Using Blind Source Separation and a Compact Microphone Array to Improve the Error Rate of Speech Recognition

Jeffrey Dean Hoffman
Portland State University

Let us know how access to this document benefits you.

Follow this and additional works at: https://pdxscholar.library.pdx.edu/open_access_etds

Part of the Electrical and Computer Engineering Commons

Recommended Citation
Hoffman, Jeffrey Dean, "Using Blind Source Separation and a Compact Microphone Array to Improve the Error Rate of Speech Recognition" (2016). Dissertations and Theses. Paper 3367.

10.15760/etd.5258

This Thesis is brought to you for free and open access. It has been accepted for inclusion in Dissertations and Theses by an authorized administrator of PDXScholar. For more information, please contact pdxscholar@pdx.edu.
Using Blind Source Separation and a Compact Microphone Array to Improve the Error Rate of Speech Recognition

by

Jeffrey Dean Hoffman

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science
in
Electrical and Computer Engineering

Thesis Committee:
James McNames, Chair
Y. C. Jenq
Fu Li

Portland State University
2016
Abstract

Automatic speech recognition has become a standard feature on many consumer electronics and automotive products, and the accuracy of the decoded speech has improved dramatically over time. Often, designers of these products achieve accuracy by employing microphone arrays and beamforming algorithms to reduce interference. However, beamforming microphone arrays are too large for small form factor products such as smart watches. Yet these small form factor products, which have precious little space for tactile user input (i.e. knobs, buttons and touch screens), would benefit immensely from a user interface based on reliably accurate automatic speech recognition.

This thesis proposes a solution for interference mitigation that employs blind source separation with a compact array of commercially available unidirectional microphone elements. Such an array provides adequate spatial diversity to enable blind source separation and would easily fit in a smart watch or similar small form factor product. The solution is characterized using publicly available speech audio clips recorded for the purpose of testing automatic speech recognition algorithms. The proposal is modelled in different interference environments and the efficacy of the solution is evaluated. Factors affecting the performance of the solution are identified and their influence quantified. An expectation is presented for the quality of separation as well as the resulting improvement in word error rate that can be achieved from decoding the separated speech estimate versus the mixture obtained from a single unidirectional microphone element. Finally, directions for future work are proposed, which have the potential to improve the performance of the solution thereby making it a commercially viable product.
Acknowledgments

First, I want to thank each of my instructors at Portland State University for imparting to me their knowledge. It is only by standing on their shoulders that I am equipped to support this thesis. Second, I want to thank my thesis committee, particularly my advisor Dr. James McNames. His careful review, thoughtful guidance, and high standards propelled me toward a destination far beyond my own myopic limitations. Third, I want to thank my coworkers for their steadfast support and encouragement. The process was more bearable having people in my corner cheering me on. Most of all, I want thank my wife Meg for the many sacrifices and accommodations that she made as I pursued my dream. If I have achieved anything at all useful here, a large share of the credit I owe to her.
# Table of Contents

Abstract ................................................................................................................................. i

Acknowledgments................................................................................................................. ii

List of Tables ........................................................................................................................ iv

List of Figures ....................................................................................................................... vi

Chapter 1. Introduction ........................................................................................................ 1

Chapter 2. Literature Review ............................................................................................... 8

Chapter 3. Methods ............................................................................................................. 42

Chapter 4. Results and Discussion .................................................................................... 53

Chapter 5. Conclusions ..................................................................................................... 131

References .......................................................................................................................... 136
List of Tables

Table 1. Overall improvement in SIR with the small room model.............................. 56

Table 2. ANOVA on SIR improvement by speaker using the natural gradient IVA algorithm and SSL prior with the small room model. .................................................. 64

Table 3. ANOVA on SIR improvement by speaker using the natural gradient IVA algorithm and Liang’s prior with the small room model. .................................................. 69

Table 4. ANOVA on SIR improvement by speaker using the fixed-point IVA algorithm and SSL prior with the small room model. .................................................. 74

Table 5. ANOVA on SIR improvement by speaker using the fixed-point IVA algorithm and Liang’s prior with the small room model. .................................................. 78

Table 6. ANOVA on SIR improvement by speaker using the real time IVA algorithm and SSL prior with the small room model. .................................................. 83

Table 7. ANOVA on SIR improvement by speaker using the real time IVA algorithm and Liang’s prior with the small room model. .................................................. 88

Table 8. ANOVA on SIR improvement by speaker using the auxiliary function IVA algorithm and SSL prior with the small room model. .................................................. 93

Table 9. ANOVA on SIR improvement by speaker using the auxiliary function IVA algorithm and Liang’s prior with the small room model. .................................................. 98

Table 10. Overall improvement in SDR with the small room model. .......................... 99

Table 11. Overall change in SAR with the small room model. ................................. 102

Table 12. Overall improvement in WER by algorithm using the small room model.... 106

Table 13. Overall improvement in SIR with the large room model. .......................... 110

Table 14. Overall improvement in SDR with the large room model.......................... 112

Table 15. Overall change in SAR with the large room model................................. 114

Table 16. Overall improvement in WER by algorithm using the large room model..... 118

Table 17. Overall improvement in SIR with the corner room model. ......................... 121

Table 18. Overall improvement in SDR with the corner room model.......................... 123

Table 19. Overall change in SAR with the corner room model................................. 125
Table 20. Overall improvement in WER by algorithm using the corner room model. ... 129
List of Figures

Figure 1. Diagram of the physical mixing and blind source separation processes with two speakers and two microphones. ................................................................. 3

Figure 2. Amplitude histogram of the clean speech phrase “although trees that chance to stand alone outside the groves sweep forth long curved branches producing a striking contrast to the ordinary grove form.” ................................................................. 9

Figure 3. Diagram of the frequency domain separation process for convolved mixtures. 21

Figure 4. Spectrogram of American-English phrase “although trees that chance to stand alone outside the groves sweep forth long curved branches producing a striking contrast to the ordinary grove form.” ................................................................. 22

Figure 5. Plots of (a) discrete time sampled 4.0, 4.5 and 5.0 Hz sinusoids and (b) their magnitude spectrum at 1.0 Hz resolution. ................................................................. 25

Figure 6. Plots (a) and (c) are discrete time samples of 1 second rectangular and Hann windows. Plots (b) and (d) are their magnitude spectrums at 0.5 Hz resolution. ............ 26

Figure 7. Plots of independent and spherically symmetric bivariate Laplacian distributions. The white curves show conditional probability $p_{s0s1} = 1$. ................................................................. 30

Figure 8. The effect of sparsity parameter $m$ on the $\ell_p$-norm-invariant multivariate probability density function. Plot (a) shows the SSL prior. Plot (b) shows Liang’s prior with increased sparsity (a narrower peak). Plot (c) shows I. Lee’s prior with even greater sparsity than Liang’s. ................................................................. 39

Figure 9. The effect of symmetry control parameter $p$ on the $\ell_p$-norm-invariant multivariate probability density function. Contour plot (a) shows linear symmetry. Contour plot (b) shows spherical symmetry. Contour plot (c) show cubic symmetry. ................................................................. 40

Figure 10. A slice of an image space showing a portion of an $x$-$y$ plane containing microphone $M$ and speaker $S$. The physical room lies adjacent to the origin and is bounded by thick lines. Images of the room unfold outward toward infinity in 3D space. Trajectory $p_1$ has a single reflection and $p_2$ has two reflections. Virtual trajectories of trajectories $p_1$ and $p_2$ are shown as dashed lines. ................................................................. 43

Figure 11. The gain of the PUM-3046L-R depends on the angle $\psi$ between the longitudinal axis $x$ and the vector to the source $s$. Gain is circularly symmetric around the longitudinal axis. The microphone is shown oriented with its longitudinal axis coincident with the $x$-axis and its face in the $y$-$z$ plane pointing in the positive $x$ direction........................................ 45
Figure 12. Relative gain vs. azimuth curves in the x-y plane for a four-element microphone array constructed from PUM-3046L-R directional microphones oriented at 90° angles in the x-y plane. ................................................................. 47

Figure 13. Relative gain surface of a four element microphone array constructed from PUM-3046L-R directional microphones oriented at 90° angles in the x-y plane. Relative gain along a speech source vector passing through a point on the surface and terminating at the origin where the microphones are located is indicated by both the color at that point and the length of the line segment between the point and the origin. .............................................. 48

Figure 14. Diagram of a small 3 m by 4 m room layout showing the locations of the microphone array (gray diamond), speaker of interest (red square), and interfering speakers (blue circles). The geometry models the use case of a microphone array equipped smart watch. ........................................................................................................ 55

Figure 15. Box plots of SIR improvement by algorithm using the SSL prior with the small room model. .......................................................................................................................... 57

Figure 16. Box plots of SIR improvement by algorithm using the Liang’s prior with the small room model. .......................................................................................................................... 57

Figure 17. Box plots of SIR improvement by speaker using the natural gradient IVA algorithm and SSL prior with the small room model. ................................................................. 59

Figure 18. Box plots of SIR improvement for speaker 1 by audio clip using natural gradient IVA and SSL prior with the small room model. ................................................................. 60

Figure 19. Box plots of SIR improvement for speaker 2 by audio clip using natural gradient IVA and SSL prior with the small room model. ................................................................. 60

Figure 20. Box plots of SIR improvement for speaker 3 by audio clip using natural gradient IVA and SSL prior with the small room model. ................................................................. 61

Figure 21. Box plots of SIR improvement for speaker 4 by audio clip using natural gradient IVA and SSL prior with the small room model. ................................................................. 61

Figure 22. Normal probability plot of SIR improvement by speaker using the natural gradient IVA algorithm and SSL prior with the small room model. ........................................ 62

Figure 23. Box plots of SIR improvement by speaker using the natural gradient IVA algorithm and Liang’s prior with the small room model. ................................................................. 65

Figure 24. Box plots of SIR improvement for speaker 1 by audio clip using natural gradient IVA and Liang’s prior with the small room model. ................................................................. 66

Figure 25. Box plots of SIR improvement for speaker 2 by audio clip using natural gradient IVA and Liang’s prior with the small room model. ................................................................. 66
Figure 26. Box plots of SIR improvement for speaker 3 by audio clip using natural gradient IVA and Liang’s prior with the small room model......................................................... 67

Figure 27. Box plots of SIR improvement for speaker 4 by audio clip using natural gradient IVA and Liang’s prior with the small room model......................................................... 67

Figure 28. Normal probability plot of SIR improvement by speaker using the natural gradient IVA algorithm and Liang’s prior with the small room model............................... 68

Figure 29. Box plots of SIR improvement by speaker using the fixed-point IVA algorithm and SSL prior with the small room model......................................................... 70

Figure 30. Box plots of SIR improvement for speaker 1 by audio clip using fixed-point IVA and SSL prior with the small room model......................................................... 71

Figure 31. Box plots of SIR improvement for speaker 2 by audio clip using fixed-point IVA and SSL prior with the small room model......................................................... 71

Figure 32. Box plots of SIR improvement for speaker 3 by audio clip using natural gradient IVA and SSL prior with the small room model......................................................... 72

Figure 33. Box plots of SIR improvement for speaker 4 by audio clip using fixed-point IVA and SSL prior with the small room model......................................................... 72

Figure 34. Normal probability plot of SIR improvement by speaker using the fixed-point IVA algorithm and SSL prior with the small room model............................... 73

Figure 35. Box plots of SIR improvement by speaker using the fixed-point IVA algorithm and Liang’s prior with the small room model......................................................... 74

Figure 36. Box plots of SIR improvement for speaker 1 by audio clip using fixed-point IVA and Liang’s prior with the small room model......................................................... 75

Figure 37. Box plots of SIR improvement for speaker 2 by audio clip using fixed-point IVA and Liang’s prior with the small room model......................................................... 75

Figure 38. Box plots of SIR improvement for speaker 3 by audio clip using natural gradient IVA and Liang’s prior with the small room model......................................................... 76

Figure 39. Box plots of SIR improvement for speaker 4 by audio clip using fixed-point IVA and Liang’s prior with the small room model......................................................... 76

Figure 40. Normal probability plot of SIR improvement by speaker using the fixed-point IVA algorithm and Liang’s prior with the small room model............................... 77

Figure 41. Box plots of SIR improvement by speaker using the real time IVA algorithm and SSL prior with the small room model......................................................... 79
Figure 42. Box plots of SIR improvement for speaker 1 by audio clip using real time IVA and SSL prior with the small room model ................................................................. 80

Figure 43. Box plots of SIR improvement for speaker 2 by audio clip using real time IVA and SSL prior with the small room model ................................................................. 81

Figure 44. Box plots of SIR improvement for speaker 3 by audio clip using natural gradient IVA and SSL prior with the small room model ................................................................. 81

Figure 45. Box plots of SIR improvement for speaker 4 by audio clip using real time IVA and SSL prior with the small room model ................................................................. 82

Figure 46. Normal probability plot of SIR improvement by speaker using the real time IVA algorithm and SSL prior with the small room model ................................................................. 82

Figure 47. Box plots of SIR improvement by speaker using the real time IVA algorithm and Liang’s prior with the small room model ................................................................. 84

Figure 48. Box plots of SIR improvement for speaker 1 by audio clip using real time IVA and Liang’s prior with the small room model ................................................................. 85

Figure 49. Box plots of SIR improvement for speaker 2 by audio clip using real time IVA and Liang’s prior with the small room model ................................................................. 86

Figure 50. Box plots of SIR improvement for speaker 3 by audio clip using natural gradient IVA and Liang’s prior with the small room model ................................................................. 86

Figure 51. Box plots of SIR improvement for speaker 4 by audio clip using real time IVA and Liang’s prior with the small room model ................................................................. 87

Figure 52. Normal probability plot of SIR improvement by speaker using the real time IVA algorithm and Liang’s prior with the small room model ................................................................. 88

Figure 53. Box plots of SIR improvement by speaker using the auxiliary function IVA algorithm and SSL prior with the small room model ................................................................. 89

Figure 54. Box plots of SIR improvement for speaker 1 by audio clip using auxiliary function IVA and SSL prior with the small room model ................................................................. 90

Figure 55. Box plots of SIR improvement for speaker 2 by audio clip using auxiliary function IVA and SSL prior with the small room model ................................................................. 91

Figure 56. Box plots of SIR improvement for speaker 3 by audio clip using natural gradient IVA and SSL prior with the small room model ................................................................. 91

Figure 57. Box plots of SIR improvement for speaker 4 by audio clip using auxiliary function IVA and SSL prior with the small room model ................................................................. 92
Figure 58. Normal probability plot of SIR improvement by speaker using the auxiliary function IVA algorithm and SSL prior with the small room model. ........................................ 93

Figure 59. Box plots of SIR improvement by speaker using the auxiliary function IVA algorithm and Liang’s prior with the small room model. ........................................ 94

Figure 60. Box plots of SIR improvement for speaker 1 by audio clip using auxiliary function IVA and Liang’s prior with the small room model. ........................................ 95

Figure 61. Box plots of SIR improvement for speaker 2 by audio clip using auxiliary function IVA and Liang’s prior with the small room model. ........................................ 96

Figure 62. Box plots of SIR improvement for speaker 3 by audio clip using auxiliary function IVA and Liang’s prior with the small room model. ........................................ 96

Figure 63. Box plots of SIR improvement for speaker 4 by audio clip using auxiliary function IVA and Liang’s prior with the small room model. ........................................ 97

Figure 64. Normal probability plot of SIR improvement by speaker using the auxiliary function IVA algorithm and Liang’s prior with the small room model. ................. 98

Figure 65. Box plots of SDR improvement by algorithm using the SSL prior with the small room model. ......................................................................................... 100

Figure 66. Box plots of SDR improvement by algorithm using the Liang’s prior with the small room model. ......................................................................................... 101

Figure 67. Box plots of the change in SAR by algorithm using the SSL prior with the small room model. ......................................................................................... 103

Figure 68. Box plots of the change in SAR by algorithm using the Liang’s prior with the small room model. ......................................................................................... 103

Figure 69. WER of decoded clean speech, microphone mixture, and separated source estimates using the small room model. ................................................................. 104

Figure 70. Box plots of WER improvement by algorithm using the SSL prior with the small room model. ......................................................................................... 105

Figure 71. Box plots of WER improvement by algorithm using Liang’s prior with the small room model. ......................................................................................... 107

Figure 72. Diagram of a large 5 m by 4 m room layout showing the locations of the microphone array (gray diamond), speaker of interest (red square), and interfering speakers (blue circles). The geometry models the use case of a microphone array equipped smart watch. ......................................................................................... 109
Figure 73. Box plots of SIR improvement by algorithm using the SSL prior with the large room model. ................................................................. 111

Figure 74. Box plots of SIR improvement by algorithm using the Liang’s prior with the large room model. ................................................................. 111

Figure 75. Box plots of SDR improvement by algorithm using the SSL prior with the large room model. ................................................................. 113

Figure 76. Box plots of SDR improvement by algorithm using the Liang’s prior with the large room model. ................................................................. 113

Figure 77. Box plots of SAR improvement by algorithm using the SSL prior with the large room model. ................................................................. 115

Figure 78. Box plots of SAR improvement by algorithm using the Liang’s prior with the large room model. ................................................................. 115

Figure 79. WER of decoded clean speech, microphone mixture, and separated source estimates using the large room model. ................................................................. 117

Figure 80. Box plots of WER improvement by algorithm using the SSL prior with the large room model. ................................................................. 119

Figure 81. Box plots of WER improvement by algorithm using Liang’s prior with the large room model. ................................................................. 119

Figure 82. Diagram of a large 5 m by 4 m room layout showing the locations of the microphone array (gray diamond), speaker of interest (red square), and interfering speakers (blue circles). The geometry models the use case of a microphone array equipped smart watch located at one corner of the room. ................................................................. 120

Figure 83. Box plots of SIR improvement by algorithm using the SSL prior with the corner room model. ................................................................. 122

Figure 84. Box plots of SIR improvement by algorithm using the Liang’s prior with the corner room model. ................................................................. 122

Figure 85. Box plots of SDR improvement by algorithm using the SSL prior with the corner room model. ................................................................. 124

Figure 86. Box plots of SDR improvement by algorithm using the Liang’s prior with the corner room model. ................................................................. 124

Figure 87. Box plots of SAR improvement by algorithm using the SSL prior with the corner room model. ................................................................. 126
Figure 88. Box plots of SAR improvement by algorithm using the Liang’s prior with the corner room model.

Figure 89. WER of decoded clean speech, microphone mixture, and separated source estimates using the corner room model.

Figure 90. Box plots of WER improvement by algorithm using the SSL prior with the corner room model.

Figure 91. Box plots of WER improvement by algorithm using Liang’s prior with the corner room model.

Figure 92. Spectrograms of clean speech, the mixture at the microphone element pointed at the speaker of interest, and the separated source estimated produced by AFIVA.
Chapter 1. Introduction

Automatic speech recognition (ASR) has become a standard feature on many consumer electronics products such as personal computers, laptops, tablets, smartphones and automotive infotainment systems. ASR technology is mature and deemed reliable enough even for use in military applications such as the F-16 and F-35 fighter aircraft [1], [2]. However, ASR systems have difficulty dealing with acoustic interference [3]. This phenomenon is known as the “cocktail party problem.” Speech that is perfectly understood in a quiet environment is rendered unintelligible in an environment with many interfering sources such as a cocktail party. This problem diminishes accuracy of speech recognition in humans and ASR equipped devices alike.

1.1 Background

Microphone arrays have been used extensively to mitigate impairment due to interference and noise. These are used along with signal processing algorithms that fall primarily into one of two categories: beamforming or blind source separation (BSS). Great results have been obtained using beamforming when target and interferer locations are known [4], [5]. A number of researchers have combined beamforming with BSS and obtained great results even when locations are unknown [6]-[8]. However, beamforming has a limitation that is problematic for small form factor devices such as smart watches and fitness bands. The speed of sound through air is roughly 340 m/s, and the bandwidth of speech is roughly 3400 Hz. For ½ wavelength phase shift at the upper end of the speech spectrum, the microphone array must have an aperture of at least 5 cm. For beamforming at the lower
end of the speech spectrum, a much larger aperture is required. Small form factor devices such as smart watches and fitness bands cannot support a microphone array of this size.

