

IVAN TASHEV

Sound Capture and Processing

PRACTICAL
APPROACHES



APPLICATION

Audio
Processing
Object



AEC

DRIVER

Speaker



Companion Website

 WILEY

TN912.3
T197

Sound Capture and Processing

Practical Approaches

Ivan J. Tashev

Microsoft Research, USA



 **WILEY**

A John Wiley and Sons, Ltd., Publication



E2010002076

This edition first published 2009
© 2009 John Wiley & Sons Ltd.,

Registered office

John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com.

The right of the author to be identified as the author of this work has been asserted in accordance with the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by the UK Copyright, Designs and Patents Act 1988, without the prior permission of the publisher.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The publisher is not associated with any product or vendor mentioned in this book. This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold on the understanding that the publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

MATLAB® is a trademark of The MathWorks, Inc., and is used with permission. The MathWorks does not warrant the accuracy of the text or exercises in this book. This book's use or discussion of MATLAB® software or related products does not constitute endorsement or sponsorship by The MathWorks of a particular pedagogical approach or particular use of MATLAB® software.

Library of Congress Cataloging-in-Publication Data

Tashev, Ivan J. (Ivan Jelev)

Sound capture and processing : practical approaches / Ivan J. Tashev.
p. cm.

Includes index.

ISBN 978-0-470-31983-3 (cloth)

1. Speech processing systems. 2. Sound-Recording and reproducing-Digital techniques. 3. Signal processing-Digital techniques. I. Title.

TK7882.S65T37 2009

621.382'8-dc22

2009011987

A catalogue record for this book is available from the British Library.

ISBN 978-0-470-31983-3 (H/B)

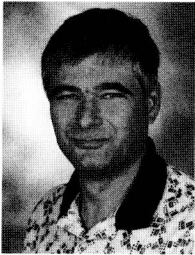
Typeset in 11/13pt Times by Thomson Digital, Noida, India.

Printed and bound in Great Britain by CPI Antony Rowe, Chippenham, Wiltshire.

Sound Capture and Processing

*To my family: the time to write
this book was taken from them*

About the Author



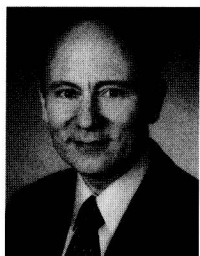
Dr Ivan Tashev took both his Engineering Diploma in Electronics and PhD in Computer Science degrees at the Technical University of Sofia, Bulgaria, in 1984 and 1990 respectively. After his graduation he worked as R&D engineer and researcher in the R&D Department of the same university. Dr Tashev became assistant professor in 1989. He created and taught two courses, “Data and signal processing” and “Programming of real-time systems” to the students of fourth and fifth year in the Department of Electronics.

Dr Tashev joined Microsoft in 1998 and held positions in various product teams until 2001 when he moved to Microsoft Research. Here he was involved in projects such as RingCam (now a Microsoft product – Round Table Device), microphone array (currently part of Windows Vista), and many others related to sound capturing devices and audio signal processing. Currently he is a member of the Speech Technology Group in Microsoft Research lab at the Microsoft headquarters in Redmond, Washington.

Dr Ivan Tashev is senior member of IEEE and IEEE Signal Processing Society, member of Audio Engineering Society and its Pacific Northwest Committee. He is reviewer for most of the audio and signal processing journals and conferences. Dr Tashev has published three books, more than fifty scientific papers and is listed as inventor of five granted U.S. patents and seventeen U.S. patent applications.

The research interests of Dr Tashev include sound capturing devices, signal processing for arrays of transducers, speech enhancement algorithms, and signal processing of audio, speech and biological signals.

