

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)

© 2009
Michael I. Mandel
Some Rights Reserved



This work is licensed under the Creative Commons Attribution-Noncommercial-No
Derivative Works 3.0 United States License. To view a copy of this license, visit
<http://creativecommons.org/licenses/by-nc-nd/3.0/us/> or send a letter to Creative
Commons, 171 Second Street, Suite 300, San Francisco, California, 94105, USA.

(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

1	Introduction	1
1.1	Motivation	1
1.1.1	Separation and localization are useful	2
1.1.2	Separation and localization are difficult	3
1.2	Overview of MESSL	4
1.3	Contributions	6
1.4	Outline	7
2	Literature review	9
2.1	Anatomy of a binaural room impulse response	9
2.1.1	Direct-path	10
2.1.2	Early echoes	11
2.1.3	Late reverberation	12
2.1.4	Summary	13
2.2	Human separation performance	13
2.2.1	Single source in reverberation	14
2.2.2	Spatial release from masking	16
2.2.2.1	In reverberant environments	16
2.2.2.2	Based on perceived location	17
2.2.3	Streaming	18
2.2.3.1	Source segmentation using spatial cues	18
2.2.3.2	Source streaming using spatial cues	19
2.2.3.3	Informational masking	20
2.2.4	Summary	21
2.3	Human localization performance	21
2.3.1	Localization of a single source in an anechoic environment	22
2.3.2	Localization in adverse conditions	24
2.3.3	Models of human sound localization	25
2.3.4	Summary	26
2.4	Separation systems	26
2.4.1	Microphone arrays	26
2.4.2	Blind source separation	27
2.4.3	Speech models and monaural speech separation	28
2.4.4	Missing data speech recognition	29
2.4.5	Binaural CASA systems	30
2.4.6	Summary	30
2.5	Localization systems	31

2.6	Summary	32
3	Statistics of interaural parameters	33
3.1	Interaural cues	33
3.1.1	Example mixture	35
3.1.2	Model of interaural cues	37
3.1.3	Interaural noise observations	39
3.1.3.1	IPD noise	40
3.1.3.2	ILD noise	41
3.1.3.3	Additive noise	43
3.1.4	W-Disjoint orthogonality	44
3.2	Ground truth separations	45
3.2.1	Oracle masks	45
3.2.2	Ideal Interaural-parameter mask	47
3.2.3	Correlations between mask points	49
3.3	Summary	50
4	Localization	53
4.1	The generalized cross-correlation	54
4.1.1	Comparison to other approaches	56
4.1.2	Generalization across time	57
4.1.3	Interaural coherence	60
4.2	Probabilistic cross-correlation	63
4.2.1	The von Mises distribution	65
4.2.2	Maximum entropy circular distributions	66
4.2.3	Top-down cross-correlation	68
4.3	Experiments	70
4.3.1	Accuracy vs duration	71
4.3.1.1	Results	71
4.3.2	Localization in a mixture	73
4.3.2.1	Results	73
4.4	Summary	74
5	Separation	77
5.1	Background	78
5.2	A localization mixture model	80
5.2.1	Parameter estimation from mixtures	81
5.2.2	Modeling reverberation	85
5.2.3	Source modeling	86
5.2.4	Converting probabilities to masks	88
5.3	Example parameter estimates	90
5.4	Speech separation experiments	92
5.4.1	Shared experimental details	93
5.4.1.1	Data sources	94
5.4.1.2	Evaluation metrics	94
5.4.1.3	Control “algorithms”	95
5.4.2	Model complexity	95
5.4.2.1	Results	97

5.4.3	Initialization	97
5.4.3.1	Results	99
5.4.4	Model extensions	99
5.4.5	Separation comparison with other algorithms	100
5.4.5.1	Results	103
5.4.6	Localization	106
5.5	Summary	108
6	Evaluation	111
6.1	Introduction	111
6.2	Shared experimental details	113
6.2.1	Details of original experiment	113
6.2.2	Details of current experiment	114
6.2.3	Algorithmic separations	115
6.3	Automatic speech recognition vs intelligibility	116
6.3.1	ASR setup	116
6.3.2	Results	117
6.4	Six sub-signal attenuations	119
6.5	Objective measures	123
6.5.1	BSS_EVAL metrics	123
6.5.2	Signal-to-noise ratio metrics	125
6.5.3	PESQ	127
6.6	Summary	129
7	Conclusion	131
7.1	Future work	132
	Bibliography	135

(This page intentionally left blank)