Yet, small form factor devices could benefit immensely from ASR. Surface area for traditional tactile user interfaces (i.e. touch screen display, buttons and knobs) is extremely limited on these devices. With the current state of the art in ASR, the user could issue nearly unlimited commands or dictate text and email messages of arbitrary content hands free. However, it is unlikely that the user will be content with ASR that only functions robustly in a quiet environment. Since the physical dimensions of these devices do not support the microphone array requirements of beamforming, a BSS alternative is an attractive option for improving the reliability of ASR.

While BSS implementations often result in beamforming, beamforming is not necessarily required. Spatial diversity is a sufficient condition for separation of statistically independent sources. Spatial diversity in this context means that the speech signals reaching different microphones from the same source arrive with different amplitudes in the case of instantaneous mixing or different spectrums in the case of convolutive mixing.

A diagram of the physical mixing and blind source separation processes for a two-speaker two-microphone system is shown in Figure 1.
Figure 1. Diagram of the physical mixing and blind source separation processes with two speakers and two microphones.

In the simplest (although not very realistic) case, speakers \( s_j \) are instantaneously mixed through unequal gains \( a_{ij} \) into microphones \( x_i \). Using linear algebra, we write this as\(^1\)

\[
\mathbf{x} = \mathbf{A}\mathbf{s}
\]  

(1)

which is equivalent to

\[
x_i = \sum_{j=1}^{n} a_{ij} s_j, \quad i = \{1, \ldots, m\}, \quad n \leq m
\]  

(2)

where \( \mathbf{x} \) is an \( m \times 1 \) vector, \( \mathbf{A} \) is an \( m \times n \) matrix, \( \mathbf{s} \) is an \( n \times 1 \) vector and in this simple case \( m = n = 2 \).

The goal of blind source separation in this simplest of cases is to estimate gains \( w_{ij} \) that form the demixing matrix \( \mathbf{W} \), which is the inverse of the mixing matrix \( \mathbf{A} \). Demixing

\(^1\) In mathematical equations throughout this document, vectors and functions returning vectors are non-italicized bold lowercase, matrices and functions returning matrices are non-italicized bold uppercase, scalars and functions returning scalars are italicized non-bold, with the exception that well-known predefined functions returning scalars such as \( \log(\cdot) \) and \( \exp(\cdot) \) are non-italicized non-bold lower case.
matrix $\mathbf{W}$ can then be used along with microphones $x_i$ to produce estimates $u_j$ of speakers $s_j$ that are optimal in some sense (e.g., least squares) without knowing the value of the mixing matrix $\mathbf{A}$. Using linear algebra, we write the final step as

$$\mathbf{u} = \mathbf{Wx} = \hat{\mathbf{A}}^{-1}\mathbf{x} = \hat{s}$$

which is equivalent to

$$u_j = \sum_{i=1}^{m} w_{ji} x_i, \; j = \{1, ..., n\}, \; n \leq m$$

where $\mathbf{u}$ is an $n \times 1$ vector and $\mathbf{W}$ is the $n \times m$ matrix that is an estimate of the inverse of $\mathbf{A}$. For a rectangular matrix $\mathbf{A}$ of rank $m$, we define its inverse $\mathbf{A}^{-1}$ as the matrix satisfying

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$$

and the estimate of its inverse

$$\mathbf{W} = \hat{\mathbf{A}}^{-1}$$

In our simple case, the $2 \times 2$ matrix $\mathbf{A}$ is invertible if and only if it is non-singular (i.e. its determinant must not equal zero). The determinant of $\mathbf{A}$ is

$$\det(\mathbf{A}) = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{21}a_{12}$$

Referring back to Figure 1 and equation (7), we see that if the gains $a_{11}$ and $a_{21}$ from speaker $s_1$ to microphones $x_1$ and $x_2$ are equal and the gains $a_{12}$ and $a_{22}$ from speaker $s_2$ to microphones $x_1$ and $x_2$ are also equal, then the determinant of $\mathbf{A}$ is zero and $\mathbf{A}$ is not invertible. Thus spatial diversity is a necessary condition for separation. In practice, problems can occur even if the matrix is invertible, but ill-conditioned. With an ill-conditioned matrix, very small changes in $\mathbf{x}$ produce large changes in $\mathbf{u}$. 


Finally, in the linear system of equation (1), we can only solve for the unknown quantities $s_j$ if the number of known quantities $x_i$ are equal to or greater than the number of $s_j$. This is a fundamental property of systems of linear equations. Hence, the limitation $n \leq m$.

In a more realistic case, there are multiple paths for audio pressure waves to reach a particular microphone. This is referred to as a reverberant environment. As the pressure waves from an audio source bounce off solid objects like a floor, ceiling or walls, they arrive at a microphone at different times and strengths depending on the absorption of the various objects and the trajectory. This can be modelled in discrete time by replacing the coefficients of mixing matrix $A$ with IIR filters. Mixing is now a convolutive process where multiplication in the instantaneous case is replaced by convolution in the reverberant case. We write this mathematically as

$$x_i[t] = \sum_{j=1}^{n} \sum_{\tau=0}^{\infty} a_{ij}[\tau] s_j[t-\tau], \quad i = \{1, ..., m\}, \quad n \leq m \quad (8)$$

In practice, it is not necessary to implement an infinite length filter because the strength of the arriving pressure wave(s) diminishes rapidly as the trajectory increases in length. In order to reduce complexity, IIR filters may be replaced by FIR filters without significant loss of fidelity. Methods for predicting reverberation time based on the physical characteristics of the environment were published by Lehmann and Johansson [9] and are frequently cited in the literature.

Separation requires the estimation of the demixing matrix $W$ consisting of the inverse of the matrix of filters in $A$. Once estimated, demixing matrix $W$ is convolved with
microphone outputs $x_i$ to produce discrete time estimates $u_j$ of sources $s_j$. We write this mathematically as

$$u_j = \sum_{i=1}^{m} \sum_{\tau=0}^{\infty} w_{ji}[\tau] x_i[t - \tau], \quad j = \{1, ..., n\}, \quad n \leq m$$

(9)

While this problem is much more complex than the instantaneous case with many more parameters to be estimated, we shall see in Chapter 2 that the same fundamental techniques can be applied.

In my search of the literature on BSS for speech signals, all of the solutions relied on arrays of omnidirectional directional microphones with sufficient spacing for spatial diversity. An array of closely spaced omnidirectional microphones that would fit in a smart watch or fitness band would result in strongly correlated signals emanating from all microphone elements in the array. The mixing matrix would be nearly singular and therefore ill-conditioned.

Fortunately, unidirectional microphones as small as $6.0 \times 3.2$ mm and costing as little as $1.17$ in quantity are commercially available today. Smaller models down to $2.56 \times 2.74$ mm and costing less than $20$ are also commercially available. These microphones can be collocated in a way that provides spatial diversity by aiming them in different directions. Increased demand due to the improved user experience that robust ASR would bring to smart watches and fitness bands would result in economies of scale. Because the trend in electronic components is almost universally toward smaller and cheaper with economies of scale, it is reasonable to expect reductions in cost of the smaller unidirectional microphones.
1.2 Contribution of This Work

The small form factors achievable with directional microphone arrays make them suitable for integration into wearables like smart watches and fitness bands. Wearables that would benefit immensely from ASR typically connect wirelessly to gateway devices such as smartphones, tablets, laptops and PCs where ASR software is commonplace. The primary contribution of this work is to characterize the ASR performance improvement that can be expected by combining state of the art BSS algorithms with a directional microphone array constructed from standard off the shelf components.

The remainder of this thesis is organized as follows: Chapter 2 provides a review of the BSS literature that is relevant to speech signals. Chapter 3 presents the methods used to model the proposed end-to-end ASR system and evaluate its performance. Chapter 4 discusses the results of the experiments conducted. Chapter 5 summarizes the work done, draws conclusions on the efficacy of the solution, discusses factors affecting performance, and suggests directions for future work.
Chapter 2. Literature Review

The body of literature on BSS for speech signals is rich and varied. This review focuses on the literature that is relevant to the state of the art BSS algorithms that are still in use today. It presents the fundamentals of BSS for speech signals beginning with the problem of instantaneous mixtures described in equation (2) where many of the key principals are established. It then proceeds through the major developments that ultimately lead to robust solutions to the problem of convolutive mixtures described in equation (8).

2.1 Independent Component Analysis (ICA)

Independent component analysis was first published by Herault and Jutten in July 1991 [10]. It is an improvement over principal component analysis (PCA) for the purpose of source separation when the sources are non-Gaussian. PCA transforms a set of correlated random variables to a set of uncorrelated random variables using an orthogonal transformation such as eigendecomposition. However, lack of correlation does not guarantee independence. ICA maximizes independence by attempting to decompose a multivariate random signal into non-Gaussian independent components. Speech signals are known to have a non-Gaussian distribution [11]. An example of the amplitude histogram of a speech signal is shown in Figure 2. It has a much narrower peak and lower shoulders than the Gaussian distribution overlaid in red having the same mean and standard deviation. Because speech signals are non-Gaussian, they are well suited for separation using ICA. Indeed, the literature on separation of speech signals using some form of ICA is rich [12]-[40].
Figure 2. Amplitude histogram of the clean speech phrase “although trees that chance to stand alone outside the groves sweep forth long curved branches producing a striking contrast to the ordinary grove form.”

2.1.1 Preprocessing

In order to improve the performance of ICA, the incoming mixed data is often preprocessed by centering and spatial whitening [12], [16], [19], [22], [24], [25], [36], [37], [40]. Measurements such as the covariance of random variables $x_1$ and $x_2$ require knowledge of their expected values.

$$\sigma(x_1, x_2) = \mathbb{E}((x_1 - \mathbb{E}(x_1))(x_2 - \mathbb{E}(x_2)))$$  \hspace{1cm} (10)

where $\mathbb{E}\{\cdot\}$ is the expectation operator. An estimate of the expected value is the sample mean.
\[
\bar{x}_i = \frac{1}{T} \sum_{t=t_0}^{t_0+T-1} x_i[t]
\]  
(11)

where \(x_i[t]\) is the \(t^{th}\) discrete time sample of the \(i^{th}\) microphone input, \(T\) is some finite number of discrete time samples, and \(t_0\) is the starting sample index for an ensemble of data to be processed. Centering is the removal of bias (i.e. an estimate of the expected value) from the incoming microphone data.

\[
\dot{x}_i[t] = x_i[t] - \bar{x}_i, \quad t_0 \leq t < t_0 + T
\]  
(12)

Note that in order to avoid confusion with the variable \(n\), which represents the number of speech sources, the variable \(t\) is used throughout this thesis to represent both continuous time and a discrete time sample index. Likewise, the variable \(T\) is used throughout to represent both an interval of continuous time and a number of discrete time samples. Parenthesis as in \(x(t)\) denote the instantaneous value of \(x\) at time \(t\). Brackets as in \(x[t]\) denote the \(t^{th}\) sampled value of \(x\).

Whitening is a linear transformation of the zero mean vector \(\mathbf{x} = [\dot{x}_1, \dot{x}_2, ..., \dot{x}_m]\), which is an instantaneous sample of the mixed speech signal from \(m\) microphones after centering, such that its components are uncorrelated and have unit variance. In other words, the covariance matrix of the transformed data is the identity matrix. This transformation is always possible with a non-zero covariance matrix and can be done using eigendecomposition [19]. The \(m \times m\) square covariance matrix \(\mathbf{C} = \mathbf{E}\{\mathbf{x}\mathbf{x}^T\}\) can be decomposed into

\[
\mathbf{C} = \mathbf{EDE}^{-1}
\]  
(13)
where \( \mathbf{E} \) is an \( m \times m \) orthonormal matrix (i.e. its rows and columns contain orthogonal unit vectors) whose columns contain the eigenvectors of \( \mathbf{C} \), and \( \mathbf{D} \) is an \( m \times m \) diagonal matrix whose diagonal elements contain the eigenvalues of \( \mathbf{C} \). In practice, \( \mathbb{E}\{\mathbf{x}\mathbf{x}^T\} \) is unknown. However, the covariance matrix can be estimated from a sufficiently large \( T \) samples of the input microphone data. Using the sample mean from (11),

\[
\hat{c}_{ij} = \frac{1}{T} \sum_{t=t_0}^{t_0+T-1} (x_i[t] - \bar{x}_i)(x_j[t] - \bar{x}_j)
\]

(14)

\[
\hat{\mathbf{C}} = \begin{bmatrix}
\hat{c}_{11} & \cdots & \hat{c}_{1m} \\
\vdots & \ddots & \vdots \\
\hat{c}_{m1} & \cdots & \hat{c}_{mm}
\end{bmatrix}
\]

(15)

Since a covariance matrix is symmetric positive semidefinite, its eigenvectors \( \mathbf{e}_i \) form an orthogonal basis for \( \mathbf{C} \), and its eigenvalues \( d_{ii} \) are the variances in those directions.

\[
\mathbf{C} \mathbf{e}_i = d_{ii} \mathbf{e}_i
\]

(16)

There are a number of procedures to choose from for finding eigenvectors and eigenvalues (the characteristic polynomial, power method and QR algorithm are three well-known examples). To use the characteristic polynomial, we begin by rearranging (16),

\[
(d_{ii} \mathbf{I} - \mathbf{C}) \mathbf{e}_i = 0
\]

(17)

Since by definition the eigenvector \( \mathbf{e}_i \neq 0 \), the matrix \( (d_{ii} \mathbf{I} - \mathbf{C}) \) must be singular (i.e. non-invertible). Therefore, its determinant must equal zero. This determinant is known as the characteristic polynomial \( p_C(\cdot) \) of the covariance matrix \( \mathbf{C} \) and its \( m \) roots \( d_{11}, d_{22}, \cdots, d_{mm} \) are the eigenvalues of \( \mathbf{C} \).

\[
p_C(d_{ii}) = \text{det}(d_{ii} \mathbf{I} - \mathbf{C})
\]

(18)
Once eigenvectors and eigenvalues have been obtained, each input vector $\mathbf{x}$ containing one discrete time sample of the mixed speech signal from the $m$ microphones can be transformed such that its covariance matrix is the identity matrix (i.e. all components are uncorrelated and of unit variance).

$$\tilde{x} = D^{-1/2} E^T \mathbf{x}$$  \hspace{1cm} (19)

$$E = [e_1 \ e_2 \ \cdots \ e_m]$$  \hspace{1cm} (20)

$$D^{-1/2} = \text{diag} \left( \frac{1}{\sqrt{d_{11}}}, \frac{1}{\sqrt{d_{22}}}, \cdots, \frac{1}{\sqrt{d_{nn}}} \right)$$  \hspace{1cm} (21)

$$E\{\tilde{x}\tilde{x}^T\} = I$$  \hspace{1cm} (22)

In addition to whitening, the eigenvalues are often used to reduce the dimensionality of the incoming mixed data. When the number of sources $n$ is less than the number of microphones $m$, the system model of equation (1) is overdetermined. In an overdetermined system where the signal to noise ratio (SNR) is high, $n$ of the eigenvalues $d_{ii}$ will have much larger values than the remaining $m - n$. Only those components with large eigenvalues carry information (i.e. have relatively high variance). Discarding the $m - n$ components with smaller eigenvalues reduces noise. The components carrying information are referred to as the principal components and the procedure consisting of eigendecomposition followed by discarding components with small eigenvalues is referred to as principal component analysis (PCA). Indeed, PCA is often an important first step of ICA.
2.1.2 Mutual Information

While the early ICA literature dealt only with instantaneous mixing [10], [12]-[25], a number of its principal developments find frequent use to this day. One of these developments is the use of mutual information as a contrast function [12]-[14], [17], [18], [21], [22], [27], [29], [31]-[36], [38]-[40]. A contrast function measures the divergence of one probability distribution from another. In this context, divergence is a measure of distance except that it is not necessarily symmetric. In other words, given two distributions defined by probability density functions \( p(u) \) and \( q(u) \), the divergence of \( p(u) \) from \( q(u) \) is not necessarily equal to the divergence of \( q(u) \) from \( p(u) \). However, this distinction is not critical to the discussion at hand, and divergence can simply be interpreted as a measure of distance.

Mutual information is a measure of the mutual dependence of a set of random variables. It is equal to the Kullback-Leibler divergence of a product of their marginal distributions from their joint distribution.

\[
I(u_1, u_2, \ldots, u_n) = D_{KL}\left(p_u(u) \parallel \prod p_{u_i}(u_i)\right) \\
= \int p_u(u) \log \frac{p_u(u)}{\prod p_{u_i}(u_i)} \, du
\]

(23)

where \( u = [u_1, u_2, \ldots, u_n]^T \) is a multivariate random vector, \( I(\cdot) \) is mutual information, \( D_{KL}(\cdot \parallel \cdot) \) is the Kullback-Leibler divergence, \( p_u(u) \) is the probability density function defining the joint distribution of the random variables, and \( \prod p_{u_i}(u_i) \) is the product of the probability density functions defining their marginal distributions. A close inspection of the integral provides some important insights. First, we know that \( p_u(u) = \prod p_{u_i}(u_i) \) if
and only if the $u_i$ are mutually independent. Second, $\int z p_z(z) dz$ is the expectation of random variable $z$. Hence, this function provides an expectation of the log difference between the joint probability and the product of marginal probabilities. In simpler terms, this function will return zero if the expectation is for independence and a value greater than zero otherwise. The larger the value returned, the more mutually dependent the random variables $u_i$ are.

We need an objective (or cost) function that can be minimized with respect to $W$ in order to maximize the independence of the source estimates $u_i$. If we rewrite mutual information in terms of the demixing matrix $W$, we will have exactly that. Papoulis [26] gives us an important property of mutual information for the invertible linear transform $u = Wx$,

$$I(u_1, u_2, ... u_n) = \sum_i H(u_i) - H(x) - \log|\det(W)|$$

where $H(\cdot)$ is entropy. Entropy is a measure of the average amount of information contained in the signal carried by the random variable and is defined mathematically as $H(z) = E\{-\log p_z(z)\}$. Since $x$ is the independent variable, $H(x)$ is not a function of $W$ and can be discarded for minimization purposes. We therefore derive the objective function

$$J(W) = -\left(\sum_i E\{\log p_{u_i}(u_i)\}\right) - \log|\det(W)|$$

that is minimized by the natural gradient optimization algorithm to be discussed next.
2.1.3 Natural Gradient Algorithm

Another principal development found in the early ICA literature that finds frequent use in the literature to this day is the natural gradient learning algorithm [14], [17], [18], [21], [27], [33]-[35], [38], [40]. In July 1996, Amari et al. published the natural gradient learning algorithm for BSS [14]. The natural gradient learning algorithm is an iterative algorithm that can minimize a non-linear objective function with asymptotic Fischer-efficiency [18]. A Fischer-efficient estimator is one that is unbiased and has minimum possible variance [42]. Like the ordinary gradient learning algorithm, natural gradient learning works by adjusting the coefficients of the demixing matrix $W$ in the direction of their natural gradient $\Delta w_{ji}$ iteratively in small increments.

$$ W^+ = W + \eta \Delta W $$ \hspace{1cm} (26)

where $\eta$ controls the step size and may be a constant or a sequence that changes on each iteration in order to speed convergence. However, it differs from the ordinary gradient in that for the parameter space of matrices, the ordinary gradient does not represent its steepest direction of ascent, whereas the natural gradient does [18].

