AUDIO SOURCE SEPARATION AND SPEECH ENHANCEMENT

EDITED BY
EMMANUEL VINCENT
TUOMAS VIRTANEN
SHARON GANNOT

WILEY

Audio Source Separation and Speech Enhancement

Audio Source Separation and Speech Enhancement

Edited by

Emmanuel Vincent
Inria
France

Tuomas Virtanen
Tampere University of Technology
Finland

Sharon Gannot
Bar-Ilan University
Israel

WILEY

This edition first published 2018

Registered Offices
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Editorial Office
The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Library of Congress Cataloging-in-Publication Data

Names: Vincent, Emmanuel (Research scientist), editor. | Virtanen, Tuomas, editor. | Gannot, Sharon, editor.
Title: Audio source separation and speech enhancement / edited by Emmanuel Vincent, Tuomas Virtanen, Sharon Gannot.
Description: Hoboken, NJ : John Wiley & Sons, 2018. | Includes bibliographical references and index. |
Identifiers: LCCN 2018013163 (print) | LCCN 2018021195 (ebook) | ISBN 9781119279884 (pdf) | ISBN 9781119279914 (epub) | ISBN 9781119279891 (cloth)
Subjects: LCSH: Speech processing systems. | Automatic speech recognition.
Classification: LCC TK7882.S65 (ebook) | LCC TK7882.S65 .A945 2018 (print) | DDC 006.4/54–dc23
LC record available at https://lccn.loc.gov/2018013163

Cover Design: Wiley
Cover Images: © 45RPM/iStockphoto;
© franckreporter/iStockphoto

Set in 10/12pt WarnockPro by SPi Global, Chennai, India

10 9 8 7 6 5 4 3 2 1

Contents

List of Authors

Shoko Araki
NTT Communication Science Laboratories
Japan

Roland Badeau
Institut Mines-Télécom
France

Alessio Brutti
Fondazione Bruno Kessler
Italy

Israel Cohen
Technion
Israel

Simon Doclo
Carl von Ossietzky-Universität Oldenburg
Germany

Jun Du
University of Science and Technology of China
China

Zhiyao Duan
University of Rochester
NY
USA

Cédric Févotte
CNRS
France

Sharon Gannot
Bar-Ilan University
Israel

Tian Gao
University of Science and Technology of China
China

Timo Gerkmann
Universität Hamburg
Germany

Emanuël A.P. Habets
International Audio Laboratories Erlangen
Germany

Elior Hadad
Bar-Ilan University
Israel

Hirokazu Kameoka
The University of Tokyo
Japan

Walter Kellermann
Friedrich-Alexander Universität Erlangen-Nürnberg
Germany

Zbyněk Koldovský
Technical University of Liberec
Czech Republic

Dorothea Kolossa
Ruhr-Universität Bochum
Germany

Antoine Liutkus
Inria
France

Michael I. Mandel
City University of New York
NY
USA

Erik Marchi
Technische Universität München
Germany

Shmulik Markovich-Golan
Bar-Ilan University
Israel

Daniel Marquardt
Carl von Ossietzky-Universität
Oldenburg
Germany

Rainer Martin
Ruhr-Universität Bochum
Germany

Nasser Mohammadiha
Chalmers University of Technology
Sweden

Gautham J. Mysore
Adobe Research
CA
USA

Tomohiro Nakatani
NTT Communication Science
Laboratories
Japan

Patrick A. Naylor
Imperial College London
UK

Maurizio Omologo
Fondazione Bruno Kessler
Italy

Alexey Ozerov
Technicolor
France

Bryan Pardo
Northwestern University
IL
USA

Pasi Pertilä
Tampere University of Technology
Finland

Gaël Richard
Institut Mines-Télécom
France

Hiroshi Sawada
NTT Communication Science
Laboratories
Japan

Paris Smaragdis
University of Illinois at
Urbana-Champaign
IL
USA

Piergiorgio Svaizer
Fondazione Bruno Kessler
Italy

Emmanuel Vincent
Inria
France

Tuomas Virtanen
Tampere University of Technology
Finland

Shinji Watanabe
Johns Hopkins University
MD
USA

Felix Weninger
Nuance Communications
Germany

Preface

Source separation and speech enhancement are some of the most studied technologies in audio signal processing. Their goal is to extract one or more source signals of interest from an audio recording involving several sound sources. This problem arises in many everyday situations. For instance, spoken communication is often obscured by concurrent speakers or by background noise, outdoor recordings feature a variety of environmental sounds, and most music recordings involve a group of instruments. When facing such scenes, humans are able to perceive and listen to individual sources so as to communicate with other speakers, navigate in a crowded street or memorize the melody of a song. Source separation and speech enhancement technologies aim to empower machines with similar abilities.

