


Fig. 4.17 Amplitude modulation spectrogram of a vowel *I* (fundamental frequency approx. 110 Hz) in comparison to a modulation spectrogram of speech-simulating noise. The modulation spectrum in each audio frequency band is displayed as color (or greyscale, respectively) in the two-dimensional plane given by audio center frequency versus modulation frequency

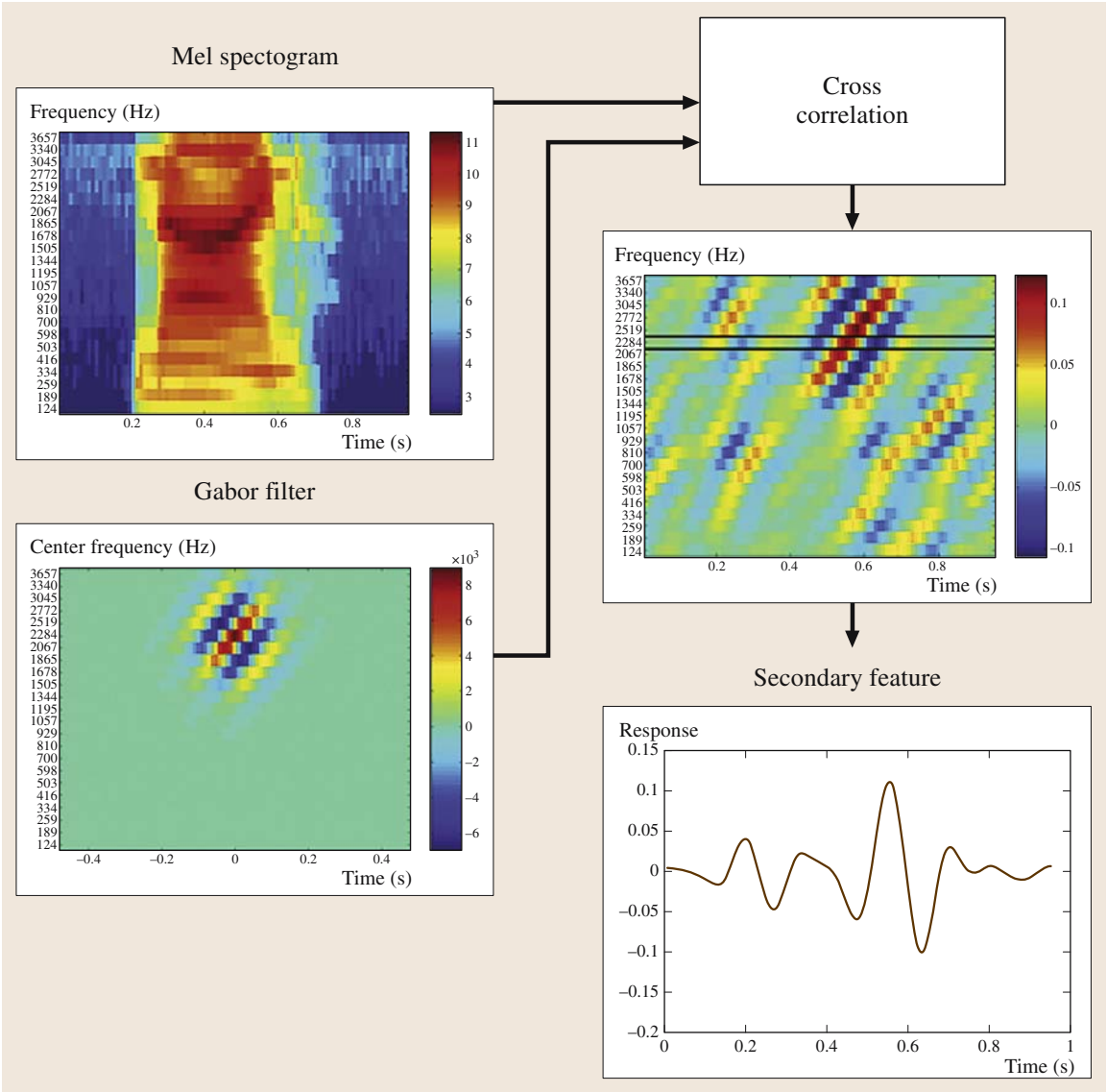


Fig. 4.18 Sample representation of a Gabor feature that detects a certain speech feature. The two-dimensional Gabor feature (*lower left panel*) that extends both in the time and frequency domain is cross-correlated with the mel spectrogram (*upper-left panel*) to yield the temporal and spatial position of a best match (*middle panel*). In each audio frequency band, the time-dependent output of the cross correlation is used as the input feature to an automatic speech recognizer [4.38]

