

UCSF

UC San Francisco Electronic Theses and Dissertations

Title

Functional Organization of Speech Processing Areas and A Systematic Approach to the Cocktail Party Problem

Permalink

<https://escholarship.org/uc/item/77q5w40n>

Author

Hullett, Patrick W.

Publication Date

2013

Peer reviewed|Thesis/dissertation

**Functional Organization of Human Speech Areas and A Systematic Approach
to the Cocktail Party Problem**

by

Patrick W. Hullett

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Bioengineering

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

AND

UNIVERSITY OF CALIFORNIA, BERKELEY

Dedication

To my mother.

Acknowledgements

I would like to acknowledge my parents for their guidance through the years. In particular, my mother's unending support has been incredibly beneficial along my journey in life and the 10% of her determination and willpower I inherited has proven to be invaluable.

I would like to thank and acknowledge Christoph Schreiner. I cannot be grateful enough. There are a lot of factors in a research environment that contribute to one's development as a scientist, many of which are only realized in hindsight. While these factors are different depending on the individual, it is hard for me to imagine a better environment for my training and I feel very fortunate to have spent this critical period in my research career under his guidance.

I would like to acknowledge my dissertation committee: Michael Brainard, Frederic Theunissen, and Christoph Schreiner. Although committee meetings are by nature sparse, they were among the most insightful moments I had during graduate school.

I would like to acknowledge my fellow labmates: Craig Atencio, Jonathan Shi, Bryan Seybold, Brian Malone and Ralph Beitel. It is amazing how much one learns during the course of graduate school and I was lucky to have been surrounded by very smart and talented people to learn from. In particular, Craig and Jonathan were in the lab during my entire course of graduate school and were constant sources of knowledge and insight.

Finally, I would like to acknowledge Eddie Chang and Nima Mesgarani. Our work together on the functional organization of human superior temporal gyrus was incredibly rewarding and I am continually amazed by the questions being asked and the progress being made in the Chang lab.

Abstract

Functional Organization in Human Speech Areas and A Systematic Approach to the Cocktail Party Problem

by

Patrick W. Hullett

Doctor of Philosophy in Bioengineering

University of California, San Francisco

and

University of California, Berkeley

Professor Christoph E. Schreiner, Chair

The brain is a physical system that can perform intelligent computations. We are interested in nature of those computations to understand how the brain does intelligent things. To that end we have focused on two particularly fruitful questions that were tractable given the current state of knowledge and resources: What is the organization of processing in human speech centers? And, how does the brain solve the cocktail party problem?

To address the first question, we recorded superior temporal gyrus activity in awake human subjects passively listening to speech stimuli using electrocorticography. The high spatial and temporal resolution of this recording technique combined with maximally informative dimension analysis made it possible to compute high density spectrotemporal receptive field maps in a

region of the brain specialized for speech perception. Based on these maps, we found that human superior temporal gyrus has a strong modulotopic organization - a higher order analog of tonotopic organization that has not been previously identified in any human or non-human auditory area.

To investigate the mechanisms by which neural systems solve the cocktail party problem, we created animals that are specialists at extracting vocalization information in the face of by noise-rearing rats and testing them behaviorally to show specialization. Through single unit recordings from primary auditory cortex, we identified a subpopulation of neurons that can extract vocalization information in the face of noise. Although the prevalence of these neurons is the same in both groups of animals, neurons from specialized animals extract information at significantly higher rates. Further receptive field analysis will give insight to the underlying mechanism of this ability. This work demonstrates the ability to create animals specialized at solving the cocktail party problem and a method to identify neurons that contribute to this specialization. This approach can be applied to different classes of noise to generate and refine models of cocktail party processing.

Table of Contents

Introduction.....	1
Part 1	
Chapter 1: Speech Based Modulotopic Organization in Human Superior Temporal Gyrus.....	6
Chapter 2: Speech Based Analysis of Tonotopic Organization in Human Superior Temporal Gyrus	36
Part 2	
Chapter 3: A Systematic Approach to Investigating Neural Solutions to the Cocktail Party Problem	62
Concluding Remarks.....	97

Table of Figures

Chapter 1

Figure 1. Experimental Approach and the STRF.....	10
Figure 2. Patient EC6 Cortical STRF Map.....	12
Figure 3. Spectrotemporal modulation tuning is a higher-order analog of frequency tuning.	13
Figure 4. Modulotopic Organization in Human STG.	16
Figure 5. Spectrotemporal Modulation Content of Phonemes.....	19
Figure 6. Stimulus Transformations Along the Auditory Hierarchy	21

Chapter 2

Figure 1 Identification of best frequency.	41
Figure 2. Best frequency distribution along human STG	42
Figure 3. Identifying multi-peaked spectral receptive fields.....	43
Figure 4. Multi-peak corrected best frequency maps.....	44
Figure 5. STG best frequency distribution and the spectral content in speech.....	45
Figure 6. Organization of Response Latency	47

Chapter 3

Figure 1. Natural and synthesized "naturalistic" noise.....	66
Figure 2 Vocalization stimuli use for behavioral training.....	68
Figure 3. Vocalization recognition performance in naturalistic noise.....	69
Figure 4. Visual stimulus recognition performance in naturalistic noise.	71
Figure 5. Identification of cocktail party neurons (cp-neurons).....	73
Figure 6. Cp-neuron prevalence and information transmission rates.....	75

Introduction

The brain is a remarkable system. Things that seem very intangible and are hard to even describe, like the fleeting sensation of happiness or love, have a purely physical basis in the brain. For many the idea that consciousness can somehow be cast in terms of a population of neurons firing in the brain seems almost inconceivable and is what motivates many to study the brain. Today more than ever people with widely diverse backgrounds and skill sets are focusing their research efforts on the brain and a large government effort, analogous to the Human Genome Project, is underway to focus American efforts on unraveling brain function (Alivisatos et al., 2012) . If lucky humans will understand the inner workings of the brain within our lifetime, and although this may remove some of the allure of the brain, akin to a trick being revealed by a magician, what we will have gained is knowledge on the essence of intelligence. Knowledge that can be used to design intelligent systems that will change how humans interact with the world beyond what one can imagine.

Although we are far from understanding the essence of intelligence, we can make headway on how the brain performs certain intelligent computations. We focus on problems the nervous system seems to solve trivially, due to the subconscious nature in which they are solved, but which have been challenging to implement in artificial systems. Although we do not claim to solve these problems in the course of this work, the intention was to contribute to the

understanding of these problems alone.

The first problem we studied concerns the nature of speech processing. Although the eventual goal will be to understand the underlying computations involved in speech perception, there is a sparsity of knowledge detailing the basic organization of speech processing in areas specialized for speech perception in human auditory cortex. In part 1 of this work, we characterize the organization of speech processing in human superior temporal gyrus (STG) from the perspective that a description of functional organization gives a broad view of the transformation a given area performs on the stimulus. By understanding this transformation, and the transformations performed at lower areas we will be able to gain insight on the processing that gives rise to our perceptual abilities.

To characterize the organization of speech processing we use high-resolution local field potential recordings during passive speech perception in awake humans. We find that human superior temporal gyrus has a strong modulotopic organization - a higher order analog of tonotopic organization. In addition, we find that phonemes are systematically organized in modulation space such that vowels cluster at one end of the spectrotemporal modulation axis and consonants at the other. The implications of this work are twofold. First, this provides support for a unified framework of what seemed to be very disparate forms of processing in lower tonotopically organized areas and non-

tonotopically organized STG. The system is performing a Fourier-like decomposition in an iterative manner as stimuli propagate up the auditory hierarchy. Second, this work implies an organization of phonetic tuning across human STG which could facilitate a better understanding of phoneme representation and perception.

The second problem we studied concerns the brain's ability to extract information about signals of interest that are embedded in noise. This is commonly referred to as the cocktail party problem. (Cherry, 1953; McDermott, 2009). In this work we demonstrate a systematic approach to investigate this problem by creating animals that are specialized at solving the cocktail party problem, and then identifying subpopulations of neurons that can extract vocalization information in the face of noise. To create specialized rats, we reared them in noise. This serves as a natural training paradigm that occurs continuously, has natural behavior-reward contingencies, and starts at infancy when the brain is most plastic (Keuroghlian & Knudsen, 2007; Zhang, Bao, & Merzenich, 2001; de Villiers-Sidani, Chang, Bao, & Merzenich, 2007). We then tested animals behaviorally and show noise-rearing facilitates the development of superior abilities to extract vocalization information in the presence of noise. Using a method to identify neurons with the ability to encode vocalization information in the face of noise, we find neurons in primary auditory cortex of noise-reared animals with the ability to extract significantly higher rates of vocalization information compared to naive animals. Additional receptive field analysis will give insight on the underlying mechanism that allows for this increased ability. This work demonstrates a systematic approach to studying the

cocktail party problem and can be applied to other classes of noise to generate and refine models of cocktail party processing.

References

- Alivisatos, a P., Chun, M., Church, G. M., Greenspan, R. J., Roukes, M. L., & Yuste, R. (2012). The brain activity map project and the challenge of functional connectomics. *Neuron*, *74*(6), 970–4. doi:10.1016/j.neuron.2012.06.006
- Cherry, E. (1953). Some experiments on the recognition of speech, with one and two ears. *The Journal of the acoustical society of America*, *25*, 975–979.
- Keuroghlian, A. S., & Knudsen, E. I. (2007). Adaptive auditory plasticity in developing and adult animals. *Progress in neurobiology*, *82*(3), 109–21. doi:10.1016/j.pneurobio.2007.03.005
- McDermott, J. H. (2009). The cocktail party problem. *Current biology : CB*, *19*(22), R1024–7. doi:10.1016/j.cub.2009.09.005
- Zhang, L. I., Bao, S., & Merzenich, M. M. (2001). Persistent and specific influences of early acoustic environments on primary auditory cortex. *Nature neuroscience*, *4*(11), 1123–30. doi:10.1038/nn745

de Villers-Sidani, E., Chang, E. F., Bao, S., & Merzenich, M. M. (2007). Critical period window for spectral tuning defined in the primary auditory cortex (A1) in the rat. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 27(1), 180–9. doi:10.1523/JNEUROSCI.3227-06.2007

Chapter 1

Speech Based Modulotopic Organization in Human Superior Temporal Gyrus

Abstract

Despite knowing human superior temporal gyrus is critical for speech perception, the functional organization for speech processing in this region is poorly understood. Here we use high-resolution local field potential recordings during passive speech perception to characterize the underlying organization of speech processing in human superior temporal gyrus. We find that human superior temporal gyrus has a strong modulotopic organization - a higher order analog of tonotopic organization. In addition, we find that phonemes are systematically organized in modulation space such that vowels cluster at one end of the spectrotemporal modulation axis and consonants at the other.