These technologies are already present in our lives today. Beyond "clean" single-source signals recorded with close microphones, they allow the industry to extend the applicability of speech and audio processing systems to multi-source, reverberant, noisy signals recorded with distant microphones. Some of the most striking examples include hearing aids, speech enhancement for smartphones, and distant-microphone voice command systems. Current technologies are expected to keep improving and spread to many other scenarios in the next few years.

Traditionally, *speech enhancement* has referred to the problem of segregating speech and background noise, while *source separation* has referred to the segregation of multiple speech or audio sources. Most textbooks focus on one of these problems and on one of three historical approaches, namely sensor array processing, computational auditory scene analysis, or independent component analysis. These communities now routinely borrow ideas from each other and other approaches have emerged, most notably based on deep learning.

This textbook is the first to provide a comprehensive overview of these problems and approaches by presenting their shared foundations and their differences using common language and notations. Starting with prerequisites (Part I), it proceeds with single-channel separation and enhancement (Part II), multichannel separation and enhancement (Part III), and applications and perspectives (Part IV). Each chapter provides both introductory and advanced material.

We designed this textbook for people in academia and industry with basic knowledge of signal processing and machine learning. Thanks to its comprehensiveness, we hope it will help students select a promising research track, researchers leverage the acquired cross-domain knowledge to design improved techniques, and engineers and developers choose the right technology for their application scenario. We also hope that it will be useful for practitioners from other fields (e.g., acoustics, multimedia, phonetics, musicology) willing to exploit audio source separation or speech enhancement as a pre-processing tool for their own needs.

May 2017 *Emmanuel Vincent, Tuomas Virtanen, and Sharon Gannot*

Acknowledgment

We would like to thank all the chapter authors, as well as the following people who helped with proofreading: Sebastian Braun, Yaakov Buchris, Emre Cakir, Aleksandr Diment, Dylan Fagot, Nico Gößling, Tomoki Hayashi, Jakub Janský, Ante Jukić, Václav Kautský, Martin Krawczyk-Becker, Simon Leglaive, Bochen Li, Min Ma, Paul Magron, Zhong Meng, Gaurav Naithani, Zhaoheng Ni, Aditya Arie Nugraha, Sanjeel Parekh, Robert Rehr, Lea Schönherr, Georgina Tryfou, Ziteng Wang, and Mehdi Zohourian

May 2017 *Emmanuel Vincent, Tuomas Virtanen, and Sharon Gannot*

Notations

Linear algebra

x	scalar
$\mathbf{x}$	vector
$[x_i]_i$	vector with entries x_i
$(\mathbf{x})_i$	ith entry of vector $\mathbf{x}$
$\mathbf{0}_I$	$I \times 1$ vector of zeros
$\mathbf{1}_I$	$I \times 1$ vector of ones
$\mathbf{X}$	matrix
$[x_{ij}]_{ij}$	matrix with entries x_{ij}
$(\mathbf{X})_{ij}$	(i, j)th entry of matrix $\mathbf{X}$
$\mathbf{I}_I$	$I \times I$ identity matrix
$\mathcal{X}$	tensor/array (with three or more dimensions) or set
$\{x_{ijk}\}_{ijk}$	tensor with entries x_{ijk}
$\mathrm{Diag}(\mathbf{x})$	diagonal matrix whose entries are those of vector $\mathbf{x}$
$\mathbf{X} \circ \mathbf{Y}$	entrywise product of matrices $\mathbf{X}$ and $\mathbf{Y}$
$\mathrm{tr}(\mathbf{X})$	trace of matrix $\mathbf{X}$
$\det(\mathbf{X})$	determinant of matrix $\mathbf{X}$
$\mathbf{x}^T$	transpose of vector $\mathbf{x}$
$\mathbf{x}^H$	conjugate-transpose of vector $\mathbf{x}$
x^*	conjugate of scalar x
$\Re(x)$	real part of scalar x
$\jmath$	imaginary unit

Statistics

$p(x)$	probability distribution of continuous random variable x
$p(x \mid y)$	conditional probability distribution of x given y
$P(x)$	probability value of discrete random variable x
$P(x \mid y)$	conditional probability value of x given y
$\mathbb{E}\{x\}$	expectation of random variable x
$\mathbb{E}\{x \mid y\}$	conditional expectation of x
$\mathbb{H}\{x\}$	entropy of random variable x
$\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$	real Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$
$\mathcal{N}_c(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$	complex Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$
$\widehat{x}$	estimated value of random variable x (e.g., first-order statistics)