the frequency domain as well as a ripple frequency analysis for combined temporal and spectral modulations. Such a temporal-spectral ripple analysis is motivated by physiological findings of the auditory receptive fields in ferrets [4.39] as well as psychoacoustical findings by *Kaernbach* [4.40] who demonstrated a sensitivity towards combinations of spectral variations and temporal variations. An elegant way to formalize the sensitivity to joint temporal and spectral energy variations is the Gabor feature concept [4.38] that considers features with a limited spectro-temporal extent tuned to a certain combina-

Table 4.3 Recognition rates (in percent correct) for human speech recognition (HSR) at a signal-to-noise ratio (SNR) of −10 dB, compared to automatic speech recognition (ASR) accuracies at several signal-to-noise ratios. The recognition task for ASR and for human listeners was to classify the middle phoneme in simple nonsense words, which were combinations of either consonant–vowel–consonant or vowel–consonant–vowel. The average rates are broken down into consonant and vowel recognition. At +10 dB SNR, ASR reaches an overall performance that is comparable to HSR at −10 dB SNR. If the same SNR of −10 dB is employed for ASR, error rates are almost 50% higher than for HSR

Condition		Average	Consonants	Vowels
HSR	−10 dB	74.5	67.7	80.5
ASR				
	clean	80.4	85.2	76.2
	15 dB	76.1	77.7	74.6
	10 dB	74.6	75.6	73.7
	5 dB	69.8	69.5	70.0
	0 dB	59.2	55.4	62.5
	−5 dB	49.8	41.0	57.5
	−10 dB	28.4	20.8	35.0

tion of temporal modulation frequency and spectral ripple frequency (Figure 4.18).

The advantage of such a second-order receptive field (i. e., the sensitivity to a certain combination of a spectral and a temporal cue) is the ability to detect specific spectro-temporal structures, e.g., formant glides or changes of fundamental frequency. It can also be considered as a generalization of the concepts outlined in this section. Even though this approach has successfully been implemented to improve the robustness of automatic speech recognizers [4.38], it has not yet been used to model human speech perception.

4.3.4 Man–Machine Comparison

Despite enormous technical advances in recent years, automatic speech recognition (ASR) still suffers from a lack of performance compared to human speech recognition (HSR), which prevents this technology from being widely used. Recognition accuracies of machines drop dramatically in acoustically adverse conditions, i. e., in the presence of additive or convolutive noise, which clearly demonstrates the lack of robustness. For complex tasks such as the recognition of spontaneous speech, ASR error rates are often an order of magnitude higher than those of humans [4.28]. If no high-level grammatical information can be exploited (as in a simple phoneme recognition task), the difference in performance gets smaller, but still remains very noticeable. For example, the HSR consonant recognition rate derived from the confusion matrix in Table 4.2 is 67.7%. The ASR score for the very same task (i. e., the same speech signals at an SNR of −10 dB), obtained with a common recognizer is 20.8%, which corresponds to a relative increase of error rates of 144% (Table 4.3).

This large gap underlines that current state-of-the-art ASR technology is by far not as capable as the human auditory system to recognize speech. As a consequence, the fields of ASR and speech perception modeling in humans may benefit from each other. Since the human auditory system results from a long biological evolution process and seems to be optimally adjusted to perform robust speech recognition, ASR may profit from auditory front ends which are based upon physiological findings and incorporate principles of our hearing system. Ideally, the feature matrix extracted from a speech sound which is used to classify the respective speech element should resemble the internal representation of that speech sound in our brain as closely as possible. Since this internal representation can be approximated by an auditory model, such an auditory model seems to be a good preprocessing stage for a speech recognizer [4.35].

On the other hand, models of the signal processing in the human auditory system can be evaluated using ASR, because–under ideal conditions–human speech perception and its model realization as anthropomorphic ASR system should yield a similar recognition performance and error pattern in well-defined acoustical conditions. Thus, modeling human speech perception can benefit from the computational methods developed in ASR.