Introduction

Human superior temporal gyrus (STG) is a tertiary auditory area above primary auditory cortex in the human auditory hierarchy. Since the description of Wernicke's area, in which stroke-induced lesions cause language comprehension deficits, STG has been considered an important region for speech processing and perception (Wernicke, 1874). However, despite considerable interest in its role in speech processing, we do not know the functional organization of spectrotemporal processing across STG. To answer this question, we compute high-density spectrotemporal receptive field (STRF) maps based on responses to speech in awake human subjects. Until recently, experimental and analytical limitations have prevented this type of analysis. Traditional noninvasive methods for recording neural activity in humans are limited in either temporal or spatial resolution (George et al., 1995; Kim et al., 1997; Dale and Halgren, 2001). We use Electrocorticography (ECoG), an alternative method restricted to rare clinical settings,

which can record neural activity with the spatial and temporal resolution necessary to generate high-density STRF maps.

In addition to high-resolution neural recordings, stimuli are needed which engage the spectrotemporal processing of interest in STG. While simple noise-like stimuli are tractable for computing receptive fields, they are often ineffective at driving activity in higher order areas like STG (Démonet et al., 1992; Binder et al., 1996, 1997, 2000). In addition, receptive fields estimated with simple stimuli often fail to characterize processing of more natural stimuli due to the nonlinearity of the system and the effects of adaptation (Theunissen et al., 2000; Sharpee et al., 2006; Christianson et al., 2008). We therefore used natural speech stimuli to characterize the functional organization of speech processing in human superior temporal gyrus.

Although speech is an ideal stimulus to characterize processing in STG, unbiased STRFs cannot be computed using natural signals with traditional techniques (Chichilnisky, 2001; Theunissen et al., 2001; Sharpee et al., 2004). To compute STRFs using speech stimuli, we used Maximally Informative Dimension (MID) analysis, an information based method designed to compute unbiased receptive fields with natural signals (Sharpee et al., 2004; Atencio et al, 2008). This method lifts the requirement of using statistically tractable noise-like stimuli for the estimation of receptive fields and allows the use of more complex natural stimuli for computing STRFs.

By using electrocorticography to record speech driven activity in STG, we were able to compute high resolution STRF maps using MID analysis and characterized the

organization of speech processing in this higher order region of human auditory cortex. Using this approach we find robust modulotopic organization - a higher order analog of tonotopic organization. In addition, we find that phonemes are systematically organized in modulation space such that vowels cluster at one end of the spectrotemporal modulation axis and consonants at the other.

RESULTS

STRF Maps

Six patients undergoing a surgical procedure for the treatment of epilepsy were implanted with ECoG arrays spanning the temporal and parietal lobe. All patients had epileptic foci outside our regions of interest. After placement of ECoG arrays and recovery from the surgical procedure, patients passively listened to 15 – 25 minutes of natural speech for the purpose of identifying the functional organization of speech processing in human STG. The speech stimulus consisted of prerecorded sentences drawn randomly from the TIMIT speech corpus with one-second silent intervals between each sentence (Garofolo et al., 1993). Speech stimuli were balanced for male and female speakers, different speakers produced each sentence, and the lexical content of each sentence was unique.

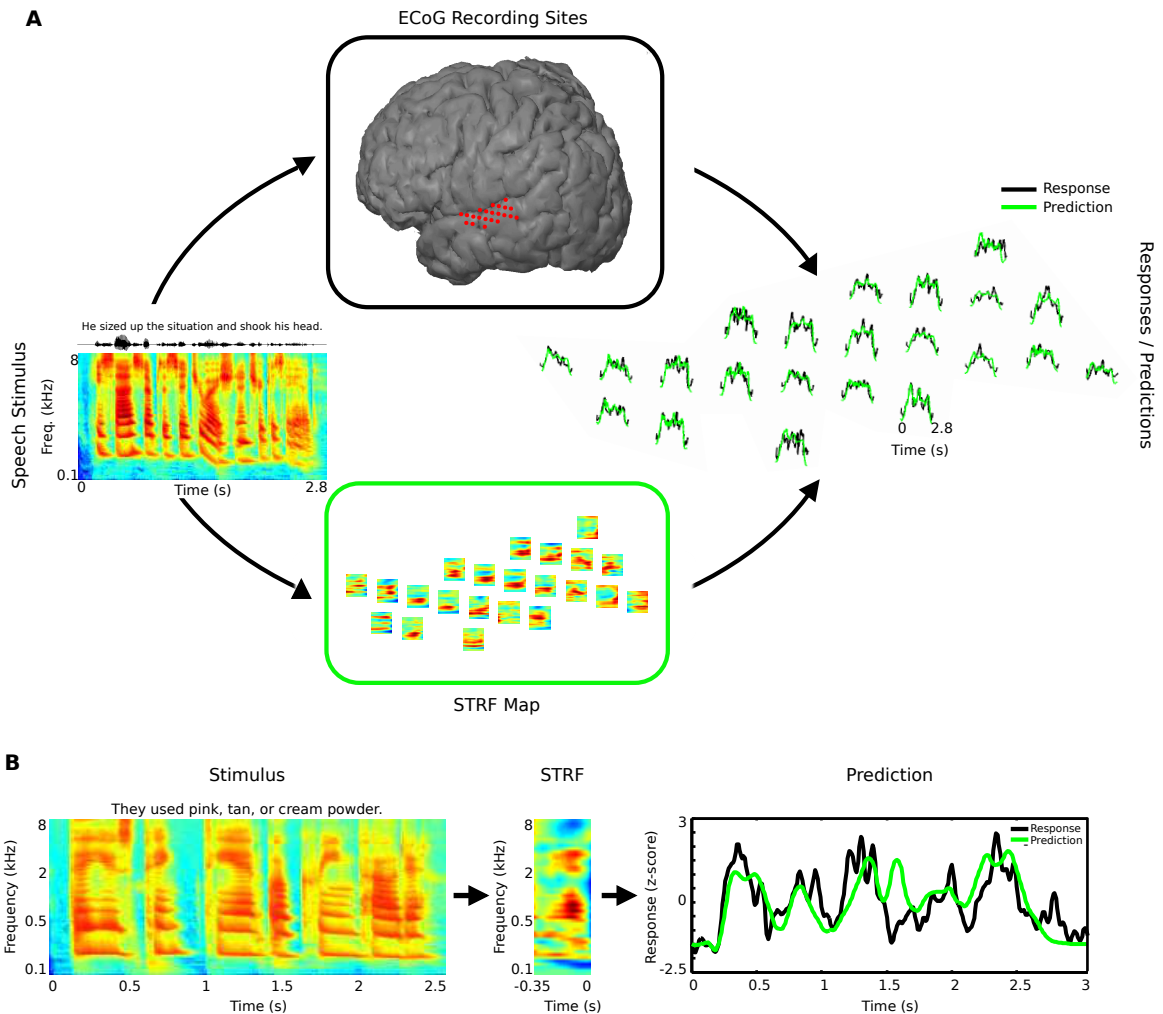


Figure 1. Experimental Approach and the STRF.

(A) Experimental approach. A spectrotemporal receptive field (STRF) was computed for each ECoG site to generate a STRF map for each patient. The measured and predicted responses for the sentence “He sized up the situation and shook his head,” are shown. The computed STRFs are predictive of neural responses and therefore characterize spectrotemporal processing at each site and how spectrotemporal processing varies across STG. **(B)** The STRF. An STRF describes the spectrotemporal envelope structure in the stimulus that drives activity. Predicted responses are obtained by convolving the stimulus with the STRF and are proportional to the similarity between the spectrotemporal content in the stimulus and the STRF.

To investigate functional organization of spectrotemporal processing in STG, an STRF was computed for each cortical site to generate an STRF map for each patient (Figure 1). STRFs were computed using the high-gamma band of ECoG

recordings which correlates with tuning properties based on multi-unit spiking in auditory cortex (Crone et al., 2001; Steinschneider et al., 2008). We estimated STRFs with two methods specifically designed for use with natural signals: maximally informative dimension (MID) analysis, and normalized reverse correlation (Theunissen et al., 2001; Sharpee et al., 2004). We found MID analysis consistently gave better predictions and therefore use MID based STRFs for the remainder of analysis (supplementary Figure 1). MID based STRFs were predictive of neural responses with an average correlation coefficient between the predicted and actual response of 0.42 ± 0.12 (mean \pm std). STRFs that are significantly predictive of the neural response characterize the spectrotemporal processing taking place at each site. The cortical location of predictive STRFs for patient EC6 is depicted in Figure 2. Given that each STRF characterizes the local spectrotemporal processing, a map of STRFs characterizes how spectrotemporal processing varies across cortex, which we use to investigate the presence of functional organization in STG.

A

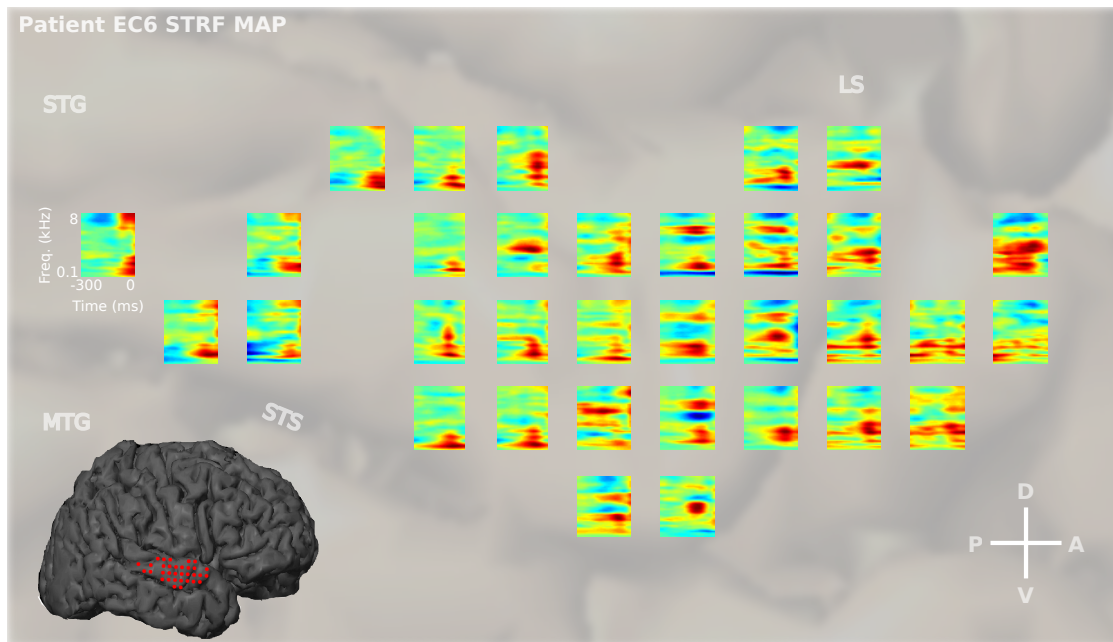


Figure 2. Patient EC6 Cortical STRF Map

(A) STRF map for patient EC6. All STRFs included in maps are predictive of neural response (mean correlation coefficient between the predicted and actual response: 0.42 ± 0.12 std). STRFs near each other tend to be similar, suggesting spectrotemporal processing changes gradually across the surface of STG. LS = lateral sulcus; STS = superior temporal sulcus; STG = superior temporal gyrus; MTG = medial temporal gyrus, CS = central sulcus.