σ_x^2	variance of random variable x
$\widehat{\sigma}_x^2$	estimated second-order statistics of random variable x
$\mathbf{\Sigma}_\mathbf{x}$	autocovariance of random vector $\mathbf{x}$
$\widehat{\mathbf{\Sigma}}_\mathbf{x}$	estimated second-order statistics of random vector $\mathbf{x}$
$\mathbf{\Sigma}_\mathbf{xy}$	covariance of random vectors $\mathbf{x}$ and $\mathbf{y}$
$\widehat{\mathbf{\Sigma}}_\mathbf{xy}$	estimated second-order statistics of random vectors $\mathbf{x}$ and $\mathbf{y}$
$C^{\text{cost}}(\boldsymbol{\theta})$	cost function to be minimized w.r.t. the vector of parameters $\boldsymbol{\theta}$
$\mathcal{M}^{\text{objective}}(\boldsymbol{\theta})$	objective function to be maximized w.r.t. the vector of parameters $\boldsymbol{\theta}$
$\mathcal{Q}(\boldsymbol{\theta}, \cdot)$	auxiliary function to be minimized or maximized, depending on the context

Common indexes

I	number of microphones or channels
i	microphone or channel index in $\{1, \dots, I\}$
J	number of sources
j	source index in $\{1, \dots, J\}$
T	number of time-domain samples
t	sample index in $\{0, \dots, T-1\}$
L	time-domain filter length
τ	tap index in $\{0, \dots, L-1\}$
N	number of time frames
n	time frame index in $\{0, \dots, N-1\}$
F	number of frequency bins
f	frequency bin index in $\{0, \dots, F-1\}$
ν_f	frequency in Hz corresponding to frequency bin f
$x(t)$	time-domain signal x
$x(n,f)$	complex-valued STFT coefficient of signal x

Signals

x_i	input signal recorded at microphone i
$\mathbf{x}$	$I \times 1$ multichannel input signal, e.g. $\mathbf{x}(t) = [x_1(t), \dots, x_I(t)]^T$
$\mathbf{X}$	matrix of input signals, e.g. $\mathbf{X} = [x_i(t)]_{it}$ or $\mathbf{X} = [x(n,f)]_{fn}$
$\lvert\mathbf{X}\rvert$	input magnitude spectrogram, i.e. $\lvert\mathbf{X}\rvert = [\lvert x(n,f)\rvert]_{fn}$
$\mathcal{X}$	tensor/array/set of input signals, e.g. $\mathcal{X} = [x_i(n,f)]_{ifn}$
s_j	point source signal
$\mathbf{s}$	$J \times 1$ vector of source signals, e.g. $\mathbf{s}(t) = [s_1(t), \dots, s_J(t)]^T$
$\mathbf{S}$	matrix of source signals, e.g. $\mathbf{S} = [s_j(t)]_{jt}$
c_{ij}	spatial image of source j as recorded on microphone i
$\mathbf{c}_j$	$I \times 1$ spatial image of source j on all microphones
$\mathcal{C}$	tensor/array/set of spatial source image signals, e.g. $\mathcal{C} = [c_{ij}(n,f)]_{ijfn}$
a_{ij}	acoustic impulse response (or transfer function) from source j to microphone i
$\mathbf{a}_j$	$I \times 1$ vector of acoustic impulse responses (or transfer functions) from source j, mixing vector

$\mathbf{A}$	$I \times J$ matrix of acoustic impulse responses (or transfer functions), mixing matrix
$\mathbf{u}$	$I \times 1$ noise signal

Filters

$\star$	convolution operator
w	single-output single-channel filter (mask), e.g. $\hat{s} = w^* x$
$\mathbf{w}$	single-output multichannel filter (beamformer), e.g. $\hat{s} = \mathbf{w}^H \mathbf{x}$
$\mathbf{W}$	multiple-output multichannel filter, e.g. $\hat{\mathbf{s}} = \mathbf{W}^H \mathbf{x}$

Nonnegative matrix factorization

$\mathbf{b}_k$	kth nonnegative basis spectrum
$\mathbf{B}$	matrix of nonnegative basis spectra
$h_k(n)$	kth activation coefficient in time frame n
$\mathbf{h}(n)$	vector of activation coefficients in time frame n
$\mathbf{H}$	matrix of activation coefficients

Deep learning

H	number of layers
h	layer index in $\{1, \dots, H\}$
K_h	number of neurons in layer h
k	neuron index in $\{1, \dots, K_h\}$
$\mathbf{Z}_h$	matrix of weights and biases in layer h
g_h	activation function in layer h
$g_{\mathcal{Z}}$	multivariate nonlinear function encoded by the full DNN

Geometry

$\mathbf{m}_i$	3D location of microphone i with respect to the array origin
$\ell_{ii'}$	distance between microphones i and i'
$\mathbf{p}_j$	3D location of source j with respect to the array origin
r_{ij}	distance between source j and microphone i
θ_j	azimuth of source j
φ_j	elevation of source j
c	speed of sound in air
$\Delta_{ii'j}$	time difference of arrival of source j between microphones i and i'