The natural gradient is computed by taking the partial derivative of the objective function with respect to $W$ and multiplying it by $W^T W$ (the proof is given in [18]).

$$ \Delta W = \frac{\partial J}{\partial W} W^T W = [I - \mathbb{E}\{f(u)u^T\}]W $$ \hspace{1cm} (27)

where $f(u)$ is a score function that quantifies the sensitivity of the log likelihood to the source estimate. It is the derivative of the log likelihood of the last source estimate.
The likelihood of the source estimates $p_{u_i}(u_i)$ derives from the source prior which models the probability density function of the sources $p_{s_i}(s_i)$. Simply said, $f(u)$ quantifies in log scale the slope of the source probability density function evaluated at the latest source estimate, which is a function of the parameters $w_{ji}$. Much of the early literature revolved around the choice of a source prior, and this remains an active area of research to this day, as we shall see later.

2.1.4 Fixed-Point Algorithm

Another principal development found in the early ICA literature that finds frequent use in the literature to this day is the fixed-point algorithm [16], [19], [22], [24], [25], [36], [40]. In October 1997, Hyvärinen and Oja published the fixed-point algorithm for BSS [16]. If the iteration $z_{i+1} = f(z_i), \ i = 1, 2, ..., n$ converges on the point $z_n = f(z_n)$, it is said to be a fixed-point algorithm. Newton’s method of finding a minimum or maximum of a non-linear function is an example of a fixed-point algorithm. If the function $f(z)$ is twice differentiable, the Newton iteration

$$z_{i+1} = z_i - \frac{f'(z_i)}{f''(z_i)}$$

converges on the point where $f'(z) = 0$, which is a global or local minimum or maximum of $f(\cdot)$. 

\[
 f(u) = \begin{bmatrix}
 \frac{d}{du_0} \log p_{u_0}(u_0) \\
 \frac{d}{du_1} \log p_{u_1}(u_1) \\
 \vdots \\
 \frac{d}{du_n} \log p_{u_n}(u_n)
 \end{bmatrix}
\]
In their original paper, Hyvärinen and Oja describe a fixed-point algorithm for ICA by minimization or maximization of kurtosis. Kurtosis is a measure of the “peakedness” of a probability distribution relative to the Gaussian distribution. A Gaussian distribution has a kurtosis value of zero. Distributions with narrower peaks and heavier tails than Gaussian are said to be super-Gaussian and have positive kurtosis. Distributions with wider peaks and lighter tails than Gaussian are said to be sub-Gaussian and have negative kurtosis. Since the distribution of speech signals has a very narrow peak, speech was thought to be a good candidate for ICA by maximization of kurtosis. However, a very attractive property of an estimator is its robustness against outliers, and kurtosis is sensitive to outliers. In later papers Hyvärinen discourages ICA by maximization of kurtosis for super-Gaussian distributions such as speech because it is so sensitive to outliers [19], [22], [24].

Instead, Hyvärinen defines a contrast function that is an approximation of negentropy. Negentropy is the difference between the entropy of the distribution in question and a Gaussian distribution with the same covariance matrix.

\[
J(z) = H(z_{\text{Gaussian}}) - H(z)
\]  

(30)

Because Gaussian random variables have the largest entropy among all random variables with the same variance [41], negentropy has a positive value for all non-Gaussian distributions. From the Central Limit Theorem, we know that the sum of independent random variables with identical mean and variance tends toward a Gaussian distribution regardless of their underlying distributions. Given that speech signals are known to be super-Gaussian [11], the purer the demixed estimate of speech source \( \hat{s}_j = u_j = w_j^T x \), where \( w_j^T \) is the \( j^{\text{th}} \) row of demixing matrix \( W \), the less Gaussian its distribution.
The objective function that Hyvärinen develops in [43] is an approximation of negentropy

\[ J(w_j) = [E(G(w_j^T x)) - E(G(v))]^2 \]  \hspace{1cm} (31)

where \( G(\cdot) \) is the log-likelihood function, which is based on prior knowledge of the source distribution, \( v \) is any Gaussian random variable with zero mean and unit variance, and the source estimate \( w_j^T x \) is constrained to zero mean and unit variance. This objective function \( J(\cdot) \) draws a contrast between the likelihood of the source estimate at \( w_j \) and the likelihood of a Gaussian random variable with the same mean and variance. The purer the demixed estimate, the greater that contrast will be. The task then is to maximize \( \sum_{j=1}^{n} J(w_j) \) with respect to \( w_j \) under the constraint of decorrelation (i.e. \( E((w_j^T x)(w_k^T x)) = \delta_{jk} \)). Referring back to (25), we see that the two objectives are roughly equivalent. Finding directions where mutual information is minimized is roughly equivalent to finding directions where negentropy is maximized.

Using Newton’s method with objective function (31), Hyvärinen [19], [22] derives the fixed-point iteration for the \( j^{th} \) row of the demixing matrix, which produces source estimate \( \hat{s}_j = u_j = w_j^T x \)

**Step 1:**

\[ w_j^+ = E(x g(w_j^T x)) - E(g'(w_j^T x))w_j \]

**Step 2:**

\[ w_j^+ = \frac{w_j^+}{\|w_j^+\|} \]  \hspace{1cm} (32)

where \( g(\cdot) \) is the derivative of log-likelihood \( G(\cdot) \) with respect to parameters \( w_j \) and \( g'(\cdot) \) the derivative of \( g(\cdot) \). In cases where the Newton method fails to converge, adding a step
size parameter $\mu$ less than unity can help convergence at the expense of increased iteration count [19].

Step 1: 
\[
\mathbf{w}_j^+ = \mathbf{w}_j - \mu \frac{\mathbb{E}\{xg(\mathbf{w}_j^T \mathbf{x})\} - \beta \mathbf{w}_j}{\mathbb{E}\{g'(\mathbf{w}_j^T \mathbf{x})\} - \beta}, \quad \beta = \mathbb{E}\{\mathbf{w}_j^T x g(\mathbf{w}_j^T \mathbf{x})\}
\]  
(33)

Step 2: 
\[
\mathbf{w}_j^+ = \frac{\mathbf{w}_j^+}{\|\mathbf{w}_j^+\|}
\]

Whitening of microphone data can be avoided by incorporating the covariance matrix $\mathbf{C} = \mathbb{E}\{\mathbf{x}\mathbf{x}^T\}$ into (32)

Step 1: 
\[
\mathbf{w}_j^+ = \mathbf{C}^{-1} \mathbb{E}\{xg(\mathbf{w}_j^T \mathbf{x})\} - \mathbb{E}\{g'(\mathbf{w}_j^T \mathbf{x})\} \mathbf{w}_j
\]

Step 2: 
\[
\mathbf{w}_j^+ = \frac{\mathbf{w}_j^+}{\sqrt{(\mathbf{w}_j^+)^T \mathbf{C} \mathbf{w}_j^+}}
\]  
(34)

and (33) [19]

Step 1: 
\[
\mathbf{w}_j^+ = \mathbf{w}_j - \mu \frac{\mathbb{E}\{xg(\mathbf{w}_j^T \mathbf{x})\} - \beta \mathbf{w}_j}{\mathbb{E}\{g'(\mathbf{w}_j^T \mathbf{x})\} - \beta}
\]

Step 2: 
\[
\mathbf{w}_j^+ = \frac{\mathbf{w}_j^+}{\sqrt{(\mathbf{w}_j^+)^T \mathbf{C} \mathbf{w}_j^+}}
\]  
(35)

Each row $\mathbf{w}_j^T$ of demixing matrix $\mathbf{W}$ is used to estimate one source $\mathbf{s}_j = \mathbf{u}_j = \mathbf{w}_j^T \mathbf{x}$.

In practice, a single row may be sufficient for most wearable applications where the speech source of interest is the wearer and the microphone element pointing toward to the wearer is known. In this case, once voice activity has been detected in the direction of the wearer, the $\mathbf{w}_{ji}$ element corresponding to the wearer oriented microphone element can be initialized to unity and all others to zero before starting the fixed-point algorithm for a single $\mathbf{w}_j$. 

19
However, if separation of other sources is also desired, it is necessary to decorrelate the components \( u_j \) of vector \( \mathbf{u} = \mathbf{Wx} \) after each fixed-point iteration in order to prevent different rows of demixing matrix \( \mathbf{W} \) from converging to the same value. Hyvärinen suggests three methods of doing so.

The first is a deflation scheme based on a Gram-Schmidt-like decorrelation. For each iteration of the fixed-point algorithm for \( \mathbf{w}_{p+1} \), the projections of the previously estimated \( p \) vectors are subtracted and the result renormalized.

Step 1:
\[
\mathbf{w}_{p+1} = \mathbf{w}_{p+1} - \sum_{j=1}^{p} \mathbf{w}_{p+1}^T \mathbf{C} \mathbf{w}_j \mathbf{w}_j
\]

Step 2:
\[
\mathbf{w}_{p+1} = \frac{\mathbf{w}_{p+1}}{\sqrt{\mathbf{w}_{p+1}^T \mathbf{C} \mathbf{w}_{p+1}}}
\]

The second option is to compute all \( \mathbf{w}_j \) for one fixed-point iteration then decorrelate symmetrically using the matrix square root.

\[
\mathbf{W} = (\mathbf{WCW}^T)^{-1/2} \mathbf{W}
\]

The inverse square root can be obtained using eigendecomposition of \( \mathbf{WCW}^T = \mathbf{EDE}^T \) then \( (\mathbf{WCW}^T)^{-1/2} = \mathbf{ED}^{-1/2} \mathbf{E}^T \). This is repeated after each fixed-point iteration.

The third option is to compute all \( \mathbf{w}_j \) for one fixed-point iteration then normalize demixing matrix \( \mathbf{W} \)

\[
\mathbf{W} = \frac{\mathbf{W}}{\sqrt{\| \mathbf{WCW}^T \|}}
\]

and iterate
\[ \mathbf{W}^+ = \frac{3}{2} \mathbf{W} - \frac{1}{2} \mathbf{WCW}^T \mathbf{W} \] (39)

to convergence. Note that if microphone data has been spatially whitened, \( \mathbf{C} = \mathbf{I} \).

### 2.2 ICA in the Frequency Domain

In 1998, Paris Smaragdis [27] proposed using ICA in the frequency domain to address the problem of convolutive mixing (8). Figure 3 illustrates the approach. Speech signals \( x_i(t) \) are received from \( m \) microphones. Each signal is windowed and short-time Fourier transformed (STFT) to produce complex coefficients \( X_i(\tau,f) \) at \( k \) orthogonal frequencies. The STFT is repeated at short time intervals \( (t_{\tau-1}, t_{\tau}] \) where \( \tau = \{0,1,2,\ldots,\infty\} \) to produce a spectrogram for each of the \( m \) microphones such as the one shown in Figure 4.

![Diagram of the frequency domain separation process for convolved mixtures.](image)

Figure 3. Diagram of the frequency domain separation process for convolved mixtures.
Orthogonality is a property of the discrete Fourier transform (DFT). The DFT is utilized in the STFT such that the $X_i(\tau, f)$ are orthogonal along the frequency axis.

$$\sum_{l=0}^{k-1} \left\{ X_i(\tau, f_a) e^{j2\pi f_a l} X_i^*(\tau, f_b) e^{-j2\pi f_b l} \right\} = \delta_{ab} = \begin{cases} 0 & \text{if } a \neq b \\ \frac{1}{k} |X_i(\tau, f_a)|^2 & \text{if } a = b \end{cases}$$

(40)

where $\delta_{ab}$ is the Kronecker delta and $X_i(\tau, f)$ is the spectral coefficient at frequency $f$ of the signal produced by the $i$th microphone over time interval $(t_{\tau-1}, t_{\tau}]$. Since the $X_i(\tau, f)$ are orthogonal along the frequency axis, ICA can be performed independently at each frequency along the time axis to produce an $n \times m$ unmixing matrix $W_f$ for each of the $k$ frequencies. The unmixed estimates of the $n \leq m$ sources $u(t) = \hat{s}(t)$ are obtained from separated spectrums $U_j(\tau, f)$ using the inverse short time Fourier transform (ISTFT).
2.2.1 The Short Time Fourier Transform

The short time Fourier transform gives a local spectrum of a signal whose spectrum varies with time. In continuous time, the transform is

\[ X(\tau, f) = \int_{-\infty}^{\infty} x(t) g(t - \tau) e^{-j2\pi ft} dt \]  

(41)

where \( g(\cdot) \) is a window having non-zero value for a finite interval around zero and \( X(\tau, f) \) is a complex coefficient giving the magnitude and phase of the signal \( x(t) \) at frequency \( f \) and time \( \tau \). In practice, we use uniform discrete time sampling to capture the input signal. The corresponding short time discrete Fourier transform is

\[ X(\tau, f) = \sum_{t=-\infty}^{\infty} x[t] g[t - \tau] e^{-j2\pi ft/T} \]  

(42)

where \( t \) and \( \tau \) are now integer sample indexes, \( x[t] \) is the value of the \( t \)th sample of the input signal, \( g[t - \tau] \) is the value of the window at sample \( t \), \( f \) is an integer frequency index often referred to as the frequency bin, and \( T \) is the number of non-zero values in the window.

The choice of window is an important consideration. First, in order to accurately transform the room impulse response \( h(t) \) to frequency response \( H(f) \), the window must be long enough to capture the many trajectories from speaker to microphone. Second, the window should not be so long as to blend phonemes (i.e. perceptually distinct units of sound that distinguish one word from another) and silence intervals together. The silence intervals between phonemes give speech its super-Gaussian distribution. This super-Gaussian distribution is exploited in the contrast function. Without it, there is no contrast. Figure 4 above shows the spectrogram of a 24-word American-English phrase. Phoneme
utterances are the intervals of strong harmonic content separated by short intervals of silence.

The discrete Fourier transform suffers from a phenomenon called “leakage.” This problem can be mitigated by using windows designed to shape leakage. Leakage occurs in the DFT because only frequencies that correspond to discrete frequency bins are orthogonal. Frequencies between those discrete bins project onto them (i.e. “leak” into them). Discrete frequency bins are spaced at intervals of

$$ f_R = \frac{f_S}{T} \quad (43) $$

where $f_R$ is the spacing between frequency bins (i.e. the frequency resolution of the DFT), $f_S$ is the sample rate in samples-per-second, and $T$ is the number of samples in the window.

Plots of (a) a 16-sample window of 4.0, 4.5 and 5.0 Hz sinusoids sampled at 16 samples-per-second and (b) the DFT of 4.0, 4.5 and 5.0 Hz sinusoids with 1.0 Hz resolution are shown in Figure 5. The blue 4.0 Hz sinusoid and green 5.0 Hz sinusoid transform into impulses at the ±4.0 and ±5.0 Hz frequency bins respectively. However, the red 4.5 Hz sinusoid “leaks” into all of the frequency bins.
Figure 5. Plots of (a) discrete time sampled 4.0, 4.5 and 5.0 Hz sinusoids and (b) their magnitude spectrum at 1.0 Hz resolution.

The reason for this can be seen in Figure 6. Plot (a) shows a 16 sample 1 second rectangular window and plot (b) its DFT with 0.5 Hz resolution. The Fourier transform of the rectangle window is the sinc function. Nulls can be seen at the integer frequencies, which are spaced at the reciprocal of the window length (i.e. 1 Hz) intervals. Frequencies between the nulls follow the envelope of the sinc function. Multiplication of the rectangular window by the signal is equivalent to the convolution of their transforms $G \ast X(f)$.

$$g(t)x(t) \overset{\mathcal{F}}{\Rightarrow} (G \ast X)(f) = \sum_{\xi = -\infty}^{\infty} G(\xi)X(f - \xi)$$  \hspace{1cm} (44)

Thus any discrete frequency bin $f$, includes contributions from all frequencies that are not multiples of 1 Hz (i.e. the reciprocal of the window width).
Figure 6. Plots (a) and (c) are discrete time samples of 1 second rectangular and Hann windows. Plots (b) and (d) are their magnitude spectrums at 0.5 Hz resolution.

On the other hand, Figure 6 (c) shows a 16 sample 1 second Hann window and (d) plots its DFT with 0.5 Hz resolution. The convolution of the signal spectrum with a Hann window’s frequency response results in contributions from only frequencies local to \( f \) being included in frequency bin \( f \). Thus “leakage” is not eliminated, but reshaped to provide the frequency response local to bin \( f \). Windows have been designed to provide various main lobe widths and side lobe suppressions. The Hann window (a.k.a. raised cosine window) can be described mathematically as

\[
g(t) = \frac{1}{2} \left[ 1 - \cos \left( 2\pi \frac{t}{T} \right) \right]
\]  

With -18 dB per octave side lobe roll-off, the Hann window is the most popular choice in the literature [31]-[33], [35]-[38].
2.2.2 Permutation Ambiguity

ICA in the frequency domain has advantages over time domain approaches for reverberant environments [44]. It is computationally less complex because convolutive mixtures in the time domain are transformed to instantaneous mixtures in the frequency domain. Convergence is also faster due to fewer parameters to be adjusted.

However, ICA in the frequency domain suffers from permutations in the coefficients of the demixing matrices from frequency bin to frequency bin. ICA’s estimate of the inverse of the mixing matrix is scaled and permuted

\[ W = DP\hat{A}^{-1} \]  

(46)

where \( D \) is a diagonal scaling matrix and \( P \) is a permutation matrix having a single unitary element per row and column and the remaining elements zero. Scaling can be controlled by normalizing the rows of \( W \), but permutation is arbitrary. Suppose that \( S_i(f) \) are the short time frequency coefficients of source \( i \), and \( U_j(f) \) are its estimates using ICA. There is a one-to-one mapping of \( i \) to \( j \), but \( i \) does not necessarily equal \( j \). Furthermore, because ICA is performed independently for each frequency bin \( f \), there is no guarantee that the coefficients in the \( j^{th} \) row of the demixing matrix \( W_{fa} \) produce an estimate of the same source as those in the \( j^{th} \) row of the demixing matrix \( W_{fb} \) where \( f_a \neq f_b \). In fact, it has been observed that these permutations between the rows of \( W_{fa} \) and those of \( W_{fb} \) do indeed occur. If this ambiguity in the mapping of source \( i \) to estimate \( j \) is not resolved, the ISTFT will produce meaningless results.

A number of solutions to this problem have been published with varying degrees of success [28]-[33]. The common factor in all of these solutions is that they take advantage
of dependency between frequency bins. Speech consists of a sequence of phonemes separated by intervals of silence. Figure 4 above demonstrates the dependency between frequency bins. It shows the spectrogram of a 24-word American-English phrase. Intervals of phoneme utterance show harmonics spanning the speech bandwidth. These phoneme utterance intervals are separated by intervals of silence. Therefore, if any frequency bin has energy, it is quite likely that many frequency bins have energy. Exploitation of this dependency led to the development of independent vector analysis, which the next section covers.

2.3 Independent Vector Analysis (IVA)

Solutions to the permutation problem of ICA in the frequency domain led researchers to view the independence of source spectrums in a different way. Rather than treating the random source variables $S_j(\tau, f)$ as independent over frequency, they exploited the inherent dependency of speech signals over frequency and developed a class of algorithms that came to be referred to as independent vector analysis.

2.3.1 Natural Gradient IVA Algorithm

In January of 2007, T. Kim published a natural gradient IVA algorithm [35] that utilized the spherically symmetric Laplacian (SSL) multivariate source prior model

$$p_{\tilde{S}_j}(\tilde{S}_j) \propto \exp \left( - \sqrt{ \sum_f |S_j(f)|^2 } \right)$$

(47)

where $\tilde{S}_j$ refers to the discrete frequency spectrum of the $j^{th}$ source, which is a vector. The Laplacian is a widely used source prior for separation of super-Gaussian sources such as
speech [35]. The SSL distribution more accurately models the speech spectrogram than an independent multivariate joint Laplacian because it models the dependence of sources across frequency.