Modulotopic Organization of Human STG

The basilar membrane and early auditory processing perform a Fourier decomposition of the acoustic waveform in terms of spectral frequency to generate a spectrographic representation of sound. Any spectrogram representation of sound can be further decomposed in terms of its Fourier components, which are spectrotemporal modulations (Figure 3A, Singh and Theunissen, 2003; Woolley et al., 2005). A spectrotemporal modulation is the

two-dimensional analog of spectral frequency and takes the form of a ripple in time-frequency space (Figure 3B). Analogous to tonotopic organization, which reflects an organized decomposition of sound in terms of spectral frequency, we looked for the presence of modulotopic organization, which would reflect the organized decomposition of sound in terms of spectrotemporal modulations. In present, this would imply the auditory system is performing a stacked Fourier decomposition of sound in which an organized spectral decomposition is performed by early auditory processing and an organized spectrotemporal modulation based decomposition is performed in higher order areas within the auditory pathway.

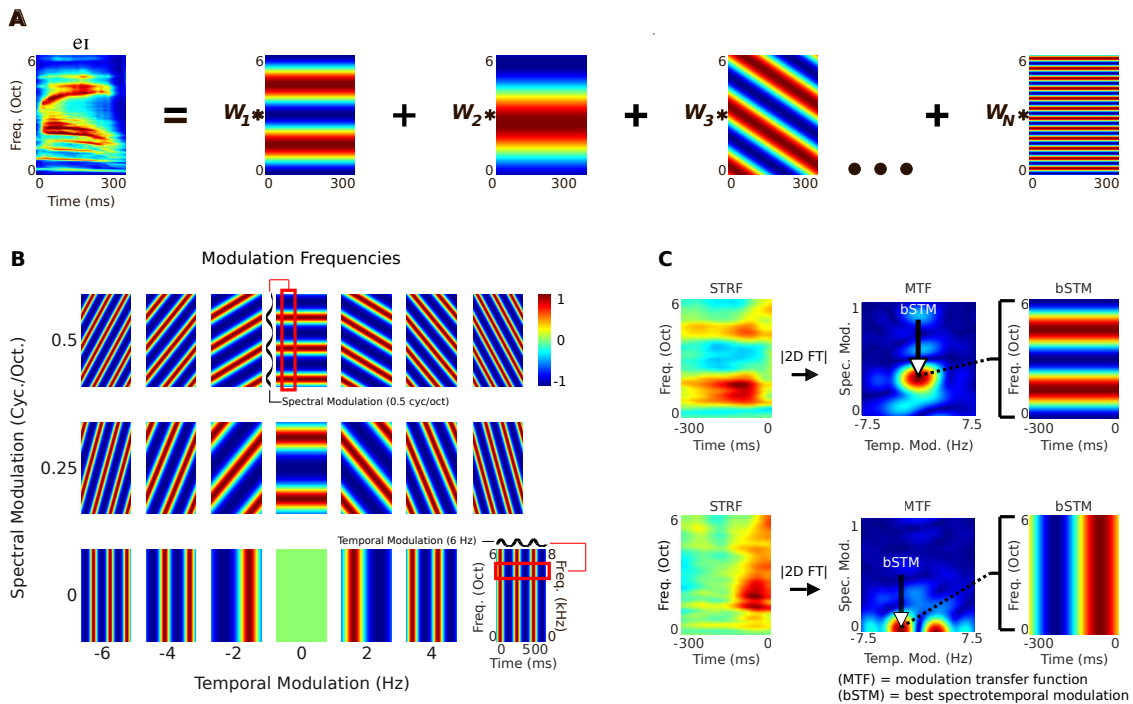


Figure 3. Spectrotemporal modulation tuning is a higher-order analog of frequency tuning.

(A) Any spectrogram can be decomposed in terms of its spectrotemporal modulation content. The spectrotemporal modulation components of the phoneme /e/ are shown. Like a weighted sum of sinusoids will reconstruct an acoustic waveform, a weighted sum of spectrotemporal modulations will reconstruct the envelope of a phoneme exactly. **(B)** The Fourier components of a spectrogram are spectrotemporal modulations or “ripples”. A bank of ordered spectrotemporal modulations are shown. This is the Fourier basis of a spectrogram. Each spectrotemporal modulation is specified by its’ spectral and temporal modulation frequency which characterize the periodicity with which energy is modulated along the spectral and temporal axis. **(C)** Analogous to a frequency tuning curve, the tuning for spectrotemporal modulations is characterized by the modulation transfer function (MTF). The MTF is derived as the magnitude of the two-dimensional Fourier transform of an STRF. Like the best frequency of a frequency tuning curve, the peak of the MTF is called the best spectrotemporal modulation (bSTM).

To investigate presences of modulotopic organization, we characterize spectrotemporal modulation tuning at each ECoG site. As with a frequency tuning curve, which characterizes spectral frequency tuning, the modulation transfer function (MTF) characterizes spectrotemporal modulation tuning and is given by taking the magnitude of the two-dimensional Fourier transform of each STRF (Figure 3C). Furthermore, like the best frequency of a tuning curve, the peak of the modulation transfer function is called the best spectrotemporal modulation (bSTM) and specifies the spectrotemporal modulation that drives activity at a given site most strongly (Figure 3C).

From the modulation transfer functions, we examined the distribution of spectrotemporal modulation tuning along STG. Figure 4A shows the distribution of best spectrotemporal modulations (bSTMs) across all patients. As seen in other auditory areas, bSTMs show a spectral-temporal tradeoff in which they fall along a hyperbolic curve that runs from high-spectral modulation frequencies to high-temporal modulation

frequencies (Miller et al., 2002; Schönwiesner and Zatorre, 2009; Rodríguez et al., 2010). We fit this relationship with an exponential function, which is converted into a color map. Actual bSTM data points were then mapped to the closest point on the curve and assigned the corresponding color. This color is plotted for each cortical location to generate an undistorted map of spectrotemporal modulation tuning across STG (Figure 4B).

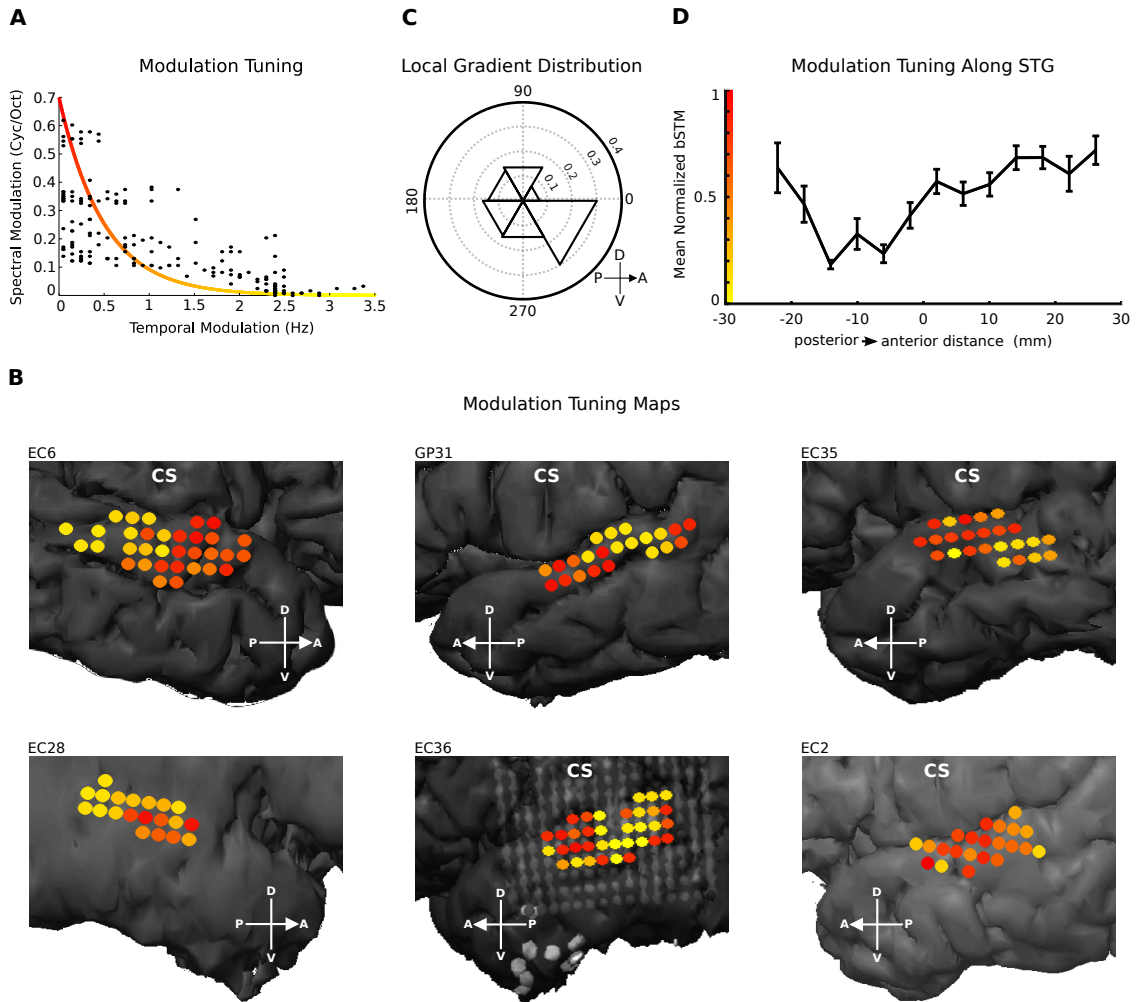


Figure 4. Modulotopic Organization in Human STG.

(A) Distribution of best spectrotemporal modulation (bSTM) tuning across patients. The distribution of bSTMs shows a spectral-temporal tradeoff. We fit this distribution with a curve which is used as a color map. (B) Modulotopic organization. Best spectrotemporal modulation values in (A) were projected onto the color map, and each cortical site was assigned the corresponding color. Within each map there is a significant degree of neighborhood similarity in which sites near each other tend to have similar values ($p < 0.005$, patients EC6, EC28, GP31, EC36, EC35; $p < 0.05$ patient EC2; permutation test). In addition high temporal modulation tuned regions (yellow) transition into high spectral modulation tuned regions (red) thus forming an organized distribution of modulation tuning along the major axis of STG (C) Posterior to anterior gradient. A local gradient was computed at each site. The rose plot depicts the distribution of local gradient

directions for all sites and shows a dominant gradient direction from the posterior to anterior direction along STG ($p < 0.05$, permutation test). **(D)** Average normalized modulation tuning as a function of posterior to anterior distance. Maps were aligned by their map centers (see methods) and average modulation tuning as a function of distance from the map center was computed. Across patients there is a high temporal modulation tuned region which transitions into high spectral modulation tuned regions in the anterior and posterior direction.

