2017

Physiology-based model of multi-source auditory processing

BOSTON UNIVERSITY

COLLEGE OF ENGINEERING

Dissertation

PHYSIOLOGY-BASED MODEL OF MULTI-SOURCE AUDITORY PROCESSING

by

JUNZI DONG

B.S., Washington University in St. Louis, 2011
M.S., Boston University, 2014

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2017

Approved by

First Reader

H. Steven Colburn, Ph.D.
Professor of Biomedical Engineering

Second Reader

Kamal Sen, Ph.D.
Associate Professor of Biomedical Engineering

Third Reader

Barbara G. Shinn-Cunningham, Ph.D.
Professor of Biomedical Engineering

Fourth Reader

Gerald D. Kidd, Jr., Ph.D.
Professor of Speech, Language and Hearing Sciences
Sargent College of Health and Rehabilitation Sciences

Fifth Reader

Virginia A. Best, Ph.D.
Research Associate Professor of Speech, Language and Hearing Sciences
Sargent College of Health and Rehabilitation Sciences

ACKNOWLEDGMENTS

I worked around a group of very kind and intelligent people who also happened to be wonderful teachers and role models. They've helped me grow as a person and a researcher, and I'm very grateful for the help and support they've given me.

I thank my co-advisors, Drs. Steven Colburn and Kamal Sen, for their patient guidance. I think very few PhD students feel the kind of genuine care and support that I received from Steve and Kamal, and I was able to grow out of a nervous and anxious first-year student in the open and nurturing environment that they created.

The companionship that my lab mates Jing Mi and Nathan Spencer provided was a source of tremendous support in lab, and I feel very lucky to have worked in length with them. I want to thank Jing especially, for being a colleague, friend, and confidant.

I want to thank my committee members Drs. Barbara Shinn-Cunningham, Gerald Kidd, and Virginia Best for their generosity and kindness in sharing their intellectual thoughts and time. I'm especially grateful to have Barb and Gin on the committee, who showed me inspiring models of passionate and kind female academics.

Outside of lab, the unwavering love, support, and understanding of my fiancée, Shoko Ryu, was more than I could ask for. My roommate, Bing Xia, was generous in sharing his time and technical knowledge, be it critiquing my presentation or teaching me how to use the cluster.

I thank the people who generously shared their work with me and made my projects possible: Dr. Ross Maddox, for obtaining the neural data that laid the foundation of my work; and Dr. Brian Fischer and Jose Pena, for sharing their model with me.

There were periods of time when I questioned my decision to pursue a PhD, but meeting and getting to know all of you has supported me through the toughest times, and I'm extremely grateful of taking the path that led me here.

PHYSIOLOGY-BASED MODEL OF

MULTI-SOURCE AUDITORY PROCESSING

JUNZI DONG

Boston University, College of Engineering, 2017

Major Professors: H. Steven Colburn, Ph.D., Professor of Biomedical Engineering

and

Kamal Sen, Ph.D., Associate Professor of Biomedical Engineering

ABSTRACT

Our auditory systems are evolved to process a myriad of acoustic environments. In complex listening scenarios, we can tune our attention to one sound source (e.g., a conversation partner), while monitoring the entire acoustic space for cues we might be interested in (e.g., our names being called, or the fire alarm going off). While normal hearing listeners handle complex listening scenarios remarkably well, hearing-impaired listeners experience difficulty even when wearing hearing-assist devices. This thesis presents both theoretical work in understanding the neural mechanisms behind this process, as well as the application of neural models to segregate mixed sources and potentially help the hearing impaired population.

On the theoretical side, auditory spatial processing has been studied primarily up to the midbrain region, and studies have shown how individual neurons can localize sounds using spatial cues. Yet, how higher brain regions such as the cortex use this information to process multiple sounds in competition is not clear. This thesis demonstrates a physiology-based spiking neural network model, which provides a

mechanism illustrating how the auditory cortex may organize up-stream spatial information when there are multiple competing sound sources in space.

Based on this model, an engineering solution to help hearing-impaired listeners segregate mixed auditory inputs is proposed. Using the neural model to perform sound-segregation in the neural domain, the neural outputs (representing the source of interest) are reconstructed back to the acoustic domain using a novel stimulus reconstruction method.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CRM ..... Coordinated response measure

CSII ..... Coherence-based speech intelligibility index

EPSC ..... Excitatory post-synaptic conductance

HRTF ..... Head-related transfer functions

ILD ..... Internaural level difference

IPSC ..... Inhibitory post-synaptic conductance

ITD ..... Internaural time difference

MLd ..... Midbrain

MSE ..... Mean-squared-error

NCM ..... Normalized coherence metric

PESQ ..... Perceptual evaluation of speech quality

SNR ..... Signal-to-noise ratio

SRM ..... Spatial release from masking

STD ..... Standard deviation

STOI ..... Short-time objective intelligibility

STRF ..... Spectral-temporal receptive field

TMR ..... Target-to-masker ratio

XCorr ..... Cross-correlation between spectrogram

CHAPTER ONE: INTRODUCTION

1.1 Using a modeling approach to understand auditory spatial processing in the physiological brain

1.1.1 Two modes of spatial hearing

Animals and humans alike rely on their spatial hearing abilities for survival. The most commonly talked about type of spatial hearing by auditory neuroscientists is the "cocktail party effect", which describes our amazing ability to listen in complex, multi-source listening environments (such as a cocktail party), and focus only on the sound source of interest to us. Considering the problem in a spatial setting, solving the "cocktail

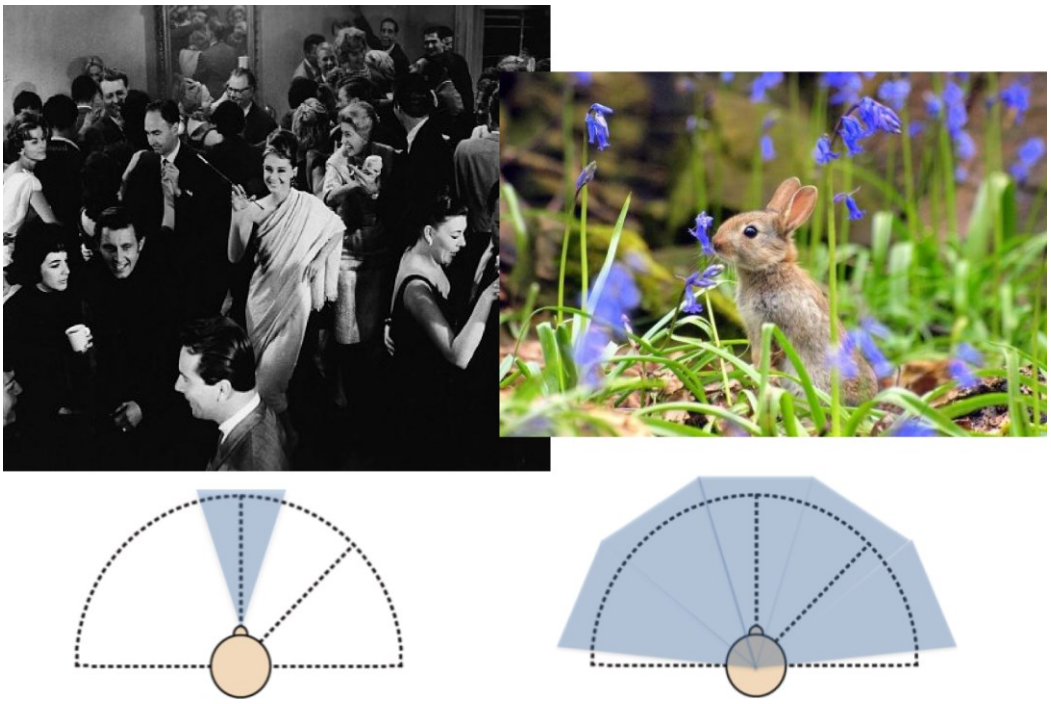

Figure 1 Two types of spatial hearing illustrated. Left: Cocktail party environments require a narrowly tuned auditory receptive field to separate the source of interest in space. Right: Monitoring the entire listening space to detect key events requires coverage of a broad field.

party problem" engages our brain's ability to focus on the narrow region of interest while ignoring the broader acoustic space. A less discussed type of spatial hearing is our constant monitoring of the entire listening space. Unlike the visual system, our auditory system is always on. This enables us to instantly become aware of the presence and location of novel acoustic stimuli, such as a car approaching from behind on a dark road.

What is especially remarkable is that we can switch between the two listening modes with extreme ease. In this doctoral dissertation project, I set out to model the neural circuitry behind these two modes of spatial listening, and build the physiology-based model into a sound-source segregation algorithm to help hearing-impaired listeners.

1.1.2 Current knowledge of spatial hearing

Spatial hearing faces unique challenges compared to spatial processing in the visual system, since spatial information is not inherent in the signals received by our ears. The signals arriving at our left and right ears are simply small mechanical vibrations representing the left and right, or binaural, acoustic signals. To identify the location of the sound source, the two most important spatial cues are interaural time difference (ITD) and interaural level difference (ILD). ITD is created when a sound arrives at the more proximal ear earlier than the more distal ear, while ILD is created when the head shadows the more distal ear, decreasing the loudness compared to the more proximal ear.

Physiological and modeling studies have revealed that auditory neurons in the midbrain are sensitive to ITD and ILD, therefore encoding for the spatial location of a single sound source in space. Fewer studies have looked at: 1) how higher brain regions

use the spatial information from the midbrain areas, and 2) how neurons in higher brain regions encode for multiple sources in space (e.g., in a cocktail party). A recent study by Maddox and colleagues (Maddox et al., 2012) demonstrated bimodal responses of bird cortical neurons to (1) single-source stimuli versus (2) two competing stimuli, suggesting that the same neuron can adopt a more selective spatial response only when a second competing source is present. Combining knowledge from midbrain spatial processing studies and the Maddox study of new observations in the cortex, the first part of this thesis project provides a mechanistic integrate-and-fire neural network model for spatial processing between the midbrain and cortex.

1.2 Spatial processing applications in hearing aids and cochlear implants

Hearing-assist devices, namely hearing aids and cochlear implants, have improved the speech communication abilities of hearing impaired and deaf users. Nevertheless, current devices are not effective when used in cocktail party-type environments, which are the most challenging hearing environments for users (Edwards, 2007). Recent improvements in wireless communication and computing power will soon enable the addition of spatial pre-processing in cochlear implants and hearing aids. With this clinical need and technical opportunity, I hope to apply the physiology-inspired spatial processing model to provide spatial pre-processing in hearing-assist devices.

Physiological auditory spatial processing is robust and versatile in both quiet environments and cocktail party scenarios. In contrast, the existing spatial processing solution in advanced hearing aids—usually a form of beamforming—lacks the versatility of switching freely between the two scenarios. When beamforming mode is switched on,

hearing aids are tuned to a narrow frontal field while blocking out sounds from all other areas. Even in the absence of any sound source from the frontal field, the user cannot hear stimuli outside the beamforming field. This is different from our natural hearing abilities, when we are always aware of sounds outside our region of interest, even when our attention is tuned to the frontal location. In addition to hindering ease of use, not being aware of surrounding sounds can be dangerous in certain real-life situations. For these reasons, a flexible, physiology-inspired spatial processing solution that can achieve both modes of processing would be an advantageous alternative. In the second and third parts of my thesis, I built an engineering solution for spatial processing based on others' and my physiological neural models, tested the speech processing capabilities of this engineering solution in both modes of hearing: single-sound source and competing sources (cocktail party), and demonstrate that the engineering solution achieves good performance in both scenarios.

1.3 Thesis project specific aims

The following project aims are reproduced from my prospectus document. The next section (**1.4 Thesis Organization**) indirectly describes the successful completion of the aims by summarizing the content of this thesis.

1.3.1 Aim 1: Construct spatial sound source segregation neural network inspired by physiological data

To better understand the central neural processing mechanisms behind spatial sound source segregation, the first goal of this project is to construct a neural network whose output matches the behaviors of recorded neurons. This network model will be

based on general knowledge of the early steps (midbrain) of spatial processing and designed to match the phenomenological cortical responses seen in the Maddox study (Maddox et al., 2012), or specifically, the following: 1) broad tuning to single target sounds; 2) emergence of spatially sensitive "hotspots" in response to target in competing masker, such that targets can be "extracted" from maskers at certain locations. Following the construction of the network model, specific testable physiological predictions will be made to encourage further physiological studies in this area. The success of Aim 1 is determined by the extent that the model is able to represent the available neural recordings with parameters adjusted for each neural recording.

1.3.2 Aim 2: Stimulus reconstruction for converting neural responses back to acoustic stimuli

To use the network for segregating sound sources in realistic applications, the segregated single-source neural response, which is the output of the network model, must be reconstructed back to acoustic waveforms. This can be done using stimulus reconstruction—a technique for mapping neural responses back to the stimulus waveform (Gabbiani and Koch, 1999, Mesgarani and Chang, 2012). The segregated and reconstructed stimulus can then be compared to the reconstructed multi-source, original multi-source, and reconstructed single-source stimuli to determine whether the network provides benefits in spatial hearing.

1.3.3 Aim 3: Peripheral model

The last aim is to develop an appropriate peripheral model that will provide spatially sensitive neural responses as inputs to the network model completed in Aim 1. The general processing behind this type of model, which generates spatially sensitive neural responses extracted from binaural cues, is widely understood. The peripheral model will be modified from a previously published model of spatially sensitive neurons in animal midbrains to fit the needs of the network model. We are looking at the Fischer model (Fischer et al., 2009). This aim will be considered complete when a model providing appropriate inputs to the network model whose spatial responses also matches known physiology is implemented.

1.4 Thesis organization

This thesis is mainly composed of two manuscripts. Chapter Two reproduces the manuscript of the physiology-based cortical network model of spatial processing published in eNeuron, summarizing the work done in Aim 1. This chapter describes the cortical network model, demonstrates that the model matches the behavior of the recorded neurons in the original physiology study in being able to achieve both broad encoding coverage without competition and selective tuning with competition, and lays out the plan for implementing the cortical network model to provide a bimodal spatial processing algorithm for hearing-assist devices. Chapter Three is the manuscript in preparation of the bimodal engineering solution for spatial processing. This chapter covers the architecture and key steps of the engineering solution, including midbrain localization, connection to the cortical network model, and the stimulus reconstruction

step to convert neural responses back to acoustic stimuli for human listeners. The chapter also includes the results of bimodal spatial simulations, demonstrating that the engineering solution can indeed cover the entire listening space without competition while blocking out unwanted direction when competition arises. This chapter describes the combined work of Aims 2 and 3.

The last chapter looks at future work in both physiology-based modeling towards understanding the auditory system and the applications of auditory spatial processing models. In the first area, this chapter will briefly talk about a new electrophysiology study in collaboration with another group to test and verify the cortical network model, and how to improve upon and modify the current model given potential experimental outcomes. In the applied direction, several ways to improve the engineering solution are discussed.