Foreword



Just a couple of decades ago we would think of “sound capture and processing” as the problems of designing microphones for converting sounds from the real world into electrical signals, as well as amplifying, editing, recording, and transmitting such signals, mostly using analog hardware technologies. That’s because our intended applications were mostly analog telephony, broadcasting, and voice and music recording. We have come a long way: small digital audio players have replaced bulky portable cassette tape players, and people make voice calls mostly via digital mobile phones and voice communication software in their com-

puters. Thanks to the evolution of digital signal processing technologies, we now focus mostly on processing sounds not as analog electrical signals, but rather as digital files or data streams in a computer or digital device. We can do a lot more with digital sound processing, such as transcribe speech into text, identify persons speaking, recognize music from humming, remove noises much more efficiently, add special effects, and so much more. Thus, today we think of sound capture as the problem of digitally processing the signals captured by microphones so as to improve their quality for best performance in digital communications, broadcasting, recording, recognition, classification, and other applications.

This book by Ivan Tashev provides a comprehensive yet concise overview of the fundamental problems and core signal processing algorithms for digital sound capture, including ambient noise reduction, acoustic echo cancellation, and reduction of reverberation. After introducing the necessary basic aspects of digital audio signal processing, the book presents basic physical properties of sound and propagation of sound waves, as well as a review of microphone technologies, providing the reader with a strong understanding of key aspects of digitized sounds. The book discusses the fundamental problems of noise reduction, which are usually solved via techniques based on statistical models of the signals of interest (typically voice) and of interfering signals. An important discussion of properties of the human auditory system is also presented; auditory models can play a very important role in algorithms for enhancing audio signals in communication and recording/playback applications, where the final destination is the human ear.

Microphone arrays have become increasingly important in the past decade or so. Thanks to the rapid evolution and reduction in cost of analog and digital electronics in recent years, it is inexpensive to capture sound through several channels, using an array of microphones. That opens new opportunities for improving sound capture, such as detecting the direction of incoming sounds and applying spatial filtering techniques. The book includes two excellent

chapters whose coverage goes from the basics of microphone array configurations and delay-and-sum beamforming, to modern sophisticated algorithms for high-performance multichannel signal enhancement.

Acoustic echoes and reverberation are the two most important kinds of signal degradations in many sound capture scenarios. If you're a professional singer, you probably don't mind holding a microphone or wearing a headset with a microphone close to your mouth, but most of us prefer microphones to be invisible, far away from our mouths. That means microphone will capture not only our own voices, but also reverberation components because of sound reflections from nearby walls, as well as echoes of signals that are being played back from loudspeakers. Removing such undesirable artifacts presents significant technical challenges, which are well addressed in the final two chapters, which present modern algorithms for tackling them.

A key quality of this book is that it presents not only fundamental theoretical analyses, models, and algorithms, but it also considers many practical aspects that are very important for the design of real-world engineering solutions to sound capture problems. Thus, this book should be of great appeal to both students and engineers.

I have had the pleasure of working with Ivan on research and development of sound capture systems and algorithms. His enthusiasm, deep engineering and mathematical knowledge, and pragmatic approaches were all contagious. His work has had significant practical impact, for example the introduction of multichannel sound capture and processing modules in the Microsoft Windows operating system. I have learned a considerable amount about sound capturing and processing from my interactions with Ivan, and I am sure you will, as well, by reading this book. Enjoy!

Henrique Malvar
Managing Director
Microsoft Research
Redmond Laboratory

Preface

Capturing and processing sounds is critical in mobile and handheld devices, communication systems, and computers using automatic speech recognition. Devices and technologies for proper conversion of sounds to electric signals and removing unwanted parts, such as noise and reverberation, have been used since the first telephones. They evolved, becoming more and more complex. In many cases the existing algorithms exceed the abilities of typical processors in these devices and computers to provide real-time processing of the captured signal.

This book will discuss the basic principles for building an audio processing stack, sound capturing devices, single-channel speech-enhancement algorithms, and microphone arrays for sound capture and sound source localization. Further, algorithms will be described for acoustic echo cancellation and de-reverberation – building blocks of a sound capture and processing stack for telecommunication and speech recognition. Wherever possible the various algorithms are discussed in the order of their development and publication. In all cases the aim is to try to give the larger picture – where the technology came from, what worked and what had to be adapted for the needs of audio processing. This gives a better perspective for further development of new audio signal processing algorithms.