Within these maps there is a significant amount of structure with strong neighborhood similarity in which sites near each other have similar modulation tuning ($p < 0.005$, patients EC6, GP31, EC28, EC35, EC36; $p < 0.05$ patient EC2; permutation test). Regions tuned to high-temporal modulation frequencies (yellow) transition smoothly into regions tuned to high-spectral modulation frequencies (red). We quantified the spatial organization of modulation tuning by computing the direction of the local gradient at each site and then examined the distribution of local gradient directions across all patients (Figure 4C). Consistent with the visual appearance of the maps, the dominant direction of local gradients is in the posterior to anterior direction along the major axis of STG ($p < 0.05$, permutation test). We further quantified spatial organization by computing the modulation tuning map values as function of distance along the posterior to anterior axis (Figure 4D). As can be seen, there is a high-temporal modulation tuning centrally, just posterior to central sulcus, which transitions to high-spectral modulation tuning in the anterior and posterior directions. These data reveal the presence of a modulotopic organization in human STG in which tuning varies from high temporal modulation tuning in central STG to high-spectral modulation tuning in anterior STG, and to some extent posterior STG as well.

Modulation Content of Phonemes

In light that this region of the auditory system is functionally organized for modulation tuning, we analyzed the spectrotemporal modulation content of phonemes to examine the relationship between modulotopic organization and speech sound representation. We focused on the informative modulation content in phonemes (see methods), which is the spectrotemporal modulation content in each phoneme that is different from the average modulation content of all phonemes. This modulation information potentially can be used for phoneme discrimination and is computed by subtracting the average modulation spectra of all phonemes from each individual phoneme. For brevity we will refer to the informative modulation spectrum simply as the modulation spectrum. Figure 5A shows the spectrograms and modulation spectra for three phonemes that occupy different regions in spectrotemporal modulation space. To assess the distribution of modulation content across phonemes, the peak of each modulation spectrum is plotted along the same curve fit to the modulation tuning data in figure 4A (Figure 5B). The modulation content across phonemes has a similar spectral-temporal tradeoff as the modulation tuning of neural sites in STG. In addition, the organization of phonemes along this curve is related to their linguistic classifications such that vowels and consonants cluster at opposite ends of the modulation dimension. Vowels and semivowels cluster at the high-spectral modulation end of the curve. This reflects the general formant structure of vowels in which energy fluctuates over frequency (high spectral modulation), but is fairly constant over time (low temporal modulation). By contrast, consonants cluster at the high-temporal modulation end. This reflects the general structure of many consonants, which are composed of broadband energy that fluctuates rapidly over time (high temporal modulation), but is relatively constant over frequency (low spectral modulation). These data reveal a non-random organization of phonemes in modulation space that correlates with their linguistic

groupings. In conjunction with modulotopic organization these results suggest an organization of phonetic information on the STG in which vowels are predominantly represented in anterior regions and consonants in posterior regions.

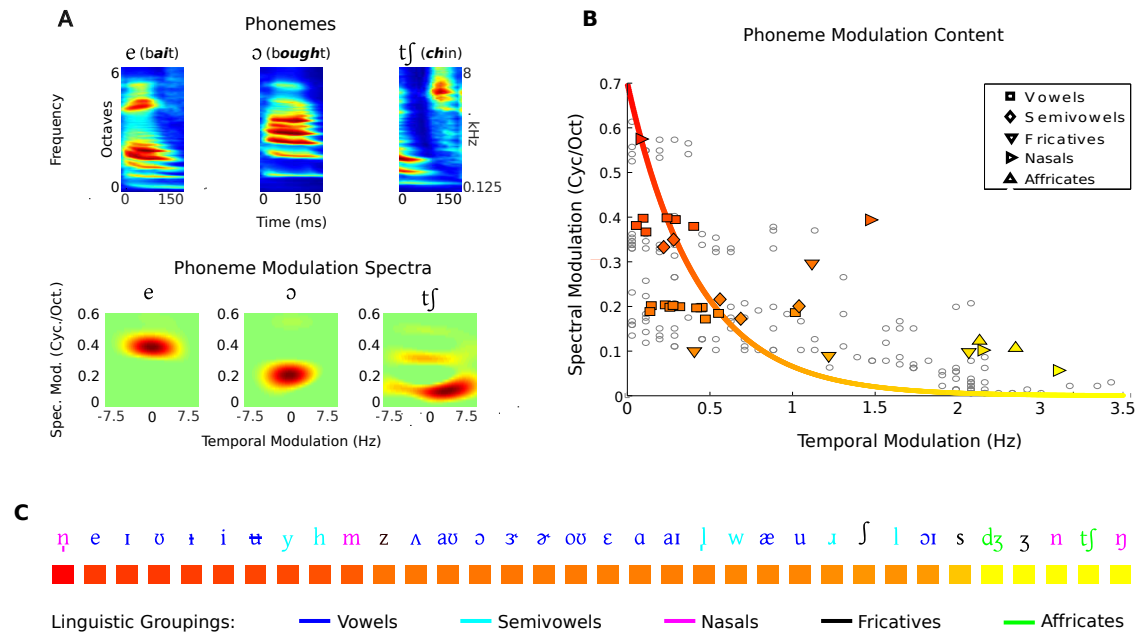


Figure 5 Spectrotemporal Modulation Content of Phonemes

(A) Phonemes and their modulation spectra. Phonemes from the words *bait*, *bought*, and *chin* with their corresponding modulation spectra are shown. The high-energy regions of the modulation spectra occupy different regions in modulation space ranging from high-spectral low-temporal modulation for *bait* to high-temporal low-spectral modulation for *chin*. (B) Phoneme modulation spectra peaks show a spectral-temporal tradeoff. The peak of each phoneme’s modulation spectrum is plotted along with bSTM values and the color map from figure 4A. The modulation spectra peaks show a spectral-temporal tradeoff similar to modulation tuning. (C) Phonemes show organization along the modulation frequency axis. Phonemes were mapped to the closest point on the color map in (B) and given the corresponding color. Vowels and semivowels cluster at the high spectral modulation end of the modulation frequency axis, while consonants cluster at the high temporal modulation end of modulation frequency axis.

Discussion

We demonstrated here the presence of a modulotopic organization in human superior temporal gyrus. These findings are evidence of a stacked Fourier decomposition in the human auditory pathway in which modulotopic organization reflects a second order Fourier decomposition of the signal, analogous to the frequency decomposition performed by the basilar membrane and areas lower in the auditory hierarchy. Figure 6 shows a simplified model of this stacked decomposition in which the acoustic waveform is initially decomposed into its spectral frequency components to generate a spectrogram representation of the signal. This time-frequency representation is then decomposed into its spectrotemporal modulation components as reflected by the modulotopic organization. This overall stacked decomposition framework has been implemented in biologically inspired computational models of auditory processing as well as automatic speech recognition algorithms (Yang et al., 1992; Chi et al., 1999; Kleinschmidt and Gelbart, 2002; Mesgarani et al., 2006; Meyer and Kollmeier, 2011). In these automatic speech recognition algorithms, modulation frequency based representation allows for more robust word recognition despite variation in speech due to changes in speaker, pitch, or loudness (Meyer and Kollmeier, 2011). In addition, modulation frequency based representation increases the robustness of these algorithms in the presence of background noise (Kleinschmidt and Gelbart, 2002; Mesgarani et al., 2006). This robustness arises because speech and non-speech often have different modulation content. Even if they overlap spectrally, a modulation based representation can separate them due to differences in modulation content. These results suggest modulotopic organization and the resulting modulation based representation are important for generating invariant representations of speech stimuli and contribute to

our ability to understand speech in the face of background noise.

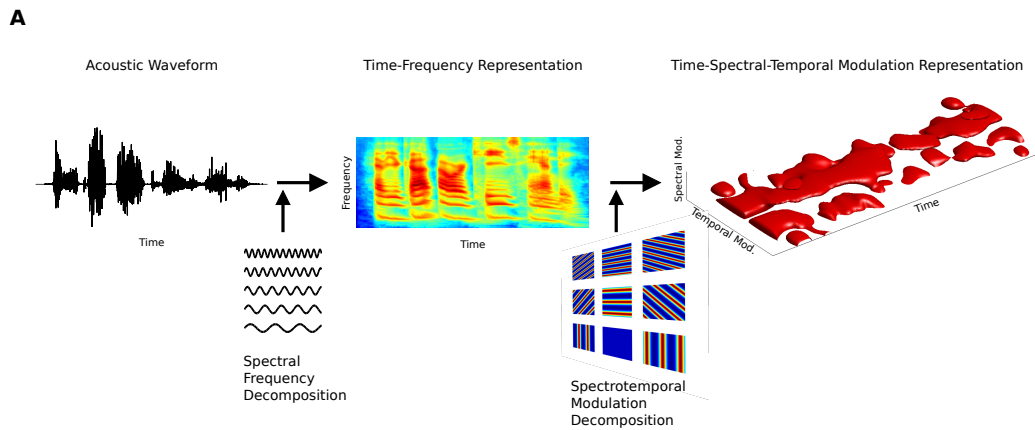


Figure 6. Stimulus Transformations Along the Auditory Hierarchy

(A) Schematic representation of a stacked frequency decomposition suggested by modulotopic organization. Early auditory processing decomposes an acoustic waveform into its frequency components to generate a spectrographic time-frequency representation. Modulotopic areas decompose this time-frequency representation in terms of its spectrotemporal modulation content at higher-level areas of the auditory system.

Given the role of STG in speech processing, we examined the relationship between modulation tuning and the modulation content present in phonemes. The modulation content of phonemes exhibits the same spectral-temporal tradeoff present in the distribution of modulation tuning, generally spanning a similar range of modulation frequencies. In addition, vowels and semi-vowels clustered at the high spectral modulation end of the axis while consonants clustered at the high temporal modulation end. This characterization further illuminates the relationship between the functional organization of human STG and the acoustic properties of speech sounds and suggests the presence of a coarse organization in representation of phonetic information across human STG in which vowels are represented anteriorly in high spectral modulation tuned regions and consonants posteriorly in more temporally tuned regions. In support of this, our data is consistent with previous work showing selective activation of anterior

STG by vowel sounds (Obleser et al., 2006). If a more detailed map of phonetic tuning across STG exists, modulotopic organization could explain such an organization based on the modulation content in different phoneme classes.

How does the distribution of modulation tuning across the STG compare to the modulation content critical for word recognition? The distribution of best spectrotemporal modulations frequencies seen in STG span the range of 0 - 0.7 cyc/oct and 0 - 3.5 Hz. Elliot et al. showed the modulation content most critical for word recognition are spectral modulations below 1 cyc/oct and temporal modulations below 7.75 Hz (Elliott and Theunissen, 2009). This shows modulation tuning in STG spans a similar spectral modulation range and roughly half the temporal modulation range critical for word recognition. Although the distribution of bSTMs identified in STG does not span the full spectrotemporal modulation range identified by Elliot et al., there are potential reasons for an imperfect match. First, it is possible that the surface of STG only represents part of the neural area devoted to processing this critical window of modulation frequencies. As noted, some phonemes are not included in figure 5B because their temporal modulation content extend beyond the range of best temporal modulation tuning seen in STG. Most notably, stop consonant modulation spectral peaks have temporal modulation values above 3.5 Hz. It is possible there are areas critical for speech perception which have higher temporal modulation tuning, but lay outside STG - possibly in the lateral or superior temporal sulcus. An additional caveat to comparing modulation tuning in STG and the modulation content critical for word recognition is uncertainty in the borders of this critical modulation content window defined by Elliot et al. The boundaries tested to define critical spectrotemporal modulation content in speech were coarsely spaced due to the required psychophysical testing at each boundary condition. It is possible that the upper boundary of this

window is actually lower than 7.75 Hz. Finally, bSTM values only represent peak tuning. Sites with bSTM values close to 3.5 Hz have a considerable response for temporal modulations above 3.5 Hz due to the fairly broad bandwidth of the modulation filters. Overall, the window of critical modulation content defined by Elliot et al. is small given the full range of spectrotemporal modulation content present in speech (Singh and Theunissen, 2003). The fact that modulation tuning in STG falls within this window and spans a considerable range of it suggests STG is matched to the statistics of speech stimuli and supports the idea that STG is a specialized structure for speech processing.