CHAPTER TWO: CORTICAL TRANSFORMATION OF SPATIAL PROCESSING FOR SOLVING THE COCKTAIL PARTY PROBLEM—A COMPUTATIONAL MODEL

2.1 Abstract

In multi-source, "cocktail party" sound environments, human and animal auditory systems can use spatial cues to effectively separate and follow one source of sound over competing sources. While mechanisms to extract spatial cues such as interaural time differences (ITDs) are well understood in pre-cortical areas, how such information is reused and transformed in higher cortical regions to represent segregated sound sources is not clear. We present a computational model describing a hypothesized neural network that spans spatial cue detection areas and the cortex. This network is based on recent physiological findings that cortical neurons selectively encode target stimuli in the presence of competing maskers based on source locations (Maddox et al., 2012). We demonstrate that key features of cortical responses can be generated by the model network, which exploits spatial interactions between inputs via lateral inhibition, enabling the spatial separation of target and interfering sources while allowing monitoring of a broader acoustic space when there is no competition. We present the model network along with testable experimental paradigms as a starting point for understanding the transformation and organization of spatial information from midbrain to cortex. This network is then extended to suggest engineering solutions that may be useful for hearing-assist devices in solving the cocktail party problem.

2.2 Significance statement

Spatial cues are known to be critical for human and animal brains when following specific sound sources in the presence of competing sounds, but the exact mechanism by which this happens is not clear. The role of spatial cues in localizing single sound sources in the midbrain is well documented, but how these extracted cues are used downstream in the cortex to separate competing sources is not clear. We present a computational neural network model based on recent recordings to bridge this gap. The model identifies specific candidate physiological mechanisms underlying this process and can be extended to construct engineering solutions that may be useful for hearing-assist devices for coping with the cocktail party problem.

2.3 Introduction

The problem of recognizing and processing individual auditory objects in complex listening environments, the "cocktail party problem", was recognized more than fifty years ago (Cherry, 1953); however, its neural mechanism remains poorly understood. Human and animal auditory systems selectively segregate and follow a selected sound source in the presence of competition to make sense of multiple-source environments (Bregman, 1994). Spatial cues enable listeners to segregate and follow individual sources, as demonstrated by human and animal studies (Hine et al., 1994, Dent et al., 1997, Darwin and Hukin, 1998, Arbogast et al., 2002, Dent et al., 2009). While pre-cortical neurons have been extensively shown to be selectively tuned to spatial cues such as interaural time difference (ITD) (Knudsen and Konishi, 1978, Yin and Chan, 1990, Pena and Konishi, 2001, Köppl and Carr, 2008, Devore et al., 2009), how spatial

information from spatial cue detection areas is relayed to and used in higher cortical areas is not clear (Vonderschen and Wagner, 2014). Recent experiments on cortical responses revealed that while spatial tuning for single sound sources is broad, simultaneous competing sources increase spatial selectivity (Maddox et al., 2012, Middlebrooks and Bremen, 2013). Although these findings shed light on the spatial encoding capabilities of the cortex, neural mechanisms capable of generating such capabilities remain unknown. The goal of this study is to provide a computational model consistent with existing physiological evidence to describe the transformation between pre-cortical areas and the cortex, which can selectively encode target stimuli when presented with competing sources in space. Specifically, we present a model network that replicates the spatial responses observed in a study by Maddox and colleagues (Maddox et al., 2012), providing a mechanistic solution to the spatial segregation of independent sources.

Maddox and colleagues demonstrated that, while the coding of song identity is not strongly impacted by stimulus location in quiet, location does have a significant effect on neural coding when there is a competing masker. In their experiments, two birdsongs were first presented independently from one of four stimulus locations (Fig. 2a). The neuron's spatial performance was studied using the discriminability index, a metric quantifying the neural coding of song identity at each location. A larger difference in neural responses to the two songs gives higher song discriminability, indicating a location where birdsong is more "intelligible" to the neuron. For the target song alone ("clean") case, similar discriminability across locations (Fig. 2a) indicates broad spatial tuning, where all spatial locations are similarly encoded within this neuron. In the masked

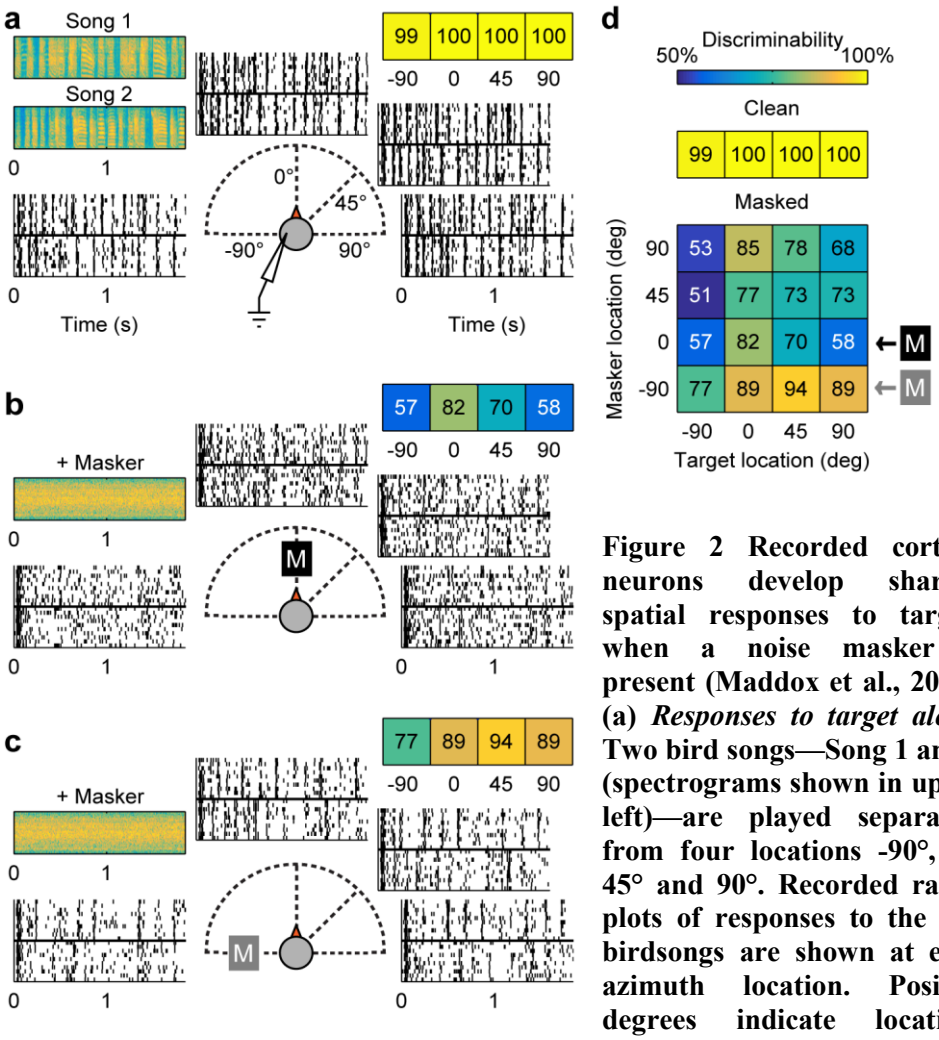

Figure 2 Recorded cortical neurons develop sharper spatial responses to targets when a noise masker is present (Maddox et al., 2012). (a) *Responses to target alone.* Two bird songs—Song 1 and 2 (spectrograms shown in upper left)—are played separately from four locations -90°, 0°, 45° and 90°. Recorded raster plots of responses to the two birdsongs are shown at each azimuth location. Positive degrees indicate locations contralateral to recording site. The color-coded discriminability values for each location are shown in the horizontal grid on the upper right. (Color map for all panels is shown in the top row of Panel d.) (b, c) *Responses to target with masker.* Masker and one target song are played concurrently from one (co-located) or two (separated) of the four stimulus locations. Panels b and c show a masker fixed at 0° or -90°, indicated by black or grey boxed M's, respectively, while the target song is played at one of the locations shown. As in Panel a, recorded raster responses from each target location are shown, and discriminability values are shown in the colored grid of values (upper right of panel). (d) *Discriminability values for all location combinations.* The top grid (single row) of numbers are the discriminability values for the "clean" (target-alone) conditions. In the lower, spatial discriminability grid, each block indicates a target and masker location combination. The rows indicated by black or grey boxed M's are cases where the masker is fixed at 0° or -90°. Blocks in all grids are colored according to the color scale given at the top of this panel.

conditions illustrated in Fig. 2, Panels b and c, a noise masker is played concurrently with a target, and the two are co-varied in location for all possible combinations. A spatial discriminability grid of responses to all recorded target and masker location combinations (Fig. 2d) shows that for this unit, discriminability is better at a few "hotspots" shaded in lighter colors. These patterns indicate a sharpened spatial preference for encoded song stimuli in the presence of a competing masker at these locations.

In this paper, our goals are to construct a model network capable of replicating key features of the experimentally observed cortical responses: i) similar discriminability for target songs in quiet at any location, indicating broad tuning and the ability of neurons to monitor the entire acoustic space in quiet; and ii) the emergence of "hotspots" where coding of song identity is enhanced at select stimulus locations in the presence of a second competing sound (the masker). The network can be adjusted to model a diverse range of spatial responses, demonstrated by fitting the population of neurons reported in the Maddox study. Finally, we propose a way to extend this network to design engineering solutions that may be useful for achieving spatial stream segregation in hearing-assist devices.

2.4 Methods

2.4.1 Network model overview

The network is composed of a three-layer structure, where the bottom layer receives pre-cortical input, and the final layer provides the cortical output, which is then compared to the recordings. The model architecture, model mechanisms and parameters, and simulated pre-cortical input are explained in separate sections below.

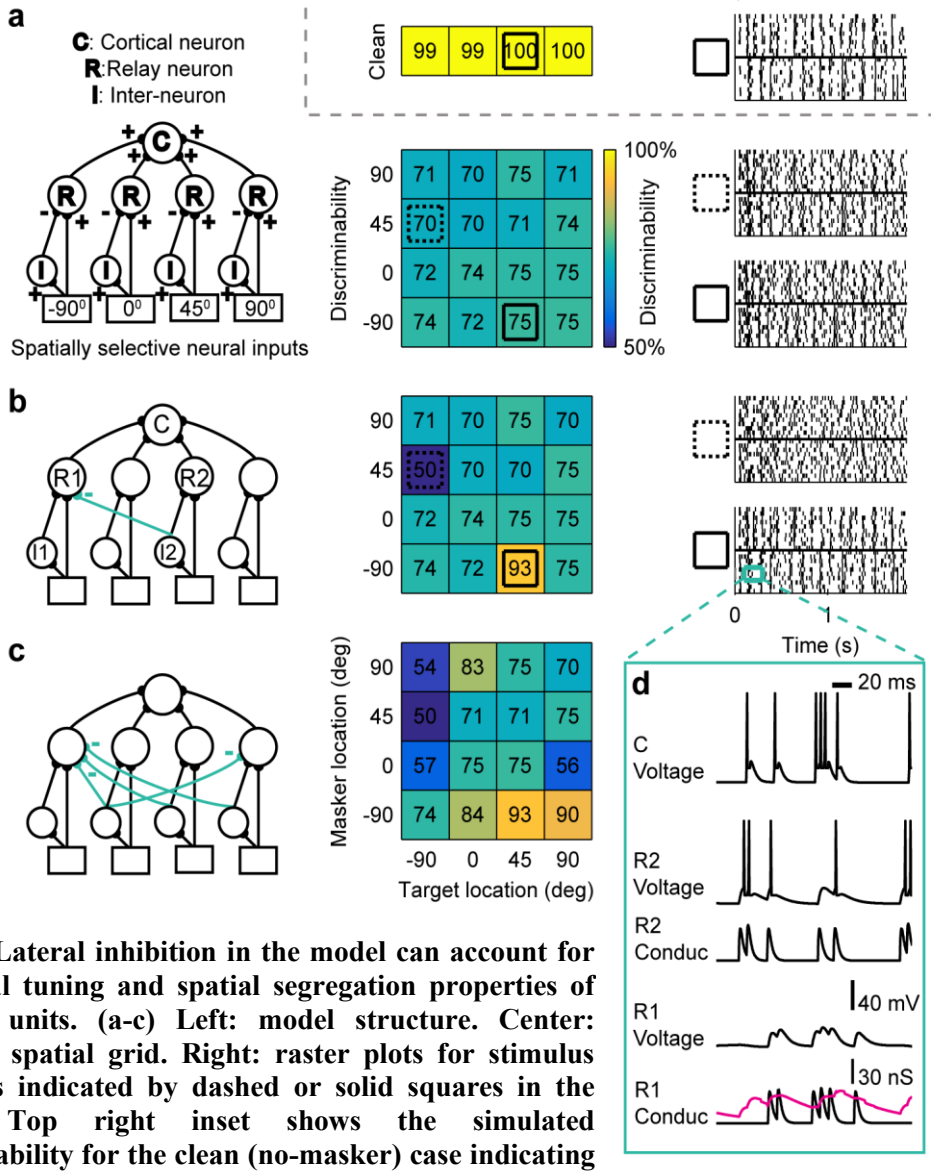

Figure 3 Lateral inhibition in the model can account for the spatial tuning and spatial segregation properties of recorded units. (a-c) Left: model structure. Center: simulated spatial grid. Right: raster plots for stimulus conditions indicated by dashed or solid squares in the grid. Top right inset shows the simulated discriminability for the clean (no-masker) case indicating broad spatial tuning. This clean case is not impacted by the addition of lateral inhibition, and is identical for all networks shown. (a) Basic model structure with no lateral inhibitory connections. Simulated multi-source spatial grid in model without lateral inhibition lacks the spatial diversity observed in the data. (b) Spatial grid produced by the model with one inhibitory connection between 0° and -90°, shows an increase in discriminability when target and masker are presented at 0° and -90°, respectively. (c) Model with additional inhibitory connections simulates the spatial response of the recorded unit shown in Figure 2d. (d) Sub-threshold responses of relay and cortical neurons, R1, R2, and C (Panel b, left), for the labeled time segment (Panel b, right) of one trial when target is presented at 0° and masker at -90°. Direct excitatory currents to R1 (R1 Conduc: black curve) are offset by inhibitory currents from I2 (R1 Conduc: magenta curve), and R1 is unable to reach spiking threshold, as seen in its voltage trace (R1 Voltage: black curve). In contrast, R2 is able to relay its temporal information to C, whose spiking pattern (C Voltage) resembles that of R2 (R2 Voltage).

2.4.2 Network model architecture

The structure of the model, which was custom written in Matlab, can be seen in Fig. 3 (Panels a, b, & c). The basic architecture consists of an input layer with four spatial input channels corresponding to -90, 0, 45 and 90 degrees to mirror the experimental design of Maddox et al., 2012, and an intermediate layer of processing that includes excitatory relay neurons (R) and inhibitory neurons (I), and an output cortical neuron (C). The detailed network connectivity is determined by the inhibitory connections as illustrated in Fig. 3. Our goal was to match the response of the output cortical neuron C in the model to the main features of the neurons recorded in the experiments by Maddox and colleagues.