The difference between an independent multivariate joint Laplacian distribution and SSL is illustrated in Figure 7. Under the independent Laplacian distribution, conditional probabilities $p(s_0|s_1)$ and $p(s_1|s_0)$ have the same peaked Laplacian shape as marginal probabilities $p(s_0)$ and $p(s_1)$ for all values of $s_0$ and $s_1$. Under the SSL distribution, conditional probabilities $p(s_0|s_1)$ and $p(s_1|s_0)$ only retain the Laplacian shape at $p(s_0|s_1 = 0)$ and $p(s_1|s_0 = 0)$. For all non-zero values of $s_0$ and $s_1$, the distribution is less sharply peaked. The white curve corresponding to the conditional probability $p(s_0|s_1 = 1)$ in each plot illustrates this point. If source spectrum coefficient $s_1$ is non-zero, then it is less likely that source spectrum coefficient $s_0$ is zero under the SSL distribution than it is under the independent distribution.
Figure 7. Plots of independent and spherically symmetric bivariate Laplacian distributions. The white curves show conditional probability $p(s_0|s_1 = 1)$. 
Kim used the SSL multivariate source prior with the natural gradient algorithm discussed in Section 2.1.3 above to maximize the independence of the estimated source spectrums $\tilde{U}_j = \hat{S}_j$. As with ICA in the frequency domain, natural gradient descent is performed at each frequency on whitened and centered microphone spectrum coefficients $X_i(f)$ using a score function derived from the SSL prior

$$\Delta W_f = [I - E\{\Phi(f, z)z^H\}]W_f, \quad z = [\tilde{U}_1, ..., \tilde{U}_n]^T$$  \hspace{1cm} (48)

$$\Phi(f, z) = \begin{bmatrix} \varphi(f, U_1(f_1), ..., U_1(f_k)) \\ \vdots \\ \varphi(f, U_j(f_1), ..., U_j(f_k)) \\ \vdots \\ \varphi(f, U_n(f_1), ..., U_n(f_k)) \end{bmatrix}$$  \hspace{1cm} (49)

$$\varphi(f, U_j(f_1), ..., U_j(f_k)) = \frac{\partial}{\partial U_j(f)} \sqrt{\sum_{\xi = f_1}^{f_k} |U_j(\xi)|^2} = \frac{U_j(f)}{\sqrt{\sum_{\xi = f_1}^{f_k} |U_j(\xi)|^2}}$$  \hspace{1cm} (50)

Kim found that his algorithm consistently outperformed two frequency domain ICA algorithms with permutation mitigation schemes based on interfrequency dependency [28] and interfrequency dependency combined with direction of arrival [33].

### 2.3.2 Fixed-point IVA Algorithm

Around the same time that Kim published his natural gradient IVA algorithm, I. Lee published a fixed-point IVA algorithm [36] using the SSL source prior. The log-likelihood function based on the SSL source prior is
\[ G(\vec{U}) = \sqrt{\sum_f |U_j(f)|^2} \]  

(51)

After deriving the non-linear score functions and substituting them into the fixed-point iteration (32) for one row of the demixing matrix assuming centered and spatially whitened microphone spectrums, one obtains

\[ w_{j,f}^+ = E \left\{ \frac{1}{\left( \sum_f |U_j(f)|^2 \right)^{1/2}} + \frac{|U_j(f)|^2}{\left( \sum_f |U_j(f)|^2 \right)^{3/2}} \right\} w_{j,f} \]

\[ \quad - E \left\{ \frac{U_j^*(f)[X_0(f), ..., X_m(f)]}{\left( \sum_f |U_j(f)|^2 \right)^{1/2}} \right\} \]

(52)

Step 2:

\[ w_{j,f}^+ = \frac{w_{j,f}^+}{\|w_{j,f}^+\|} \]

where \( w_{j,f} \) denotes the \( j^{th} \) row of the \( f^{th} \) demixing matrix \( W_f \). After computing all rows for one fixed-point iteration, the rows of the demixing matrix are decorrelated using symmetric decorrelation.

\[ W_f = (W_f W_f')^{-1/2} W_f \]  

(53)

Lee also evaluated other non-linear score functions not derived from the SSL source prior. However, later research by Liang et al. [40] showed that a score function based on a spherically symmetric sparse prior delivered the best performance. We will review this research in Section 2.3.5 below.
2.3.3 IVA in Real Time

In July of 2010, T. Kim published an IVA algorithm that could be executed in real time on a low power 24-bit DSP running at only 7.68 MHz and utilizing less than 24 kB of data memory and 36 kB of program memory [38]. This is a significant result since the type of wearable that would benefit from BSS enhanced ASR will also need to conserve power. If the wearer is only one speaker of interest, but there are many microphone elements, the power saved in transmitting a single separated speech signal rather than the mixed signals from all microphone elements, may be greater than the power required to do the separation on the wearable. This would result in a net savings in total power consumption.

The natural gradient and fixed-point algorithms discussed in Sections 2.3.1 and 2.3.2 above could be done in real time using a block-wise batch approach. This approach would require buffering an ensemble of spectrums in order to compute the expectations in (48) and (52). Buffering this ensemble of spectrums would require a large amount of memory, which is antithetical to low power. Kim chose instead an online algorithm based on the natural gradient, but with a few modifications.

First, spatial whitening is a complex algorithm requiring the computation of expectations, which we want to avoid. Kim’s online algorithm uses unwhitened microphone spectrum data.

Second, the expectation in (48) is replaced by the instantaneous scored correlation $R_f(\tau)$. 
\[ \Delta W_f(\tau) = \left( I - R_f(\tau) \right) W_f(\tau), \]
\[ R_f(\tau) = \Phi(f, z(\tau))z^H(\tau), \] 
\[ z(\tau) = [\tilde{U}_1(\tau), \ldots, \tilde{U}_n(\tau)]^T \] 

Third, the gradient in (54) goes to zero whenever \( R_f(\tau) = I \). However, Kim’s algorithm does not spatially whiten the input data and uses instantaneous estimates for \( \tilde{U}_j(\tau) \) rather than their expectations. Therefore, a large change in average amplitude of the spectral coefficients \( \tilde{U}_j(\tau) \) results in a large gradient causing the demixing matrices to diverge. This condition occurs frequently in speech due to the silence intervals. Kim’s solution is to introduce a nonholonomic constraint [21].

\[ \Delta W_f(\tau) = \left( \text{diag}\{R_f(\tau)\} - R_f(\tau) \right) W_f(\tau) \] 

(55)

This constrains \( w_{il, f} = 0 \) and prevents the rows of \( W_f \) from diverging. The constraint is nonholonomic because at any point \( W_f \) in the \( n^2 \)-dimensional space of \( W_f \), \( \Delta W_f \) is constrained in \( (n^2 - n) \) directions, but trajectories can reach any point in the entire \( n^2 \)-dimensional space.

In addition to stabilizing convergence, Kim’s constraint eliminates the need to adjust the diagonal of the demixing matrix altogether. Instead, the mixing matrix is initialized to \( I \) and the diagonal elements remain always at unity. This is justified by the fact that scaling the rows of the mixing matrix does not change the quality of separation. As shown in (46), all scalings of the rows of \( W \) are in fact equivalent with respect to separation.
Finally, the replacement of the expectation in (48) with the instantaneous scored correlation in (54) and the absence of spatial whitening results in noisy gradients, which degrade convergence. To mitigate this, Kim introduces a normalization factor $\xi_f(\tau)$ to the demixing matrix update equation.

$$W_f(\tau + 1) = W_f(\tau) + \eta \frac{\Delta W_f(\tau)}{\sqrt{\xi_f(\tau)}}$$

$$\xi_f(\tau) = \beta \xi_f(\tau - 1) + (1 - \beta) \frac{1}{m} \sum_{i=1}^{m} |X_i(\tau, f)|^2$$

$\xi_f(\tau)$ can be viewed as the low pass IIR filtered value of each spectral coefficient RMS averaged over all microphone elements. Like the learning rate $\eta$, the smoothing factor $\beta$ is adjusted empirically for fast and robust convergence.

### 2.3.4 IVA Based on the Auxiliary Function Technique

In October of 2011, Ono published an IVA algorithm [39] based on the auxiliary function technique for the minimization of the objective function $J(W)$ given in (25) with respect to $W$. The auxiliary function technique (also known as the majorization minimization technique) is a method for iteratively minimizing an objective function by introducing an auxiliary function and variables. Let the auxiliary function $Q(W, V)$ be defined such that

$$J(W) = Q(W, V_0) < Q(W, V) \forall V \neq V_0,$$

$$V_0 = \arg\min_V \{Q(W, V)\}$$

We then minimize the objective function indirectly by minimizing the auxiliary function with respect to alternately $V$ then $W$. 


Step 1: \[ \mathbf{V}^+ = \arg\min_{\mathbf{V}} \{ Q(\mathbf{W}, \mathbf{V}) \} \] (58)

Step 2: \[ \mathbf{W}^+ = \arg\min_{\mathbf{W}} \{ Q(\mathbf{W}, \mathbf{V}^+) \} \]

The monotonic decrease in \( J(\mathbf{W}) \) is guaranteed as steps 1 and 2 are repeated until \( J(\mathbf{W}) \) converges to a minimum (i.e. \( \mathbf{W} \) converges to a fixed-point).

Ono defines the auxiliary function

\[ Q(\mathbf{W}, \mathbf{V}) = \sum_{f} Q_{f}(\mathbf{W}_f, \mathbf{V}_f) \] (59)

\[ Q_{f}(\mathbf{W}_f, \mathbf{V}_f) = \frac{1}{2} \sum_{j} \mathbf{w}_{j,f} \mathbf{V}_{j,f} \mathbf{w}_{j,f}^H - \log|\det(\mathbf{W}_f)| + R \] (60)

\[ \mathbf{V}_{j,f} = \mathbb{E} \left\{ \frac{G_R(\tau_j)}{\tau_j} \mathbf{z}(f) \mathbf{z}^H(f) \right\}, \quad \mathbf{z}(f) = [X_1(f), \cdots, X_m(f)]^T \] (61)

\[ \tau_j = \| \mathbf{U}_j \|_2 = \sqrt{\sum_f |U_{j,f}(f)|^2} \] (62)

where \( \mathbf{w}_{j,f} \) is the \( j^{th} \) row of the \( f^{th} \) demixing matrix \( \mathbf{W}_f \) and \( \mathbf{V}_{j,f} \) is a weighted covariance matrix of the \( f^{th} \) vector of microphone spectral coefficients. Ono proves in [39] that his auxiliary function satisfies equation (57) when the contrast function \( G_R(\tau_j) = \tau_j \), which is based on the SSL prior of equation (47).

The update rule of step 1 in equation (58) is easily obtained by substituting (62) into (61). The update rule of step 2 in equation (58) has no closed form solution. However, a closed form solution for each row \( \mathbf{w}_{j,f} \) of the demixing matrix \( \mathbf{W}_f \) is available. If the vector of microphone spectral coefficients \( \mathbf{z}(f) = [X_1(f), \cdots, X_m(f)]^T \) is spatially whitened and the \( U_{j,f} = \mathbf{w}_{j,f} \mathbf{z}(f) \) are independent, then we can write
\[
\mathbf{w}_{j,f} \mathbf{V}_{j,f} \mathbf{w}_{j,f}^H = 1 \tag{63}
\]
\[
\mathbf{w}_{l,f} \mathbf{V}_{j,f} \mathbf{w}_{j,f}^H = 0 \quad \forall \ l \neq j \tag{64}
\]
which leads to
\[
\mathbf{W}_f \mathbf{V}_{j,f} \mathbf{w}_{j,f}^H = \begin{bmatrix}
\mathbf{w}_{1,f} \\
\vdots \\
\mathbf{w}_{j-1,f} \\
\mathbf{w}_{j,f} \\
\mathbf{w}_{j+1,f} \\
\vdots \\
\mathbf{w}_{n,f}
\end{bmatrix} \mathbf{V}_{j,f} \mathbf{w}_{j,f}^H = \mathbf{e}_j = \begin{bmatrix}
0 \\
\vdots \\
0 \\
1 \\
0 \\
\vdots \\
0
\end{bmatrix} \tag{65}
\]
where \( \mathbf{e}_j \) denotes the vector with the \( j \)th element unity and all others zero. We now obtain the update equation
\[
\mathbf{w}_{j,f}^{H+} = \left[ \mathbf{W}_f \mathbf{V}_{j,f} \right]^{-1} \mathbf{e}_j \tag{66}
\]
and finally normalize to satisfy (63)
\[
\mathbf{w}_{j,f}^+ = \frac{\mathbf{w}_{j,f}^{H+}}{\sqrt{\mathbf{w}_{j,f}^{H+} \mathbf{V}_{j,f} \mathbf{w}_{j,f}^{H+}}} \tag{67}
\]
To summarize, the step 1 update is to substitute (62) into (61) then evaluate. The step 2 update is to evaluate (66) then use (67) to normalize.

2.3.5 On the Assumption of a Spherically Symmetric Sparse Source Prior

In order to validate the assumption that the multivariate source spectrum \( \tilde{\mathbf{S}}_j = [S_j(f_1), S_j(f_2), \ldots, S_j(f_k)] \) has a spherically symmetric distribution, I. Lee evaluated the family of multivariate spherically symmetric source priors to which SSL belongs [37].
\[ p_{\tilde{S}_j}(\tilde{S}_j) \propto \exp \left( -\frac{1}{\sigma} \left( \sum_f |S_j(f)|^p \right)^{\frac{1}{pm}} \right) \] (68)

He refers to these as $l^p$-norm-invariant multivariate probability density functions. The parameter $m$ controls the sparsity of the distribution as shown in Figure 8, and the parameter $p$ controls the symmetry as shown in Figure 9. The sparser the distribution, the higher its density around zero. When $p = 2$ and $m = 1$, $p_{\tilde{S}_j}(\tilde{S}_j)$ is the SSL probability density function.

Lee evaluated these priors over a range of $p$ and $m$. He obtained a global optimum at $p = 1.9$ and $m = 7$, which validates the assumption of spherical symmetry. Later research by Liang et al. also validated the assumption of spherical symmetry [40]. However, their results showed that $p = 2$ and $m = 3$ outperformed Lee’s prior in most cases. Since Lee did not publish results at $m = 3$, it is unclear whether or not he evaluated it. Liang et al. also measured the performance of their proposed source prior when used with natural gradient, fixed-point and auxiliary function algorithms. It outperformed the original SSL prior in almost all cases.
Figure 8. The effect of sparsity parameter $m$ on the $l^p$-norm-invariant multivariate probability density function. Plot (a) shows the SSL prior. Plot (b) shows Liang’s prior.
with increased sparsity (a narrower peak). Plot (c) shows I. Lee’s prior with even greater sparsity than Liang’s.

Figure 9. The effect of symmetry control parameter $p$ on the $l_p$-norm-invariant multivariate probability density function. Contour plot (a) shows linear symmetry. Contour plot (b) shows spherical symmetry. Contour plot (c) show cubic symmetry.

2.4 Literature Review Summary

As stated in Section 1.2 above, the primary contribution of this work is to characterize the ASR performance improvement that can be expected by combining state of the art BSS algorithms with a directional microphone array constructed from standard off the shelf components. IVA is the state of the art in BSS. The natural gradient and fixed-point learning algorithms for IVA are widely used and well established. Their performance has been verified and published by independent researchers [35], [36], [40]. Therefore, they make a good baseline by which to characterize the proposed microphone array. In addition, the auxiliary function technique is a more recent development that has been verified by independent researchers to perform similarly to natural gradient and fixed-point learning algorithms [39], [40]. Finally, the real time algorithm, which is an adaptation of the natural gradient learning algorithm, affords an opportunity to characterize the proposed
microphone array with an algorithm that can be executed in real time by a low-cost and low-power DSP [38]. BSS systems based on the proposed microphone array will be characterized using all four of these algorithms along with both the SSL and Liang’s prior [37], [40]. The next chapter describes in detail the methods used to model the microphone array and characterizes the BSS systems.
Chapter 3. Methods

This chapter describes the methods used to model the reverberant room environment and the array of directional microphone elements. It also describes the clean speech signal sources and the algorithms used to separate mixtures. Finally, it describes the metrics used to evaluate the effectivity of the proposed separation algorithms and compare them quantitatively.

3.1 Reverberant Room Model

The reverberant room model relies on the image-source method first introduced in 1979 by Allen & Berkley [45] and later refined by Lehmann and Johansson [46], [47]. This method has proven effective and seen much use in the recent literature [35]-[38], [40].

A 2D slice of an image-source space showing a portion of an $x$-$y$ plane containing microphone $M$ and speaker $S$ is diagramed in Figure 10. The physical room lies adjacent to the origin and is bounded by thick lines. Images of the room unfold outward toward infinity in 3D space. In addition to the direct-trajectory from the physical-source to microphone, each image-source has a virtual-trajectory to the microphone passing through one of the physical-walls and possibly one or more image-walls. Each virtual-trajectory represents a physical-trajectory of equal length and having reflections at the same points of contact with a wall. For example, physical-trajectory $p_1$ has a single reflection at the point where its virtual-trajectory passes through the north wall. Physical-trajectory $p_2$ has two reflections at the points where its virtual-trajectory passes through the south physical-wall and the east image-wall. Physical-trajectories are shown as solid lines and virtual trajectories are shown as dashed lines.
Figure 10. A slice of an image space showing a portion of an x-y plane containing microphone M and speaker S. The physical room lies adjacent to the origin and is bounded by thick lines. Images of the room unfold outward toward infinity in 3D space. Trajectory $p_1$ has a single reflection and $p_2$ has two reflections. Virtual trajectories of trajectories $p_1$ and $p_2$ are shown as dashed lines.

If the walls are rigid, the mathematical model for the room’s impulse response is

$$h(t, s, m) = \sum_{p=1}^{8} \sum_{r=-\infty}^{\infty} \frac{1}{4\pi|\mathbf{r}_p + \mathbf{r}_r|} \delta \left[ t - \frac{|\mathbf{r}_p + \mathbf{r}_r|}{c} \right]$$

(69)

where $s = (x_s, y_s, z_s)$ is the location of the physical-source, $m = (x_m, y_m, z_m)$ is the location of the microphone, $\mathbf{r}_p = (x_s \pm x_m, y_s \pm y_m, z_s \pm z_m)$ are the eight displacement vectors indexed by $p$ from the sources in rooms adjacent to the origin (one physical-source and seven image-sources) and the microphone, $\mathbf{r}_r = 2(nx_w, ly_w, kz_w)$ are the infinite displacement vectors indexed by $r$ unfolding outward toward infinity in even multiples of
the room dimensions (i.e. \( n, l, \) and \( k \) belong to the set of all integers), and \( c \) is the speed of sound. In short, the impulse response is the sum of all trajectories from a source (physical or image) to the microphone.

In reality, walls are not always rigid. They have a coefficient of reflection. We label these coefficients \( \beta_{x1}, \beta_{x2}, \beta_{y1}, \beta_{y2}, \beta_{z1}, \) and \( \beta_{z2} \) where \( x1 \) corresponds to the south wall, \( x2 \) the north wall, \( y1 \) the west wall, \( y2 \) the east wall, \( z1 \) the floor, and \( z2 \) the ceiling. The mathematical model for the impulse response including reflection coefficients is

\[
h(t, s, m) = \sum_{p=1}^{8} \sum_{r=\infty}^{\infty} \frac{\beta_{x1}^{n-w}\beta_{x2}^{l-v}\beta_{y1}^{l-v}\beta_{y2}^{l-v}\beta_{z1}^{k-u}\beta_{z2}^{k-u}}{4\pi |r_p + r_r|} \delta \left[ t - \frac{|r_p + r_r|}{c} \right]
\]  

where we rewrite the eight displacement vectors indexed by \( p \) as \( r_p = (x_s - x_m + 2wx_m, y_s - y_m + 2vy_m, z_s - z_m + 2uz_m) \) with \( u = \{0, 1\}, v = \{0, 1\} \) and \( w = \{0, 1\} \).