Even the best equations and signal processing algorithms are not worth anything before being implemented and verified by processing of real data. That is why, in this book, stress is placed on experimenting with recorded sounds and implementation of the algorithms. In practice, frequently a simpler model with fewer parameters to estimate works better than a more precise but more complex model with a larger number of parameters. With the latter one has either to sacrifice estimation precision or to increase the estimation time. This balance of simplicity, precision, and reaction time is critical for real-time systems, where on top of everything we have to watch out for parameters such as latency, consumed memory, and CPU time.

Most of the algorithms and approaches described in this book are based on statistical models. In mathematics, a single example cannot prove but can disprove a theorem. In statistical signal processing, a single example is . . . just a sample. What matters is careful evaluation of the algorithms with a good corpus of speech or audio signals, distributed in their signal-to-noise ratios, type of noise, and other parameters – as close as possible to the real problem we are trying to solve.

The solution of practically any signal processing problem can be improved by tuning the parameters of the algorithm, provided we have a proper criterion for optimality. There are always adaptation time constants, thresholds, which cannot be estimated and their values have to be adjusted experimentally. The mathematical models and solutions we use are usually

optimal in one or another way. If they reflect properly the nature of the process they model, then we have a good solution and the results are satisfactory. In all cases it is important to remember that we do not want a “minimum mean-square error solution,” or a “maximum-likelihood solution,” or even a “log minimum mean-square error solution.” We do not want to improve the signal-to-noise ratio. What we want is for listeners to perceive the sound quality of the processed signal as better – improved – compared to the input signal. From this perspective, the final judge of how good is an algorithm is the human ear, so use it to verify the solution. Hearing is an important sense for humans and animals. In many places in this book are provided examples of how humans and animals hear and localize sounds – this explains better some signal processing approaches and brings biology-inspired designs for sound capture and processing systems.

In many cases the signal processing chain consists of several algorithms for sound capture and speech enhancement. The practice shows us that a sequence of separately optimized algorithms usually provides suboptimal results. Tuning and optimization of the designed sound capturing system end-to-end is a must if we want to achieve best results.

For further information please visit http://www.wiley.com/go/tashev_sound

Ivan Tashev
Redmond, WA
USA

Acknowledgements

I want to thank the Book Program in MathWorks and especially Dee Savageau, Naomi Fernandes, and Meg Vulliez for the help and responsiveness. The MATLAB[®] scripts, part of this book, were tested with MATLAB[®] R2007a, provided as part of this program.

I am grateful to my colleagues from Microsoft Research Alex Acero, Amitav Das, Li Deng, Dinei Florencio, Cormac Herley, Zicheng Liu, Mike Seltzer, and Cha Zhang. They read the chapters of this book and provided valuable feedback.

And last, but not least, I want to express my great pleasure working with the nice and helpful people from John Wiley & Sons, Ltd. During the long process from proposal, through writing, copyediting, and finalizing the book with all the details, they were always professional, understanding, and ready to suggest the right solution. I was lucky enough to work with Tiina Ruonamaa, Sarah Hinton, Sarah Tilley, and Catlin Flint – thank you all for everything you did during the process of writing this book!

Contents

About the Author	xv
Foreword	xvii
Preface	xix
Acknowledgements	xxi
1 Introduction	1
1.1 The Need for, and Consumers of, Sound Capture and Audio Processing Algorithms	1
1.2 Typical Sound Capture System	2
1.3 The Goal of this Book and its Target Audience	3
1.4 Prerequisites	4
1.5 Book Structure	4
1.6 Exercises	5
2 Basics	7
2.1 Noise: Definition, Modeling, Properties	7
2.1.1 Statistical Properties	7
2.1.2 Spectral Properties	9
2.1.3 Temporal Properties	11
2.1.4 Spatial Characteristics	11
2.2 Signal: Definition, Modeling, Properties	12
2.2.1 Statistical Properties	13
2.2.2 Spectral Properties	16
2.2.3 Temporal Properties	17
2.2.4 Spatial Characteristics	18
2.3 Classification: Suppression, Cancellation, Enhancement	19
2.3.1 Noise Suppression	19
2.3.2 Noise Cancellation	20
2.3.3 Active Noise Cancellation	20
2.3.4 De-reverberation	21
2.3.5 Speech Enhancement	21
2.3.6 Acoustic Echo Reduction	21
2.4 Sampling and Quantization	23
2.4.1 Sampling Process and Sampling Theorem	23