Biological rationale. The convergence architecture was hypothesized based on physiological data showing selected spatial tuning responses in the midbrain (Knudsen and Konishi, 1978, Yin and Chan, 1990, Köppl and Carr, 2008), in contrast to the broad tuning observed in the cortex (Stecker et al., 2005, Higgins et al., 2010). The spectro-temporal response properties of the input layer neurons were modeled after experimentally measured spectro-temporal receptive fields (STRFs) of neurons in the avian midbrain (Amin el al., 2010; see section *Network model input* in Methods below). We modeled four spatial input channels as described above. In the biological system, there could be more input channels tuned to different locations at a finer spatial resolution. The spatial tuning of zebra finch midbrain neurons remains unknown. We began with the simplest assumption that there were no interactions across spatial input channels, and later relaxed this assumption to allow spatial overlap between the input

channels and demonstrated that the model remains robust over a range of spatial overlaps (see ***2.4.8 Spatial tuning width at the input stage*** in Methods below, also Fig. 5 of Results).

This model architecture is consistent with the inhibitory (and relay) neurons being located anywhere in the processing stream between the input (midbrain) neurons and the output cortical neuron. It is possible that the inhibitory (and relay) neurons are located in the thalamus. Inhibitory neurons have been found at the thalamic level in birds (Pinaud and Mello 2007) and some mammals (Winer, 1992). Alternatively, inhibitory (and relay) neurons might be located within cortex prior to the output cortical neuron. There is extensive evidence supporting the presence of inhibitory neurons at the cortical level, both in birds and mammals (Pinaud and Mello, 2007; Oswald et al., 2006).

2.4.3 Model Neurons

All neurons in the model are integrate-and-fire neurons. Specific parameters used are described below. Resting potential was -60 mV, spiking threshold was -40 mV, and the reversal potential for excitatory currents was 0 mV for all neurons. In relay neurons, the reversal potential for inhibitory currents was -70 mV. In inter-neurons, excitatory post-synaptic conductance (EPSC) was modeled as an alpha function with a time constant of one millisecond. In relay neurons, both EPSC and inhibitory post-synaptic conductance (IPSC) were modeled as the difference of a rising and a falling exponential, where rise and fall time constants were 1 and 3 ms, and 4 and 50 ms, respectively. An absolute refractory period of 3 ms was enforced in all neurons. These values are physiologically plausible (Froemke et al., 2007). In the cortical neuron, spike-rate

adaptation was implemented by a hyperpolarizing conductance term that increases after firing and then recovers to zero exponentially (Dayan and Abbott, 2001). The adaptation time-constant was 400 ms, and the strengths of the adaptation conductance for simulated neural units are shown in Table 1. Input synapses to the cortical neuron also have synaptic depression, which were modeled as described in Varela et al. (Varela et al., 1997). Although this quantitative formulation was applied to visual cortical synapses in

Table 1 Spectral temporal receptive fields (STRFs) input and adaptation conductance used for each simulated neural unit. STRF input and adaptation conductance were fit to best match the firing characteristics of each neuron recorded in the Maddox study, while other neuron modeling parameters were fixed as reported above.

STRF #	Neural Units	Adaptation conductance
1	3, 6, 9, 10, 11, 13, 21, 23	0.025
	14, 22	0.04
2	15	0
3	29	0.12
	27	0.1
4	7	0.07
5	19	0.06
6	2	0.06
7	5	0.2
	25	0.16
8	20, 32	0.07
	1, 12,, 33	0.08
9	16, 23	0.09
10	8	0.09
11	4	0.09
12	26, 28, 31	0.03
13	17	0.01
	18	0.03

Varela et al., synaptic depression is also observed in auditory thalamocortical circuits (Oswald et al., 2006, Rose and Metherate, 2005, Atzori et al 2001, Levy and Reyes

2012). We used a single synaptic depression component with fixed time course of 80 ms, and synaptic depression factor of 0.95, to model the experimental data in Maddox et al., 2012. Both adaptation and synaptic depression were implemented in the simulations shown in Fig. 3 and 5 for all modeled neurons.

2.4.4 Parameter fitting

Parameters were held constant throughout all simulations, except for the synaptic strengths and the strength of neural adaptation. To fit each recorded neuron, we first fit the general neural dynamics and baseline discriminability values by adjusting the strength of neural adaptation and the synaptic strengths without lateral inhibition. The specific values of neural adaptation used can be found in Table 1. The feed-forward synaptic weights (Input to Relay neuron) were then adjusted to match the discriminability values for clean and co-located cases at each azimuth, while other parameters were held the same. For lateral inhibition, the synaptic strength of each inhibitory connection was chosen to model the recorded discriminability of its corresponding song and masker location. Our goal in this study was to fit the spatial discriminability grids observed experimentally.

2.4.5 Network model input

The model input is composed of four spatial input channels corresponding to the stimulus locations used in the experiment by Maddox and colleagues (as shown in Fig. 3). Each channel receives simulated spike train responses of neurons at midbrain level as input. Input responses were simulated with spectro-temporal receptive fields (STRFs)

modeled after typical STRFs obtained from the midbrain (MLd) of zebra finch songbirds (Amin et al., 2010). The input generation process is illustrated in Fig. 4 and explained in detail below. For the majority of simulations, the azimuth response field for each modeled neuron was simulated with a Gaussian function and across the population there was minimal overlap between response fields. (Fig. 5, bottom left illustration). This no-overlap assumption effectively means that for the azimuth locations used in the experiment, neighboring sources are outside the spatial receptive field, and each input channel will only respond to stimuli from its corresponding location. The effect of wider spatial tuning was also studied by running separate simulations with wider, overlapping Gaussian inputs (see Results).

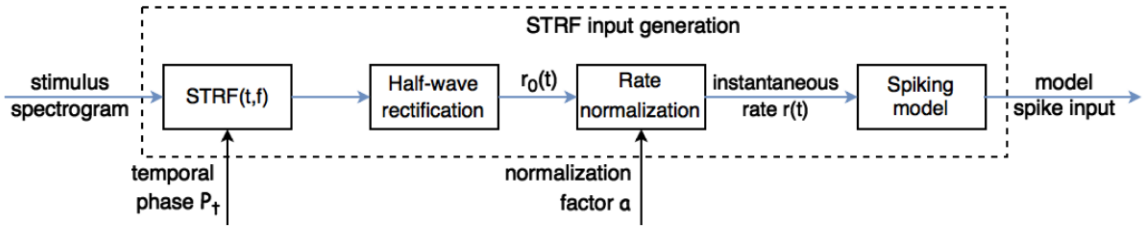

Figure 4 Illustration of model input generation process. The stimulus spectrogram was convolved with STRFs modeled after midbrain neurons, followed by half-wave rectification, then rate normalization to generate an instantaneous output firing rate. This firing rate was then used to generate spikes using a spiking model (see Methods for details). The values of temporal phase P_t and normalization factor a used were reported in Table 2.

2.4.6 Model input using spectral temporal receptive fields (STRFs)

STRFs were used to simulate input responses. These STRFs were modeled using the product of Gabor functions in the time and frequency domain (Qiu et al., 2003):

$STRF(t,f) = G(f) \cdot H(t)$, where

$G(f) = e^{-0.5[(f-f_0)/\sigma_f]^2} \cdot \cos[2\pi \cdot \Omega_f \cdot (f - f_0)]$, and

$$H(t) = e^{-0.5[(t-t_0)/\sigma_t]^2} \cdot \cos[2\pi \cdot \Omega_t \cdot (t-t_0) + P_t].$$

The frequency range is determined by f_0, the best frequency; σ_f, the spectral bandwidth; and Ω_f, the best spectral modulation frequency, which were chosen and fixed at 4300 Hz, 2000 Hz, and 50 μs, respectively, to generate a broadband STRF for all simulations based on physiological ranges reported in the midbrain (MLd) of zebra finch songbirds by Amin and colleagues (Amin et al., 2010). Temporal parameters t_0, the temporal latency; σ_t, the temporal bandwidth; and Ω_t, the best temporal modulation frequency, were assigned 7 ms, 4.5 ms, and 56 Hz, respectively, based on recorded physiological values (Amin et al., 2010).

Table 2 Parameters used for each type of input model STRF. Temporal phase P_t and normalization factor were adjusted to match the recorded responses of the corresponding neurons, while other temporal and spectral parameters are held fixed and reported above.

STRF #	Normalization factor	P_t (rad)
1	0.08	1.4608
2	0.1	1.4923
3	0.07	1.508
4	0.1	
5	0.12	1.5237
6	0.1	1.5394
7	0.07	1.5425
8	0.087	
9	0.15	1.5582
10	0.05	
11	0.08	
12	0.16	1.5598
13	0.17	1.5708

The normalization factor and temporal phase (P_t) were varied to match the neuron-specific raster responses seen in the neural recordings of the Maddox study. Other STRF parameters were largely fixed for simplicity, but the model is robust to variations in these parameters. Specific values of used parameters are shown in Table 2.

2.4.7 STRF modeled input spike trains

As shown in Fig. 4, STRFs were first converted to firing rates by convolving the stimulus spectrogram with the model STRF and half-wave rectifying so that rate outputs were positive. For each simulated neuron, the firing rate was normalized by factor a to adjust the final mean firing rate: $r(t) = a \cdot r_0(t)$. Finally, a Poisson spike model with a refractory period of 6 ms generated the neural response spikes used as the network model inputs, consistent with the instantaneous rates.

2.4.8 Spatial tuning width at the input stage

Spatial tuning width at the midbrain level varies across species, and is notably broader relative to the behavioral tuning for some mammals (Vonderschen and Wagner, 2014). To investigate whether the network model is functionally feasible with broader spatial tuning, the effect of spatial tuning width variation was studied by running simulations on an example neural unit and its neural network. The spatial tuning curves of input neurons were assumed to be Gaussian functions with varying standard deviations, as shown in Fig. 5. Tuning widths (twice the standard deviation σ) of 15° or smaller result in no crosstalk between the input channels separated by 45°, as implemented in the main experiment. For the model unit used to test the effect of overlap

(Unit 2 in Table 1), the tuning was then increased to show differences in model responses.

2.4.9 Discriminability index: evaluating stimulus encoding and spatial tuning

The discriminability index calculates the level of dis-similarity between spike trains generated in response to two songs (Wang et al., 2007). For both sets of ten spike trains recorded from the same neuron, a random spike train from each song is chosen as a template, and the remaining spike trains are assigned to the closest template based on the van Rossum spike distance metric, which measures discrimination between two spike trains (van Rossum, 2001). This yields a perfect discriminability of 100% for an ideal response pair, and a chance discriminability of 50% for an indiscriminate response pair.

2.4.10 Quantifying goodness of fit

To assess the fit of the model to individual units from the original study, we calculated the average deviation and correlation coefficient between the discriminability values for clean and masked responses of the data and that of the simulation. The average deviation is the mean value of the absolute difference between each corresponding discriminability value.

2.5 Results

2.5.1 Cross-channel lateral inhibition enables the network to match experimentally observed neural responses

As described in Methods, a multi-layer network model (Fig. 3a, left) of integrate-and-fire neurons was constructed to replicate selective spatial responses to competing

sound sources. Input layer neurons represent neurons at the spatial cue detection level, and receive inputs generated by the model in Fig. 4 when a stimulus is presented at the corresponding location (see Methods). Thus, there are four input "channels" corresponding to each speaker location in the experiment. The four input units excite four corresponding channels of relay neurons and inter-neurons in the middle-layer, which inherit their spatial tuning. Relay neurons converge to excite the cortical neuron (Fig. 3a, left), making it broadly tuned to stimuli from all directions in the clean (i.e., no masker) case (discriminability grid shown in Fig. 3a inset), as observed in the data (Fig. 2a; also see ***2.4.2 Network Model Architecture*** in Methods). However, in this network (Fig. 3a, left), the spatial discriminability grid is relatively uniform (Fig. 3a, center column), unlike that observed in the data (Fig. 2d). Thus, this basic network replicates the broad response in the target alone case, but fails to produce the configuration-dependent "hotspots" observed in the data.

Introducing lateral inhibition from inter-neurons across spatial channels allows the target response to suppress the masker response when presented at the tuned locations, generating a "hotspot" of performance for a given target and masker location combination (Fig. 3b). Figure 3d depicts the sub-threshold conductance and voltage changes in the relay and cortical neurons in the expanded time segment. While neuron R2 spikes predictably in response to increases in excitatory post-synaptic conductance (EPSC), R1 is unable to spike following its EPSC input due to long-lasting suppression by lateral inhibition as seen in the increase in inhibitory post-synaptic conductance (IPSC) (Fig. 3d, bottom, magenta trace) from I2. In this case, the voltage response of the

cortical neuron resembles that of R2 and the 0° target input (Fig. 3d). This is seen in the raster plots for the same stimulus paradigm, which resembles the target alone condition (Fig 3b, bottom right), indicating that the cortical neuron is able to follow the target and largely ignore the masker. Note that when the locations of target and masker are reversed, discriminability decreases due to the masking of target by noise (Fig. 3b, center and top right). The preferred spatial location combinations in the recorded unit (Fig. 2d) can be modeled by introducing additional lateral inhibitory connections as shown in Fig. 3c.

By adjusting model parameters, we were able to satisfactorily fit 32 out of 33 units recorded in the original study. The model was largely robust in the parameter ranges we tested (see Methods for details). We used two parameters to assess the closeness of fit between each unit and its model simulation. Average deviation measures the closeness of the discriminability values of the simulation compared to the data in units of discriminability percentage, and was 3.39±0.97% for all simulated units. The correlation coefficient ranging from -1 to 1 measures how closely the pattern of the simulated grid agrees with the experimental grid, and was 0.94±0.04 for the simulated units. The neural unit that did not have an overall satisfactory fit had a spatial grid that was very uniform, where discriminability variations within the grid were small and random. As a result, the simulated fit had a deviation value within the normal range, but a very low cross-correlation coefficient.

It is noteworthy that the model network without lateral inhibition showed a relatively uniform spatial grid (Fig. 3a, center column), unlike the experimental data. This network did include adaptation and synaptic depression (see Methods section ***2.4.3***

Model neurons). Thus, without lateral inhibition, adaptation and synaptic depression are not sufficient to explain the experimentally observed "hotspots" in the spatial grid.

2.5.2 Spatial tuning

The sharpness of spatial tuning curves was varied to test whether the model can describe the data with broader spatial input at the midbrain level. In the initial simulations, we assumed no crosstalk between spatial channels, which corresponds approximately to a Gaussian spatial tuning curve of width 2σ (twice the standard deviation σ) less than 15°, where σ is the standard deviation of the Gaussian function. As the width increased, more and more overlap occurred between channels, as shown in the left column of Fig. 5.