In practice, it is unnecessary to carry out the second summation to infinity. Impulse energy decays with distance and number of reflections. One need only carry out the summation to a displacement \( r_r \) where \( h(t, s, m) \) is insignificant. Furthermore, for discrete time sampled \( h(\tau, s, m) \), computing frequency response \( H(f, s, m) \) allows representation of delays that are not necessarily integer multiples of the sampling period [46], [47]. Impulse response is easily obtained from frequency response using the inverse Fourier transform \( h(\tau, s, m) = \mathcal{F}^{-1}\{H(f, s, m)\} \). MATLAB® code for computing room impulse response using the image-source method can be downloaded from MATLAB® Central [48] or directly from Eric Lehman’s website [49].
3.2 Microphone Array Model

The microphone array model is based on the PUM-3046L-R directional microphone from PUI Audio Inc. This is a cylindrical microphone 3 mm in height and 6 mm in diameter that can be purchased for $1.17 in quantity from Digi-Key Electronics. Because of its small size and low cost, multiple microphones could be embedding in a small form factor wearable such as a smart watch.

The gain of the PUM-3046L-R depends on the angle between the longitudinal axis of the microphone labelled \( x \) and the vector to the source labelled \( s \) as shown in Figure 11.

![Diagram of microphone array model](image)

Figure 11. The gain of the PUM-3046L-R depends on the angle \( \psi \) between the longitudinal axis \( x \) and the vector to the source \( s \). Gain is circularly symmetric around the longitudinal axis. The microphone is shown oriented with its longitudinal axis coincident with the \( x \)-axis and its face in the \( y-z \) plane pointing in the positive \( x \) direction.

Based on an analysis of characterization data obtain from PUI Audio Inc., relative gain \( \rho \) can be approximated by the equation
\[
\rho = \begin{cases} 
\cos^2 \left( \frac{3}{5} \psi \right), & -\frac{5}{6} \pi < \psi < \frac{5}{6} \pi \\
0, & \text{otherwise}
\end{cases}
\] (71)

Using trigonometry, \( \psi \) can be expressed in terms of elevation \( \theta \) and azimuth \( \varphi \).

\[
\psi = \tan^{-1} \left( \frac{\sqrt{\cos^2(\theta) \sin^2(\varphi) + \sin^2(\theta)}}{\cos(\theta) \cos(\varphi)} \right) \quad (72)
\]

Constructing a four-element coplanar microphone array from PUM-3046L-R directional microphones oriented at 90° angles in the \( x-y \) plane results in the directivity pattern plotted in Figure 12. With one element pointing toward the speaker of interest at a distance of one foot and the other three elements pointing away and to either side in the \( x-y \) plane, the geometry is compatible with the use case of an ASR equipped smart watch. Figure 13 shows the relative gain surface in three-dimensional space. Relative microphone gain along a speech source vector passing through a point on the surface and terminating at the origin where the microphones are located is indicated by both the color at that point and the length of the line segment between the point and the origin.
Figure 12. Relative gain vs. azimuth curves in the x-y plane for a four-element microphone array constructed from PUM-3046L-R directional microphones oriented at 90° angles in the x-y plane.
Figure 13. Relative gain surface of a four element microphone array constructed from PUM-3046L-R directional microphones oriented at 90° angles in the x-y plane. Relative gain along a speech source vector passing through a point on the surface and terminating at the origin where the microphones are located is indicated by both the color at that point and the length of the line segment between the point and the origin.

3.3 BSS Algorithms

The BSS algorithms tested are the natural gradient IVA [35], fixed-point IVA [36], real-time IVA [38], and auxiliary function IVA [39] algorithms. As stated in Section 2.4 above, these represent the state of the art in BSS. The natural gradient and fixed-point learning
algorithms were selected because they are widely used and well established. They make a good baseline by which to characterize a BSS system based on the proposed microphone array. The auxiliary function algorithm was selected because it is a newer algorithm that has been independently verified to perform similarly to natural gradient and fixed-point algorithms. The real time algorithm was selected because it affords an opportunity to test the proposed microphone array with an algorithm that can be executed in real time by a low-power low-cost DSP. Each of these are tested with both the SSL and $l^p$-norm-invariant multivariate prior recommended by Liang [40]. MATLAB code for natural gradient and fixed-point IVA algorithms can be downloaded from Taesu Kim’s home page [50], [51]. The other algorithms were implemented in MATLAB based on information provided in the referenced literature.

3.4 Speech Signal Sources

The speech signals used to characterize the systems described above were selected from the VoxForge (http://www.voxforge.org) speech corpus. VoxForge is a website that was set up to collect transcribed speech for use with free and open source ASR engines such as CMU Sphinx, ISIP, Julius and HTK. Visitors to the website can upload recordings of their own speech and download the recordings of others under a GNU General Public License (GPL). Each recording comes with a transcript making it relatively easy to evaluate the error rate of a large vocabulary ASR system.

VoxForge was selected because of the large size of the corpus. In order for a large vocabulary ASR system to function well, the speakers should all speak the same language and dialect that the ASR system was trained on. American English is the language and
dialect chosen for this experiment, and VoxForge has a large number of American English speakers to choose from. In order to gain high confidence in our characterization of ASR performance improvement, a large number of sample recordings should be available from each speaker, and VoxForge has dozens from each speaker.

The selected recordings are of two male and two female American English speakers reading American classic literature. They are grouped into fours, one from each speaker, according to length from 2 to 14 seconds. All of the speakers are rotated through the speaker of interest position and the positions of the interfering speakers are randomized. Having both male and female speakers represented in both speaker of interest and interfering speaker positions characterizes sensitivity of the system to speaker sex. Randomizing the positions of the interfering speakers characterizes the sensitivity of the system to spatial orientation.

3.5 Performance Evaluation

The performance of each system model described above is evaluated in two ways. First, the improvement in source to interference ratio (SIR), source to distortion ratio (SDR), and source to artifact ratio (SAR) between the mixed speech signal produced by the microphone pointed at the speaker of interest and the separated source estimate is evaluated using a toolbox for performance measurement in (blind) source separation developed by E. Vincent [52]. Second, the improvement in word error rate (WER) between the mixed and separated speech is measured using the commercial ASR software package Nuance® Dragon Naturally Speaking.
The methods used to measure SIR, SDR, and SAR are described by Vincent et al in their conference paper [53]. In the following discussion, the vector notation $\tilde{x}_i$ is used to mean the ensemble of uniform time samples making up the signal $x_i(t)$ in order to avoid confusion with the vector $\mathbf{x}$ consisting of the $m$ microphone outputs. The mixture $x_i$ at the microphone pointed at the speaker of interest is broken down into components

$$x_i = s_{\text{target}} + e_{\text{interferers}} + e_{\text{artifacts}}$$  \hspace{1cm} (73)

where $s_{\text{target}}$ is the component of $x_i$ due to the clean speech source from the speaker of interest, $e_{\text{interferers}}$ is the sum of the error due to interfering speakers, and $e_{\text{artifacts}}$ is the error due to artifacts introduced by signal processing. These components are then used in the calculation of the performance metrics

$$\text{SIR} = 10 \log_{10} \frac{\|\tilde{s}_{\text{target}}\|^2}{\|\tilde{e}_{\text{interferers}}\|^2}$$  \hspace{1cm} (74)

$$\text{SAR} = 10 \log_{10} \frac{\|\tilde{s}_{\text{target}} + \tilde{e}_{\text{interferers}}\|^2}{\|\tilde{e}_{\text{artifacts}}\|^2}$$  \hspace{1cm} (75)

$$\text{SDR} = 10 \log_{10} \frac{\|\tilde{s}_{\text{target}}\|^2}{\|\tilde{e}_{\text{interferers}} + \tilde{e}_{\text{artifacts}}\|^2}$$  \hspace{1cm} (76)

where $\|\cdot\|^2$ is the $L^2$-norm.

$s_{\text{target}}$ is the orthogonal projection of the clean speech signal $s_j$ on $x_i$,

$$s_{\text{target}} = \tilde{x}_i^T \tilde{s}_j \frac{s_j}{\|\tilde{s}_j\|^2}$$  \hspace{1cm} (77)

$e_{\text{interferers}}$ is the sum of the orthogonal projections of the interfering speech signals $s_{j' \neq j}$ on $x_i$,  

51
\[ e_{\text{interferers}} = \sum_{j' \neq j} \tilde{x}_i^T \tilde{s}_{j'} \frac{S_{j'}}{\| \tilde{s}_{j'} \|} \]  

(78)

and \( e_{\text{artifacts}} \) is the remaining error term after subtraction of \( s_{\text{target}} \) and \( e_{\text{interferers}} \) from mixture \( x_i \).

\[ e_{\text{artifacts}} = x_i - s_{\text{target}} - e_{\text{interferers}} \]  

(79)

The calculations are then repeated replacing the mixture \( x_i \) with the separated estimate \( u_j \) and the improvements recorded.
Chapter 4. Results and Discussion

This chapter describes the experiments that are conducted using the methods described in Chapter 3 and the results thereby obtained. Three room models of two different sizes and three different spatial orientations with respect to the microphone array are used. With each of these room models, four speakers, two males and two females, simultaneously read passages from a work of American classic literature. The mixture produced by the microphone array model is separated using the natural gradient (NG), fixed-point (FP), real time (RT), and auxiliary function (AF) IVA BSS algorithms described in Section 2.3. The separated speech signals are then analyzed using the performance toolbox described in Section 3.5. Finally, the separated speech signals are processed using Nuance® Dragon Naturally Speaking, the resulting transcription is compared against the original prompt, and WER is recorded.

In each of the three room models, one speaker (the speaker of interest) talks to their microphone array equipped smart watch, which is held in front of them at a distance of 0.3 m. The locations of the other three speakers (interferers) are randomized. Each of the four speakers takes a turn in the speaker of interest position. For each speaker of interest, five repetitions are performed using the same audio clips, but with the locations of the interferers randomized. This produces 80 sets of results: four speakers by five repetitions by four algorithms. The SIR, SDR, SAR and WER of the separated speech belonging to the speaker of interest is compared to the mixture produced by the single directional microphone element pointed toward him or her. The change in SIR, SDR, SAR and WER is recorded.
The experiment described above is performed on each of twenty sets of four passages from works of American classic literature. This produces a total of 1600 sets of SIR, SDR, SAR and WER results for each of the three room models. These results are statistically analyzed to determine expectations for the quality of separation as well as the factors effecting separation. The entire experiment is conducted with both the SSL and \( l^p \)-norm-invariant multivariate prior recommended by Liang [40]. Results are presented below organized by room model.

4.1 Small Room Model Results

The model of a small conference room measuring 3 m by 4 m horizontally and 2.5 m vertically is shown in Figure 14. Absorption coefficients are set to model drywall walls, carpeted floor, and sound absorbing ceiling tiles. The speaker of interest indicated by the red square is located 1 m inward from the center of a 4 m wall. The microphone array indicated by the gray diamond is located 0.3 m inward from the speaker of interest. The five possible locations for the three interfering speakers are indicated by blue circles. With each experiment, the locations of the three interfering speakers are selected at random from the set of five possibilities. This geometry models the use case of a microphone array equipped smart watch.
Figure 14. Diagram of a small 3 m by 4 m room layout showing the locations of the microphone array (gray diamond), speaker of interest (red square), and interfering speakers (blue circles). The geometry models the use case of a microphone array equipped smart watch.

4.1.1 SIR Improvement with the Small Room Model

Table 1 shows the overall improvement in SIR between the mixture $x_i$ at the microphone pointed at the speaker of interest and the separated source estimate $u_j$ of the speaker of interest. The data are collected over 400 experiments with each algorithm. As expected, the natural gradient and fixed-point algorithms showed similar performance, which is consistent with the literature. The auxiliary function algorithm showed markedly superior performance only when used with the $l_P$-norm-invariant multivariate prior recommended by Liang [40], while the real time algorithm suffered under Liang’s prior. While showing
underwhelming mean and median SIR improvement, the real time algorithm did excel in one area: in every case SIR was improved using the real time algorithm.

Table 1. Overall improvement in SIR with the small room model.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Median (dB)</th>
<th>Mean (dB)</th>
<th>Std. (dB)</th>
<th>Improved (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NGIVA w/SSL Prior</td>
<td>7.545</td>
<td>7.504</td>
<td>4.397</td>
<td>94.75</td>
</tr>
<tr>
<td>FPIVA w/SSL Prior</td>
<td>7.980</td>
<td>7.785</td>
<td>4.152</td>
<td>94.50</td>
</tr>
<tr>
<td>RTIVA w/SSL Prior</td>
<td>2.600</td>
<td>2.734</td>
<td>0.843</td>
<td>100.00</td>
</tr>
<tr>
<td>AFIVA w/SSL Prior</td>
<td>8.100</td>
<td>7.847</td>
<td>5.294</td>
<td>94.00</td>
</tr>
<tr>
<td>NGIVA w/Liang's Prior</td>
<td>7.795</td>
<td>7.410</td>
<td>5.391</td>
<td>93.00</td>
</tr>
<tr>
<td>FPIVA w/Liang's Prior</td>
<td>8.165</td>
<td>8.093</td>
<td>4.095</td>
<td>96.25</td>
</tr>
<tr>
<td>RTIVA w/Liang's Prior</td>
<td>2.180</td>
<td>2.246</td>
<td>0.773</td>
<td>100.00</td>
</tr>
<tr>
<td>AFIVA w/Liang's Prior</td>
<td>9.595</td>
<td>9.770</td>
<td>3.865</td>
<td>99.75</td>
</tr>
</tbody>
</table>

Figure 15 shows box plots of SIR improvement by algorithm using the SSL prior with the small room model. Figure 16 shows the same box plots using Liang’s prior. The central horizontal red line is the median improvement. The blue box spans vertically the 2\textsuperscript{nd} and 3\textsuperscript{rd} quartiles. The red crosses are outliers, which are defined as being beyond $\pm 2.7$ standard deviations from the mean. The black dashed lines referred to as the whiskers extend to the limits of the data that are not considered outliers. One notable observation is that the real time algorithm, while inferior to the others in terms of performance, is much more stable in the sense of having shorter whiskers and few outliers. Another observation is that the auxiliary function algorithm was unstable at times when used with the SSL prior. Similarly, the natural gradient algorithm was unstable at times when used with Liang’s prior.
Figure 15. Box plots of SIR improvement by algorithm using the SSL prior with the small room model.

Figure 16. Box plots of SIR improvement by algorithm using the Liang’s prior with the small room model.
In the following sections, we take a closer look at SIR improvement with each combination of algorithm and prior. For each of the eight combinations listed in Table 1 above, the four speakers are rotated through the speaker of interest position, 100 experiments are run with each speaker using five repetitions of twenty audio clips, and the positions of the interferer speakers are randomized on each experiment. The collected data is first grouped by speaker then further grouped by audio clip to determine their effects on SIR improvement.

4.1.1.1 SIR Improvement with Natural Gradient Algorithm using SSL Prior

Figure 17 shows box plots of SIR improvement by speaker using the natural gradient algorithm with the SSL prior. It is evident from the box plots that both the median and interquartile range depend on the speaker.
Figure 17. Box plots of SIR improvement by speaker using the natural gradient IVA algorithm and SSL prior with the small room model.

Figure 18 though Figure 21 show box plots of SIR improvement by audio clip for each of the four speakers using the natural gradient algorithm with the SSL prior. The sample ensemble for each box plot consist of five repetitions with randomized interferer location. It is evident from the box plots that the choice of audio clip has as much or more effect on SIR improvement as the locations of the interferers.
Figure 18. Box plots of SIR improvement for speaker 1 by audio clip using natural gradient IVA and SSL prior with the small room model.

Figure 19. Box plots of SIR improvement for speaker 2 by audio clip using natural gradient IVA and SSL prior with the small room model.
Figure 20. Box plots of SIR improvement for speaker 3 by audio clip using natural gradient IVA and SSL prior with the small room model.

Figure 21. Box plots of SIR improvement for speaker 4 by audio clip using natural gradient IVA and SSL prior with the small room model.
Figure 22 shows the normal probability plot of SIR improvement by speaker using the natural gradient IVA algorithm and SSL prior with the small room model. This plot shows how well the actual distribution of the SIR improvement expressed in dB fits a normal distribution under the combined effects of interferer location and audio clip. It is evident from the plot that with the exception of some divergence in the tails the four speakers have SIR improvement distributions that are approximately normal, but differ in both mean and variance.

![Normal Probability Plot of ΔSIR by Speaker using NGIVA w/SSL Prior](image)

Figure 22. Normal probability plot of SIR improvement by speaker using the natural gradient IVA algorithm and SSL prior with the small room model.
Because the SIR improvement distributions are approximately normal, we use an \( F \)-test to determine how significantly the speaker effects SIR improvement. The \( F \)-test statistic is the ratio of the between group variability to the within group variability.

\[
F = \frac{\text{between-group variability}}{\text{within-group variability}}
\]

\[
= \frac{MS_{\text{group}}}{MS_{\text{error}}}
\]

\[
= \frac{SS_{\text{group}}/df_{\text{group}}}{SS_{\text{error}}/df_{\text{error}}}
\]  

\[
= \frac{\sum_i n_i (\bar{Y}_i - \bar{Y})^2/(K - 1)}{\sum_{ij} (Y_{ij} - \bar{Y}_i)^2/(N - K)}
\]

where \( MS \) stands for mean-squares, \( SS \) sum of squares, \( df \) degrees of freedom, \( Y_{ij} \) is the random variable, \( i \) indexes the group, \( j \) indexes the observation within a group, \( K \) is the number of groups, and \( N \) is the total number of observations. When applying the \( F \)-test to determine if the choice of speaker has a significant effect on SIR improvement, \( i \) is the speaker of interest, \( j \) is the experiment on the \( i \)th speaker of interest, \( Y_{ij} \) is the SIR improvement measured during the experiment, \( \bar{Y}_i \) is the mean SIR improvement measured over all \( n_i \) experiments on the \( i \)th speaker of interest, \( K \) is the number of speakers, \( N \) is the number of experiments in total, and \( \bar{Y} \) is the overall mean SIR improvement.

Analysis of variance (ANOVA) is a common procedure for analyzing the variance between and within groups and testing the hypotheses

\[ H_0: \bar{Y}_1 = \bar{Y}_2 = \cdots = \bar{Y}_K. \]

\[ H_1: \bar{Y}_i \neq \bar{Y}_j. \]  

(81)
where $H_0$ is referred to as the null hypothesis and $H_1$ means that at least one group mean does not equal the others. It produces an $F$-test statistic as well as a $p$-value, which is the probability under the assumption of the null hypothesis that the actual $F$ could be greater than or equal to the observed $F$.

Table 2 shows the ANOVA results when SIR improvement data are grouped by speaker of interest. It is evident from the large $F$-test statistic and small $p$-value that the mean SIR improvement with different speakers is not equal. Therefore, we conclude that the amount of SIR improvement is significantly affected by the choice of speaker.

Table 2. ANOVA on SIR improvement by speaker using the natural gradient IVA algorithm and SSL prior with the small room model.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>$F$</th>
<th>Prob&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>1402.1</td>
<td>3</td>
<td>467.37</td>
<td>29.318</td>
<td>3.88E-17</td>
</tr>
<tr>
<td>Error</td>
<td>6312.8</td>
<td>396</td>
<td>15.94</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>7714.9</td>
<td>399</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4.1.1.2 SIR Improvement with Natural Gradient Algorithm using Liang’s Prior

Figure 23 shows box plots of SIR improvement by speaker using the natural gradient algorithm with Liang’s prior. Similarities are evident between these box plots and the same box plots for the SSL prior. As is the case with the SSL prior, it is evident from the box plots that both the median and interquartile range depend on the speaker. Also, separation for speaker 1 is superior to that for the other three speakers. However, unlike the SSL prior, Liang’s prior produced a number of outlying cases where the algorithm did not converge. In the four cases of SIR loss greater than 10 dB, the speech estimate was dominated by artifacts.
Figure 23. Box plots of SIR improvement by speaker using the natural gradient IVA algorithm and Liang’s prior with the small room model.