The characterization of modulotopic organization was simplified by the spectral-temporal tradeoff seen in modulation tuning. To what extent is this a general feature of auditory processing? This spectral-temporal tradeoff has also been seen in other areas within the lemniscal auditory pathway including inferior colliculus, the ventral division of the medial geniculate nucleus, and primary auditory cortex of cats and primary auditory cortex of humans (Miller et al., 2002; Schönwiesner and Zatorre, 2009; Rodríguez et al., 2010). In addition, sparse coding models of speech processing learn STRFs with modulation tuning properties that show a similar spectral-temporal tradeoff (Carlson et al., 2012). Collectively, this shows that the spectral-temporal tradeoff in modulation tuning is a consistent property within ascending areas of the auditory pathway and suggests the computational role of this tradeoff as an important property of auditory processing.

As stimulus information propagates up the sensory hierarchy it is successively transformed into new representations. These transformations eventually give rise to our

perceptual abilities. Our results are consistent with a stimulus transformation that decomposes auditory information in terms of spectrotemporal modulations, which forms the basis of modulotopic organization. Are the stimulus transformations at successive stages of processing drastically different or similar in nature? Given the direct relationship between spectral frequency decomposition and modulation frequency decomposition, this would suggest that at some level certain stimulus transformations are similar in nature at successive levels of the hierarchy. We hope that concise descriptions of processing provided by functional organization studies at each level of the auditory hierarchy will continue to provide insight to the stimulus transformations taking place as information ascends the auditory hierarchy and help illuminate the different processing steps which give rise to our perceptual abilities.

METHODS

Patients and Neural Recordings

Subdural ECoG arrays (interelectrode distance: 4mm) were placed over the temporal and parietal lobes unilaterally in six patients (2 right hemisphere, 4 left hemisphere) undergoing a neurosurgical procedure for the treatment of epilepsy. All patients were native English speakers with no hearing or communication deficits. All experimental protocols were approved by the University of California, San Francisco Institutional Review Boards and Committees on Human Research. Array placement was determined by clinical criteria. After array placement, patients were allowed to recover and the monitoring phase of the procedure started, during which time patients were asked to passively listen to 15 – 25 minutes of natural speech while cortical activity was recorded. Signals were amplified and sampled at 3052 Hz. After rejection of electrodes with excessive noise or artifacts, signals were referenced to a common average and the

high gamma band (75 - 150 Hz) was extracted using the Hilbert Transform (Crone et al., 2001). The resulting signal for each electrode was z-scored based on the mean and standard deviation of spontaneous activity.

Stimuli

Speech stimuli were delivered binaurally through calibrated ear insets at 70 dB average sound pressure level. The frequency power spectrum of stimuli spanned 0 - 8000 Hz. The stimulus set consisted of prerecorded (2 – 4 second) sentences drawn randomly without replacement from the phonetically transcribed TIMIT speech corpus with one-second silent intervals between each sentence presentation (Garofolo et al., 1993). Each patient was presented 484-499 sentences with the exception of patient GP36 who was presented 246 sentences. The speech corpus was balanced for male and female speakers, a different speaker produced each sentence, and the lexical content of each sentence was unique. Spectrogram representations of speech stimuli were based on a cochlear model of auditory processing (Yang et al., 1992).

Analysis

STRFs

STRFs were computed with two different methods designed specifically for use with natural signals- maximally informative dimension analysis (MID) and normalized reverse correlation (Theunissen et al., 2001; Sharpee et al., 2004). To compute STRFs using MID analysis a gradient ascent procedure was used to search for the receptive field that maximizes the KL divergence between STRF-Stimulus projection values and STRF-stimulus projection values weighted by the magnitude of the response. STRF estimates based on normalized reverse correlation were computed using open source

code available at: <http://strfpak.berkeley.edu/>. STRFs were computed with both methods on the same estimation set (90% of the total data) and cross-validated on the same test set which was withheld from the estimation process (10% of the data). MID based STRFs showed better prediction performance and were therefore used to analyze functional organization in STG. To investigate functional organization of spectrotemporal processing, we restricted our analysis to STRFs which predicted at least five percent of the total variance in the response to prevent clouding our view of functional organization with poor STRF models.

Modulation Tuning

To characterize modulation tuning based organization, the modulation transfer function (MTF) for each site was computed by taking the magnitude of the two-dimensional Fourier transform ($\mathfrak{S}_2\{\bullet\}$) of each STRF:

$$MTF(\omega_t, \omega_s) = \left| \mathfrak{S}_2\{STRF(t, f)\} \right|$$

Where (t, f) are the time and frequency and (ω_t, ω_s) are the temporal and spectral modulation respectively. The best spectrotemporal modulation (bSTM) is defined as the peak of the MTF. We take the absolute value of the temporal modulation frequency of each bSTM to maximize the number of data points in a single quadrant for analysis of the distribution of bSTMs as function of spectral and temporal modulation.

To generate modulation tuning maps, we fit the bSTM distribution with an exponential function which is converted into a color map. Actual best bSTM data points were then mapped to the closest point on the curve and assigned the corresponding color which is then plotted at the corresponding cortical location. To fit the data and assign data points a color, we first normalized the spectral and temporal modulation values to have a range [0-1] so temporal and spectral modulation values are weighted equally. An exponential function was then fit to the data by minimizing mean squared error and a color map was generated through a linear mapping between color and arc length along the fit exponential function. bSTM values were then mapped to the closest point (defined by Euclidean distance) on the color map and assigned a color.

Spatial Analysis

A permutation test was used to assess the significance of neighborhood similarity. For each site a neighborhood similarity metric is computed by taking absolute value of the difference between each sites modulation tuning value and its neighbor and averaging over all directly adjacent neighbors. Site neighborhood similarity indices were then averaged together to compute the mean neighborhood similarity index for the map. The sites in the map are then randomly permuted 10,000 times and the map neighborhood similarity index is recomputed on each permutation to generate the distribution of possible neighborhood similarity indices for a randomly organized map. The true neighborhood similarity index is compared to the random map neighborhood similarity index distribution to access the level of significance.

To analyze spatial organization we characterized the distribution of local gradients as well as modulation tuning as a function of distance along the posterior to anterior axis of each map. To determine the distribution of local gradients, the numeric

gradient at each ECoG site was computed (Cheung et al., 2001). In order to combine gradients across patients, gradient vectors in left hemisphere subjects were flipped across the ordinate to generate their mirror image. The distribution of gradient directions was then computed and displayed in the form of a rose plot. Significance was determined by using a two-sample Kolmogorov-Smirnov test between this distribution and the null hypothesis distribution which was determined by randomly permuting each map prior to gradient calculation and repeating this procedure 10,000 times to obtain an accurate estimate.

To quantify modulation tuning as a function of distance, modulation tuning values were collapsed along the major axis of each map which runs parallel to the long axis of STG. This quantifies modulation tuning as a function of posterior to anterior distance. Each modulation tuning distance function displayed a characteristic peak and trough defined by the maximum and minimum average modulation tuning values. The center of each modulation tuning distance function is defined as the midpoint between its peak and trough. To combine data across patients, data from left hemisphere patients was flipped across the ordinate. Modulation tuning distance functions were then aligned by their centers and averaged to quantify modulation tuning as a function of distance across patients.

Analysis of Phoneme Modulation Content

The modulation spectrum for each phoneme was estimated by taking the squared magnitude of the two-dimensional Fourier transform ($\mathfrak{F}_2\{\bullet\}$) of each phoneme instance and averaging over instances:

$$MS(\omega_t, \omega_s) = \frac{1}{N} \sum_i^N \left| \mathfrak{F}_2\{Phn_i(t, f)\} \right|^2$$

Instances of each phoneme were zero meaned prior to calculation of the modulation spectra, and analysis of phoneme instances between 100 -150 ms in duration were used for modulation spectra estimation. Due to the statistics of speech and natural signals in general, power is concentrated at low modulation frequencies and decreases as function of frequency according to $\frac{1}{\omega^\alpha}$ where ω is either spectral or temporal modulation frequency and α ranges between 0.78 and 2.26 depending on the class of sounds analyzed (Field, 1987; Singh and Theunissen, 2003). Due to this, the majority of the power in all phonemes is concentrated at very low modulation frequencies making the raw modulation spectra of all phonemes very similar and relatively uninformative. A more informative characterization of the modulation spectra of each phoneme would highlight the modulation content specific to that phoneme. We therefore compute the informative modulation spectrum of each phoneme by subtracting the ensemble phoneme modulation spectrum from the each individual phoneme modulation spectrum to identify modulation content in each phoneme that is different from the average modulation content across phonemes. To compute the ensemble phoneme modulation spectrum we normalized the modulation spectrum for each phoneme to have a total power of 1, and average across phonemes with each phoneme weighted equally. The Informative modulation content of each phoneme is then computed by subtracting the ensemble phoneme modulation spectrum from the individual phoneme modulation spectrum (both normalized to have a total power of 1) to identify the modulation content which differentiates the phoneme of interest from other phonemes. This is the modulation content that could potentially be used by the nervous system to differentiate phonemes.

References

- Atencio, C.A., Sharpee, T.O., and Schreiner, C.E.(2008) Cooperative nonlinearities in auditory cortical neurons. *Neuron* 58, 956-966.
- Binder, J.R., Frost, J. a, Hammeke, T. a, Bellgowan, P.S., Springer, J. a, Kaufman, J.N., and Possing, E.T. (2000). Human temporal lobe activation by speech and nonspeech sounds. *Cerebral Cortex (New York, N.Y. : 1991)* 10, 512–528.
- Binder, J.R., Frost, J. a, Hammeke, T. a, Cox, R.W., Rao, S.M., and Prieto, T. (1997). Human brain language areas identified by functional magnetic resonance imaging. *The Journal of Neuroscience : the Official Journal of the Society for Neuroscience* 17, 353–362.
- Binder, J.R., Frost, J. a, Hammeke, T. a, Rao, S.M., and Cox, R.W. (1996). Function of the left planum temporale in auditory and linguistic processing. *Brain : a Journal of Neurology* 119 (Pt 4, 1239–1247.
- Carlson, N., Ming, V., and DeWeese, M. (2012). Sparse codes for speech predict spectrotemporal receptive fields in the inferior colliculus. *PLoS Computational Biology* 8, e1002594.
- Cheung, S., Bedenbaugh, P.H., Nagarajan, S.S., and Schreiner, C.E. (2001). Functional organization of squirrel monkey primary auditory cortex: responses to pure tones. *Journal of Neurophysiology* 85, 1732–1749.