References

4.1 H. Fastl, E. Zwicker: *Psychoacoustics: Facts and Models* (Springer, Berlin-Heidelberg 2005)

4.2 E. Zwicker, G. Flottorp, S.S. Stevens: Critical bandwidth in loudness summation, J. Acoust. Soc. Am. **29**, 548 (1957)

4.3 T. Dau, B. Kollmeier, A. Kohlrausch: Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers, J. Acoust. Soc. Am. **102**, 2892–2905 (1997)

4.4 M.R. Schroeder: *Computer Speech: Recognition, Compression, Synthesis* (Springer, Berlin-Heidelberg 2005)

4.5 B.C.J. Moore, R.D. Patterson: *Auditory Frequency Selectivity* (Plenum, New York 1986)

4.6 A.J. Houtsma: Pitch perception. In: *Handbook of Perception and Cognition: Hearing*, ed. by B.C.J. Moore (Academic, London 1995) pp. 267–295

4.7 J. Verhey, D. Pressnitzer, I.M. Winter: The psychophysics and physiology of co-modulation masking release, Exp. Brain Res. **153**, 405–417 (2003)

4.8 R. Beutelmann, T. Brand: Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners J. Acoust. Soc. Am. **120**(1), 33–42 (2006)

4.9 H.S. Colburn, N.I. Durlach: Models of binaural interaction. In: *Handbook of Perception*, Vol. 4 (Academic, New York 1978) pp. 467–518

4.10 J.P. Penrod: Speech threshold and word recognition/ discrimination testing. In: *Handbook of Clinical Audiology, 4th edn*, ed. by J. Katz (Williams and Wilkins, Baltimore 1994) pp. 147–164

4.11 R. Plomp, A. Mimpen: Improving the reliability of testing the speech-reception threshold for sentences, Audiology **18**, 43–52 (1979)

4.12 B. Hagerman: Sentences for testing speech intelligibility in noise, Scand. Audiol. **11**, 79–87 (1982)

4.13 M. Nilsson, S.D. Soli, J.A. Sullivan: Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise, J. Acoust. Soc. Am. **95**(2), 1085–1099 (1994)

4.14 B. Kollmeier, M. Wesselkamp: Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment, J. Acoust. Soc. Am. **102**, 2412–2421 (1997)

4.15 K. Wagener, V. Kühnel, B. Kollmeier: Entwicklung und Evaluation eines Satztests für die deutsche Sprache I: Design des Oldenburger Satztests (Development and evaluation of a German sentence test I: Design of the Oldenburg sentence test), Zeitschrift für Audiologie **38**, 4–15 (1999)

4.16 K. Wagener, J.L. Josvassen, R. Ardenkjaer: Design, optimization and evaluation of a Danish sentence test in noise, Int. J. Audiol. **42**(1), 10–17 (2003)

4.17 T. Brand, B. Kollmeier: Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests, J. Acoust. Soc. Am. **111**(6), 2801–2810 (2002)

4.18 A. Bronkhorst: The cocktail party phenomenon: a review of research on speech intelligibility in multiple-talker conditions, Acustica **86**, 117–128 (2000)

4.19 N.R. French, J.C. Steinberg: Factors governing the intelligibility of speech sounds, J. Acoust. Soc. Am. **19**, 90–119 (1947)

4.20 H. Fletcher, R.H. Galt: The perception of speech and its relation to telephony, J. Acoust. Soc. Am. **22**, 89–151 (1950)

4.21 ANSI: *Methods for the calculation of the articulation index, ANSI S3.5-1969* (American National Standards Institute, New York 1969)

4.22 ANSI: *Methods for calculation of the speech intelligibility index, ANSI S3.5-1997* (American National Standards Institute, New York 1997)

4.23 T. Houtgast, H.J.M. Steeneken: A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria, J. Acoust. Soc. Am. **77**, 1069–1077 (1985)

4.24 IEC: Sound system equipment – Part 16: Objective rating of speech intelligibility by speech transmission index. INTERNATIONAL STANDARD 60268-16 Second edition 1998-03 (1998)