List of Figures

1.1	Overview of the Model-based EM Source Separation and Localization system (MESSL).	4
2.1	The first 64 ms of two example binaural room impulse responses (BRIRs). .	10
2.2	An example utterance convolved with different parts of a binaural room impulse response.	12
2.3	The geometry of localization.	22
3.1	The causes of interaural time, phase, and level differences.	34
3.2	Example observations in the time-frequency domain at the left and right ears. .	35
3.3	Example observations of the interaural spectrogram: IPD and ILD.	36
3.4	Two-dimensional histograms of IPD and ILD of the target at 0° and the masker at 75°	39
3.5	Two-dimensional histograms of IPD and IPD residual of a source at 75°	41
3.6	Two-dimensional histograms of ILD of a source at 75°	42
3.7	Two-dimensional histograms of interaural parameters for a direct-path source at 75° azimuth with additive, speech-shaped noise.	43
3.8	Histogram of the difference in energy between two simultaneous speakers at corresponding time-frequency points.	44
3.9	Example kernel density estimates in dB of (a) target energy, (b) masker energy, (c) energy ratio between target and masker	48
3.10	Correlations between time-frequency points in the DP-oracle mask.	49
4.1	Example of cross-correlation taken over axis-aligned Gaussian regions around each time-frequency point in a spectrogram.	59
4.2	Example of coherence taken over axis-aligned Gaussian regions around each time-frequency point in a spectrogram.	61
4.3	The three different spectrograms involved in the calculation of the normalized cross-correlation at $\tau = 0$ for figure 4.2(e).	62
4.4	Example of the multimodality of cross-correlation in τ , using axis-aligned Gaussian regions around each time-frequency point in a spectrogram. . . .	63
4.5	A histogram of measured IPD residual angles in a reverberant environment. . . .	64
4.6	The histogram of figure 4.5 plotted with a von Mises distribution with matching moments.	66
4.7	The histogram of figure 4.5 plotted with successive maximum entropy approximations to it.	67
4.8	A caricature of our probabilistic model of interaural phase difference (IPD) as a function of frequency.	69

4.9	Mean absolute localization error using maximum entropy circular distributions evaluated in reverberant (right) and anechoic plus speech-shaped noise (left) conditions.	72
4.10	Mean absolute localization error for reverberant multiple-source mixtures.	74
5.1	The ILD prior captures the dependence of ILD on ITD.	86
5.2	Histograms of MESSL mask values weighted by target and masker energies	88
5.3	Sigmoid functions of equation (5.38) with different parameters.	89
5.4	Contour plot of signal-to-noise ratio improvement (in dB) as a function of α and β parameters in the sigmoidal nonlinearity.	90
5.5	Contribution of IPD and ILD to MESSL mask using frequency-dependent parameters, garbage source, and probability-to-mask warping.	91
5.6	Example ILD for two sources at 0° (upper lines) and 75° (lower lines) in reverberation.	92
5.7	Interaural phase differences (IPD) as a function of frequency for two speakers at 0° and 75° in reverberation.	93
5.8	Signal-to-distortion ratio for complexity experiment, two speakers in reverberation.	96
5.9	SDR as a function of angle for extended variants of MESSL.	100
5.10	Example masks from the algorithms compared in section 5.4.5	102
5.11	SDR as a function of angle for four conditions.	103
5.12	PESQ as a function of angle for four conditions.	105
5.13	Mean absolute localization error as a function of angle for four conditions.	107
6.1	Localization as a function of target-to-masker ratio.	115
6.2	Automatic speech recognition performance as a function of target-to-masker after applying ground truth separation masks.	118
6.3	Automatic speech recognition performance as a function of target-to-masker ratio after applying algorithmic separation masks.	119
6.4	Initial levels of the six different types of energy:	120
6.5	DERM analysis of masking separations. Masking attenuation of six different types of energy:	121
6.6	BSS_EVAL evaluation of ground truth and algorithmic masking systems as a function of target-to-masker ratio.	124
6.7	SNR evaluation of ground truth and algorithmic masking systems as a function of target-to-masker ratio.	127
6.8	SNR_{Li} evaluation of ground truth and algorithmic masking systems as a function of target-to-masker ratio.	127
6.9	PESQ evaluation of ground truth and algorithmic masking systems as a function of target-to-masker ratio.	128

List of Tables

2.1	A comparison of some of the characteristics of the three components of a binaural room impulse response.	11
3.1	The six oracle masks used in this work.	46
5.1	SDR for different model complexities, separating two speakers in reverberation.	97
5.2	SDR for MESSL with different initializations and no garbage source or ILD prior.	99
5.3	SDR for extended variants of MESSL.	101
5.4	SDR results (in dB) comparing systems in anechoic and reverberant environments with 2 and 3 speakers.	104
5.5	PESQ results comparing systems in anechoic and reverberant environments with 2 and 3 speakers.	106
5.6	Mean absolute localization error (in μs) for algorithms under comparison.	108
6.1	Comparison of two SNR metrics for extreme masks.	125

(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
DERM	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
MESL	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 Biadisy, 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.