Figure 24 though Figure 27 show box plots of SIR improvement by audio clip for each of the four speakers using the natural gradient algorithm with the Liang’s prior. Again, similarities are evident between these box plots and the same box plots for the SSL prior. As is the case with the SSL prior, it is evident from the box plots that the choice of audio clip has as much or more effect on SIR improvement as the positions of the interferers. Also, with the exception of outliers, SIR improvement is directly correlated between the two priors both by speaker and audio clip. After removal of the four outliers discussed above, the Pearson correlation coefficient of mean SIR improvement between the two priors by speaker is 0.98 and by audio clip ranges from 0.91 to 0.99 depending on speaker.
Figure 24. Box plots of SIR improvement for speaker 1 by audio clip using natural gradient IVA and Liang’s prior with the small room model.

Figure 25. Box plots of SIR improvement for speaker 2 by audio clip using natural gradient IVA and Liang’s prior with the small room model.
Figure 26. Box plots of SIR improvement for speaker 3 by audio clip using natural gradient IVA and Liang’s prior with the small room model.

Figure 27. Box plots of SIR improvement for speaker 4 by audio clip using natural gradient IVA and Liang’s prior with the small room model.
Figure 28 shows the normal probability plot of SIR improvement by speaker using the natural gradient IVA algorithm and Liang’s prior with the small room model. As is the case with the SSL prior, it is evident from the plot that with the exception of some divergence in the tails the four speakers have SIR improvement distributions that are approximately normal, but differ in both mean and variance. Also, the extreme SIR loss cases are clearly seen as outliers here based on their deviation from the best fit normal probability line.

Table 3 shows the ANOVA results when SIR improvement data are grouped by speaker of interest. As was the case with the SSL prior, the large $F$-test statistic and small
p-value lead us to conclude that the amount of SIR improvement is significantly affected by the choice of speaker.

Table 3. ANOVA on SIR improvement by speaker using the natural gradient IVA algorithm and Liang’s prior with the small room model.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>Prob&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>951.9</td>
<td>3</td>
<td>317.30</td>
<td>11.803</td>
<td>2.0239E-07</td>
</tr>
<tr>
<td>Error</td>
<td>10646.1</td>
<td>396</td>
<td>26.88</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>11598.0</td>
<td>399</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4.1.1.3 SIR Improvement with Fixed-point Algorithm using SSL Prior

Figure 29 shows box plots of SIR improvement by speaker using the fixed-point algorithm with the SSL prior. Similarities are evident between these box plots and the same box plots for the natural gradient algorithm. As is the case with natural gradient algorithm, it is evident from the box plots that both the median and interquartile range depend on the speaker. Also, separation for speaker 1 is superior to that for the other three speakers.
Figure 29. Box plots of SIR improvement by speaker using the fixed-point IVA algorithm and SSL prior with the small room model.

Figure 30 though Figure 33 show box plots of SIR improvement by audio clip for each of the four speakers using the fixed-point algorithm with the SSL prior. There are striking similarities between these box plots and the same box plots for the natural gradient algorithm. As was the case with the natural gradient algorithm, it is evident from the box plots that the choice of audio clip has as much or more effect on SIR improvement as the positions of the interferers. Also, SIR improvement is directly correlated between the two algorithms both by speaker and by audio clip. The Pearson correlation coefficient for mean SIR improvement between fixed-point and natural gradient algorithms using the SSL prior by speaker is 0.99 and by audio clip ranges from 0.93 to 0.98 depending on speaker.
Figure 30. Box plots of SIR improvement for speaker 1 by audio clip using fixed-point IVA and SSL prior with the small room model.

Figure 31. Box plots of SIR improvement for speaker 2 by audio clip using fixed-point IVA and SSL prior with the small room model.
Figure 32. Box plots of SIR improvement for speaker 3 by audio clip using natural gradient IVA and SSL prior with the small room model.

Figure 33. Box plots of SIR improvement for speaker 4 by audio clip using fixed-point IVA and SSL prior with the small room model.
Figure 34 shows the normal probability plot of SIR improvement by speaker using the fixed-point IVA algorithm and SSL prior with the small room model. As was the case with the natural gradient algorithm, it is evident from the plot that with the exception of some divergence in the tails the four speakers have SIR improvement distributions that are approximately normal, but differ in both mean and variance.

Table 4 shows the ANOVA results when SIR improvement data are grouped by speaker of interest. As was the case with the natural gradient algorithm, the large $F$-test statistic and small $p$-value lead us to conclude that the amount of SIR improvement is significantly affected by the choice of speaker.
Table 4. ANOVA on SIR improvement by speaker using the fixed-point IVA algorithm and SSL prior with the small room model.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>Prob&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>1678.1</td>
<td>3</td>
<td>559.37</td>
<td>42.598</td>
<td>7.06E-24</td>
</tr>
<tr>
<td>Error</td>
<td>5200.1</td>
<td>396</td>
<td>13.13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>6878.2</td>
<td>399</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 4.1.1.4 SIR Improvement with Fixed-point Algorithm using Liang’s Prior

Figure 35 shows box plots of SIR improvement by speaker using the fixed-point algorithm with Liang’s prior. Similarities are evident between these box plots and the same box plots for the SSL prior. As is the case with the SSL prior, it is evident from the box plots that both the median and interquartile range depend on the speaker. Also, separation for speaker 1 is superior to that for the other three speakers.

Figure 35. Box plots of SIR improvement by speaker using the fixed-point IVA algorithm and Liang’s prior with the small room model.
Figure 36 though Figure 39 show box plots of SIR improvement by audio clip for each of the four speakers using the fixed-point algorithm with the Liang’s prior. Again, there are striking similarities between these box plots and the same box plots for the SSL prior and natural gradient algorithm. As was the case with the others, it is evident from the box plots that the choice of audio clip has as much or more effect on SIR improvement as the positions of the interferers. Also, SIR improvement obtained using the fixed-point algorithm with SSL and Liang’s prior is directly correlated by both speaker and audio clip. The Pearson correlation coefficient for mean SIR improvement between the two priors by speaker is 0.999 and by audio clip ranges from 0.97 to 0.99 depending on speaker.

![Box Plots of ΔSIR by Audio Clip for Speaker 1 using FPIVA w/Liang’s Prior](image)

Figure 36. Box plots of SIR improvement for speaker 1 by audio clip using fixed-point IVA and Liang’s prior with the small room model.
Figure 37. Box plots of SIR improvement for speaker 2 by audio clip using fixed-point IVA and Liang’s prior with the small room model.

Figure 38. Box plots of SIR improvement for speaker 3 by audio clip using natural gradient IVA and Liang’s prior with the small room model.
Figure 39. Box plots of SIR improvement for speaker 4 by audio clip using fixed-point IVA and Liang’s prior with the small room model.

Figure 40 shows the normal probability plot of SIR improvement by speaker using the fixed-point IVA algorithm and Liang’s prior with the small room model. As was the case with the natural gradient algorithm and SSL prior, it is evident from the plot that with the exception of some divergence in the tails the four speakers have SIR improvement distributions that are approximately normal, but differ in both mean and variance.
Figure 40. Normal probability plot of SIR improvement by speaker using the fixed-point IVA algorithm and Liang’s prior with the small room model.

Table 5 shows the ANOVA results when SIR improvement data are grouped by speaker of interest. As was the case with the SSL prior, the large $F$-test statistic and small $p$-value lead us to conclude that the amount of SIR improvement is significantly affected by the choice of speaker.

Table 5. ANOVA on SIR improvement by speaker using the fixed-point IVA algorithm and Liang’s prior with the small room model.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>$F$</th>
<th>Prob&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>1597.6</td>
<td>3</td>
<td>532.52</td>
<td>41.395</td>
<td>2.7439E-23</td>
</tr>
<tr>
<td>Error</td>
<td>5094.3</td>
<td>396</td>
<td>12.86</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>6691.9</td>
<td>399</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4.1.1.5 SIR Improvement with Real Time Algorithm using SSL Prior

Figure 41 shows box plots of SIR improvement by speaker using the real time algorithm with the SSL prior. Similarities are evident between these box plots and the same box plots for the natural gradient algorithm and fixed-point algorithms. As is the case with natural gradient and fixed-point algorithms, it is evident from the box plots that both the median and interquartile range depend on the speaker. Also, separation for speaker 1 is superior to that for the other three speakers. However, unlike the natural gradient and fixed-point algorithms, clusters of positive outliers are observed not just on speaker two, but on three and four as well.

Figure 41. Box plots of SIR improvement by speaker using the real time IVA algorithm and SSL prior with the small room model.
Figure 42 though Figure 45 show box plots of SIR improvement by audio clip for each of the four speakers using the real time algorithm with the SSL prior. As was the case with the previous algorithms, it is evident from the box plots that the choice of audio clip has as much or more effect on SIR improvement as the positions of the interferers. However, correlation of SIR improvement between real time and natural gradient algorithms by speaker and audio clip is much weaker than it is between fixed-point and natural gradient algorithms. The Pearson correlation coefficient of mean SIR improvement between real time and natural gradient algorithms using the SSL prior by speaker is 0.64 and by audio clip ranges from 0.59 to 0.69 depending on speaker.

![Box Plots of ΔSIR by Audio Clip for Speaker 1 using RTIVA w/SSL Prior](image)

Figure 42. Box plots of SIR improvement for speaker 1 by audio clip using real time IVA and SSL prior with the small room model.
Figure 43. Box plots of SIR improvement for speaker 2 by audio clip using real time IVA and SSL prior with the small room model.

Figure 44. Box plots of SIR improvement for speaker 3 by audio clip using natural gradient IVA and SSL prior with the small room model.
Figure 45. Box plots of SIR improvement for speaker 4 by audio clip using real time IVA and SSL prior with the small room model.

Figure 46 shows the normal probability plot of SIR improvement by speaker using the real time IVA algorithm and SSL prior with the small room model. As with the natural gradient and fixed-point algorithms, both mean and variance of SIR improvement depend on speaker. However, distributions of SIR improvement obtained using the real time algorithm with the SSL prior show more positive skew than those obtained using the natural gradient and fixed-point algorithms particularly for speakers two, three and four.
Figure 46. Normal probability plot of SIR improvement by speaker using the real time IVA algorithm and SSL prior with the small room model.

While the $F$-test assumes normality within a group and equal variance between groups, it has been shown that the statistic is robust to violations of these assumptions [54]. Therefore, we will use the $F$-test here to determine the effect of speaker on SIR improvement without regard to normality or variance. Table 6 shows the ANOVA results when SIR improvement data are grouped by speaker of interest. As was the case with the preceding algorithms, the large $F$-test statistic and small $p$-value lead us to conclude that the amount of SIR improvement is significantly affected by the choice of speaker.

Table 6. ANOVA on SIR improvement by speaker using the real time IVA algorithm and SSL prior with the small room model.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>$F$</th>
<th>Prob&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>121.60</td>
<td>3</td>
<td>40.535</td>
<td>99.256</td>
<td>6.35E-48</td>
</tr>
<tr>
<td>Error</td>
<td>161.72</td>
<td>396</td>
<td>0.408</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>283.33</td>
<td>399</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4.1.1.6 SIR Improvement with Real Time Algorithm using Liang’s Prior

Figure 47 shows box plots of SIR improvement by speaker using the real time algorithm with Liang’s prior. These box plots correlate directly to those taken using the SSL prior, but with lower medians and ranges. As with all previous cases, it is evident from the box plots that both the median and interquartile range depend on the speaker. Also, separation for speaker 1 is superior to that for the other three speakers.

![Box Plots of SIR Improvement by Speaker Using Real Time Algorithm and Liang’s Prior](image)

Figure 47. Box plots of SIR improvement by speaker using the real time IVA algorithm and Liang’s prior with the small room model.

Figure 48 though Figure 51 show box plots of SIR improvement by audio clip for each of the four speakers using the real time algorithm with the Liang’s prior. As with all previous cases, it is evident from the box plots that the choice of audio clip has as much or more
effect on SIR improvement as the positions of the interferers. Also, SIR improvement obtained using the real time algorithm with SSL and Liang’s prior is directly correlated by both speaker and audio clip. The Pearson correlation coefficient of mean SIR improvement between the two priors by speaker is 0.96 and by audio clip ranges from 0.82 to 0.96 depending on speaker.

Figure 48. Box plots of SIR improvement for speaker 1 by audio clip using real time IVA and Liang’s prior with the small room model.
Figure 49. Box plots of SIR improvement for speaker 2 by audio clip using real time IVA and Liang’s prior with the small room model.

Figure 50. Box plots of SIR improvement for speaker 3 by audio clip using natural gradient IVA and Liang’s prior with the small room model.
Figure 51. Box plots of SIR improvement for speaker 4 by audio clip using real time IVA and Liang’s prior with the small room model.

Figure 52 shows the normal probability plot of SIR improvement by speaker using the real time IVA algorithm and Liang’s prior with the small room model. As with the real time algorithm using the SSL prior, it is evident from the plot that SIR improvement distributions have positive skew particularly for speakers two, three and four.
Figure 52. Normal probability plot of SIR improvement by speaker using the real time IVA algorithm and Liang’s prior with the small room model.

Table 7 shows the ANOVA results when SIR improvement data are grouped by speaker of interest. As was the case with the SSL prior, the large $F$-test statistic and small $p$-value lead us to conclude that the amount of SIR improvement is significantly affected by the choice of speaker.

Table 7. ANOVA on SIR improvement by speaker using the real time IVA algorithm and Liang’s prior with the small room model.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>$F$</th>
<th>$Prob&gt;F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>132.21</td>
<td>3</td>
<td>44.071</td>
<td>164.25</td>
<td>3.640E-69</td>
</tr>
<tr>
<td>Error</td>
<td>106.26</td>
<td>396</td>
<td>0.268</td>
<td></td>
<td>0.268</td>
</tr>
<tr>
<td>Total</td>
<td>238.47</td>
<td>399</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4.1.1.7 SIR Improvement with Auxiliary function Algorithm using SSL Prior

Figure 53 shows box plots of SIR improvement by speaker using the auxiliary function algorithm with the SSL prior. Similarities are evident between these box plots and the same box plots for the other three algorithms. As is the case with the others, it is evident from the box plots that both the median and interquartile range depend on the speaker. Also, separation for speaker 1 is superior to that for the other three speakers. However, unlike the other algorithms when using the SSL prior, the auxiliary function algorithm produced a number of outlying cases where the algorithm did not converge. In many of these cases, the speech estimate was contaminated with strong artifacts.

![Box Plots of ΔSIR by Speaker using AFIVA w/SSL Prior](image)

Figure 53. Box plots of SIR improvement by speaker using the auxiliary function IVA algorithm and SSL prior with the small room model.
Figure 54 though Figure 57 show box plots of SIR improvement by audio clip for each of the four speakers using the auxiliary function algorithm with the SSL prior. As with all previous cases, it is evident from the box plots that the choice of audio clip has as much or more effect on SIR improvement as the positions of the interferers. Also, with the exception of speaker 4, SIR improvement is directly correlated between the auxiliary function and other algorithms both by speaker and by audio clip. The Pearson correlation coefficient for mean SIR improvement between auxiliary function and natural gradient algorithms using the SSL prior by speaker is 0.85 and by audio clip is 0.87, 0.82 and 0.91 for speakers 1, 2 and 3 respectively.

![Box Plots of ΔSIR by Audio Clip for Speaker 1 using AFIVA w/SSL Prior](image)

Figure 54. Box plots of SIR improvement for speaker 1 by audio clip using auxiliary function IVA and SSL prior with the small room model.
Figure 55. Box plots of SIR improvement for speaker 2 by audio clip using auxiliary function IVA and SSL prior with the small room model.

Figure 56. Box plots of SIR improvement for speaker 3 by audio clip using natural gradient IVA and SSL prior with the small room model.
Figure 57. Box plots of SIR improvement for speaker 4 by audio clip using auxiliary function IVA and SSL prior with the small room model.

Figure 58 shows the normal probability plot of SIR improvement by speaker using the auxiliary function IVA algorithm and SSL prior with the small room model. As was the case with the natural gradient and fixed-point algorithms, it is evident from the plot that with the exception of some divergence in the tails the four speakers have SIR improvement distributions that are approximately normal, but differ in both mean and variance.
Figure 58. Normal probability plot of SIR improvement by speaker using the auxiliary function IVA algorithm and SSL prior with the small room model.

Table 8 shows the ANOVA results when SIR improvement data are grouped by speaker of interest. As was the case with the previous three algorithms, the large $F$-test statistic and small $p$-value lead us to conclude that the amount of SIR improvement is significantly affected by the choice of speaker.

Table 8. ANOVA on SIR improvement by speaker using the auxiliary function IVA algorithm and SSL prior with the small room model.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>$F$</th>
<th>Prob&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>1833.2</td>
<td>3</td>
<td>611.07</td>
<td>25.886</td>
<td>2.607E-15</td>
</tr>
<tr>
<td>Error</td>
<td>9348.1</td>
<td>396</td>
<td>23.61</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>11181.3</td>
<td>399</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4.1.1.8 SIR Improvement with Auxiliary function Algorithm using Liang’s Prior

Figure 59 shows box plots of SIR improvement by speaker using the auxiliary function algorithm with Liang’s prior. Similarities are evident between these box plots and the same box plots for the SSL prior. As is the case with the SSL prior, it is evident from the box plots that both the median and interquartile range depend on the speaker. Also, separation for speaker 1 is superior to that for the other three speakers.

![Box Plots of ΔSIR by Speaker using AFIVA w/Liang’s Prior](image)

Figure 59. Box plots of SIR improvement by speaker using the auxiliary function IVA algorithm and Liang’s prior with the small room model.

Figure 60 though Figure 63 show box plots of SIR improvement by audio clip for each of the four speakers using the auxiliary function algorithm with the Liang’s prior. As was the case with the SSL prior, it is evident from the box plots that the choice of audio clip has as much or more effect on SIR improvement as the positions of the interferers.
Also, SIR improvement obtained using the auxiliary function algorithm with SSL and Ling’s prior is directly correlated by both speaker and audio clip. The Pearson correlation coefficient for mean SIR improvement between the two priors by speaker is 0.996 and by audio clip ranges from 0.78 to 0.89 depending on speaker.

Figure 60. Box plots of SIR improvement for speaker 1 by audio clip using auxiliary function IVA and Liang’s prior with the small room model.
Figure 61. Box plots of SIR improvement for speaker 2 by audio clip using auxiliary function IVA and Liang’s prior with the small room model.

Figure 62. Box plots of SIR improvement for speaker 3 by audio clip using auxiliary function IVA and Liang’s prior with the small room model.

96
Figure 63. Box plots of SIR improvement for speaker 4 by audio clip using auxiliary function IVA and Liang’s prior with the small room model.

Figure 64 shows the normal probability plot of SIR improvement by speaker using the auxiliary function IVA algorithm and Liang’s prior with the small room model. As was the case with the SSL prior, it is evident from the plot that with the exception of some divergence in the tails the four speakers have SIR improvement distributions that are approximately normal, but differ in both mean and variance.
Table 9 shows the ANOVA results when SIR improvement data are grouped by speaker of interest. As was the case with the previous three algorithms, the large $F$-test statistic and small $p$-value lead us to conclude that the amount of SIR improvement is significantly affected by the choice of speaker.

Table 9. ANOVA on SIR improvement by speaker using the auxiliary function IVA algorithm and Liang’s prior with the small room model.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>$F$</th>
<th>Prob&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>1600.9</td>
<td>3</td>
<td>533.62</td>
<td>48.484</td>
<td>1.0425E-26</td>
</tr>
<tr>
<td>Error</td>
<td>4358.5</td>
<td>396</td>
<td>11.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>5959.3</td>
<td>399</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4.1.2 SDR Improvement with the Small Room Model

Table 10 shows the overall improvement in SDR between the mixture $x_i$ at the microphone pointed at the speaker of interest and the separated source estimate $u_j$ of the speaker of interest. The data are collected from the same experiments as are SIR data above, and the results are similar, but lower in magnitude. The difference between SIR and SDR can be seen in equation (76) above. SDR includes artifacts in the denominator of the ratio. Therefore, the value of SDR will normally be lower than that of SIR, but directly correlated. Comparing Table 1 to Table 10, we see that the Pearson correlation coefficient between mean SIR and SDR improvement by algorithm is 0.85.