4.25 H. Müsch, S. Buus: Using statistical decision theory to predict speech intelligibility. I. Model structure, J. Acoust. Soc. Am. **109**, 2896–2909 (2001)

4.26 H. Müsch, S. Buus: Using statistical decision theory to predict speech intelligibility, II. Measurement and prediction of consonant-discrimination performance J. Acoust. Soc. Am. **109**, 2910–2920 (2001)

4.27 I. Holube, B. Kollmeier: Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model, J. Acoust. Soc. Am. **100**, 1703–1716 (1996)

4.28 R. Lippmann: Speech recognition by machines and humans, Speech Commun. **22**, 1–15 (1997)

4.29 S. Greenberg, W.A. Ainsworth, A.N. Popper: Speech Processing in the Auditory System. In: *Handbook of Auditory research*, Vol. 18, ed. by R.R. Fay (Springer, New York 2004)

4.30 M.R. Schroeder, H.W. Strube: Flat spectrum speech, J. Acoust. Soc. Am. **79**, 1580–1583 (1986)

4.31 R.V. Shannon, F.G. Zeng, V. Kamth, J. Wygonsky, M. Ekelid: Speech recognition with primarily temporal cues, Science **270**, 303–304 (1995)

4.32 R. Jakobson, C.G.M. Fant, M. Halle: *Preliminaries to speech analysis: the distinctive features and their correlates* (MIT Press, Cambridge 1963)

4.33 G.A. Miller, P.E. Nicely: An analysis of perceptual confusions among some english consonants, J. Acoust. Soc. Am. **27**, 338–352 (1955)

4.34 M.D. Wang, R.C. Bilger: Consonant confusions in noise: a study of perceptual features, J. Acoust. Soc. Am. **54**, 1248–1266 (1973)
4.35 J. Tchorz, B. Kollmeier: A model of auditory perception as front end for automatic speech recognition, J. Acoust. Soc. Am. **106**(4), 2040–2050 (1999)
4.36 M. Hansen, B. Kollmeier: Objective modeling of speech quality with a psychoacoustically validated auditory model, J. Audio Eng. Soc. **48**(5), 395–408 (2000)
4.37 C.E. Schreiner, G. Langner: Periodicity coding in the inferior colliculus of the cat II. Topographical organization, J. Neurophys. **60**, 1823–1840 (1988)
4.38 M. Kleinschmidt: Methods for capturing spectro-temporal modulations in automatic speech recognition, Acustica united with Acta Acustica **88**(3), 416–422 (2002)
4.39 D.A. Depireux, J.Z. Simon, D.J. Klein, S.A. Shamma: Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex, J. Neurophysiol. **85**(3), 1220–1234 (2001)
4.40 C. Kaernbach: The Memory of Noise, Exp. Psychol. **1**(4), 240–248 (2004)

5. Speech Quality Assessment

V. Grancharov, W. B. Kleijn

In this chapter, we provide an overview of methods for speech quality assessment. First, we define the term *speech quality* and outline in Sect. 5.1 the main causes of degradation of speech quality. Then, we discuss subjective test methods for quality assessment, with a focus on standardized methods. Section 5.3 is dedicated to objective algorithms for quality assessment. We conclude the chapter with a reference table containing common quality assessment scenarios and the corresponding most suitable methods for quality assessment.

5.1 **Degradation Factors Affecting Speech Quality** 84
5.2 **Subjective Tests** 85
5.2.1 Single Metric (Integral Speech Quality) 85
5.2.2 Multidimensional Metric (Diagnostic Speech-Quality) 87
5.2.3 Assessment of Specific Quality Dimensions 87
5.2.4 Test Implementation 88
5.2.5 Discussion of Subjective Tests 89
5.3 **Objective Measures** 90
5.3.1 Intrusive Listening Quality Measures 90
5.3.2 Non-Intrusive Listening Quality Measures 93
5.3.3 Objective Measures for Assessment of Conversational Quality 94
5.3.4 Discussion of Objective Measures 94
5.4 **Conclusions** 95
References 96

The rapid deployment of speech processing applications increases the need for speech quality evaluation. The success of any new technology (e.g., network equipment, speech codec, speech synthesis system, etc.) depends largely on end-user opinion of perceived speech quality. Therefore, it is vital for the developers of a new service or speech processing application to assess its speech quality on a regular basis.