2.4.2	Quantization	25
2.4.3	Signal Reconstruction	27
2.4.4	Errors During Real Discretization	29
2.4.4.1	Discretization with a Non-ideal Sampling Function	29
2.4.4.2	Sampling with Averaging	30
2.4.4.3	Sampling Signals with Finite Duration	31
2.5	Audio Processing in the Frequency Domain	32
2.5.1	Processing in the Frequency Domain	32
2.5.2	Properties of the Frequency Domain Representation	33
2.5.3	Discrete Fourier Transformation	35
2.5.4	Short-time Transformation, and Weighting	36
2.5.5	Overlap-add Process	37
2.5.6	Spectrogram: Time-Frequency Representation of the Signal	40
2.5.7	Other Methods for Transformation to the Frequency Domain	42
2.5.7.1	Lapped Transformations	42
2.5.7.2	Cepstral Analysis	43
2.6	Bandwidth Limiting	45
2.7	Signal-to-Noise-Ratio: Definition and Measurement	48
2.8	Subjective Quality Measurement	49
2.9	Other Methods for Quality and Enhancement Measurement	50
2.10	Summary	52
	Bibliography	53
3	Sound and Sound Capturing Devices	55
3.1	Sound and Sound Propagation	55
3.1.1	Sound as a Longitudinal Mechanical Wave	55
3.1.2	Frequency of the Sound Wave	56
3.1.3	Speed of Sound	58
3.1.4	Wavelength	60
3.1.5	Sound Wave Parameters	61
3.1.5.1	Intensity	61
3.1.5.2	Sound Pressure Level	61
3.1.5.3	Power	62
3.1.5.4	Sound Attenuation	63
3.1.6	Huygens' Principle, Diffraction, and Reflection	63
3.1.7	Doppler Effect	65
3.1.8	Weighting Curves and Measuring Sound Pressure Levels	66
3.2	Microphones	68
3.2.1	Definition	68
3.2.2	Microphone Classification by Conversion Type	69
3.3	Omnidirectional and Pressure Gradient Microphones	70
3.3.1	Pressure Microphone	70
3.3.2	Pressure-gradient Microphone	71
3.4	Parameter Definitions	73
3.4.1	Microphone Sensitivity	73
3.4.2	Microphone Noise and Output SNR	74

3.4.3	Directivity Pattern	74
3.4.4	Frequency Response	75
3.4.5	Directivity Index	75
3.4.6	Ambient Noise Suppression	77
3.4.7	Additional Electrical Parameters	77
3.4.8	Manufacturing Tolerances	78
3.5	First-order Directional Microphones	82
3.6	Noise-canceling Microphones and the Proximity Effect	84
3.7	Measurement of Microphone Parameters	87
3.7.1	Sensitivity	87
3.7.2	Directivity Pattern	87
3.7.3	Self Noise	90
3.8	Microphone Models	92
3.9	Summary	92
	Bibliography	93
4	Single-channel Noise Reduction	95
4.1	Noise Suppression as a Signal Estimation Problem	96
4.2	Suppression Rules	96
4.2.1	Noise Suppression as Gain-based Processing	96
4.2.2	Definition of A-Priori and A-Posteriori SNRs	97
4.2.3	Wiener Suppression Rule	98
4.2.4	Artifacts and Distortions	99
4.2.5	Spectral Subtraction Rule	100
4.2.6	Maximum-likelihood Suppression Rule	100
4.2.7	Ephraim and Malah Short-term MMSE Suppression Rule	102
4.2.8	Ephraim and Malah Short-term Log-MMSE Suppression Rule	103
4.2.9	More Efficient Solutions	103
4.2.10	Exploring Other Probability Distributions of the Speech Signal	105
4.2.11	Probability-based Suppression Rules	108
4.2.12	Comparison of the Suppression Rules	111
4.3	Uncertain Presence of the Speech Signal	115
4.3.1	Voice Activity Detectors	115
4.3.1.1	ROC Curves	116
4.3.1.2	Simple VAD with Dual-time-constant Integrator	118
4.3.1.3	Statistical-model-based VAD with Likelihood Ratio Test	122
4.3.1.4	VAD with Floating Threshold and Hangover Scheme with State Machine	123
4.3.2	Modified Suppression Rule	124
4.3.3	Presence Probability Estimators	126
4.4	Estimation of the Signal and Noise Parameters	126
4.4.1	Noise Models: Updating and Statistical Parameters	126
4.4.2	A-Priori SNR Estimation	127
4.5	Architecture of a Noise Suppressor	130
4.6	Optimizing the Entire System	137
4.7	Specialized Noise-reduction Systems	139