Chi, T., Gao, Y., Guyton, M.C., Ru, P., and Shamma, S. (1999). Spectro-temporal modulation transfer functions and speech intelligibility. *The Journal of the Acoustical Society of America* 106, 2719–2732.

Chichilnisky, E.J. (2001). A simple white noise analysis of neuronal light responses. *Network: Computation in Neural Systems* 12, 199–213.

Christianson, G.B., Sahani, M., and Linden, J.F. (2008). The consequences of response nonlinearities for interpretation of spectrotemporal receptive fields. *The Journal of Neuroscience : the Official Journal of the Society for Neuroscience* 28, 446–455.

Crone, N.E., Boatman, D., Gordon, B., and Hao, L. (2001). Induced electrocorticographic gamma activity during auditory perception. *Clinical Neurophysiology* 112, 565–582.

Dale, a M., and Halgren, E. (2001). Spatiotemporal mapping of brain activity by integration of multiple imaging modalities. *Current Opinion in Neurobiology* 11, 202–208.

Démonet, J.F., Chollet, F., Ramsay, S., Cardebat, D., Nespoulous, J.L., Wise, R., Rascol, a, and Frackowiak, R. (1992). The anatomy of phonological and semantic processing in normal subjects. *Brain : a Journal of Neurology* 115 (Pt 6, 1753–1768.

Elliott, T.M., and Theunissen, F.E. (2009). The modulation transfer function for speech intelligibility. *PLoS Computational Biology* 5, e1000302.

Field, D.J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America. A, Optics and Image Science* 4, 2379–2394.

Garofolo, J.S., Lamel, L., Fisher, W., JG, F., and Pallet, D. (1993). Acoustic-phonetic continuous speech corpus. (Linguistic Data Consortium, Philadelphia).

George, J., Aine, C., and Mosher, J. (1995). Mapping function in the human brain with magnetoencephalography, anatomical magnetic resonance imaging, and functional magnetic resonance imaging. *Journal of Clinical Neurophysiology* 12, 406–431.

Kim, S., Richter, W., and Ugurbil, K. (1997). Limitations of temporal resolution in functional MRI. *Magnetic Resonance in Medicine* 37, 631–638.

Kleinschmidt, M., and Gelbart, D. (2002). Improving word accuracy with Gabor feature extraction. *Proc. ICSLP*.

Mesgarani, N., Slaney, M., and Shamma, S.A. (2006). Discrimination of Speech From Nonspeech Based on Multiscale Spectro-Temporal Modulations. *14*, 920–930.

Meyer, B.T., and Kollmeier, B. (2011). Robustness of spectro-temporal features against intrinsic and extrinsic variations in automatic speech recognition. *Speech Communication* 53, 753–767.

Miller, L.M., Escabí, M. a, Read, H.L., and Schreiner, C.E. (2002). Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. *Journal of Neurophysiology* 87, 516–527.

Obleser, J., Boecker, H., Drzezga, A., Haslinger, B., Hennenlotter, A., Roettinger, M., Eulitz, C., and Rauschecker, J.P. (2006). Vowel sound extraction in anterior superior temporal cortex. *Human Brain Mapping* 27, 562–571.

Rodríguez, F. a, Read, H.L., and Escabí, M. a (2010). Spectral and temporal modulation tradeoff in the inferior colliculus. *Journal of Neurophysiology* 103, 887–903.

Schönwiesner, M., and Zatorre, R.J. (2009). Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. *Proceedings of the National Academy of Sciences of the United States of America* 106, 14611–14616.

Sharpee, T., Rust, N.C., and Bialek, W. (2004). Analyzing neural responses to natural signals: maximally informative dimensions. *Neural Computation* 16, 223–250.

Sharpee, T.O., Sugihara, H., Kurgansky, A.V., Rebrik, S.P., Stryker, M.P., and Miller, K.D. (2006). Adaptive filtering enhances information transmission in visual cortex. *Nature* 439, 936–942.

Singh, N.C., and Theunissen, F.E. (2003). Modulation spectra of natural sounds and ethological theories of auditory processing. *The Journal of the Acoustical Society of America* 114, 3394.

Steinschneider, M., Fishman, Y.I., and Arezzo, J.C. (2008). Spectrotemporal analysis of evoked and induced electroencephalographic responses in primary auditory cortex (A1) of the awake monkey. *Cerebral Cortex* (New York, N.Y. : 1991) 18, 610–625.

Theunissen, E., Sen, K., and Doupe, A.J. (2000). Spectral-Temporal Receptive Fields of Nonlinear Auditory Neurons. *20*, 2315–2331.

Theunissen, F.E., David, S.V., Singh, N.C., Hsu, a., Vinje, W.E., and Gallant, J.L. (2001). Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *Network: Computation in Neural Systems* 12, 289–316.

Wernicke, C. (1874). *Der Aphasische Symptomencomplex: Eine psychologische Studie auf anatomischer Basis*. Breslau: Cohn and Weigert.

Woolley, S.M.N., Fremouw, T.E., Hsu, A., and Theunissen, F.E. (2005). Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds. *Nature Neuroscience* 8, 1371–1379.

Yang, X., Wang, K., and Shamma, S. a. (1992). Auditory representations of acoustic signals. *IEEE Transactions on Information Theory* 38, 824–839.

|

Chapter 2

Speech Based Analysis of Tonotopic Organization in Human Superior Temporal Gyrus

Abstract

Tonotopic organization is a prominent form of functional organization from the periphery up through the primary auditory cortex. Superior temporal gyrus is a tertiary auditory area in human auditory cortex specialized for speech processing. Using electrocorticography, we examined the organization of spectral processing in superior temporal gyrus during natural speech perception in awake humans. No significant tonotopic organization was found suggesting that tonotopic organization is not a dominant organizing principle in human superior temporal gyrus during speech processing.

Introduction

The basilar membrane performs a spectral decomposition on incoming acoustic signals that gives rise to frequency tuned inner hair cells. Inner hair cell frequency preferences shift from high frequency to low frequency as a function of distance along the basilar membrane. This orderly array of frequency tuning is called tonotopic organization and is a fundamental organizing principle in the auditory system from the sensory epithelia, up through the primary auditory cortex (Aitkin and Webster, 1972; Merzenich and Reid, 1974; Merzenich and Roth, 1975). Tonotopic organization has also been observed in non-primary belt areas, however, organization can be less robust and not all regions show tonotopic organization (Reale and Imig, 1980; Bizley et al., 2005a; Petkov et al., 2006). Within the auditory hierarchy of the humans, tonotopic organization has been shown as high as primary auditory cortex (Formisano et al., 2003; Talavage

et al., 2004a; Woods et al., 2009; Striem-Amit et al., 2011; Moerel et al., 2012) .

Human superior temporal gyrus (STG) is a higher order auditory area within human auditory cortex with projections from A1, and a specialized role in speech processing. Recent studies using functional magnetic resonance imaging and tonal stimuli have shown evidence of tonotopic organization in human STG, however, work using electrocorticography and tonal stimuli showed only a weak organization that was present in a subset of subjects (Talavage et al., 2004b; Humphries et al., 2010a; Striem-Amit et al., 2011; Moerel et al., 2012; Nourski et al., 2012). Due to the specialized role of STG in speech processing, we approach the question of tonotopic organization from a unique perspective and ask, within the context of natural speech processing, does human STG show tonotopic organization?

This is a difficult question to ask for a number of reasons. Because the system is nonlinear, spectral processing characterized using tonal stimuli may not generalize to spectral processing of speech stimuli (Theunissen et al., 2000; Christianson et al., 2008). Therefore, to determine if spectral processing of speech stimuli is organized tonotopically, it is necessary to use natural speech. This is difficult to address with traditional noninvasive recording techniques and analytical methods. Because the spectral content of speech changes rapidly, it is necessary to record responses with high temporal resolution to capture rapid changes in neural activity that reflect rapid changes in spectral content. In addition, it is necessary to record neural activity with high spatial resolution to characterize how spectral tuning changes as a function of position across human

STG. Traditional noninvasive recording techniques have either low temporal (e.g. fMRI) or low spatial (e.g. MEG) resolution (George et al., 1995; Kim et al., 1997; Dale and Halgren, 2001). To overcome this limitation, we use electrocorticography (ECoG), a method restricted to rare clinical settings that has the spatial and temporal resolution necessary to investigate tonotopic organization in the context of natural speech perception.

Although ECoG provides the necessary spatiotemporal resolution to investigate tonotopic organization in the context of speech processing, computing receptive fields with natural stimuli as opposed to statistically tractable noise-like stimuli has only recently been developed. Traditional approaches to compute receptive fields will produce biased estimates with natural stimuli (Chichilnisky, 2001; Theunissen et al., 2001; Sharpee et al., 2004). We therefore use a more computationally intensive approach designed specifically for use with natural signals called maximally informative dimension (MID) analysis (Sharpee et al., 2004). By using high spatial and temporal resolution recordings, in conjunction with MID analysis, we are able to characterize spectral processing across STG and investigate the presence of tonotopic organization in the context of natural speech perception in awake humans.

RESULTS

Six patients undergoing a surgical procedure for the treatment of epilepsy

were implanted with ECoG arrays spanning the temporal and parietal lobe. All patients had epileptic foci outside our regions of interest. After placement of ECoG arrays and recovery from the surgical procedure, patients passively listened to 15 – 25 minutes of natural speech for the purpose of identifying the functional organization in human STG. The speech stimulus consisted of prerecorded sentences drawn randomly from the TIMIT speech corpus with one-second silent intervals between each sentence (Garofolo et al., 1993). Speech stimuli were balanced for male and female speakers, different speakers produced each sentence, and the lexical content of each sentence was unique.

Tonotopic Organization in Human STG

To investigate functional organization of spectral processing in the context of natural speech stimuli, a spectrotemporal receptive field (STRF) was computed for each cortical site using the MID method (see Methods). STRFs were summed along the temporal axis to isolate the purely spectral component of processing called the spectral receptive field (SRF) (Figure 1). The peak of each spectral receptive field is defined as the best frequency (BF).

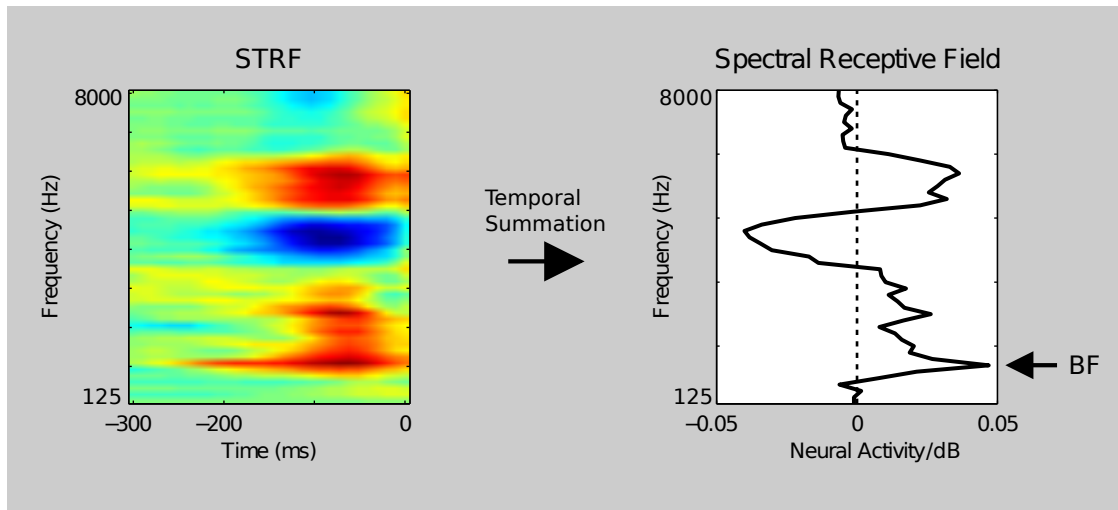


Figure 1. Identification of best frequency. Spectrotemporal receptive fields were summed along the temporal dimension to generate the spectral receptive field (SRF). The peak of the spectral receptive field is defined as best frequency (BF).