In addition to its role for services and speech processing, speech quality evaluation is of critical importance in the areas of clinical hearing diagnostics and psychoacoustical research. Although this chapter addresses speech quality mainly from the viewpoint of telecommunication applications, it is also of general interest for researchers dealing with speech quality assessment methods.

When the speech signal reaches the human auditory system, a speech perception process is initiated. This process results in an *auditory event*, which is internal and can be measured only through a description by the listener (the subject). The subject then establishes a relationship between the perceived and expected auditory event. Thus, the speech quality is a result of a perception and assessment process.

Since the quality of a speech signal does not exist independently of a subject, it is a *subjective* measure. The most straightforward manner to estimate speech quality is to play a speech sample to a group of listeners, who are asked to rate its quality. Since subjective quality assessment is costly and time consuming, computer algorithms are often used to determine an *objective* quality measure that approximates the subjective rating. Section 5.2 provides an overview of subjective tests for speech quality assessment, while Sect. 5.3 is dedicated to objective quality assessment measures.

Speech quality has many perceptual dimensions. Commonly used dimensions are intelligibility, naturalness, loudness, listening effort, etc., while less commonly used dimensions include nasality, graveness, etc. However, the use of a *multidimensional metric* for quality assessment is less common than the use of a *single metric*, mainly as a result of cost and complexity. A single metric, such as the mean opinion score scale, gives an integral (overall) perception of an auditory event and is therefore sufficient to predict the end-user opinion of a speech communication system. However, a single metric does not in general provide sufficient detail for system designers. Multidimensional-metric tests

are discussed in Sect. 5.2.2 and single-metric tests are discussed throughout the remainder of Sect. 5.2.

In some applications, it is desirable or historically accepted to measure only specific quality dimensions, such as *intelligibility*, *listening effort*, *naturalness*, and *ability for talker recognition*. The most popular among these measures are covered in Sect. 5.2.3.

The *true* speech quality is often referred to as *conversational quality*. Conversational tests usually involve communication between two people, who are questioned later about the quality aspects of the conversation, see Sect. 5.2.1. However, the most frequently measured quantity is *listening quality*, which is the focus of Sect. 5.2.1. In the listening context, the speech quality is mainly affected by speech distortion due to speech codecs, background noise, and packet loss. One can also distinguish *talking quality*, which is mainly affected by echo associated with delay and sidetone distortion.

The distorted (processed) signal or its parametric representation is always required in an assessment of speech quality. However, based on the availability of the original (unprocessed) signal, two test situations are possible: *reference based* and *not reference based*. This classification is common for both the subjective and objective evaluation of speech quality. The absolute category rating (ACR) procedure, popular in subjective tests, does not require the original signal, while in the degradation category rating (DCR) approach the original signal is needed. In objective speech quality assessment, the historically accepted terms are *intrusive* (with original) and *non-intrusive* (without original). These two test scenarios will be discussed throughout the chapter.

5.1 Degradation Factors Affecting Speech Quality

The main underlying causes of degradation of speech quality in modern speech communication systems are delay (latency), packet loss, packet delay variation (jitter), echo, and distortion introduced by the codec. These factors affect psychological parameters such as intelligibility, naturalness, and loudness, which in turn determine the overall speech quality.

In this section, we briefly list the most common impairment factors. We divide them into three classes:

1. factors that lead to listening difficulty
2. factors that lead to talking difficulty
3. factors that lead to conversational difficulty

The reader can find more-detailed information in International Telecommunication Union, Telecommunication Standardization Sector (ITU-T) Rec. G.113 [5.1]. The effect of transmission impairments on users is discussed in ITU-T Rec. P.11 [5.2].

Degradation factors that cause an increase in *listening difficulty* include packet loss, distortion due to speech codecs, speech clipping, and listener echo. *Packet loss* corresponds to the percentage of speech frames that do not reach their final destination. If no protective measures are taken, a packet loss rate of 5% results in significant degradation of the speech quality. Bursts of packet loss also affect speech quality. In systems without error concealment, *speech clipping* occurs at any time when the transmitted signal is lost. Speech clipping may temporarily occur when the connection suffers from packet loss or when voice activity detectors are used. *Listener echo* refers to a transmission condition in which the main speech signal arrives at the listener's end of the connection accompanied by one or more delayed versions (echoes) of the signal. The intelligibility decreases as the *loudness loss* increases. On the other hand, if the loudness loss decreases too much, customer satisfaction decreases because the received speech is too loud.