4.7.1	Adaptive Noise Cancellation	139
4.7.2	Psychoacoustic Noise Suppression	142
4.7.2.1	Human Hearing Organ	142
4.7.2.2	Loudness	143
4.7.2.3	Masking Effects	144
4.7.2.4	Perceptually Balanced Noise Suppressors	149
4.7.3	Suppression of Predictable Components	150
4.7.4	Noise Suppression Based on Speech Modeling	157
4.8	Practical Tips and Tricks for Noise Suppression	158
4.8.1	Model Initialization and Tracking	158
4.8.2	Averaging in the Frequency Domain	159
4.8.3	Limiting	159
4.8.4	Minimal Gain	159
4.8.5	Overflow and Underflow	160
4.8.6	Dealing with High Signal-to-Noise Ratios	160
4.8.7	Fast Real-time Implementation	161
4.9	Summary	161
	Bibliography	162
5	Sound Capture with Microphone Arrays	165
5.1	Definitions and Types of Microphone Array	165
5.1.1	Transducer Arrays and their Applications	165
5.1.2	Specifics of Array Processing for Audio Applications	169
5.1.3	Types of Microphone Arrays	171
5.1.3.1	Linear Microphone Arrays	171
5.1.3.2	Circular Microphone Arrays	172
5.1.3.3	Planar Microphone Arrays	173
5.1.3.4	Volumetric (3D) Microphone Arrays	173
5.1.3.5	Specialized Microphone Arrays	174
5.2	The Sound Capture Model and Beamforming	174
5.2.1	Coordinate System	174
5.2.2	Sound Propagation and Capture	176
5.2.2.1	Near-field Model	176
5.2.2.2	Far-field Model	177
5.2.3	Spatial Aliasing and Ambiguity	178
5.2.4	Spatial Correlation of the Microphone Signals	181
5.2.5	Delay-and-Sum Beamformer	182
5.2.6	Generalized Filter-and-Sum Beamformer	187
5.3	Terminology and Parameter Definitions	188
5.3.1	Terminology	188
5.3.2	Directivity Pattern and Directivity Index	190
5.3.3	Beam Width	192
5.3.4	Array Gain	193
5.3.5	Uncorrelated Noise Gain	194
5.3.6	Ambient Noise Gain	194
5.3.7	Total Noise Gain	195