For the simulations shown in Fig. 5, spatial tuning width 2σ was increased to 40°, 80°, and 120°, respectively, while keeping all other parameters identical. The results of broadened tuning widths are shown in Fig. 5b, c. The goodness of fit, as quantified by deviation and cross-correlation coefficient, diminished as tuning width was broadened. The mean and standard deviation of these two measures calculated from the population of simulated units, is plotted as dotted lines and shaded areas in Fig. 5b for reference. In the 40° case, both deviation and cross-correlation coefficient remain within the range for the population of simulated units. The spatial tuning grid for 40° seen in Fig. 5c (lower panel) also maintains the general features of the data (Fig. 2d) and the original minimum overlap simulation (Fig. 3c, center). Therefore, this network model remains robust when spatial tuning width is increased to 40°. Even at a spatial tuning width of 80°, which corresponds to a fairly large overlap, the correlation coefficient remains relatively high at

0.91 and the deviation relatively low at 6.42% (Figure 5b). Thus, the model remains robust for spatially overlapping tuning curves, degrading gracefully at very high overlaps (e.g., 120°, Fig 5c upper panel).

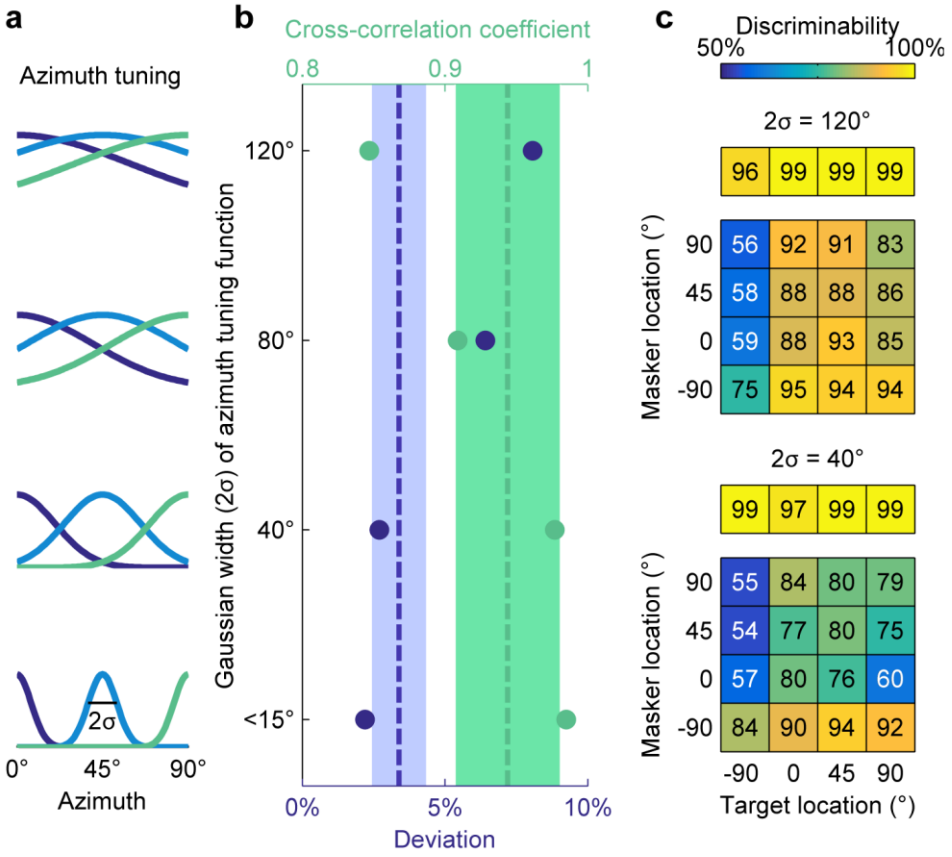

Figure 5 Network performance is robust to broader spatial tuning of inputs, as shown by extended simulations on the example unit previously displayed in Fig. 3. (a) Illustrations of Gaussian spatial tuning curves of varying widths, defined by twice the standard deviation (2σ). (b) Results of spatial grid simulations for broadened input tuning with 2σ at 40°, 80°, and 120°, compared to the no-overlap case (<15°) on the bottom. The cross-correlation coefficient and deviation of the simulated results are plotted in green and purple, respectively, on separate horizontal axes. On the cross-correlation coefficient axis (top), larger values (closer to unity) indicate a better fit, while the deviation axis (bottom) shows better fits at smaller values closer to 0%. For reference, dotted lines and shaded areas indicate the mean and standard deviation of cross-correlation coefficient and deviation values, for the original simulated population using non-overlapping inputs. As the spatial tuning of input units was broadened from <15° to 120°, the correlation coefficient (green dots) and the deviation (purple dots) degraded gracefully. The correlation coefficient remained above 0.8 and the deviation remained below 10% for the broadest tuning width. (c) Illustrations of simulated spatial grids with input widths of 40° and 120°. The 40° spatial grid can be compared to the no overlap spatial grid shown in Fig. 3c. The two grids show a similar visual pattern, which is quantified by the similar deviation and cross-correlation coefficient values shown in Panel b. The 120° grid maintains the general pattern but has overall higher discriminability throughout.

2.5.3 Extending the model network to potential engineering solutions for segregating spatial sound sources

The network can be extended to provide an engineering solution to the problem of segregating target from noise in space for the maximal number of locations on the grid. Figure 6a demonstrates a network where good discriminability is obtained for all conditions with target location to the right of masker location. This network, together with a complementary network with high performance for grid positions above the diagonal, allows the segregation of non-colocated sources for any azimuth, while maintaining consistently high intelligibility when only one source is present. An alternative engineering solution is demonstrated in Fig. 6b, where one channel acts as a beamformer by inhibiting all other channels. In this case, similar networks beaming at other directions will enable a user to selectively listen to any direction of interest.

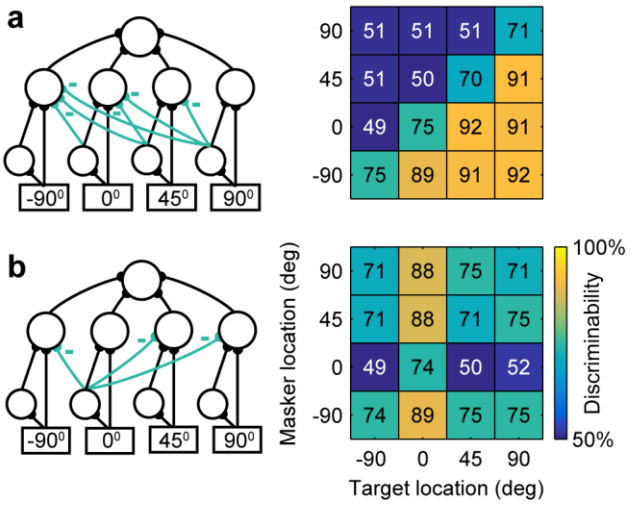

Figure 6 Engineering solutions. (a) Left: "contralateral-dominance" model network where all channels contralateral to the dominant channel are inhibited; Right: simulation results of this structure achieve the maximum number of spatially separable target and masker locations, where all targets contralateral to masker can be segregated. (b) Left: "beamformer" model network where the channel tuned to the front (0°) inhibits all other channels; Right: the simulated spatial grid illustrating the segregation of the frontal target source.

2.6 Discussion

The network model used here provides an explicit way of generating neural responses that replicate the key features of the cortical neurons recorded by Maddox and colleagues (Maddox et al., 2012), and provides a neural strategy for transforming information into selective coding for sound sources in the presence of multiple sources. The network uses information from input neurons through individual spatial channels and matches the key experimental features through convergent excitation and lateral inhibition across spatial channels.

2.6.1 Predictions and implications

2.6.1.1 Lateral inhibition

The model suggests that lateral inhibition plays an important role in spatial sound source segregation. While lateral inhibition is a widely known mechanism in the brain, to our knowledge this study is the first to demonstrate how it can be exploited in the context of the cocktail party problem. Inhibition is present in Field L as well as the mammalian primary auditory cortex (Müller and Scheich, 1988, Wehr and Zador, 2003). Recently, there has been evidence of suppression by spatially separated stimuli in the cortex of marmoset monkeys (Zhou and Wang, 2012, 2013), which could be a manifestation of the lateral inhibition postulated in the model.

Given this network, we propose a physiological experiment that may provide additional insights. One can experimentally test the nature and source of inhibition by locally blocking GABA receptors and measuring the spatial grid under the same

experimental setup. If the recorded spatial grid becomes less spatially sensitive, the proposed lateral inhibitory connections are most likely local.

2.6.1.2 Exploring alternate mechanisms for spatial sound source segregation

The above simulations show that the sharpened spatial tuning in the presence of multiple sources, which allows for spatial stream segregation, can be achieved via lateral inhibition across spatial channels. An alternate mechanism for spatial streaming, proposed in a recent study (Middlebrooks and Bremen, 2013) is forward masking. Candidate neural mechanisms underlying forward masking are adaptation and synaptic depression. The network model used here incorporated both of these mechanisms to model the temporal dynamics of the cortical responses. Our simulations indicate that while these mechanisms are important in determining the temporal dynamics of neural responses, they alone fail to produce the diverse spatial grids seen in the Maddox study due to a lack of cross-channel spatial interactions (Fig. 3a, middle panel). In particular, without lateral inhibition, the model does not replicate the hotspots seen in the experimentally observed spatial grid. Thus, lateral inhibition involving interactions *across* spatial channels is necessary in the model for replicating the spatial properties in the observed data.

2.6.1.3 Response to multiple maskers

For each recorded unit, looking at its single-masker spatial grid response provides predictions for how it might respond to multiple maskers. In Fig. 3c, for example, the simulated neuron is robust to maskers presented from both -90° and 90° (independently) when the target is located at 0°. This is achieved in the model network by inhibitory

connections from 0° to -90° and 90°, which means that target stimuli at 0° could mask 2 simultaneous noise sources from -90° and 90°. Consistent with this intuition, our simulated network for this unit was robust to simultaneous maskers from -90° and 90°. It should be possible to test such predictions by performing two-masker experiments physiologically, and comparing the results to those of single-masker cases for each neuron.

2.6.1.4 Potential engineering solution to the cocktail party problem

The engineering solution visualized in Fig 6b is robust to simultaneous maskers in all channels other than the target (in this case three simultaneous maskers at -90°, 45° and 90°), making this a particularly attractive design option in the context of hearing-assist devices in the presence of multiple speakers.

We plan to use the proposed engineering solution networks in Fig. 6 to segregate mixed-source acoustic stimuli by building a system that can take mixed-source acoustic inputs and output a single desired acoustic source. This will require two additional processing steps. First, a peripheral model that converts acoustic stimuli into neural representations consistent with the network input is needed. This will be a model where neurons selectively respond to a preferred direction using interaural cues, similar to previous neural models of spatial tuning (Fischer et al., 2009). Second, the neural network output, i.e., spike trains representing the single desired source, needs to be converted back into acoustic waveforms. This can be done using stimulus reconstruction (Mesgarani and Chang, 2012). We are working on both steps with the long-term goal of

ultimately testing the segregation capabilities of the model on normal and hearing-impaired listeners.

2.6.2 Spatial tuning of inputs and applicability of model to spatial processing in birds and mammals

For the majority of simulations, input neurons are assumed to have non-overlapping Gaussian spatial tuning curves centered at azimuths corresponding to those used in the experiment. A separate set of simulations showed that the model network remains robust when the spatial tuning curves are broadened to have significant overlap.

Spatially selective neurons found in the owl midbrain (Knudsen and Konishi, 1978, Pena and Konishi, 2001) and chicken hindbrain (Köppl and Carr, 2008) demonstrate ITD sensitivity within the physiological range. Although spatial tuning of midbrain neurons in the zebra finch remains unknown, it is likely that the auditory periphery contains similarly spatially sensitive neurons like other avian species, as spatial tuning appears to follow an evolutionary divide across species (Schnupp and Carr, 2009, Ashida and Carr, 2011). An outstanding question is whether the model will hold for species whose midbrain neurons show broader spatial sensitivity, such as small-headed mammals where tuning curves span an entire hemisphere or more (Vonderschen and Wagner, 2014). As we tested, the selective mechanism remains robust when spatial tuning is widened up to 40° (Fig. 5), comparable to some azimuth ITD tuning functions recorded in the rabbit IC by Day and colleagues (Day et al., 2012).

In species that show broad spatial tuning in the midbrain, spatial tuning may be further sharpened within the cortical level. One possibility is that broad spatially tuned

pre-cortical inputs are sharpened by a high threshold at the cortical level. A second possibility is that the spatial tuning of cortical neurons is sharpened during active engagement in a task (Lee and Middlebrooks, 2011). In this case, the authors proposed a top-down activation of inhibitory mechanisms as a potential mechanism. The Maddox experiments were in an anesthetized preparation, so lacking top-down activation, but it is possible that sharpening of tuning via lateral inhibition can be elicited by top-down activation (e.g., during active engagement), or bottom-up activation (e.g., in the multiple source condition). A third possibility is that for neurons with broad spatial tuning, the hypothesized spatially tuned inputs may be achieved through population coding, i.e., computations based on effective pooling across input neurons.

The neurons in the experiments by Maddox and colleagues (Maddox et al., 2012) were recorded in field L of the zebra finch, the analog of mammalian primary auditory cortex. Although the strict homology between auditory areas in birds and mammals is still debated, the functional properties of Field L neurons, e.g., spectro-temporal receptive fields, are similar to those observed in mammalian auditory cortex (Sen et al., 2001). In addition, the trend of less spatial specificity for single sources from primary spatial cue detection areas to higher cortical areas appears common across mammalian and bird species (Vonderschen and Wagner, 2014), for which this study provides a possible explanation. Thus, the model described here may explain some of the general properties of cortical neurons in other systems.

2.6.3 Population coding and readout

The network presented here suggests that in the presence of multiple sound sources, cortical neurons can "selectively listen" to particular target sources, which correspond to "hotspots" of performance on the spatial grid. A population of such neurons, for different locations in space, would enable spatial streaming over a range of locations. This is consistent with the diversity of spatial grids with hot spots at different locations observed in the experimental data (Maddox et al., 2012). The experimental data were obtained in anesthetized animals, suggesting that such a population representation is "pre-attentive". Attention may facilitate the proper readout from this cortical population by selecting the appropriate neuron(s) for given target and masker locations.

2.7 Concluding remarks

In this study we presented a computational model describing how the auditory cortex may transform spatial representations to solve a key aspect of the cocktail party problem. The computational model is based on physiological data (Maddox et al., 2012) and makes two key predictions that can be tested experimentally. First, the model predicts that lateral inhibition is a core mechanism underlying spatial sound source segregation. It would be interesting to further elucidate the nature and the location of such inhibition in similar experiments by pharmacologically blocking local GABA receptors. Second, the model predicts that some cortical neurons will remain robust when additional maskers are added in select locations predicted by the model. This can be tested in experiments on spatial selectivity of cortical neurons with three or more sound sources.