BFs are plotted as a function of cortical position in figure 2. The distribution of BFs across STG appears random in nature with no clear form of organization. Tonotopically organized areas by nature have a degree of neighborhood similarity in which cortical sites near each other have similar best frequencies. Non-significant neighborhood similarity is consistent with lack of tonotopic organization. To test for significant neighborhood similarity we compute the average absolute difference between the BF of a given site and its directly adjacent neighbors to give the site neighborhood similarity index. This value is then averaged over sites within a map to give the overall map neighborhood similarity index. This map neighborhood similarity index is compared to the distribution of map neighborhood similarity indices for random permutations of the map sites to determine statistical significance. On each permutation, the shape of the map is the same, but each BF value is randomly assigned to a new site. None of the maps exhibited a statistically significant degree of

neighborhood similarity consistent with a lack of tonotopic organization in human STG (p-values > 0.05, permutation test).

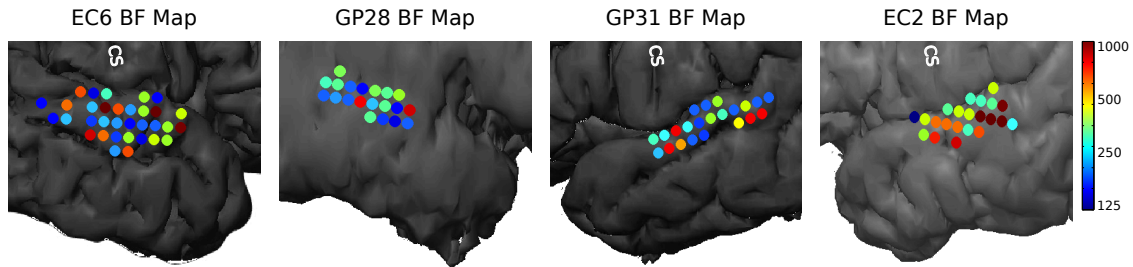


Figure 2. Best frequency distribution along human STG. (A) Best frequency is plotted as a function of position along human superior temporal gyrus in patients EC6, EC28, GP31, and EC2. No apparent tonotopic organization is present and tests for neighborhood similarity were nonsignificant (p-values > 0.05, permutation test).

Although STG appears to lack tonotopic organization based on raw BF values, a number of spectral receptive fields had multiple spectral peaks of approximately the same magnitude. Across patients, the distribution of BFs was heavily biased toward frequencies below 500 Hz. It is possible there is organization of spectral tuning in this low frequency range and the occasional high frequency BF value from multi-peaked SRFs could be obscuring identification of this organization. To account for the complexity of these multi-peaked spectral receptive fields and to use the lowest spectral frequency peak when more than one prominent peak of similar magnitude is present, we generated multi-peak corrected BF maps. To identify true multi-peaked SRFs we fit each SRF with a periodic spectral modulation function (Figure 3). Briefly, the spectral modulation function is the modulation frequency and corresponding phase identified by the peak of the spectral modulation spectrum of each SRF (see methods). The total SRF spans 6 octaves. Spectral modulation functions with

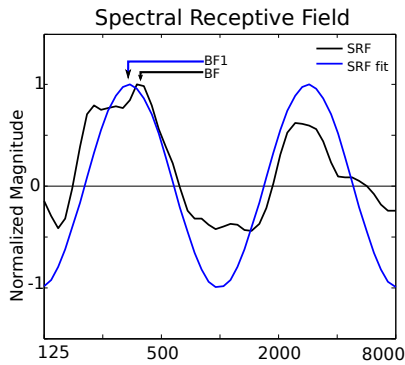


Figure 3. Identifying multi-peaked spectral receptive fields. Many spectral receptive fields were multi-peaked. We fit each spectral receptive field with its spectral modulation function (see methods). Spectral modulation functions with spectral modulation frequencies greater than 0.33 cycles/octave identify true multi-peaked spectral receptive fields, which make up 18% of sites. BF1 is defined as the lowest spectral frequency peak in a multi-peaked receptive field and is used as best frequency in the multi-peak corrected best frequency maps (Figure 4).

a peak spectral modulation frequency above 0.33 cyc/oct identify SRFs with at least two cycles over the 6-octave frequency range, and thus have two dominant peaks. For true multi-peaked SRFs identified in this way, the lowest spectral frequency peak of the fit, which we call BF1, is used for mapping. Of SRFs across patients, 18% were multi-peaked. Figure 4 shows multi-peak corrected best frequency maps. Again, there does not appear to be organization in spectral tuning across STG and none of the maps have a significant degree of neighborhood similarity (p-values > 0.05, permutation test). Despite only using the low frequency peak to characterize spectral tuning for multi-peaked SRFs there still appears to be a lack of tonotopic organization during speech processing in human superior temporal gyrus.

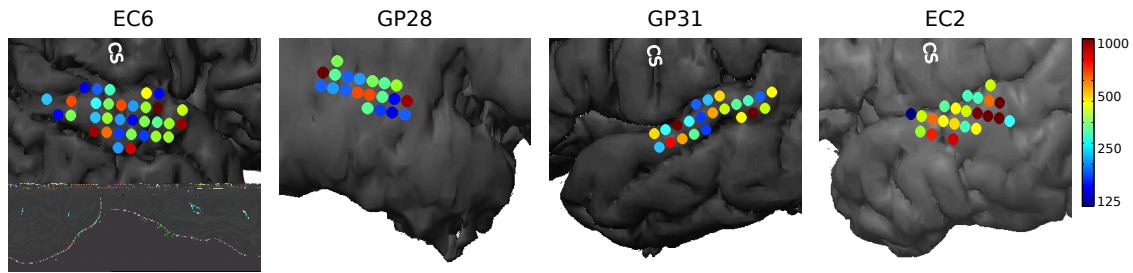


Figure 4. Multi-peak corrected best frequency maps. For true multi-peaked spectral receptive fields, BF1 was used as best frequency. For single peaked spectral receptive fields raw BFs are used. Values are plotted as a function of position along human superior temporal gyrus in patients EC6, GP28, GP31, and EC2. No apparent tonotopic organization is present and tests for neighborhood similarity were nonsignificant (pvalues > 0.05, permutation test).

In addition to characterizing the spatial organization of spectral processing, we compared the distribution of BF values to the spectral content present in speech (Figure 4). First, the BF distribution is not spread evenly along the range of spectral frequency content in speech. Rather BFs are highly concentrated in the low frequency region of the speech power spectrum. This is a rather narrow range of BF tuning relative to the spectral frequency content in speech. Second, the BF distribution is concentrated in a region of relatively low power relative to power at other frequencies in the speech power spectrum. If an organized spectral decomposition is the primary function of STG, one might expect a more even distribution of BFs across the range of spectral content present in speech.

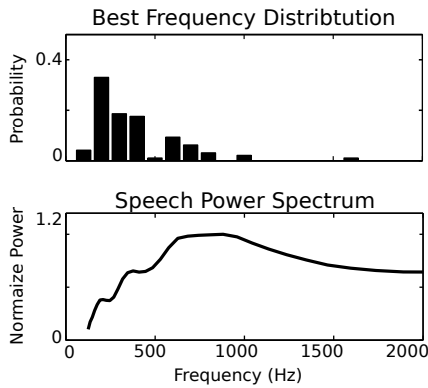


Figure 5. STG best frequency distribution and the spectral content in speech. The distribution for best frequency is shown. BFs are not distributed evenly along the range of spectral frequency content in speech, but rather, are concentrated in a narrow frequency range of the full speech power spectrum. (Note: although we only show the power spectrum extending up to 2 kHz, there is significant spectral energy up to 8 kHz). Furthermore, the area of concentration is centered on spectral frequencies with the lowest power in speech.

Lastly, we investigated latency organization of speech-driven responses in superior temporal gyrus. We were interested in determining if other organizing principles found in lower tonotopically areas are present in STG despite the apparent lack of tonotopic organization (Mendelson et al., 1997). In addition, speech based latency values characterized the timing with which speech information reaches STG and can be used to generate hypotheses about the flow of information and the structure of the auditory hierarchy. Latency is defined as the first time point after sentence onset in which the response is significantly different from baseline activity (p value $< 1 \times 10^{-6}$, two-sample Kolmogorov-Smirnov test). Latency maps showed nonrandom organization with the lowest latencies concentrated on the superior temporal gyrus posterior to the central sulcus and with increasing latencies extending anteriorly. In the case of patient GP31, latencies increase posteriorly as well (Figure 7A; functional organization p

values < 0.01 for patients EC6 and GP31; p value < 0.05 for patient EC2; permutation test). Although we were interested in the latency of responses to natural speech stimuli, it is possible organization of responses latencies obtained with more traditional stimuli, such as broadband clicks, may be different. We investigated this possibility by computing latency values in patient EC2 using TORC stimuli, which have a more stereotyped broadband onset (similar to broadband clicks). (Klein et al., 2000). Patient EC2 was the only subject in which responses to TORC stimuli were recorded. A similar overall pattern of organization was found, as demonstrated by the significant level of correlation between latency values across electrodes (Figure 7B, corr. coef. = 0.60, p value < 0.05). Overall, latency analysis shows a systematic order in the timing of speech information reaching the superior temporal gyrus in which central regions receive information first followed by more anterior and posterior regions.