5.3.8	IDOA Space Definition	195
5.3.9	Beamformer Design Goal and Constraints	197
5.4	Time-invariant Beamformers	198
5.4.1	MVDR Beamformer	198
5.4.2	More Realistic Design – Adding the Microphone Self Noise	201
5.4.3	Other Criteria for Optimality	202
5.4.4	Beam Pattern Synthesis	203
5.4.4.1	Beam Pattern Synthesis with the Cosine Function	203
5.4.4.2	Beam Pattern Synthesis with Dolph–Chebyshev Polynomials	205
5.4.4.3	Practical Use of Beam Pattern Synthesis	207
5.4.5	Beam Width Optimization	207
5.4.6	Beamformer with Direct Optimization	210
5.5	Channel Mismatch and Handling	213
5.5.1	Reasons for Channel Mismatch	213
5.5.2	How Manufacturing Tolerances Affect the Beamformer	215
5.5.3	Calibration and Self-calibration Algorithms	218
5.5.3.1	Classification of Calibration Algorithms	218
5.5.3.2	Gain Self-calibration Algorithms	219
5.5.3.3	Phase Self-calibration Algorithm	222
5.5.3.4	Self-calibration Algorithms – Practical Use	222
5.5.4	Designs Robust to Manufacturing Tolerances	223
5.5.4.1	Tolerances as Uncorrelated Noise	223
5.5.4.2	Cost Functions and Optimization Goals	224
5.5.4.3	MVDR Beamformer Robust to Manufacturing Tolerances	225
5.5.4.4	Beamformer with Direct Optimization Robust to Manufacturing Tolerances	225
5.5.4.5	Balanced Design for Handling the Manufacturing Tolerances	230
5.6	Adaptive Beamformers	231
5.6.1	MVDR and MPDR Adaptive Beamformers	231
5.6.2	LMS Adaptive Beamformers	231
5.6.2.1	Widrow Beamformer	232
5.6.2.2	Frost Beamformer	232
5.6.3	Generalized Side-lobe Canceller	233
5.6.3.1	Griffiths–Jim Beamformer	233
5.6.3.2	Robust Generalized Side-lobe Canceller	235
5.6.4	Adaptive Algorithms for Microphone Arrays – Summary	236
5.7	Microphone-array Post-processors	236
5.7.1	Multimicrophone MMSE Estimator	237
5.7.2	Post-processor Based on Power Densities Estimation	238
5.7.3	Post-processor Based on Noise-field Coherence	240
5.7.4	Spatial Suppression and Filtering in the IDOA Space	241
5.7.4.1	Spatial Noise Suppression	242
5.7.4.2	Spatial Filtering	244

5.7.4.3	Spatial Filter in Side-lobe Canceller Scheme	247
5.7.4.4	Combination with LMS Adaptive Filter	248
5.8	Specific Algorithms for Small Microphone Arrays	250
5.8.1	Linear Beamforming Using the Directivity of the Microphones	251
5.8.2	Spatial Suppressor Using Microphone Directivity	254
5.8.2.1	Time-invariant Linear Beamformers	255
5.8.2.2	Feature Extraction and Statistical Models	256
5.8.2.3	Probability Estimation and Features Fusion	258
5.8.2.4	Estimation of Optimal Time-invariant Parameters	258
5.9	Summary	260
	Bibliography	261
6	Sound Source Localization and Tracking with Microphone Arrays	263
6.1	Sound Source Localization	263
6.1.1	Goal of Sound Source Localization	263
6.1.2	Major Scenarios	264
6.1.3	Performance Limitations	266
6.1.4	How Humans and Animals Localize Sounds	266
6.1.5	Anatomy of a Sound Source Localizer	270
6.1.6	Evaluation of Sound Source Localizers	271
6.2	Sound Source Localization from a Single Frame	272
6.2.1	Methods Based on Time Delay Estimation	272
6.2.1.1	Time Delay Estimation for One Pair of Microphones	272
6.2.1.2	Combining the Pairs	278
6.2.2	Methods Based on Steered-response Power	280
6.2.2.1	Conventional Steered-response Power Algorithms	281
6.2.2.2	Weighted Steered-response Power Algorithm	281
6.2.2.3	Maximum-likelihood Algorithm	282
6.2.2.4	MUSIC Algorithm	282
6.2.2.5	Combining the Bins	284
6.2.2.6	Comparison of the Steered-response Power Algorithms	285
6.2.2.7	Particle Filters	286
6.3	Post-processing Algorithms	291
6.3.1	Purpose	291
6.3.2	Simple Clustering	294
6.3.2.1	Grouping the Measurements	294
6.3.2.2	Determining the Number of Cluster Candidates	294
6.3.2.3	Averaging the Measurements in Each Cluster Candidate	295
6.3.2.4	Reduction of the Potential Sound Sources	296
6.3.3	Localization and Tracking of Multiple Sound Sources	296
6.3.3.1	k -Means Clustering	297
6.3.3.2	Fuzzy C-means Clustering	298
6.3.3.3	Tracking the Dynamics	299
6.4	Practical Approaches and Tips	300
6.4.1	Increasing the Resolution of Time-delay Estimates	300
6.4.2	Practical Alternatives for Finding the Peaks	301