In addition to testing these key experimental predictions, it will also be interesting to implement the engineering solutions discussed in the paper and test if the proposed circuit can successfully segregate sounds sources and improve listening performance in normal and hearing impaired listeners in cocktail-party-like settings.

CHAPTER THREE: A BIMODAL ENGINEERING SOLUTION FOR ASSISTED SPATIAL PROCESSING

3.1 Introduction

Animals and humans alike rely on spatial hearing abilities for survival. The most commonly talked about type of spatial hearing by auditory neuroscientists is the "cocktail party effect", which describes our amazing ability to listen in complex, multi-source environments (such as a cocktail party) and to focus only on the sound source of interest to us. A less discussed type of spatial hearing is our constant monitoring of the entire listening space, which we rely on for detecting novel stimuli and events. Physiological auditory spatial processing is robust and versatile in both quiet environments and cocktail party scenarios, and gives us the ability to narrow our field of hearing when needed while being aware of the entire acoustic space. In contrast, existing spatial processing solutions in advanced hearing aids—usually a form of beamforming—lacks the versatility of switching freely between the two scenarios. When beamforming mode is switched on, hearing aids are tuned to the frontal field while sounds from outside the beamforming field are attenuated or blocked out. Even in the absence of any sounds from the front, stimuli outside the beamforming field are greatly attenuated. In addition to hindering ease of use, not being aware of surrounding sounds can be dangerous in certain real-life situations. To our knowledge, no spatial processing algorithms have addressed this problem. Alternative spatial processing algorithms have been under active development, which typically use a beamforming array of microphones to separate sounds rather than

using natural binaural inputs (Desloge et al., 1997, Roverud et al., 2016). Considering these factors, a flexible, physiology-inspired spatial processing solution that can achieve both modes of processing would be an advantageous alternative. In this paper, we present a physiology-based, bimodal engineering solution for spatial processing, and demonstrate robust speech processing in both modes of hearing: single-sound source and competing sources (cocktail party).

The engineering solution explored here is based on current understandings of spatial processing from peripheral neurons to cortical brain areas. Auditory midbrain neurons are known to respond preferentially to certain preferred locations in azimuth in both birds and mammals (Knudsen and Konishi, 1978, Yin and Chan, 1990). In mammals, there is some debate (Vonderschen and Wagner, 2014) about the sharpness of ITD tuning and additional mechanisms beyond the simple coincidence-based, interaural time delay (ITD) sensitive mechanism proposed by Jeffress (Jeffress, 1948). Overall, the Jeffress model is generally accepted, especially in birds (Ashida and Carr, 2011, Vonderschen and Wagner, 2014). In the proposed engineering solution, we used a Jeffress-type, owl-based localization model (Fischer et al., 2009) to provide midbrain inputs to a song-bird-inspired, cortical network model (Dong et al., 2016). Although less is known about cortical spatial processing, our recent study (Dong et al., 2016) proposed a physiology-based cortical network model capable of encoding for the entire acoustic space in quiet, and selectively encoding preferred locations only when competition arises. Using this cortical network model as a basis for a bimodal spatial processing, we built a combined neural network model by adding a midbrain localization model as input. The

midbrain peripheral localization model (Fischer et al., 2009) provides 'physiological beamformed' neural inputs to the cortical network model, which integrates responses from the full acoustic space while selectively enhancing directions of interest. Finally, spatially processed neural responses are reconstructed back to the acoustic domain using a novel stimulus reconstruction technique.

3.2 Methods

3.2.1 Engineering solution architecture

The engineering solution takes binaural acoustic inputs and generates spatially processed acoustic signals for human listeners. The spatial processing is performed in the neural domain using a combined neural model (Fig. 7, Boxes 1 and 2) composed of midbrain localization models (Fig. 7, Box 1) and cortical network models (Fig. 7, Box 2). After the combined model, the spatially processed neural responses are reconstructed (Fig. 7, Box 3) back to acoustic signals. The neural models are specific for each frequency channel, and there are no interactions between frequencies. There are 36 combined neural models for the 36 frequencies modeled on the equivalent-rectangular bandwidth (ERB) scale, from 300 to 5000 Hz.

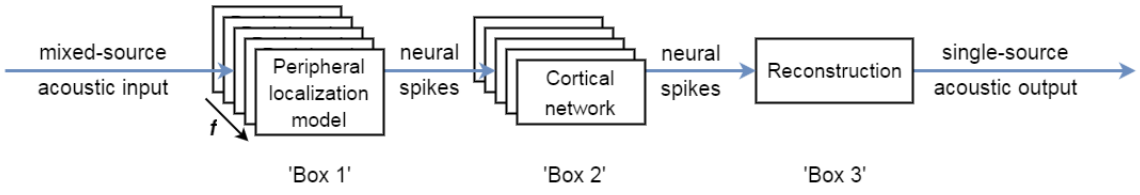

Figure 7. Using the cortical network to construct an engineering solution to spatial sound processing involves the addition of two modules. The peripheral localization model (Box 1) processes binaural inputs into spatially-sensitive neural responses, the cortical network (Box 2) generates bimodal spatial responses representing the direction of interest as described in earlier chapters, and the final reconstruction step (Box 3) converts the extracted neural code back to an acoustic waveform for human listeners.

3.2.2 Midbrain localization model

The owl midbrain is a well-studied and modeled example of the Jeffress localization mechanism. Here, we adapted a localization model based on the inferior colliculus of the barn owl (Fischer et al., 2009). The structure of the interaural time difference (ITD) and interaural level difference (ILD) detection mechanisms were kept, while the tuning parameters to ITD and ILD were modified to match the human physiology range. To modify the tuning parameters, we calculated the azimuth-specific ITD and azimuth- and frequency-specific ILD of Kemar head-related transfer functions (HRTFs) for the azimuth locations used as input to the cortical network model (-90, -45, 0, 45, and 90 degrees). For each preferred azimuth, we adjusted the ITD and ILD tuning parameters to match the ITD and ILD calculated from Kemar HRTFs for that azimuth and frequency. Matlab code for midbrain localization model was generously provided by Brian Fischer, and edited with custom modifications.

3.2.3 Cortical network model

The cortical network models shown in the top row of Fig. 8 assume that a cortical neuron integrates an array of midbrain localization neurons to encode for the entire acoustic field. To achieve spatial selectivity, inter-neurons (bottom layer of circles) of the preferred direction send lateral inhibitory connections to non-preferred directions. This results in stimuli from non-preferred directions being inhibited and silenced when there is a stimulus present at the preferred location. However, when there is no competition, all spatial locations can be encoded equally.

The cortical network model can be configured to different spatial preferences by changing the lateral inhibitory connections. In this paper, we demonstrate how changing the inhibition connectivity of the network model while using the same mixed-source inputs changes the reconstructed signal. For most of the simulations, we focus on the case

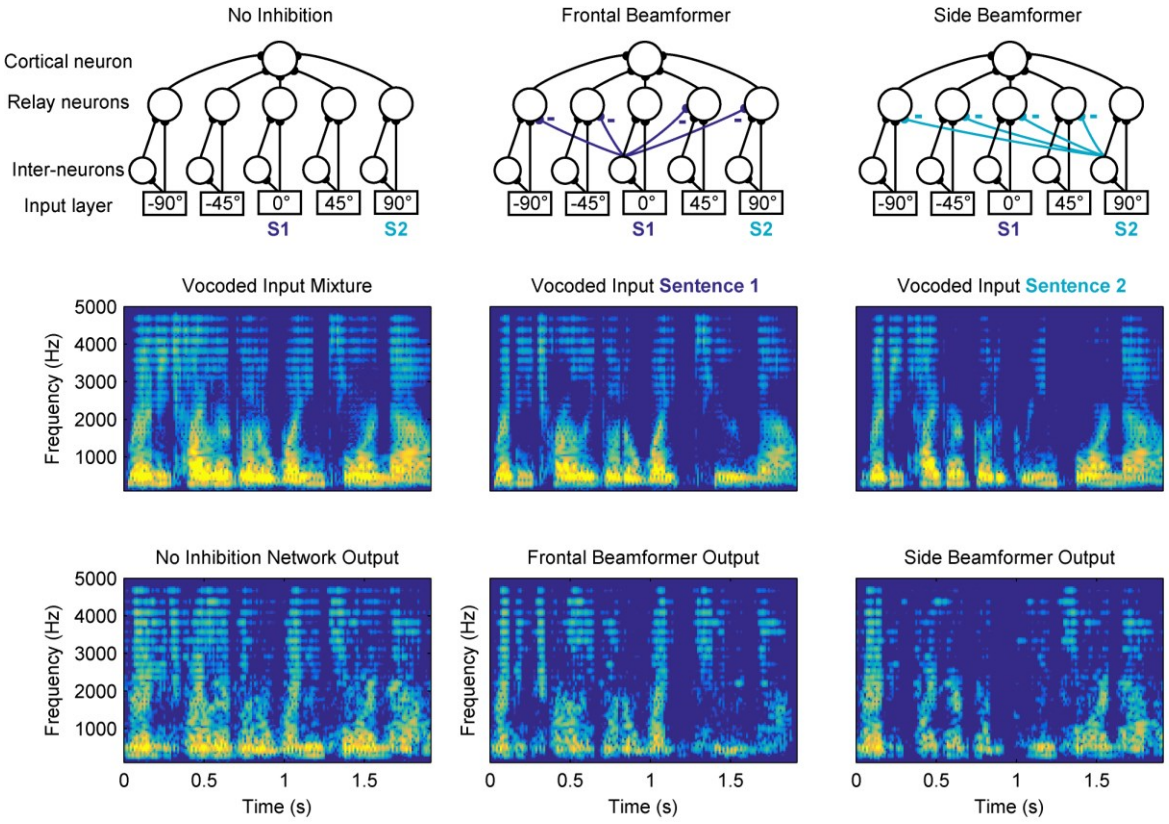

Figure 8 Outputs of engineering solution models with different network models using identical input stimuli. Top row: cortical neural network with varying inhibitory connectivity. Middle row: spectrograms of input mixture and individual input sentences. Bottom row: output of engineering solution using corresponding neural networks shown in the top row. For all simulations, two sentences were presented from 0° (S1) and 90° (S2) simultaneously. In the left column, the cortical network used had no inhibitory connections, and the reconstructed output represents a mixture of the two sentences. In the middle column, the network model used was a 'frontal beamformer' where the center spatial channel inhibits all other directions, and the output of the engineering solution represents the speech content of Sentence 1, as seen by comparing their spectrograms. In the right column, a 'side beamformer' emphasizing the 90° spatial channel was used, so the final output represents the speech content of Sentence 2 instead.

of a ‘frontal biological beamformer’, where the 0° azimuth frontal channel inhibits all other channels.

Cortical network model parameters were simplified for the engineering solution. Cortical-level adaptation and synaptic depression were removed from the complete Dong et al. model, as there was no need to match a specific recorded neuron. Additionally, the lateral inhibition time constant was increased to 1000 ms to fully suppress competition, in case the target-location input channel responses were sparse in time. The full set of parameters used is listed as follows. For all neurons, resting potential was -60 mV, spiking threshold was -40 mV, and the reversal potential for excitatory currents was 0 mV. In relay neurons, the reversal potential for inhibitory currents was -70 mV. In inter-neurons, excitatory post-synaptic conductance (EPSC) was modeled as an alpha function with a time constant of one millisecond. In relay neurons, both EPSC and inhibitory post-synaptic conductance (IPSC) were modeled as the difference of a rising and a falling exponential, where rise and fall time constants were 1 and 3 ms, and 4 and 1000 ms, respectively. An absolute refractory period of 3 ms was enforced in all neurons. Synaptic strengths were uniform across all spatial channels for the same type of synapse. The synaptic conductance between input to inter- and relay neurons were 0.11 and 0.07 nF, the synaptic conductance from relay and cortical neuron was 0.07 nF, and the lateral inhibition conductance was 0.2 nF. These value are similar to those used to fit recorded neurons (Dong et al., 2016). The cortical network model was custom written in Matlab.

3.2.4 Stimulus reconstruction

3.2.4.1 Overview

Historically used as a method of spike decoding, stimulus reconstruction assumes that the stimulus waveform can be approximated from neural spikes $x(t)$ occuring at t_i $(i = 1, 2, \cdots, n)$ using an optimal reconstruction filter $h(t)$. More specifically, one can convolve the optimal reconstruction filter with the spiking waveform to get an estimate of the original stimulus: $s_{est}(t) = \sum_{i=1}^{n} h(t - t_i)$. The reconstruction filter is derived from a known set of stimulus and neural spike responses, and can then be applied to new neural responses to predict unknown stimuli. The reconstruction filter can be found in the frequency domain using a previously-derived analytical solution: $h(\omega) = \frac{S_{sx}(\omega)}{S_{xx}(\omega)}$, where $S_{sx}(\omega)$ is the cross-spectral density of the original stimulus $s(t)$ and its neural response $x(t)$, and $S_{xx}(\omega)$ is the power spectral density of the neural response (Gabbiani and Koch, 1999).

3.2.4.2 Two-dimensional stimulus reconstruction

In addition to reconstructing from modeled rather than recorded neural responses, we made two main modifications to the traditional reconstruction method for the purpose of reconstructing acoustic signals. The first was multi-frequency reconstruction: breaking down the time-domain acoustic signal into different frequency components, and reconstructing the envelope of each frequency from the corresponding model neuron, as opposed to reconstructing the entire signal from a single neural response. After all

envelopes were reconstructed, we then used them to construct a vocoded acoustic signal.

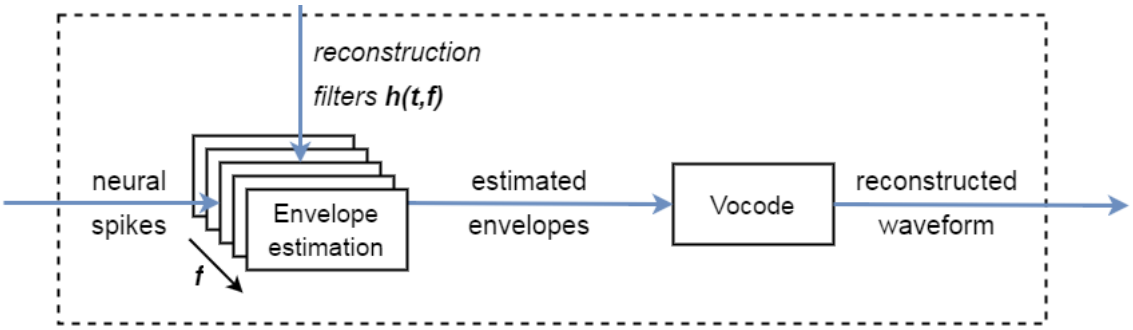

Figure 9 Stimulus reconstruction involves: 1) estimating the envelope of each frequency using neural responses across frequencies and a frequency-specific two-dimensional reconstruction filter; 2) using the estimated envelopes to construct a vocoded stimulus, by modulating sinusoids of the corresponding frequency with the envelopes and summing all frequency components.

