Binaural Model-Based Source Separation and Localization

Michael I. Mandel

Submitted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY
2010

(This page intentionally left blank)

(This page intentionally left blank)

Abstract

Binaural Model-Based Source Separation and Localization

Michael I. Mandel

When listening in noisy and reverberant environments, human listeners are able to focus on a particular sound of interest while ignoring interfering sounds. Computer listeners, however, can only perform highly constrained versions of this task. While automatic speech recognition systems and hearing aids work well in quiet conditions, source separation is necessary for them to be able to function in these challenging situations.

This dissertation introduces a system that separates more than two sound sources from reverberant, binaural mixtures based on the sources' locations. Each source is modelled probabilistically using information about its interaural time and level differences at every frequency, with parameters learned using an expectation maximization (EM) algorithm. The system is therefore called Model-based EM Source Separation and Localization (MESSL). This EM algorithm alternates between refining its estimates of the model parameters (location) for each source and refining its estimates of the regions of the spectrogram dominated by each source. In addition to successfully separating sources, the algorithm estimates model parameters from a mixture that have direct psychoacoustic relevance and can usually only be measured for isolated sources. One of the key features enabling this separation is a novel probabilistic localization model that can be evaluated at individual time-frequency points and over arbitrarily-shaped regions of the spectrogram.

The localization performance of the systems introduced here is comparable to that of humans in both anechoic and reverberant conditions, with a 40% lower mean absolute error than four comparable algorithms. When target and masker sources are mixed at similar levels, MESSL's separations have signal-to-distortion ratios 2.0 dB higher than four comparable separation algorithms and estimated speech quality 0.19 mean opinion score units higher. When target and masker sources are mixed anechoically at very different levels, MESSL's performance is comparable to humans', but in similar reverberant mixtures it only achieves 20–25% of human performance. While MESSL successfully rejects enough of the direct-path portion of the masking source in reverberant mixtures to improve energy-based signal-to-noise ratio results, it has difficulty rejecting enough reverberation to improve automatic speech recognition results significantly. This problem is shared by other comparable separation systems.

(This page intentionally left blank)

Contents

(This page intentionally left blank)

List of Figures

List of Tables

(This page intentionally left blank)

List of Abbreviations

ASR	Automatic Speech Recognition
BRIR	Binaural Room Impulse Response
BSS	Blind Source Separation
CASA	Computational Auditory Scene Analysis
DCT	Discrete Cosine Transform
DERTM	Direct-path, Early echoes, and Reverberation, of Target and Masker
DFT	Discrete Fourier Transform
DP	Direct-path
DUET	Degenerate Unmixing Estimation Technique
EM	Expectation Maximization
FFT	Fast Fourier Transform
GCC	Generalized Cross-Correlation
GMM	Gaussian Mixture Model
HRIR	Head-Related Impulse Response
HRTF	Head-Related Transfer Function
ICA	Independent Component Analysis
IID	Independent and Identically distributed
IIPM	Ideal Interaural Parameter Mask
ILD	Interaural Level Difference
IPD	Interaural Phase Difference
ITD	Interaural Time Difference
JND	Just-noticeable difference
KDE	Kernel Density Estimator
MAA	Minimum Audible Angle
MESSL	Model-based EM Source Separation and Localization
MMSE	Minimum Mean Squared Error
MOS	Mean Opinion Score
PDF	Probability Distribution Function
PESQ	Perceptual Evaluation of Speech Quality
PHAT	PHAse Transform
RMS	Root-Mean Square
SAR	Signal-to-Artifact Ratio

SDR	Signal-to-Distortion Ratio
SIR	Signal-to-Interferer Ratio
SNR	Signal-to-Noise Ratio
SNRI	Signal-to-Noise Ratio Improvement
SRT	Speech Reception Threshold
STFT	Short-Time Fourier Transform
TMR	Target-to-Masker Ratio
TRINICON	TRIple-N ICA for CONvolutive mixtures
WDO	Windowed-Disjoint Orthogonal(ity)

Acknowledgments

Many people have helped me along the way to this milestone. Foremost, my adviser, Dan Ellis, who has created a wonderful lab where I was given the freedom to pursue whatever research topics interested me. For that freedom I am grateful and for his indulgence in our lengthy discussions on and off these topics. I would also like to thank Barbara Shinn-Cunningham for welcoming me into her lab at Boston University for (part of) a summer. It was a refreshing experience and a good reality check to work with people who look at similar problems from a different perspective.

I've also had the good fortune to work with and for many inspiring mentors, starting when I wandered into the Music, Mind, and Machine group as a sophomore looking for a summer job. I was lucky to start working with Brian Whitman and Youngmoo Kim there, under the guidance of Barry Vercoe. That connection also led to my internships at Bose with Keith Martin, and at Google with Adam Berenzweig, and to my graduate work at Columbia. Another undergraduate position, with Erik Sudderth and Alex Ihler under the watchful supervision of Bill Freeman stimulated my interest in all things Bayesian, which I have yet to satisfy.

In the less academic aspects of graduate school, I owe thanks to the other members of LabROSA. Notably, Marios Athineos and Graham Poliner, who not only shared the lab with me, but also the swimming pool and the cubicle. Marios also deserves thanks for continually challenging my thinking about life, academic and otherwise, and for writing the stylesheet used for this document. Ron Weiss was a great teammate in this work and a great motivator to pursue it. All three of them are excellent cycling partners. I would also like to thank the other members of LabROSA who provided numerous interesting discussions, and Neeraj Kumar and Fadi Biadsy, who have always been willing to discuss anything and everything. Scott Bressler and the rest of the BU Auditory Neuroscience Lab also deserve thanks for stimulating and informative discussions.

I am grateful to my parents, Richard and Paula, for their support and encouragement throughout my many years in school. They have been excellent role models and are always eager to be engaged in my academic life. Also to my sister, Lauren, who is a great friend and the best sister I could hope for. She also deserves extra credit for drawing figure 2.3 for me.

Finally, I owe an enormous debt of gratitude to Johanna Devaney for her help, support, love, humor, proof-reading, and general enforcement of my sanity during the writing of this dissertation. I look forward to reciprocating in the near future.

In the preparation of this dissertation I have had help from Joanne Devaney, Marios Athineos, Ron Weiss, Lauren Mandel, Richard Mandel, Paula Mandel, and Mark Tobenkin for which I am most grateful.