Table 10. Overall improvement in SDR with the small room model.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Median (dB)</th>
<th>Mean (dB)</th>
<th>Std. (dB)</th>
<th>Improved (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NGIVA w/SSL Prior</td>
<td>3.135</td>
<td>3.205</td>
<td>4.543</td>
<td>73.75</td>
</tr>
<tr>
<td>FPIVA w/SSL Prior</td>
<td>3.205</td>
<td>3.159</td>
<td>4.425</td>
<td>74.00</td>
</tr>
<tr>
<td>RTIVA w/SSL Prior</td>
<td>2.305</td>
<td>2.411</td>
<td>0.687</td>
<td>100.00</td>
</tr>
<tr>
<td>AFIVA w/SSL Prior</td>
<td>3.640</td>
<td>3.394</td>
<td>5.343</td>
<td>76.25</td>
</tr>
<tr>
<td>NGIVA w/Liang’s Prior</td>
<td>3.850</td>
<td>3.378</td>
<td>5.694</td>
<td>78.50</td>
</tr>
<tr>
<td>FPIVA w/Liang’s Prior</td>
<td>3.330</td>
<td>3.437</td>
<td>4.377</td>
<td>79.00</td>
</tr>
<tr>
<td>RTIVA w/Liang’s Prior</td>
<td>2.050</td>
<td>2.094</td>
<td>0.690</td>
<td>100.00</td>
</tr>
<tr>
<td>AFIVA w/Liang’s Prior</td>
<td>5.590</td>
<td>5.321</td>
<td>4.231</td>
<td>89.00</td>
</tr>
</tbody>
</table>

Figure 65 shows box plots of SDR improvement by algorithm using the SSL prior with the small room model. Figure 66 shows the same plots using Liang’s prior. Like SIR, a notable observation is that the real time algorithm appears to be much more stable than the others in the sense of having shorter whiskers and few outliers. Also, the instability of the auxiliary function algorithm when used with the SSL prior is evident in Figure 65. Similarly, the instability of the natural gradient algorithm when used with Liang’s prior is
evident in Figure 66. Finally, an analysis of variance on SDR improvement by algorithm, speaker and audio clip shows the same dependencies as are observed with SIR. Like SIR, SDR improvement is significantly affected by all three factors.

Figure 65. Box plots of SDR improvement by algorithm using the SSL prior with the small room model.
4.1.3 SAR Change with the Small Room Model

Table 11 shows the overall change in SAR between the mixture $x_i$ at the microphone pointed at the speaker of interest and the separated source estimate $u_j$ of the speaker of interest. The data are collected from the same experiments as are SIR and SDR data above. The main difference is that SAR is not expected to improve. Referring back to equation (75), we expect $\|\tilde{e}_{\text{artifacts}}\|$ of the mixture $x_i$ to be nearly zero, since no signal processing (a major source of artifacts) has yet been done. On the other hand, we expect $\|\tilde{e}_{\text{artifacts}}\|$ of the separated source estimate $u_j$ to be non-zero since it resulted from signal processing. Subtracting the nearly infinite SAR of the mixture $x_i$ from the finite SAR of the separated source estimate $u_j$ usually results in a negative value.
However, separation can result in a small reduction in artifacts. The signals received by the microphones contain frequency response distortion due to multipath. When the projections of the clean speech and interferers are removed, the artifacts of frequency response distortion are left behind as shown in equation (79) above. The stronger the interference, the more artifacts of distortion remain. So a small improvement in SAR is possible purely as a result of separation.

Close inspection of Table 11 shows that in the majority of cases, the artifacts introduced by signal processing exceed those removed by separation resulting in an overall decrease in SAR. The one exception is the real time algorithm using Liang’s prior. In this case, over 75% of the experiments resulted in an increase in SAR. In other words, the artifacts removed by separation exceed those introduced by signal processing. The reason for this discrepancy is that the learning rate $\eta$ required in order to stabilize the real time algorithm when using Liang’s prior is only one tenth that used with the SSL prior. Experiments with the real time algorithm show that decreasing $\eta$ consistently increases SAR at the expense of reduced SIR and SDR.

Table 11. Overall change in SAR with the small room model.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Median (dB)</th>
<th>Mean (dB)</th>
<th>Std. (dB)</th>
<th>Improved (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NGIVA w/SSL Prior</td>
<td>-7.245</td>
<td>-7.131</td>
<td>4.655</td>
<td>6.25</td>
</tr>
<tr>
<td>FPIVA w/SSL Prior</td>
<td>-7.785</td>
<td>-7.346</td>
<td>4.580</td>
<td>6.75</td>
</tr>
<tr>
<td>RTIVA w/SSL Prior</td>
<td>-0.640</td>
<td>-0.798</td>
<td>0.762</td>
<td>11.25</td>
</tr>
<tr>
<td>AFIVA w/SSL Prior</td>
<td>-6.900</td>
<td>-6.943</td>
<td>5.179</td>
<td>10.25</td>
</tr>
<tr>
<td>NGIVA w/Liang's Prior</td>
<td>-6.290</td>
<td>-6.686</td>
<td>5.410</td>
<td>9.75</td>
</tr>
<tr>
<td>FPIVA w/Liang's Prior</td>
<td>-7.200</td>
<td>-7.083</td>
<td>4.536</td>
<td>7.25</td>
</tr>
<tr>
<td>RTIVA w/Liang's Prior</td>
<td>0.250</td>
<td>0.222</td>
<td>0.377</td>
<td>75.50</td>
</tr>
<tr>
<td>AFIVA w/Liang's Prior</td>
<td>-4.740</td>
<td>-5.146</td>
<td>4.504</td>
<td>16.00</td>
</tr>
</tbody>
</table>
Figure 67 shows box plots of the change in SAR by algorithm using the SSL prior with the small room model. Figure 68 shows the same plots using Liang’s prior. Like SIR and SDR, the instability of the auxiliary function algorithm when used with the SSL prior is evident in Figure 67. Similarly, the instability of the natural gradient algorithm when used with Liang’s prior is evident in Figure 68. Finally, an analysis of variance on SAR improvement by algorithm, speaker and audio clip shows the same dependencies as are observed with SIR and SDR. Like SIR and SDR, the change in SAR is significantly affected by all three factors.

Figure 67. Box plots of the change in SAR by algorithm using the SSL prior with the small room model.
Figure 68. Box plots of the change in SAR by algorithm using the Liang’s prior with the small room model.

### 4.1.4 WER Improvement with the Small Room Model

Figure 69 shows the WER after decoding the clean speech $s_j$, mixture $x_i$ from the microphone pointed at the speaker of interest, and the separated source estimate $u_j$ of the speaker of interest obtained using the NGIVA, FPIVA, RTIVA and AFIVA algorithms with both SSL and Liang’s prior. The data are collected from the same experiments as are SIR, SDR and SAR data above. It is clear from the bar graph that all combinations of separation algorithm and prior result in a decrease of WER relative to the mixture at the microphone pointed at the speaker of interest. The AFIVA algorithm clearly outperforms the others reaching within 5% of the mean WER of the decoded clean speech signal.
Figure 69. WER of decoded clean speech, microphone mixture, and separated source estimates using the small room model.

Table 12 shows the overall improvement in WER between the mixture $x_i$ at the microphone pointed at the speaker of interest and the separated source estimate $u_j$ of the speaker of interest. Improvement is measured as the WER of the decoded mixture $x_i$ minus the WER of the decoded separated source estimate $u_j$. Comparing Table 12 to Table 10 above, we see that the Pearson correlation coefficient between mean WER and SDR improvement by algorithm is 0.82. Comparing Table 12 to Table 1 above we see that the Pearson correlation coefficient between mean WER and SIR by algorithm is 0.70. The
weakness of correlation suggests that WER is also influenced by factors other than interference and distortion. A close inspection of Table 12 reveals that for most algorithms, the mean improvement in WER is small in comparison to the standard deviation. This might suggest low confidence that the BSS algorithm will deliver an improvement in WER. However, all BSS algorithms produced far more cases of WER improvement than degradation with AFIVA approaching 90% of cases improved and only 3% degraded.

Table 12. Overall improvement in WER by algorithm using the small room model.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Median (%)</th>
<th>Mean (%)</th>
<th>Std. (%)</th>
<th>Improved (%)</th>
<th>Unaffected (%)</th>
<th>Degraded (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NGIVA w/SSL Prior</td>
<td>12.25</td>
<td>10.52</td>
<td>22.03</td>
<td>76.75</td>
<td>8.25</td>
<td>15.00</td>
</tr>
<tr>
<td>FPIVA w/SSL Prior</td>
<td>10.50</td>
<td>8.54</td>
<td>22.35</td>
<td>71.75</td>
<td>9.00</td>
<td>19.25</td>
</tr>
<tr>
<td>RTIVA w/SSL Prior</td>
<td>5.30</td>
<td>7.19</td>
<td>9.41</td>
<td>67.75</td>
<td>22.25</td>
<td>10.00</td>
</tr>
<tr>
<td>AFIVA w/SSL Prior</td>
<td>16.70</td>
<td>19.44</td>
<td>17.09</td>
<td>87.00</td>
<td>7.25</td>
<td>5.75</td>
</tr>
<tr>
<td>NGIVA w/Liang’s Prior</td>
<td>11.75</td>
<td>8.60</td>
<td>24.45</td>
<td>73.00</td>
<td>9.50</td>
<td>17.50</td>
</tr>
<tr>
<td>FPIVA w/Liang’s Prior</td>
<td>11.50</td>
<td>8.82</td>
<td>22.64</td>
<td>74.50</td>
<td>8.00</td>
<td>17.50</td>
</tr>
<tr>
<td>RTIVA w/Liang’s Prior</td>
<td>4.20</td>
<td>4.16</td>
<td>8.79</td>
<td>57.00</td>
<td>28.75</td>
<td>14.25</td>
</tr>
<tr>
<td>AFIVA w/Liang’s Prior</td>
<td>20.75</td>
<td>21.10</td>
<td>15.45</td>
<td>89.50</td>
<td>7.25</td>
<td>3.25</td>
</tr>
</tbody>
</table>

Figure 70 shows box plots of WER improvement by algorithm using the SSL prior with the small room model. Figure 71 shows the same plots using Liang’s prior. One notable observation is the correlation between WER and SIR or SDR in terms of median and interquartile range. A notable difference between WER and SIR or SDR is the large number of negative outliers produced by the NGIVA and FPIVA algorithms regardless of the prior used. Finally, an analysis of variance on WER improvement by algorithm, speaker and audio clip shows a significant dependency on all three factors. However, there was little correlation between WER and SIR, SDR or SAR by speaker or audio clip. This is not
an entirely surprising result since ASR algorithms operate on temporal features of the speech signal (sequences of phonemes) and are influenced by factors in addition to long-term averages of interference and distortion.

Figure 70. Box plots of WER improvement by algorithm using the SSL prior with the small room model.
4.2 Large Room Model Results

The model of a large conference room measuring 5 m by 4 m horizontally and 2.5 m vertically is shown in Figure 72. Absorption coefficients are set to model drywall walls, carpeted floor, and sound absorbing ceiling tiles. The speaker of interest indicated by the red square is located 1 m inward from the center of a 4 m wall. The microphone array indicated by the gray diamond is located 0.3 m inward from the speaker of interest. The eleven possible locations for the three interfering speakers are indicated by blue circles. With each experiment, the locations of the three interfering speakers are selected at random from the set of eleven possibilities. This geometry models the use case of a microphone array equipped smart watch.

Figure 71. Box plots of WER improvement by algorithm using Liang’s prior with the small room model.
Figure 72. Diagram of a large 5 m by 4 m room layout showing the locations of the microphone array (gray diamond), speaker of interest (red square), and interfering speakers (blue circles). The geometry models the use case of a microphone array equipped smart watch.

4.2.1 SIR Improvement with the Large Room Model

Table 13 shows the overall improvement in SIR between the mixture $x_i$ at the microphone pointed at the speaker of interest and the separated source estimate $u_j$ of the speaker of interest. The data are collected over 400 experiments with each algorithm. With a Pearson correlation coefficient of 0.9994, mean SIR improvement by algorithm between large and small room models is directly correlated. However, SIR improvement is slightly lower for all but the real time algorithm with Liang’s prior.
Table 13. Overall improvement in SIR with the large room model.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Median (dB)</th>
<th>Mean (dB)</th>
<th>Std. (dB)</th>
<th>Improved (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NGIVA w/SSL Prior</td>
<td>5.265</td>
<td>5.688</td>
<td>4.348</td>
<td>91.00</td>
</tr>
<tr>
<td>FPIVA w/SSL Prior</td>
<td>5.835</td>
<td>5.879</td>
<td>4.084</td>
<td>92.50</td>
</tr>
<tr>
<td>RTIVA w/SSL Prior</td>
<td>2.370</td>
<td>2.628</td>
<td>1.008</td>
<td>100.00</td>
</tr>
<tr>
<td>AFIVA w/SSL Prior</td>
<td>6.005</td>
<td>5.882</td>
<td>5.555</td>
<td>88.50</td>
</tr>
<tr>
<td>NGIVA w/Liang's Prior</td>
<td>5.745</td>
<td>5.546</td>
<td>5.862</td>
<td>90.50</td>
</tr>
<tr>
<td>FPIVA w/Liang's Prior</td>
<td>6.230</td>
<td>6.158</td>
<td>4.128</td>
<td>94.50</td>
</tr>
<tr>
<td>RTIVA w/Liang's Prior</td>
<td>2.205</td>
<td>2.324</td>
<td>0.987</td>
<td>100.00</td>
</tr>
<tr>
<td>AFIVA w/Liang's Prior</td>
<td>7.105</td>
<td>7.039</td>
<td>5.290</td>
<td>93.75</td>
</tr>
</tbody>
</table>

Figure 73 shows box plots of SIR improvement by algorithm using the SSL prior with the large room model. Figure 74 shows the same plots using Liang’s prior. The same instability of the natural gradient algorithm when using Liang’s prior and auxiliary function algorithm when using the SSL prior that was observed with the small room model is evident in these plots for the large room model. Furthermore, an analysis of variance shows the same dependency on speaker and audio clip that was observed with the small room model. In general, the same characteristics that were observed in the small room are present in the large room.
Figure 73. Box plots of SIR improvement by algorithm using the SSL prior with the large room model.

Figure 74. Box plots of SIR improvement by algorithm using Liang’s prior with the large room model.
4.2.2 SDR Improvement with the Large Room Model

Table 14 shows the overall improvement in SDR between the mixture $x_i$ at the microphone pointed at the speaker of interest and the separated source estimate $u_j$ of the speaker of interest. The data are collected from the same experiments as are SIR data above. Except for the real time algorithm, mean SDR improvement by algorithm between large and small room models is directly correlated with a Pearson correlation coefficient of 0.99. However, SDR improvement is slightly lower for all but the real time algorithm.

Table 14. Overall improvement in SDR with the large room model.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Median (dB)</th>
<th>Mean (dB)</th>
<th>Std. (dB)</th>
<th>Improved (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NGIVA w/SSL Prior</td>
<td>0.660</td>
<td>1.159</td>
<td>4.843</td>
<td>58.25</td>
</tr>
<tr>
<td>FPIVA w/SSL Prior</td>
<td>0.675</td>
<td>1.073</td>
<td>4.772</td>
<td>57.75</td>
</tr>
<tr>
<td>RTIVA w/SSL Prior</td>
<td>2.125</td>
<td>2.302</td>
<td>0.844</td>
<td>100.00</td>
</tr>
<tr>
<td>AFIVA w/SSL Prior</td>
<td>0.675</td>
<td>1.069</td>
<td>6.163</td>
<td>56.25</td>
</tr>
<tr>
<td>NGIVA w/Liang’s Prior</td>
<td>1.520</td>
<td>1.105</td>
<td>6.643</td>
<td>62.75</td>
</tr>
<tr>
<td>FPIVA w/Liang’s Prior</td>
<td>1.100</td>
<td>1.213</td>
<td>4.877</td>
<td>58.75</td>
</tr>
<tr>
<td>RTIVA w/Liang’s Prior</td>
<td>2.050</td>
<td>2.129</td>
<td>0.864</td>
<td>100.00</td>
</tr>
<tr>
<td>AFIVA w/Liang’s Prior</td>
<td>2.495</td>
<td>2.354</td>
<td>6.094</td>
<td>64.50</td>
</tr>
</tbody>
</table>

Figure 75 shows box plots of SDR improvement by algorithm using the SSL prior with the large room model. Figure 76 shows the same plots using Liang’s prior. In general, the same characteristics that were observed in the small room are present in the large room, but with an overall loss in SDR improvement.
Figure 75. Box plots of SDR improvement by algorithm using the SSL prior with the large room model.

Figure 76. Box plots of SDR improvement by algorithm using the Liang’s prior with the large room model.
4.2.3 SAR Change with the Large Room Model

Table 15 shows the overall change in SAR between the mixture $x_i$ at the microphone pointed at the speaker of interest and the separated source estimate $u_j$ of the speaker of interest. The data are collected from the same experiments as are SIR and SDR data above. With a Pearson correlation coefficient of 0.99, mean change in SAR by algorithm between large and small room models is directly correlated. However, SAR is slightly lower for all.

Table 15. Overall change in SAR with the large room model.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Median (dB)</th>
<th>Mean (dB)</th>
<th>Std. (dB)</th>
<th>Improved (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NGIVA w/SSL Prior</td>
<td>-9.835</td>
<td>-9.534</td>
<td>4.692</td>
<td>1.00</td>
</tr>
<tr>
<td>FPIVA w/SSL Prior</td>
<td>-10.115</td>
<td>-9.768</td>
<td>4.667</td>
<td>0.50</td>
</tr>
<tr>
<td>RTIVA w/SSL Prior</td>
<td>-0.930</td>
<td>-1.131</td>
<td>1.124</td>
<td>13.75</td>
</tr>
<tr>
<td>AFIVA w/SSL Prior</td>
<td>-10.105</td>
<td>-9.707</td>
<td>5.900</td>
<td>3.50</td>
</tr>
<tr>
<td>NGIVA w/Liang’s Prior</td>
<td>-8.820</td>
<td>-9.330</td>
<td>6.168</td>
<td>1.75</td>
</tr>
<tr>
<td>FPIVA w/Liang’s Prior</td>
<td>-9.860</td>
<td>-9.687</td>
<td>4.742</td>
<td>0.25</td>
</tr>
<tr>
<td>RTIVA w/Liang’s Prior</td>
<td>-0.060</td>
<td>-0.152</td>
<td>0.576</td>
<td>45.25</td>
</tr>
<tr>
<td>AFIVA w/Liang’s Prior</td>
<td>-8.205</td>
<td>-8.386</td>
<td>5.761</td>
<td>6.25</td>
</tr>
</tbody>
</table>

Figure 75 shows box plots of the change in SAR by algorithm using the SSL prior with the large room model. Figure 76 shows the same plots using Liang’s prior. In general, the same characteristics that were observed in the small room are present in the large room, but with an overall loss of SAR.
Figure 77. Box plots of SAR improvement by algorithm using the SSL prior with the large room model.