Degradation factors that cause difficulty while *talking* are talker echo and an incorrectly set sidetone. *Talker echo* occurs when some portion of the talker's speech signal is returned with a delay sufficient (typically more than 30 ms) to make the signal distinguishable from the normal sidetone. The *sidetone* of a telephone set is the transmission of sound from the telephone microphone to the telephone receiver in the same telephone set. Too little sidetone loss causes the returned speech levels to be too loud and thus reduces customer satisfaction. Excessive sidetone loss can make a telephone set sound *dead* as one is talking. In addition, the sidetone path provides another route by which room noise can reach the ear.

Conversation difficulties are caused by a third class of degradation factors. *Delay* is defined as the time it takes for the packet to arrive at its destination. Long delays impair a conversation. *Intelligible crosstalk* occurs when the speech signal from one telephone connection is coupled to another telephone connection such that the coupled signal is audible and intelligible to one or both of the participants on the second telephone connection. The *background noise* in the environment of the tele-

phone set may have a substantial effect on the ease of carrying on a conversation.

The study of degradation factors is important in the design of a speech quality assessment test. The set of degradation factors present in a communication system determines the type of test to be performed. If the degradation factors cause only an increase in listening difficulty, it is sufficient to perform relatively inexpensive and simple tests that measure *listening quality*. If the degradation factors cause difficulty while talking, or difficulty while conversing, it is recommended to perform the more-complex *conversational quality* tests.

5.2 Subjective Tests

Speech quality is a complex psychoacoustic outcome of the human perception process. As such, it is necessarily subjective, and can be assessed through listening test involving human test subjects that listen to a speech sample and assign a rating to it. In this section, we cover the most commonly used subjective quality tests.

5.2.1 Single Metric (Integral Speech Quality)

Users of new speech processing applications are often unaware of the underlying technology. Their main criterion for assessing these applications is based on overall speech quality. Therefore, we start our discussion with *single-metric* subjective tests. In these tests, speech is played to a group of listeners, who are asked to rate the quality of this speech signal based on their *overall perception*.

Listening Quality

In an ACR test, a pool of listeners rate a series of audio files using a five-level impairment scale, as shown in Table 5.1. After each sample is heard, the listeners express an opinion, based only on the most recently heard sample. The average of all scores thus obtained for speech produced by a particular system represents its *mean opinion score* (MOS). The ACR listening quality method is standardized in [5.3], and is the most commonly used subjective test procedure in telecommunications. The main reason for the popularity of this test is its simplicity.

A good method for obtaining information on the detectability of a distortion (e.g., echo) as a function of some objective quantity (e.g., listening level) is to use the *detectability opinion scale* (Table 5.2). The decisions on a detectability scale are not equivalent to responses on a continuous scale. It is therefore recommended to use as a method of analysis the probability of response [5.3].

A disadvantage of ACR methods is that for some applications the resolution of their quality scale is not sufficient. In such cases the DCR method is appropriate. DCR methods provide a quality scale of higher resolution, due to comparison of the distorted signal with one or more reference/anchor signals. In a DCR test, the listeners are presented with the unprocessed signal as a reference before they listen to the processed signal. The task for the listener is to rate the perceived degradation by comparing the second stimulus to the first on the scale presented in Table 5.3. The quantity evaluated from the scores is referred to as the *degradation mean opinion score* (DMOS). DCR methods are also standardized in [5.3].

ABX is another popular method for speech quality assessment [5.4]. It consists of presenting the listener

Table 5.1 Grades in the MOS scale. Listeners express their opinion on the quality of the perceived speech signal (no reference presented)

Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

Table 5.2 Grades in the *detectability opinion* scale. Listeners give their opinion on the detectability of some property of a sound

Objectionable	3
Detectable but not objectionable	2
Not detectable	1

Table 5.3 Grades in the DMOS scale. Listeners are asked to describe *degradation* in the second signal in relation to the first signal

Inaudible	5
Audible but not annoying	4
Slightly annoying	3
Annoying	2
Very annoying	1