The second modification was training and using two-dimensional reconstruction filters when reconstructing a single frequency envelope. Two-dimensional reconstruction filters, as opposed to traditional one-dimensional filters, allows the reconstruction process to make use of complementary information in other frequency channels to achieve higher reconstruction accuracy.

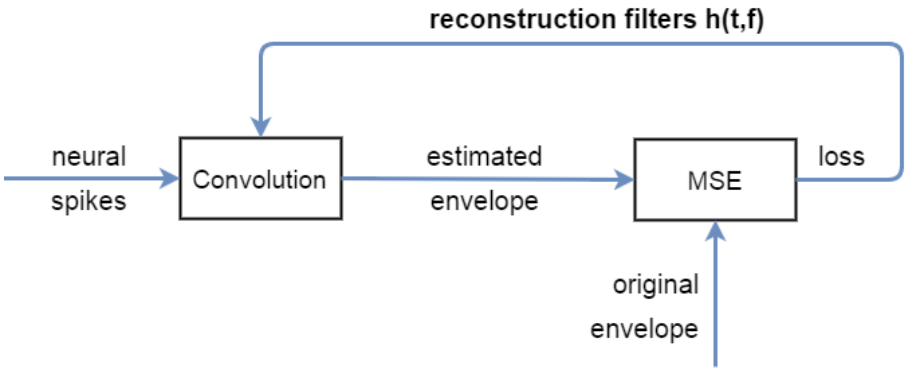

Figure 10 Calculating the two-dimensional reconstruction filter involves iteratively optimizing the filter by decreasing the mean-squared error (MSE) between the estimated and original envelopes. In each iteration, we compute the incremental change to the filter values that would give the largest decrease in MSE for the next step.

We used a combination of the traditional analytical solution and gradient descent to find optimal values for the two-dimensional optimal reconstruction filter. The analytical one-dimensional optimal filter was used to set an initial guess for the two-dimensional filter, and then gradient descent was used to find the optimal two-dimensional filter by minimizing the mean-squared-error (MSE) between the reconstruction and original envelopes, treating values of the reconstruction filter as free parameters. Initial one-dimensional reconstruction filters were calculated using Matlab code available on the website of the *Theoretical Neuroscience* textbook (Gabbiani and Koch, 1999), and two-dimensional optimization was done in Python using custom-written Theano code.

3.2.4.3 Speech Stimuli

The Coordinated Response Measure (CRM) Corpus was used to train and test the novel stimulus reconstruction technique, as well as test the segregation and reconstruction results using the engineering solution. The CRM Corpus is a large set of recorded sentences in the form of 'Ready [CALL SIGN] go to [COLOR] [NUMBER] now', where call sign, color, and number have 8, 4, and 8 variations, respectively.

To train the reconstruction filter, we extracted one instance of all call signs and color-number combinations to include all the variations for each variable, and used this nine-second duration training sentence and the corresponding combined neural model response as training stimulus and neural response. During gradient descent optimization of the reconstruction filter, we monitored the MSE of another test CRM sentence and

neural response pair to make sure that the reconstruction filter was not overfit to the training stimulus and neural responses.

To test the segregation and reconstruction quality using the engineering solution, we randomly selected 20 trios of CRM sentences, with the criterion that sentences in each trio cannot contain the same call sign, color, or number. For segregation simulations, the first sentence in each trio was designated as the target, while the remaining sentences served as symmetrical maskers. Two simulations were run for each trio to switch the location of the two masker sentences. For the target alone simulation, only the first sentence in each trio was used. During analysis, the mean and standard deviation (STD) of all stimulus sets were calculated for each assessment measure used.

3.2.5 Test simulation scenarios

To demonstrate that the engineering solution is capable of: 1) encoding for the entire azimuth in quiet, 2) selectively encoding for a preferred location while suppressing non-preferred locations when competition arises, and 3) robustly encoding for preferred locations when maskers become louder than targets, we designed the following three simulations. In the first simulation, we presented the combined neural model with a single target location from 0 to 90 degrees in azimuth, at 5-degree intervals. We then calculated assessment measures of the quality and intelligibility of the reconstructed signal compared to the original vocoded target signal. In the second experiment, we presented one sentence at the target location (0°), and roved two masker sentences symmetrically from 0 to ±90 degrees. We then calculated assessment measures of the reconstruction compared to the target and masker sentences, respectively, at all masker

locations. The last simulation was designed to test the robustness of the engineering solution at low signal-to-noise ratios (SNRs). In this simulation, the target was fixed at 0° and the maskers at ±90° respectively. The target-to-masker ratio (TMR), or the energy difference between the target and individual maskers, was then varied between -12 and 12 dB. This corresponds to SNRs from -15 to 9 dB. Stimulus direction was simulated by applying the Kemar HRTF of the desired location to raw CRM sentences to generate realistic binaural signals with appropriate ITD and ILD cues. In both simulations, the randomly selected set of 20 trios of CRM sentences was used as described in the previous section.

3.2.6 Assessment measures of segregation and reconstruction quality

We compared several computational measures of speech intelligibility including NCM, STOI, CSII, and PESQ, which calculate the similarity and intelligibility of a processed signal compared to its original unprocessed form for human subjects (Kates and Arehart, 2005, Chen and Loizou, 2011, Taal et al., 2011, Cosentino et al., 2012). A higher score indicates better intelligibility of the processed signal to human listeners, as well as more similarity to the original unprocessed signal. In our formal analysis, these intelligibility measures performed similarly, and we chose STOI as it gave comparatively the most consistent measure when conditions varied. In addition, we used the cross-correlation between the spectrograms of the reconstructed and reference waveforms (XCorr), as XCorr is a commonly used measure in this field (Mesgarani and Chang, 2012) which can easily be compared to previous results in reconstruction. Since the reconstructed signal is constructed by vocoding reconstructed envelopes, a perfect

reconstruction would produce the vocoded version of the original signal. For this reason, we used the vocoded original signal as the reference signal in most assessments. Matlab functions for intelligibility measures were generously provided by Stefano Cosentino.

3.2.7 Engineering solution performance comparison to psychoacoustic data

To compare the spatial segregation performance of the engineering solution to known psychoacoustic data, we computed the minimal TMR thresholds at which the engineering solution could extract the target for all target-masker separations in central-target (0°) and symmetrical-maskers simulations. The threshold was defined as the TMR at which at least 50% of engineering solution outputs for each spatial separation could be correctly classified as target using both STOI and XCorr. In other words, for a specific separation, the threshold is the minimal TMR at which the output is more similar to the target than masker sentences, as defined by STOI and XCorr. For example, at 0° separation, 3 dB was the lowest TMR at which for 10 or more out of the 20 (≥ 50%) example sentence trios, the model output had higher STOI and XCorr scores compared to the original target than to masker sentences. Therefore, the TMR threshold was calculated to be 3dB, which is then compared to human performance TMRs under the same spatial setting.

3.3 Results

3.3.1 Bimodal engineering solution performance

The bimodal engineering solution was able to achieve intelligible reconstruction from all directions when only one stimulus was present, as well as good segregation and assessment scores when competing maskers were presented.

To demonstrate the function of the lateral inhibitory connections in the cortical network model, as well as demonstrate the spatial segregation capabilities of the engineering solution, Fig. 8 (Page 38) shows how changing lateral inhibition connectivity in the cortical network model while providing the same stimulus inputs changed the reconstructed output of the engineering solution. In all three simulations, Sentence 1 was presented from the front (0°), and Sentence 2 was presented simultaneously from 90° azimuth. In the first column, we used a cortical network with no inhibition connectivity, and the reconstructed signal represents a mixture of Sentences 1 and 2. In the middle column, we used a 'frontal biological beamformer' where the middle (0°) channel inhibited all other spatial channels. In this case, the reconstruction output resembles Sentence 1 more than Sentence 2, as seen in the plotted spectrograms. In the last column, we used a 'side biological beamformer' where the right-most (90°) channel inhibited all other spatial channels, and the reconstruction output resembles Sentence 2 more than Sentence 1. When only one stimulus is present, the engineering solution reliably reconstructs sounds from all directions. The left plots in Fig. 11 show that when a target is roved alone in space, the reconstructed signal always resembles that target, as demonstrated by the high assessment measures (STOI and XCorr) regardless of location.

Note that stimuli falling between spatial channels show slightly lower assessment scores, as seen in the slight dips in the single-target curves. Encoding for single sources at all locations allows the listener to be alerted of any novel stimuli and events in the entire listening space.

The segregation capabilities of the engineering solution when competition is present are shown in detail in the plots on the right side of Fig. 11. Using a 'frontal biological beamformer' and presenting a target fixed at 0°, while two maskers of equal amplitude (0 TMR) are played simultaneously and symmetrically at directions anywhere between 0 and ±90 degrees, the engineering solution attempts to segregate and reconstruct the frontal target while suppressing the maskers. Calculating the assessment measures STOI and XCorr compared to the target and masker, improved segregation can be seen as the masker is moved further away from the target, as scores compared to target increases while scores compared to maskers decrease. Segregation performance saturates at around 15- to 20-degree target-and-masker separations as shown in the plots on the right side of Fig. 11. This can be compared to human psychoacoustic performance in spatial release from masking experiments, where spatial release from masking benefits have been shown to plateau at around 15° (Marrone et al., 2008, Srinivasan et al., 2016). Figure 12 demonstrates that the engineering solution is effective in situations where the target source is weaker than masker sources. Again using a 'frontal biological beamformer', and presenting the target at 0° and two maskers at ±90°, the target to masker ratio (TMR) was varied to test whether the engineering solution can segregate competing maskers at low TMRs. Figure 12 shows that the reconstructed signal more

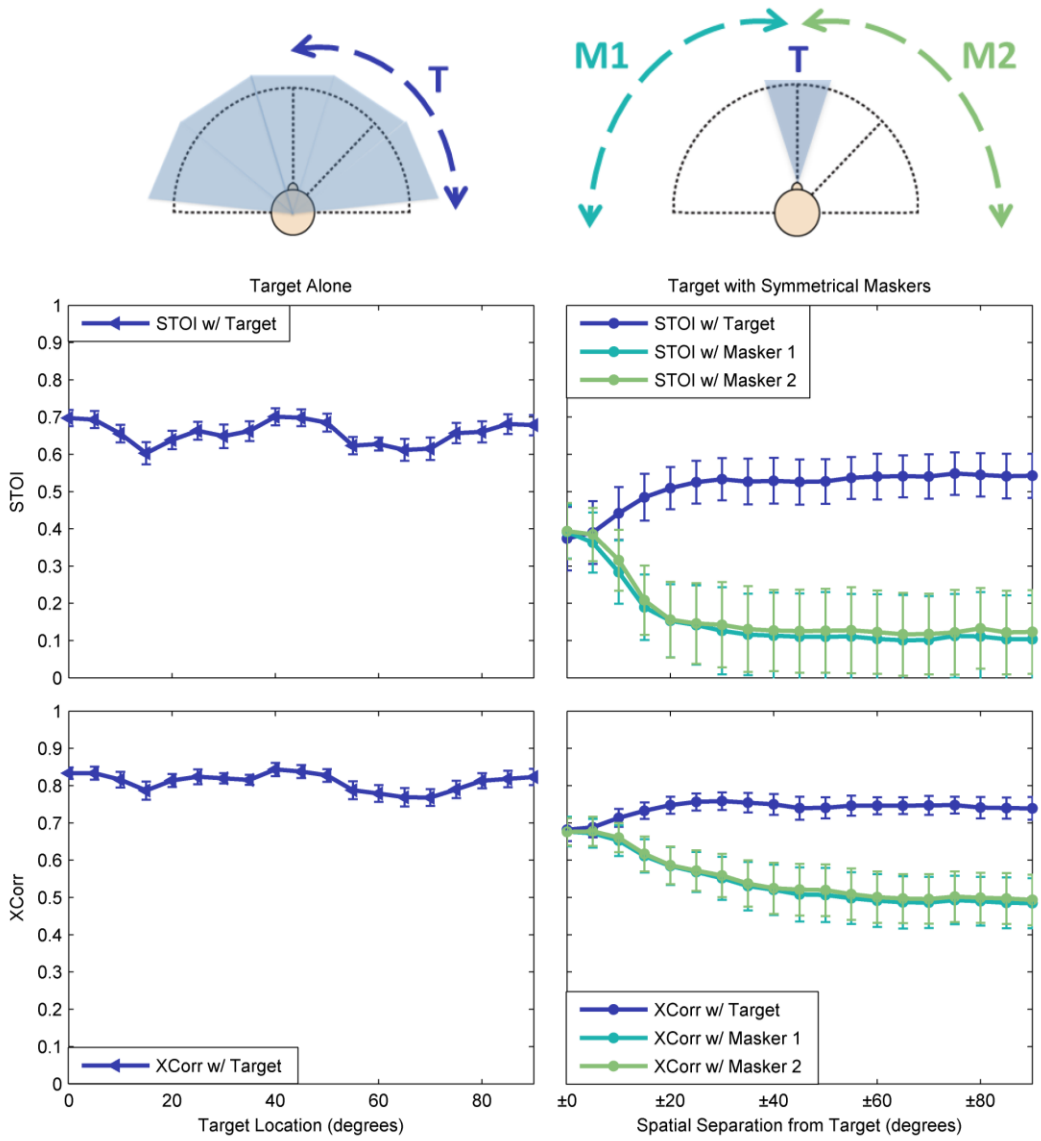

Figure 11 Performance of the engineering solution in one-source (left figures) and competing three-source (right figures) environments. The model used is a frontal beamformer where 0° inhibits all other spatial channels. In the single-source case shown in the left column, the STOI and XCorr of the reconstructed signal are compared to the vocoded target as the target is roved in space from 0 to 90 degrees. The high STOI and XCorr across all locations indicate that the engineering solution can capture single-sources in all spatial locations. In the three-source simulations shown in the right column, the target sentence is fixed in the front, while maskers are moved from 0 to ±90 degrees. The STOI and XCorr of the reconstruction output are computed in reference to the vocoded target and vocoded masker signals. As the separation between target and maskers increases, the reconstructed signal becomes a more reliable representation of the target signal, indicating that the engineering solution is effectively separating the frontal target from side maskers. The increased spatial separation benefit saturates between 15 to 20 degrees.

closely resembles the target than maskers down to around -8 dB, as seen in higher STOI and XCorr scores when comparing the reconstruction to the target than the maskers. For TMRs under -8 dB, the reconstruction ceases to robustly encode the target over maskers. This value can be compared to psychoacoustic data showing the 50% speech reception threshold of normal-hearing listeners for the same task was around -10 dB (Marrone et al., 2008).