Figure 78. Box plots of SAR improvement by algorithm using the Liang’s prior with the large room model.
4.2.4 WER Improvement with the Large Room Model

Figure 79 shows the WER after decoding the clean speech $s_j$, mixture $x_i$ at the microphone pointed at the speaker of interest, and the separated source estimate $u_j$ of the speaker of interest obtained using the NGIVA, FPIVA, RTIVA and AFIVA algorithms with both SSL and Liang’s prior. The data are collected from the same experiments as are SIR, SDR and SAR data above. As is the case with the small room model, all combinations of separation algorithm and prior result in a decrease of mean WER relative to the mixture at the microphone pointed at the speaker of interest. Also like the small room model, the AFIVA algorithm outperforms the others approaching to within 4% of the mean WER of the decoded clean speech signal. However, the difference between AFIVA and the others is not as great in the large room as it is in the small room. Also, the WER of the unseparated mixture is much lower in the large room. This is expected since the trajectories between interferers and microphone are on average longer in the large room than in the small room resulting in weaker interference levels.
Figure 79. WER of decoded clean speech, microphone mixture, and separated source estimates using the large room model.

Table 16 shows the overall improvement in WER between the mixture $x_i$ at the microphone pointed at the speaker of interest and the separated source estimate $u_j$ of the speaker of interest. Comparing Table 16 to Table 12 above, we see strong correlation in mean WER improvement by algorithm between large and small room models with a Pearson correlation coefficient of 0.98. Also, as was the case with the small room model, all algorithms produced far more cases of WER improvement than degradation. However, the benefit is much lower in the large room.
Table 16. Overall improvement in WER by algorithm using the large room model.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Median (%)</th>
<th>Mean (%)</th>
<th>Std. (%)</th>
<th>Improved (%)</th>
<th>Unaffected (%)</th>
<th>Degraded (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NGIVA w/SSL Prior</td>
<td>4.50</td>
<td>4.55</td>
<td>17.60</td>
<td>59.50</td>
<td>20.25</td>
<td>20.25</td>
</tr>
<tr>
<td>FPIVA w/SSL Prior</td>
<td>4.50</td>
<td>4.86</td>
<td>17.72</td>
<td>59.50</td>
<td>21.25</td>
<td>19.25</td>
</tr>
<tr>
<td>RTIVA w/SSL Prior</td>
<td>2.60</td>
<td>4.62</td>
<td>9.16</td>
<td>50.25</td>
<td>35.75</td>
<td>14.00</td>
</tr>
<tr>
<td>AFIVA w/SSL Prior</td>
<td>5.30</td>
<td>8.24</td>
<td>15.49</td>
<td>66.00</td>
<td>17.50</td>
<td>16.50</td>
</tr>
<tr>
<td>NGIVA w/Liang's Prior</td>
<td>4.80</td>
<td>3.75</td>
<td>21.13</td>
<td>61.25</td>
<td>21.00</td>
<td>17.75</td>
</tr>
<tr>
<td>FPIVA w/Liang's Prior</td>
<td>4.50</td>
<td>4.78</td>
<td>18.21</td>
<td>61.00</td>
<td>20.25</td>
<td>18.75</td>
</tr>
<tr>
<td>RTIVA w/Liang's Prior</td>
<td>0.00</td>
<td>2.69</td>
<td>8.65</td>
<td>44.00</td>
<td>35.75</td>
<td>20.25</td>
</tr>
<tr>
<td>AFIVA w/Liang's Prior</td>
<td>6.30</td>
<td>9.02</td>
<td>15.59</td>
<td>67.25</td>
<td>21.50</td>
<td>11.25</td>
</tr>
</tbody>
</table>

Figure 80 shows box plots of WER improvement by algorithm using the SSL prior with the large room model. Figure 81 shows the same plots using Liang’s prior. Comparing these box plots to the same for the small room model, we see good correlation in terms of medians and interquartile ranges. However, there are in general more outliers in the large room and WER improvement is lower.
Figure 80. Box plots of WER improvement by algorithm using the SSL prior with the large room model.

Figure 81. Box plots of WER improvement by algorithm using Liang’s prior with the large room model.
4.3 Corner Room Model Results

The corner room model is a model of the large conference room discussed above with the speaker of interest located at one corner as shown in Figure 82. The speaker of interest indicated by the red square is located 1 m inward from both walls. The microphone array indicated by the gray diamond is located 0.3 m inward diagonally from the speaker of interest. The eleven possible locations for the three interfering speakers are indicated by blue circles. With each experiment, the locations of the three interfering speakers are selected at random from the set of eleven possibilities. This geometry models the use case of a microphone array equipped smart watch located at one corner of the room.

Figure 82. Diagram of a large 5 m by 4 m room layout showing the locations of the microphone array (gray diamond), speaker of interest (red square), and interfering speakers (blue circles). The geometry models the use case of a microphone array equipped smart watch located at one corner of the room.
4.3.1 SIR Improvement with the Corner Room Model

Table 17 shows the overall improvement in SIR between the mixture $x_i$ at the microphone pointed at the speaker of interest and the separated source estimate $u_j$ of the speaker of interest. The data are collected over 400 experiments with each algorithm. With a Pearson correlation coefficient of 0.94, mean SIR improvement by algorithm between large and corner room models is directly correlated. However, SIR improvement is slightly lower for all but the real time algorithm.

Table 17. Overall improvement in SIR with the corner room model.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Median (dB)</th>
<th>Mean (dB)</th>
<th>Std. (dB)</th>
<th>Improved (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NGIVA w/SSL Prior</td>
<td>4.370</td>
<td>4.616</td>
<td>4.535</td>
<td>83.25</td>
</tr>
<tr>
<td>FPIVA w/SSL Prior</td>
<td>4.610</td>
<td>4.519</td>
<td>4.320</td>
<td>86.25</td>
</tr>
<tr>
<td>RTIVA w/SSL Prior</td>
<td>2.635</td>
<td>2.837</td>
<td>1.151</td>
<td>100.00</td>
</tr>
<tr>
<td>AFIVA w/SSL Prior</td>
<td>5.235</td>
<td>5.381</td>
<td>5.241</td>
<td>85.75</td>
</tr>
<tr>
<td>NGIVA w/Liang's Prior</td>
<td>5.180</td>
<td>4.852</td>
<td>5.377</td>
<td>85.00</td>
</tr>
<tr>
<td>FPIVA w/Liang's Prior</td>
<td>4.960</td>
<td>4.935</td>
<td>4.145</td>
<td>88.75</td>
</tr>
<tr>
<td>RTIVA w/Liang's Prior</td>
<td>2.950</td>
<td>2.952</td>
<td>1.251</td>
<td>100.00</td>
</tr>
<tr>
<td>AFIVA w/Liang's Prior</td>
<td>7.010</td>
<td>6.638</td>
<td>4.782</td>
<td>93.00</td>
</tr>
</tbody>
</table>

Figure 83 shows box plots of SIR improvement by algorithm using the SSL prior with the corner room model. Figure 84 shows the same plots using Liang’s prior. The same instability of the natural gradient algorithm when using Liang’s prior and auxiliary function algorithm when using the SSL prior that was observed with the small and large room models is evident in these plots for the corner room model. Furthermore, an analysis of variance shows the same dependency on speaker and audio clip that was observed with the small and large room models. In general, the same characteristics that were observed in the small and large rooms are present in the corner room.
Figure 83. Box plots of SIR improvement by algorithm using the SSL prior with the corner room model.

Figure 84. Box plots of SIR improvement by algorithm using the Liang’s prior with the corner room model.
4.3.2 SDR Improvement with the Corner Room Model

Table 18 shows the overall improvement in SDR between the mixture $x_i$ at the microphone pointed at the speaker of interest and the separated source estimate $u_j$ of the speaker of interest. The data are collected from the same experiments as are SIR data above. With a Pearson correlation coefficient of 0.93, mean SDR improvement by algorithm between large and corner room models is directly correlated. However, SDR improvement is slightly lower for all but the real time algorithm with Liang’s prior.

Table 18. Overall improvement in SDR with the corner room model.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Median (dB)</th>
<th>Mean (dB)</th>
<th>Std. (dB)</th>
<th>Improved (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NGIVA w/SSL Prior</td>
<td>-0.760</td>
<td>-0.090</td>
<td>5.043</td>
<td>46.50</td>
</tr>
<tr>
<td>FPIVA w/SSL Prior</td>
<td>-1.125</td>
<td>-0.401</td>
<td>5.008</td>
<td>44.75</td>
</tr>
<tr>
<td>RTIVA w/SSL Prior</td>
<td>2.330</td>
<td>2.495</td>
<td>0.952</td>
<td>100.00</td>
</tr>
<tr>
<td>AFIVA w/SSL Prior</td>
<td>0.150</td>
<td>0.635</td>
<td>6.041</td>
<td>50.50</td>
</tr>
<tr>
<td>NGIVA w/Liang's Prior</td>
<td>0.370</td>
<td>0.384</td>
<td>6.164</td>
<td>53.00</td>
</tr>
<tr>
<td>FPIVA w/Liang's Prior</td>
<td>-0.350</td>
<td>-0.030</td>
<td>4.905</td>
<td>47.25</td>
</tr>
<tr>
<td>RTIVA w/Liang's Prior</td>
<td>2.710</td>
<td>2.715</td>
<td>1.113</td>
<td>100.00</td>
</tr>
<tr>
<td>AFIVA w/Liang's Prior</td>
<td>2.465</td>
<td>2.104</td>
<td>5.606</td>
<td>60.75</td>
</tr>
</tbody>
</table>

Figure 85 shows box plots of SDR improvement by algorithm using the SSL prior with the corner room model. Figure 86 shows the same plots using Liang’s prior. In general, the same characteristics that were observed in the small and large rooms are present in the corner room.
Figure 85. Box plots of SDR improvement by algorithm using the SSL prior with the corner room model.

Figure 86. Box plots of SDR improvement by algorithm using Liang’s prior with the corner room model.
4.3.3 SAR Change with the Corner Room Model

Table 15 shows the overall change in SAR between the mixture $x_t$ at the microphone pointed at the speaker of interest and the separated source estimate $u_j$ of the speaker of interest. The data are collected from the same experiments as are SIR and SDR data above. With a Pearson correlation coefficient of 0.996, mean change in SAR by algorithm between large and corner room models is directly correlated. However, SAR is slightly lower for all but the real time algorithm.

Table 19. Overall change in SAR with the corner room model.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Median (dB)</th>
<th>Mean (dB)</th>
<th>Std. (dB)</th>
<th>Improved (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NGIVA w/SSL Prior</td>
<td>-11.780</td>
<td>-10.972</td>
<td>5.394</td>
<td>2.25</td>
</tr>
<tr>
<td>FPIVA w/SSL Prior</td>
<td>-12.180</td>
<td>-11.362</td>
<td>5.380</td>
<td>1.25</td>
</tr>
<tr>
<td>RTIVA w/SSL Prior</td>
<td>-0.865</td>
<td>-0.963</td>
<td>0.883</td>
<td>15.75</td>
</tr>
<tr>
<td>AFIVA w/SSL Prior</td>
<td>-11.105</td>
<td>-10.205</td>
<td>6.456</td>
<td>7.50</td>
</tr>
<tr>
<td>NGIVA w/Liang's Prior</td>
<td>-10.745</td>
<td>-10.245</td>
<td>6.330</td>
<td>3.50</td>
</tr>
<tr>
<td>FPIVA w/Liang's Prior</td>
<td>-11.495</td>
<td>-11.026</td>
<td>5.264</td>
<td>1.00</td>
</tr>
<tr>
<td>RTIVA w/Liang's Prior</td>
<td>0.230</td>
<td>0.018</td>
<td>0.663</td>
<td>65.25</td>
</tr>
<tr>
<td>AFIVA w/Liang's Prior</td>
<td>-8.895</td>
<td>-8.669</td>
<td>6.094</td>
<td>10.00</td>
</tr>
</tbody>
</table>

Figure 87 shows box plots of SAR improvement by algorithm using the SSL prior with the corner room model. Figure 88 shows the same plots using Liang’s prior. In general, the same characteristics that were observed in the small and large rooms are present in the corner room.
Figure 87. Box plots of SAR improvement by algorithm using the SSL prior with the corner room model.

Figure 88. Box plots of SAR improvement by algorithm using the Liang’s prior with the corner room model.
4.3.4 WER Improvement with the Corner Room Model

Figure 89 shows the WER after decoding the clean speech $s_j$, mixture $x_i$ at the microphone pointed at the speaker of interest, and the separated source estimate $u_j$ of the speaker of interest obtained using the NGIVA, FPIVA, RTIVA and AFIVA algorithms with both SSL and Liang’s prior. The data are collected from the same experiments as are SIR, SDR and SAR data above. As is the case with the small and large room models, all combinations of separation algorithm and prior result in a decrease of mean WER relative to the mixture at the microphone pointed at the speaker of interest. Also like the small and large room models, the AFIVA algorithm outperforms the others approaching to within 3% of the mean WER of the decoded clean speech signal. The main difference between the corner and other room models is the lower WER of the decoded mixture at the microphone pointed at the speaker of interest. This is expected since the best line-of-sight interference trajectory in the corner room is much farther off the longitudinal axis than in the large room. This results in greater attenuation of interferers.
Figure 89. WER of decoded clean speech, microphone mixture, and separated source estimates using the corner room model.

Table 20 shows the overall improvement in WER between the mixture $x_i$ at the microphone pointed at the speaker of interest and the separated source estimate $u_j$ of the speaker of interest. Comparing Table 20 to Table 12 and Table 16 above, we see weak correlation in mean WER improvement by algorithm between corner and large or small room models with Pearson correlation coefficients of 0.78 and 0.79 respectively. The weakness of correlation is due to the relatively poor mean WER improvement seen from the NGIVA and FPIVA algorithms in the corner room. However, as was the case with the
small and large room models, all algorithms produced far more cases of WER improvement than degradation, but the benefit is least in the corner room.

Table 20. Overall improvement in WER by algorithm using the corner room model.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Median (%)</th>
<th>Mean (%)</th>
<th>Std. (%)</th>
<th>Improved (%)</th>
<th>Unaffected (%)</th>
<th>Degraded (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NGIVA w/SSL Prior</td>
<td>2.90</td>
<td>1.16</td>
<td>15.75</td>
<td>50.75</td>
<td>25.75</td>
<td>23.50</td>
</tr>
<tr>
<td>FPIVA w/SSL Prior</td>
<td>2.80</td>
<td>0.13</td>
<td>18.04</td>
<td>50.25</td>
<td>27.75</td>
<td>22.00</td>
</tr>
<tr>
<td>RTIVA w/SSL Prior</td>
<td>0.00</td>
<td>3.19</td>
<td>6.89</td>
<td>44.75</td>
<td>44.00</td>
<td>11.25</td>
</tr>
<tr>
<td>AFIVA w/SSL Prior</td>
<td>5.00</td>
<td>5.53</td>
<td>11.85</td>
<td>62.75</td>
<td>25.25</td>
<td>12.00</td>
</tr>
<tr>
<td>NGIVA w/Liang’s Prior</td>
<td>3.80</td>
<td>0.74</td>
<td>18.48</td>
<td>53.25</td>
<td>26.75</td>
<td>20.00</td>
</tr>
<tr>
<td>FPIVA w/Liang’s Prior</td>
<td>3.30</td>
<td>0.31</td>
<td>18.17</td>
<td>50.75</td>
<td>27.75</td>
<td>21.50</td>
</tr>
<tr>
<td>RTIVA w/Liang’s Prior</td>
<td>0.00</td>
<td>2.49</td>
<td>6.65</td>
<td>40.25</td>
<td>44.00</td>
<td>15.75</td>
</tr>
<tr>
<td>AFIVA w/Liang’s Prior</td>
<td>5.30</td>
<td>6.49</td>
<td>10.27</td>
<td>64.50</td>
<td>26.25</td>
<td>9.25</td>
</tr>
</tbody>
</table>

Figure 90 shows box plots of WER improvement by algorithm using the SSL prior with the large room model. Figure 91 shows the same plots using Liang’s prior. Comparing these box plots to the same for the small and large room models, we see good correlation in terms of medians and interquartile ranges. However, there are in general more outliers in the corner room and WER improvement is least.
Figure 90. Box plots of WER improvement by algorithm using the SSL prior with the corner room model.

Figure 91. Box plots of WER improvement by algorithm using Liang’s prior with the corner room model.
Chapter 5. Conclusions

With the primary contribution of this work completed (characterization of the ASR performance improvement that can be expected by combining state of the art BSS algorithms with a compact directional microphone array) we now move to drawing conclusions about the efficacy of the solution, factors effecting it, and directions for future work.

5.1 Efficacy of the Solution

In all of the interference environments modelled, and with all of the algorithms tested, the use of BSS with a compact array of directional microphone elements results in far more cases of WER improvement than degradation. In terms of both SIR and WER performance, the AFIVA algorithm used along with Liang’s prior yields superior results in all environments. In the relatively high interference small room environment, AFIVA with Liang’s prior yields an average SIR improvement of 9.77 dB with 99.75% of cases improved. WER under the same conditions is less than half that of a single directional microphone element with 89.5% of cases improved and only 3.25% of cases degraded. Even in the relatively low interference corner room environment, AFIVA with Liang’s prior yields an average SIR improvement of 6.64 dB with 93.0% of cases improved. WER under the same conditions is reduced by over one quarter from a single directional microphone element with 64.5% of cases improved and only 9.25% of cases degraded.

An example of a typical case of using AFIVA with Liang’s prior in the small room environment to improve WER is shown in Figure 92. The spectrograms of clean speech, the mixture at the microphone element pointed at the speaker of interest, and the separated
source estimate are shown in vertical tiles. The clean speech was recorded with squelch enabled explaining the noise free dark blue silence intervals. Interference is visible in the mixture at the microphone element pointed at the speaker of interest particularly in those silence intervals. This interference increased WER from 10.5% in the decoded clean speech signal to 36.8% in the decoded mixture. The interference is visibly attenuated in the source estimate after separation. The reduced interference level resulted in an improvement of WER from 36.8% in the decoded mixture to 15.8% in the decoded source estimate. This is decrease of 21.1% WER over using a single direction microphone, which is the mean improvement from Table 12 above achieved under these conditions.
Figure 92. Spectrograms of clean speech, the mixture at the microphone element pointed at the speaker of interest, and the separated source estimated produced by AFIVA.
5.2 Factors and their Effects

The interference environment has a clear effect on both separation quantified by SIR and ASR accuracy quantified by WER. Both SIR and WER improvement over a single directional microphone element are greatest in the strong interference environment of the small room model. Improvement is poorest in the weak interference environment of the corner room model. However, not all algorithms are effected in the same way by the room model. The NGIVA and FPIVA algorithms yielded exceptionally poor WER performance in the corner room model even though their SIR performance in the corner room correlated well with that of the large room.

In addition to the interference environment, analysis of variance shows that all algorithms in all environments are effected by both speaker and audio clip. Furthermore, with the exception of the RTIVA algorithm, there is good correlation between algorithms and priors with respect to SIR improvement by speaker. This suggests that tuning the BSS algorithm to the speaker may improve performance.

Finally, with respect to SIR improvement, we observe some interaction between algorithm and prior. While Liang’s prior yielded the best SIR improvement on average, the NGIVA algorithm is afflicted with negative outliers when used with Liang’s prior. The opposite is true of the AFIVA algorithm, which is afflicted with negative outliers when used with the SSL prior. This suggests that tuning the prior to the BSS algorithm may reduce outliers.
5.3 Directions for Future Work

Perhaps the most troubling outcomes of this characterization are the WER degradation outliers observed mainly with the NGIVA and FPIVA algorithms, but also to a lesser extent with the AFIVA algorithm. Even though a small number of cases result in degradation, we would like to see a solution where no harm is done. One known source of WER degradation is the insertion of artifacts into the signal by the BSS algorithm. Finding ways to mitigate this contamination will benefit WER performance. Further improvements may derive from tuning the algorithm to the speaker. Parameters such as learning rate may have different optimal settings depending on the speaker of interest. Finally, there is no reason to assume that the sparsity of the prior modelled by the multivariate probability density function shown in equation (68) is the same for all spectral coefficients. Nor is there any reason to assume that the set of probability density functions is the same for all speakers. Tuning the prior to a particular speaker’s temporal and spectral nuances may improve separation, reduce artifacts, and improve WER.
References