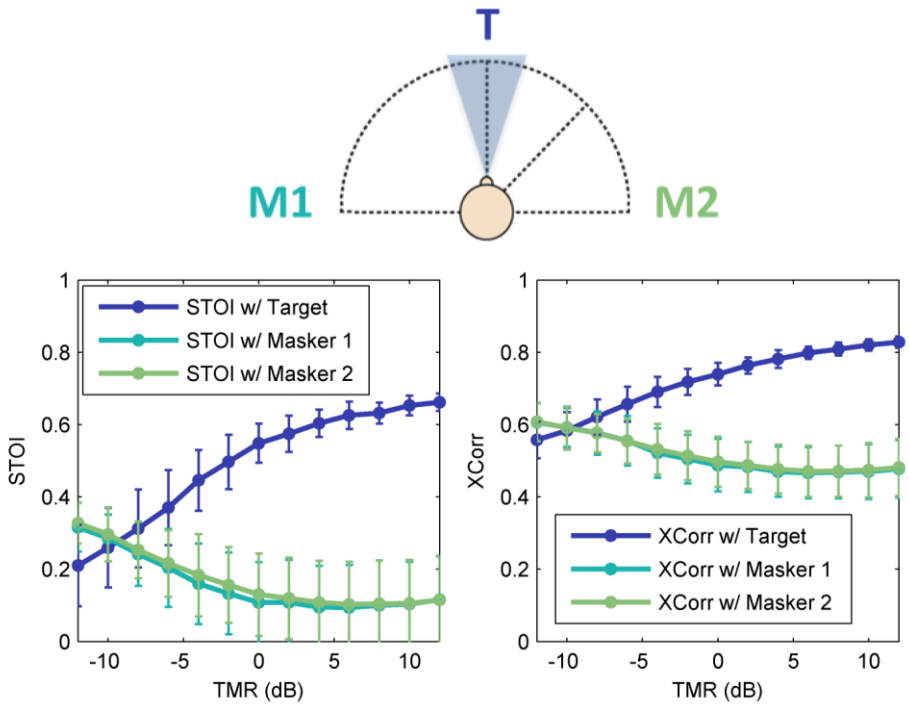

Figure 12 Performance of the engineering solution for separating a target sentence from two symmetrical maskers sources located at ±90° as target-to-masker ratio (TMR) ratio is changed. This figure shows that the engineering solution can robustly reconstruct the target while ignoring the masker down low TMRs comparable to human performance, as seen in the high STOI and XCorr of the output compared to the target and low STOI compared to the masker.

3.3.2 Improvements of two-dimensional stimulus reconstruction method over traditional stimulus reconstruction method

We found that the improvement of switching from the traditional one-dimensional reconstruction filters to optimized two-dimensional reconstruction filters was quite significant. Using the optimized two-dimensional reconstruction filter decreased MSE by 66±7% and 43±11% across all frequency envelopes for training and test inputs respectively. The decrease in MSE when two-dimensional filters are used for all frequencies can be seen in Fig. 13. Speech assessment scores (STOI and XCorr) between the reconstructed and original signals also improved significantly for both training and test sets, as shown in Table 3. Scores were calculated for one-dimensional and two-dimensional filter reconstructions, compared to either the original waveform or

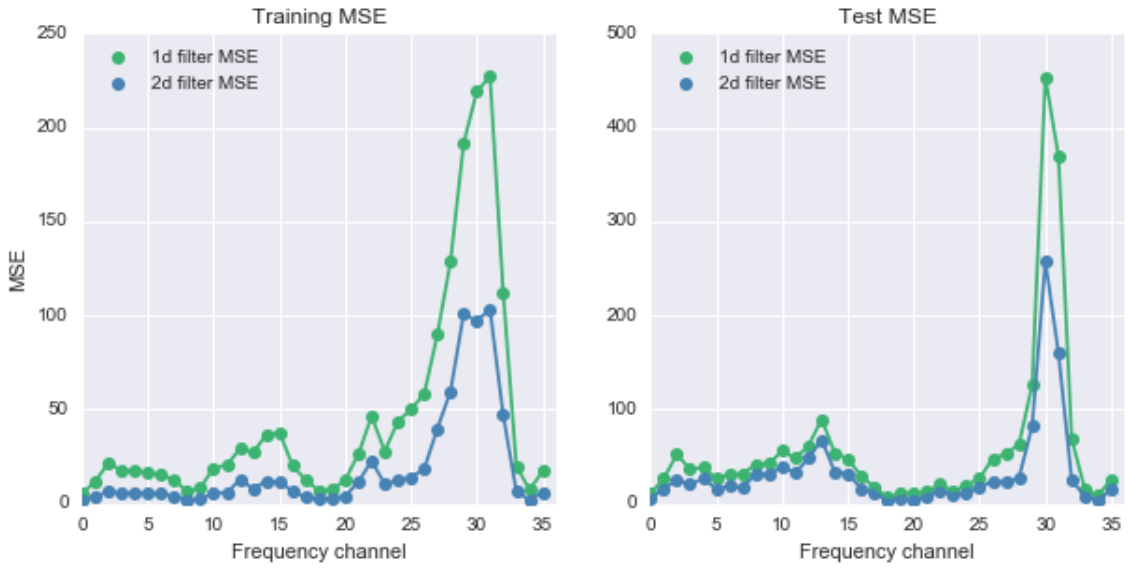

Figure 13 Mean-squared errors (MSEs) of the reconstructed envelopes for each frequency using two-dimensional and one-dimensional reconstruction filters, for training and test stimuli. For all frequencies, the two-dimensional reconstruction filter improved envelope estimation as shown by the decreased MSE for both training and test cases.

the vocoded original waveform. The two-dimensional filter reconstructions show higher STOI and XCorr scores in all cases.

Table 3 Improvements in speech assessment scores from using one- to two-dimensional optimal filters.

		1D filter reconstruction	2D filter reconstruction	1D filter reconstruction	2D filter reconstruction
		compared to original signal		*compared to vocoded signal*	
Training	STOI	0.60	0.70	0.59	0.75
	XC	0.58	0.66	0.74	0.89
Test	STOI	0.59	0.67	0.58	0.73
	XC	0.58	0.67	0.67	0.81

3.4 Discussion

3.4.1 Comparison to human performance

It's postulated that human listeners organize sounds into 'streams' (Bregman, 1994), formed by finding components of sound with coherent spatial and non-spatial cues such as pitch, volume, and other sound qualities. There has been debate about what cues are most critical for processing and separating sound mixtures (Eramudugolla et al., 2008, Kidd et al., 2005). Since the engineering solution only uses ITD and ILD cues to perform segregation, it would be interesting to see how its spatial segregation capabilities compares to human listeners, who have access to additional cues. To this end, we compared the spatial segregation performance of the engineering solution directly to human psychoacoustic data. In Fig. 11, we observed that the spatial segregation performance of the engineering solution improves with increased separation between the target and masker. This phenomenon is observed in human psychoacoustics, and the benefit of increased separation between target and masker has been termed spatial release

from masking (SRM). Psychoacoustic studies (Marrone et al., 2008, Srinivasan et al., 2016) have recorded the TMR thresholds for 50% correct human performance in listening experiments with a center (0°) target and symmetrical maskers at different spatial separations. For comparison, we calculated the 50% classification threshold based on STOI and XCorr for each target-masker separation for the same center (0°) target and symmetrical masker simulations. The 50% classification threshold for each separation is the TMR where the assessment measures (STOI and XCorr) of the engineering solution output are higher compared to the target sentence than the masker sentences for at least 50% of sentence trios.

Figure 14 compares the engineering solution threshold TMRs to those measured in psychoacoustic studies (Marrone et al., 2008, Srinivasan et al., 2016). The overall range and trend of TMRs for model performance and human data are roughly similar. Comparing the two more specifically, the psychoacoustic data show a more gradual saturation in performance improvements with increased spatial separation, where the benefits gradually stop between 15 and 45 degrees. The engineering solution performance saturates more quickly at 15°. The threshold TMRs of the engineering solution under 15° are very similar to psychoacoustic values, while those at larger spatial separations underperform humans by around 2dB. For the engineering solution, once target and maskers are sufficiently separated in space (>15°) for computing the ITD and ILD cues, there is no additional benefit to further increases in spatial separation. This threshold effect does not occur for the human listener—further increased spatial separation provides more perceptually distinct sources, making it easier to perform the task.

Therefore, the engineering solution performance may be more limited at larger separations due to a lack of integration of spatial and non-spatial cues.

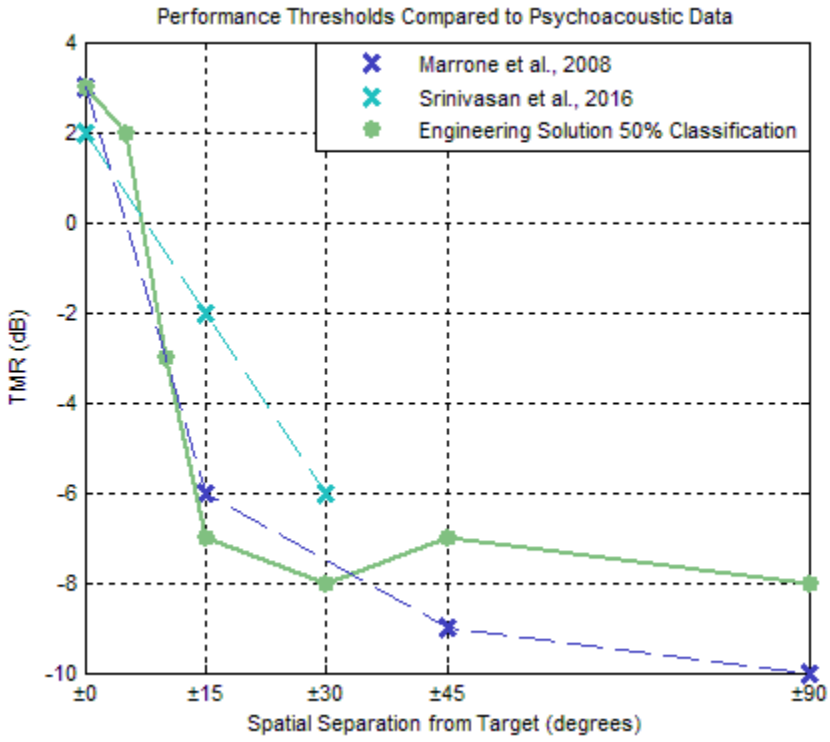

Figure 14 Performance comparison of engineering solution to psychoacoustic data. The engineering solution performance threshold is defined as the TMR at which at least 50% of reconstructed sentence examples are more similar to the target than maskers, as quantified by STOI and XCorr. TMRs under 15° are consistent for model and data. The engineering solution performance saturates more quickly than psychoacoustic data between 15 and 45 degrees, and then plateaus at higher TMR for spatial separations larger than 45°.

In addition to spatial cues, studies have shown that the temporal fine structure of speech is critical for the perceptual segregation of talkers in SRM, while envelopes are important for speech intelligibility (Swaminathan et al., 2016, Smith et al., 2002). Temporal fine structure is not reconstructed in the vocoding step of the engineering solution, but is available to the peripheral localization neural model when extracting spatial cues. It's possible that the good segregation performance of the engineering

solution is explained by the availability of binaural temporal cues to the peripheral localization model, while the intelligibility of the reconstructed signal is explained by the reconstruction of stimulus envelopes, as suggested in the study by Swaminathan and colleagues (Swaminathan et al., 2016).

3.4.2 Application

The engineering solution provides a flexible, bimodal, spatial processing scheme to assist listeners in segregating directions of interest when the need arises, while allowing the listener to monitor the entire listening space. Using the novel stimulus reconstruction method described above, we demonstrate that physiology-based auditory spatial processing models can be applied to improve the processing of human speech and achieve high intelligibility.

Although this early-phase, proof-of-concept engineering solution has achieved high speech intelligibility, the performance is limited by the vocoding step of the reconstruction process. A perfect reconstruction can only be as good as the vocoded version of the original speech, which sounds less natural. We are continuing research in this direction to improve the naturalness of the reconstruction, by attempting to restore the fine structure of the original stimulus in addition to reconstructing the envelopes.

We feel that this is an especially exciting time to create new technologies for existing hearing-assist devices such as hearing aids and cochlear implants, due to recent advances in battery life and computing power of these portable devices (Edwards, 2007), as well as our fundamental understanding of the auditory system. As we demonstrate with

our successful reconstruction of acoustic stimuli from neural responses, physiology-based computational models can and should be used to help solve real-world problems.

3.5 Concluding remarks

We have shown a physiology-based engineering solution for the spatial processing of single and multiple sound sources. The core mechanisms of the engineering solution were inspired by cortical neurons that can encode for the entire azimuth when only one stimulus is present in space, and selectively encode preferred directions when multiple competing stimuli are present simultaneously. We demonstrate that the engineering solution can take binaural speech mixtures and generate segregated sound waveforms using spatial processing mechanisms in the neural domain, and that the segregation performance is close to human normal hearing listeners.

CHAPTER FOUR: FUTURE WORK

4.1 Physiology-based modeling

The physiology-based cortical modeling work described in Chapter Two is currently being verified and tested in mice, in collaboration with Howard Gritton and Nick James from Xue Han's Lab. Currently, single-source and competing-stimuli experiments are being replicated to confirm the physiological results in mice. If the results seen in the songbird (Maddox et al., 2012) are replicated, it would be very exciting to test the new experiments proposed in Chapter Two. Specifically, we can use pharmacological reagents to suppress local inhibition, and test the location of the proposed lateral inhibitory connections between channels. Additionally, the cortical network model predicts that neurons can remain robust to more than one masker, as long as the additional maskers match the non-preferred location of lateral inhibitory connections. Based on results from these new experiments, the cortical network model can then be verified or modified.

If the results in mice do not match those seen in the songbird, we can think about how the inputs and mechanisms within the model may be different for mice. Adapting the model to work for mice may provide additional insights on alternative mechanisms for cortical processing across all species.

4.2 Engineering solution for spatial sound processing

The current engineering solution can be improved in three main ways. First, the current quality and intelligibility of reconstruction is capped by the vocoding step. More specifically, since we are reconstructing envelopes and using them to vocode the the

reconstructed stimulus, a perfect reconstruction can only be as good as the vocoded version of the original stimulus. While the current solution achieves good segregation and intelligibility, the perceptual quality of the reconstructed speech is not ideal. Work on how to either restore fine structure and phase information in the reconstructed envelopes, or reconstruct fine structure waveforms directly, would remove this reconstruction performance cap. Second, the current model only uses spatial cues, ITD and ILD, to separate competing stimuli in space. We know that humans have access to and actively use, many other types of cues to form a coherent 'image' of individual sound sources and pull them apart. Adding additional separation cues such as frequency coherence or harmonic structure could improve the segregation capabilities of the model. Three, the current engineering solution does not address some real-world scenarios such as sounds coming from the back and reverberant conditions. Making the model robust in these real-world conditions will be a huge challenge, but is necessary if the ultimate goal is to incorporate this algorithm into hearing-assist devices such as hearing aids.

BIBLIOGRAPHY

Amin N, Gill P, Theunissen FE (2010) Role of the zebra finch auditory thalamus in generating complex representations for natural sounds. Journal of Neurophysiology 104:784–798. doi:10.1152/jn.00128.2010.

Arbogast TL, Mason CR, Kidd Jr G (2002) The effect of spatial separation on informational and energetic masking of speech. Journal of the Acoustical Society of America 112:2086–2098. doi: 10.1121/1.1510141.

Ashida G, Carr CE (2011) Sound localization: Jeffress and beyond. Current Opinion in Neurobiology 21:745–751.

Atzori M, Lei S, Evans DIP, Kanold PO, Phillips-Tansey E, McIntyre O, McBain CJ (2001) Differential synaptic processing separates stationary from transient inputs to the auditory cortex. Nature Neuroscience 4(12), 1230-1237. DOI:10.1038/nn760

Bregman AS (1994) Auditory scene analysis: The perceptual organization of sound: MIT press.

Chen F, Loizou PC (2011) Predicting the intelligibility of vocoded speech. Ear and Hearing 32:331.

Cherry EC (1953) Some experiments on the recognition of speech, with one and with two ears. Journal of the Acoustical Society of America 25:975.

Cosentino S, Marquardt T, McAlpine D, Falk TH (2012) Towards objective measures of speech intelligibility for cochlear implant users in reverberant environments. In: 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA), 2012, pp 666–671: IEEE.

Darwin C, Hukin R (1998) Auditory objects of attention. Journal of the Acoustical Society of America 103:2928-2928. DOI: http://dx.doi.org/10.1121/1.422144.

Day ML, Koka K, Delgutte B (2012) Neural encoding of sound source location in the presence of a concurrent, spatially separated source. Journal of Neurophysiology 108:2612-2628. DOI: 10.1152/jn.00303.2012.

Dayan P, Abbott LF (2001) Integrate-and-fire models: spike-rate adaptation and refractoriness. In: Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems, pp 165–166: The MIT Press.

Dent ML, Larsen ON, Dooling RJ (1997) Free-field binaural unmasking in budgerigars. Behavorial Neuroscience 111:590. DOI: 10.1037/0735-7044.111.3.590.

Dent ML, McClaine EM, Best V, Ozmeral E, Narayan R, Gallun FJ, Sen K, Shinn-Cunningham BG (2009) Spatial unmasking of birdsong in zebra finches and budgerigars. Journal of Comparative Psychology 123:357. DOI: 10.1037/a0016898.

Desloge JG, Rabinowitz WM, Zurek PM (1997) Microphone-array hearing aids with binaural output. I. Fixed-processing systems. IEEE Transactions on Speech and Audio Processing 5(6), 529-542.

Devore S, Ihlefeld A, Hancock K, Shinn-Cunningham B, Delgutte B (2009) Accurate sound localization in reverberant environments is mediated by robust encoding of spatial cues in the auditory midbrain. Neuron 62:123-134.

Dong J, Colburn HS, Sen K (2016) Cortical transformation of spatial processing for solving the cocktail party problem: A computational model. eneuro 3:ENEURO. 0086-0015.2015.

Edwards B (2007) The future of hearing aid technology. Trends in Amplification 11:31–46.

Eramudugolla R, McAnally KI, Martin RL, Irvine DR, Mattingley JB (2008).The role of spatial location in auditory search. Hearing research 238(1), 139-146.

Fischer B, Anderson C, Peña J (2009) Multiplicative auditory spatial receptive fields created by a hierarchy of population codes. PLoS One 4.

Froemke RC, Merzenich MM, Schreiner CE (2007) A synaptic memory trace for cortical receptive field plasticity. Nature 450:425-429. DOI:10.1038/nature06289.

Higgins NC, Storace DA, Escabí MA, Read HL (2010) Specialization of binaural responses in ventral auditory cortices. Journal of Neuroscience, 30(43), 14522-14532. DOI: 10.1523/JNEUROSCI.2561-10.2010

Hine JE, Martin RL, Moore DR (1994) Free-field binaural unmasking in ferrets. Behavioral Neuroscience 108:196. DOI: 10.1037/0735-7044.108.1.196.

Gabbiani F, Koch C (1999) Principles of spike train analysis: Wiener Kernals and stimulus estimation. In: Methods in Neuronal Modeling (Koch, C. and Segev, I., eds), pp 343–357 Cambridge, MA: The MIT Press.

Jeffress LA "A place theory of sound localization." Journal of comparative and physiological psychology 41.1 (1948): 35.

Kates JM, Arehart KH (2005) Coherence and the speech intelligibility index. Journal of the Acoustical Society of America 117:2224–2237.

Kidd Jr G, Arbogast TL, Mason CR, Gallun FJ (2005) The advantage of knowing where to listen. The Journal of the Acoustical Society of America 118(6), 3804-3815.

Knudsen EI, Konishi M (1978) A neural map of auditory space in the owl. Science 200(4343):797–797.

Köppl C, Carr CE (2008) Maps of interaural time difference in the chicken's brainstem nucleus laminaris. Biological Cybernetics 98:541-559. DOI: 10.1007/s00422-008-0220-6.

Lee C-C, Middlebrooks JC (2011) Auditory cortex spatial sensitivity sharpens during task performance. Nature Neuroscience 14:108-114. DOI:10.1038/nn.2713.

Levy RB, Reyes AD (2012) Spatial profile of excitatory and inhibitory synaptic connectivity in mouse primary auditory cortex." Journal of Neuroscience 32.16: 5609-5619. DOI: 10.1523/JNEUROSCI.5158-11.2012

Maddox RK, Billimoria CP, Perrone BP, Shinn-Cunningham BG, Sen K (2012) Competing sound sources reveal spatial effects in cortical processing. PLoS Biology 10:e1001319.

Marrone N, Mason CR, Kidd Jr G (2008) Tuning in the spatial dimension: Evidence from a masked speech identification task. Journal of the Acoustical Society of America 124:1146–1158.

Mesgarani N, Chang E (2012) Selective cortical representation of attended speaker in multi-talker speech perception. Nature 485:233–236.

Middlebrooks JC, Bremen P (2013) Spatial stream segregation by auditory cortical neurons. Journal of Neuroscience 33:10986-11001. DOI: 10.1523/JNEUROSCI.1065-13.2013.

Müller C, Scheich H (1988) Contribution of GABAergic inhibition to the response characteristics of auditory units in the avian forebrain. Journal of Neurophysiology 59:1673-1689.

Oswald AMM, Schiff ML, Reyes AD (2006) Synaptic mechanisms underlying auditory processing. Current Opinion in Neurobiology 16(4), 371-376. DOI:10.1016/j.conb.2006.06.015

Pena JL, Konishi M (2001) Auditory spatial receptive fields created by multiplication. Science 292:249-252. DOI: 10.1126/science.1059201.

Pinaud R, Mello CV (2007) GABA immunoreactivity in auditory and song control brain areas of zebra finches. The Journal of Chemical Neuroanatomy 34(1), 1-21. DOI:10.1016/j.jchemneu.2007.03.005

Qiu A, Schreiner CE, Escabí MA (2003) Gabor analysis of auditory midbrain receptive fields: spectro-temporal and binaural composition. Journal of Neurophysiology 90:456-476. DOI: 10.1152/jn.00851.2002.

Rose HJ, Metherate R (2005) Auditory thalamocortical transmission is reliable and temporally precise. Journal of Neurophysiology 94(3), 2019-2030. DOI: 10.1152/jn.00860.2004

Roverud E, Best V, Mason CR, Streeter T, Kidd G (2016) Evaluating performance of hearing-impaired listeners with a visually-guided hearing aid in an audio-visual word congruence task. The Journal of the Acoustical Society of America 139(4), 2210-2210.

Schnupp JW, Carr CE (2009) On hearing with more than one ear: lessons from evolution. Nature Neuroscience 12:692-697. DOI:10.1038/nn.2325.

Sen K, Theunissen FE, Doupe AJ (2001) Feature analysis of natural sounds in the songbird auditory forebrain. Journal of Neurophysiology 86:1445-1458.

Smith ZM, Delgutte B, Oxenham AJ (2002). Chimaeric sounds reveal dichotomies in auditory perception. Nature 416(6876), 87-90.

Stecker GC, Harrington IA, Middlebrooks JC (2005) Location coding by opponent neural populations in the auditory cortex. PLoS Biology 3:e78. DOI: 10.1371/journal.pbio.0030078.

Srinivasan NK, Jakien KM, Gallun FJ (2016) Release from masking for small spatial separations: Effects of age and hearing lossa). Journal of the Acoustical Society of America 140:EL73–EL78.

Swaminathan J, Mason CR, Streeter TM, Best V, Roverud E, Kidd G (2016) Role of binaural temporal fine structure and envelope cues in cocktail-party listening. Journal of Neuroscience 36(31), 8250-8257.

Taal CH, Hendriks RC, Heusdens R, Jensen J (2011) An algorithm for intelligibility prediction of time–frequency weighted noisy speech. IEEE Transactions on Audio, Speech, and Language Processing 19:2125–2136.

van Rossum MC (2001) A novel spike distance. Neural Computation 13:751-763. DOI:10.1162/089976601300014321.

Varela JA, Sen K, Gibson J, Fost J, Abbott LF, Nelson SB (1997) A quantitative description of short-term plasticity at excitatory synapses in layer 2/3 of rat primary visual cortex. Journal of Neuroscience 17:7926-7940.

Vonderschen K, Wagner H (2014) Detecting interaural time differences and remodeling their representation. Trends in Neurosciences 37:289–300.

Wang L, Narayan R, Grana G, Shamir M, Sen K (2007) Cortical discrimination of complex natural stimuli: can single neurons match behavior? Journal of Neuroscience 27:582-589. DOI: 10.1523/JNEUROSCI.3699-06.2007.

Wehr M, Zador AM (2003) Balanced inhibition underlies tuning and sharpens spike timing in auditory cortex. Nature 426:442-446. DOI:10.1038/nature02116.

Winer JA (1992) The functional architecture of the medial geniculate body and the primary auditory cortex. In: The Mammalian Auditory Pathway: Neuroanatomy, pp 287: Springer-Verlag.

Yin T, Chan J (1990) Interaural time sensitivity in medial superior olive of cat. Journal of Neurophysiology 64:465–488.

Zhou Y, Wang X (2012) Level dependence of spatial processing in the primate auditory cortex. Journal of Neurophysiology 108:810-826. DOI: 10.1152/jn.00500.2011.

Zhou Y, Wang X (2013) Spatially extended forward suppression in primate auditory cortex. European Journal of Neuroscience 919-933. DOI: 10.1111/ejn.12460.

CURRICULUM VITAE

JUNZI DONG
[redacted]

EDUCATION

BOSTON UNIVERSITY, Boston, MA **Jan, 2017**
Ph.D. Candidate in Biomedical Engineering

- Cumulative GPA: 3.89/4.0
- Computational Neuroscience Graduate Training Fellowship (NIH) 2013 – 2015
- Graduate Teaching Fellow of the Year Dec, 2014

WASHINGTON UNIVERSITY in ST LOUIS, St. Louis, MO **May, 2011**
Bachelor of Science in Biomedical Engineering, Summa Cum Laude

- Cumulative GPA: 3.96/4.0
- Tau Beta Pi, Engineering Honorary Society
- Alpha Eta Mu Beta, Biomedical Engineering Honorary Society
- HHMI Undergraduate Summer Research Fellowship June – Aug, 2010

RESEARCH EXPERIENCE

BOSTON UNIVERSITY, Boston, MA **2011 – 2016**
Graduate Research Fellow and PhD Candidate

- Built a physiology-based computational network model for understanding 'the cocktail party' problem, explaining how our brains can segregate sound sources from different locations.
- Designed an engineering solution to help hearing-impaired listeners segregate mixed auditory inputs.
  - Used a neural model to perform sound-segregation in the neural domain.
  - Reconstructed the neural output representing the source of interest back to the acoustic domain using a novel stimulus reconstruction algorithm.
  - Optimized the reconstruction process by learning the relationship between neural output and desired acoustic signal using gradient descent.

PHILIPS RESEARCH, Boston, MA **June – Aug, 2016**
Research Scientist Intern

- Worked with time series clinical data to develop predictive algorithms for non-invasive healthcare monitoring.
- Built predictive regression models combining physiology and machine learning.

- Cross-validated models of different complexities, from linear regression to multi-layer perceptrons (MLP) and recurrent neural networks (RNN).
- Created automated pipelines from data imputation, feature extraction, to model evaluation.
- Communicated findings and results to collaborating physician.

TEACHING EXPERIENCE

BOSTON UNIVERSITY, Boston, MA **2012 – 2013**
Teaching Fellow for Biomedical Instrumentation Class

- Taught and helped design 9 lab sessions on medical device design.
- Gave lecture and led interactive discussion on cochlear implants.
- Received an evaluation score of 4.88/5.0 and overwhelming positive feedback from students.

WASHINGTON UNIVERSITY in ST LOUIS, St. Louis, MO **2010 – 2011**
Teaching Assistant

- Held office hours and graded exams for Biol 2970.
- Led the Engineering Help Desk for multiple engineering courses.

PUBLICATION

1. **Dong J**, Colburn, H. S., & Sen, K. (2016). Cortical transformation of spatial processing for solving the cocktail party problem: a computational model. *eNeuro*, 3(1), ENEURO-0086.

PODIUM and POSTER PRESENTATIONS

1. **Dong J**, Colburn HS, Sen K, (2016). Physiology-based engineering solution to the cocktail party problem: spatial sound-source segregation using stimulus reconstruction. Talk and moderator at *ARO Midwinter Meeting*, San Diego.
2. **Dong J**, Colburn HS, Sen K, (2015). Spatial sound-source segregation using a physiologically inspired multi-neuron network model." Poster at *COSYNE*, Salt Lake City.
3. **Dong J**, Colburn HS, Sen K, (2014). An auditory network model for spatial sound stream segregation. Talk at *Society for Neuroscience Annual Meeting*, Washington DC.
4. **Dong J**, Colburn HS, Sen K, (2013). A piece of the cocktail party puzzle: a computational model of sound source segregation in the auditory cortex. Poster at *Society for Neuroscience Annual Meeting*, San Diego.
5. **Dong J**, Colburn, H. S., & Sen, K. (2013). A computational model of spatial tuning in the auditory cortex in response to competing sound sources. *Proceedings of Meetings on Acoustics* (Vol. 19, No. 1, p. 050105). Acoustical Society of America.

LEADERSHIP ROLES

GRADUATE WOMEN in SCIENCE and ENGINEERING (GWISE) 2012 – 2015

Program Co-chair, Board Member, Officer

- Led GWISE officers to coordinate professional development events organized by GWISE.
- Organized flagship events such as the annual fall luncheon, attended by over 130 people.

BU HEARING RESEARCH CENTER (HRC), Boston, MA 2013 – 2015

- Organized weekly HRC seminars with regular weekly attendance of 20–50 people.
- Coordinated joint speakers with adjacent hearing research institute to maximize speaker budget utilization.