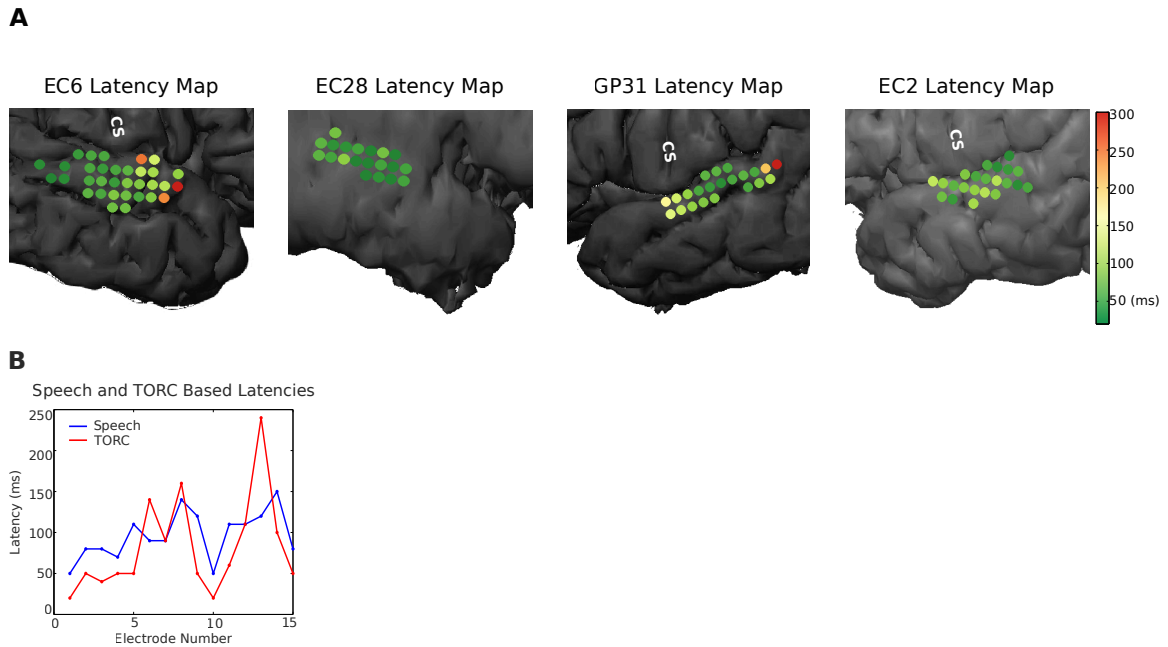


Figure 6. Organization of response latency. **(A)** Latency maps for all patients. In all patients the lowest latencies are concentrated on the superior temporal gyrus posterior to the central sulcus with increasing latencies more anterior, and in the case of patient GP31, more posterior as well. (Patients EC6, GP31: $p < 0.005$; EC2: $p < 0.05$; permutation test). **(B)** Speech versus TORC latencies in patient EC2. The absolute latency values were significantly correlated between the two stimuli (corr. coef. = 0.60, $p < 0.05$). CS = central sulcus.

Discussion

In this study we used responses to natural speech stimuli to characterize the organization of spectral processing in human STG. In contrast to previous studies, we found no clear evidence of a tonotopic organization in this region. Furthermore, best frequencies are not distributed across the full range of spectral content in speech, but are concentrated in a narrow range of frequencies below 500 Hz. Together this data is consistent with a lack of tonotopic organization in human superior temporal gyrus when processing natural speech. Because

tonotopically organized areas are generally assumed to carry out an organized spectral decomposition of the stimulus, the lack of tonotopic organization suggest a somewhat different stimulus transformation may be taking place in human superior temporal gyrus.

Although our results are consistent with a lack of tonotopic organization in human STG, our methodology for investigating tonotopic organization is somewhat different than traditional approaches. Almost all demonstrations of tonotopy in auditory cortex have been generated with multi-unit (MU) recordings in the middle layers of cortex. Our results are based on the high-gamma component of the local field potential (LFP) using surface electrodes that are likely biased for activity from superficial layers. If STG is tonotopically organized, could these differences in methodology lead to what appears to be lack of tonotopic organization? While this is possible, studies in mice do not support this notion. Recently, Guo et. al. showed that tonotopic organization characterized using multi-unit recordings in middle layers of mouse cortex, was also evident using the high-gamma component of LFP recording in superficial layers of the same region of cortex (Guo et al., 2012). In addition, tonotopy has been demonstrated in rat primary auditory cortex using the high-gamma responses from ECoG arrays (Bouchard K., personal communication). In light of this, our results are consistent with a lack of tonotopically organized spectral processing in human STG during natural speech perception. Furthermore, work using the high-gamma component of ECoG recordings with pure tone stimuli rather than speech stimuli also showed a lack of tonotopic organization in

human STG (Nourski et al., 2012). Although, a subset of patients did show some nonrandom distribution of spectral tuning consistent with a weak organization of spectral tuning.

These results are at odds with recent work showing the presence of tonotopic organization in STG using functional magnetic resonance imaging (fMRI) (Humphries, Liebenthal, & Binder, 2010; Moerel, De Martino, & Formisano, 2012; Striem-Amit, Hertz, & Amedi, 2011; Talavage et al., 2004). One possible reason for the difference is the use of population analysis. fMRI data was transformed into a standardized (MNI or Talairach) space so data could be combined across subjects. It is possible that a weak tonotopic organization, which is too noisy to identify in a single patients, is present once data is averaged across patients. Collapsing our data across patients could potentially show nonrandom organization of spectral processing. However, functional organization that is not evident within a single patient suggests alternative forms of functional organization may be present that could potentially be more informative descriptors of the processing taking place in human STG (see Chapter 1).

The distribution of latency values along STG was highly organized with the lowest latencies posterior to central sulcus and longer latencies more anterior, and in the case of GP31, at the most posterior aspect of STG as well. Although, natural speech was used to compute latency values rather than more traditional

stimuli like clicks or noise bursts, the goal of the analysis was to examine latency organization for natural speech stimuli and thus reveal the timing of speech information reaching the superior temporal gyrus during the perception of natural speech. Our data reflect a pattern in the timing of speech information reaching STG in which the central region receives information first and more anterior and posterior regions receive information later. This data is consistent with previous work showing an anterior and posterior spread of speech induced activity from central STG (Dehaene-Lambertz et al., 2006; Brauer et al., 2008; Steinschneider et al., 2011). It is unclear if this reflects a hierarchical propagation of speech information which reaches central STG first and then spreads more anteriorly and posteriorly, or if information reaches these areas in a parallel fashion with different latencies. Although STG lacks significant tonotopic organization in the context speech processing, it displays organization of response latencies- an organizing principle present in other tonotopically organized cortical areas. In primary auditory cortex response latencies are lowest in central, narrowly tuned, regions and increases dorsally and ventrally into broadly tuned regions (Schreiner and Mendelson, 1990; Mendelson et al., 1997). Further analysis will be needed to determine if latencies are correlated with spectral tuning bandwidth, although the complexity of multi-peaked SRFs in STG may require newer metrics for bandwidth quantification.

Lack of tonotopic organization in human STG – likely a parabelt region of the auditory cortex - is consistent with a trend of decreasing tonotopic organization at ascending levels of the auditory hierarchy (Reale and Imig, 1980;

Bizley et al., 2005b; Petkov et al., 2006). By contrast, strong modulotopic organization seems to be present in human superior temporal gyrus (see Chapter 1). It is intriguing to speculate on a possible tonotopic to modulotopic transition in functional organization along the auditory hierarchy. Further characterization of modulotopic organization would be needed to support this model, however, the result of such a model would be a succinct explanation for what seem to be very disparate forms of processing in lower tonotopically organized areas and higher modulotopically organized areas. The system is performing a Fourier-like decomposition in an iterative manner as stimuli propagate up the auditory hierarchy.

METHODS

Patients and Neural Recordings

Subdural ECoG arrays (interelectrode distance: 4mm) were placed over the temporal and parietal lobes unilaterally in 4 patients (2 right hemisphere, 2 left hemisphere) undergoing a neurosurgical procedure for the treatment of epilepsy. All patients were native English speakers with no hearing or communication deficits. All experimental protocols were approved by the University of California, San Francisco Institutional Review Boards and Committees on Human Research. Array placement was determined by clinical criteria. After array placement, patients were allowed to recover and the monitoring phase of the procedure started, during which time patients were asked to passively listen to 15 – 25 minutes of natural speech while cortical

activity was recorded. Signals were amplified and sampled at 3052 Hz. After rejection of electrodes with excessive noise or artifacts, signals were referenced to a common average and the high gamma band (75 - 150 Hz) was extracted using the Hilbert Transform (Crone et al., 2001). The resulting signal for each electrode was z-scored based on the mean and standard deviation of spontaneous activity.

Stimuli

Speech stimuli were delivered binaurally through calibrated ear insets at 70 dB average sound pressure level. The frequency power spectrum of stimuli spanned 0 - 8000 Hz. The stimulus set consisted of prerecorded (2 – 4 second) sentences drawn randomly without replacement from the phonetically transcribed TIMIT speech corpus with one-second silent intervals between each sentence presentation (Garofolo et al., 1993). Each patient was presented 484-499 sentences with the exception of patient GP36 who was presented 246 sentences. The speech corpus was balanced for male and female speakers, a different speaker produced each sentence, and the lexical content of each sentence was unique. Spectrogram representations of speech stimuli were based on a cochlear model of auditory processing (Yang et al., 1992).

Analysis

STRF

STRFs were computed using the high-gamma band of ECoG recordings

which correlates with tuning properties based on multi-unit spiking in auditory cortex (Crone et al., 2001; Steinschneider et al., 2008; Guo et al., 2012). STRFs were computed with two different methods designed specifically for use with natural signals- maximally informative dimension analysis (MID) and normalized reverse correlation (Theunissen et al., 2001; Sharpee et al., 2004). To compute STRFs using MID analysis a gradient ascent procedure was used to search for the receptive field that maximizes the KL divergence between the distribution of STRF-Stimulus projection values and the distribution of STRF-stimulus projection values in which contribution of each projection value to the distribution is weighted by the magnitude of the response. STRF estimates based on normalized reverse correlation were computed using open source code available at: <http://strfpak.berkeley.edu/>. STRFs were computed with both methods on the same estimation set (90% of the total data) and cross-validated on the same test set which was withheld from the estimation process (10% of the data). MID based STRFs showed better prediction performance and were therefore used to analyze functional organization in STG. To investigate functional organization we restricted our analysis to STRFs which predicted at least five percent of the total variance in the response to prevent clouding our view of functional organization with poor STRF models.

Functional Organization

To characterize tonotopic organization, each STRF was summed across its temporal axis to obtain the spectral receptive field (SRF). Best frequency is

defined as the peak of the receptive field, which was mapped onto the corresponding cortical location to investigate tonotopic organization. To obtain BF1 frequency we computed the Fourier transform of the each SRF. All values except the peak spectral modulation were set to zero and the inverse Fourier transform was computed to generate a fit of the SRF based on the peak of its spectral modulation function. The lowest spectral frequency peak of this fit is defined as BF1 frequency. Each SRF spans 6 octave, therefore, fits with modulation frequency above 0.33 cycles/octave define spectral receptive fields with two or more dominant peaks.

To investigate latency of speech responses the distribution of response values at sentence onset (time zero) for each electrode were collected and compared to the distribution of response values at each time point after sentence onset. The first time point significantly different from baseline activity was defined as the latency (Kolmogorov-Smirnov test, p value $< 1 \times 10^{-6}$). This definition of latency identifies the first time point in which the response is significantly different from baseline activity and is agnostic to the sign of the response as suppressive responses will be we treated similar to excitatory responses.

Permutation tests were used to assess functional organization of spectral tuning. Each test determined if the observed structure would be expected if the true underlying spatial organization of the parameters was random. For each site

a neighborhood similarity metric is computed which quantifies the similarity of a site and its nearest neighbors. The values are averaged across sites to compute the mean neighborhood similarity index for the map. The map is then randomly permuted 10,000 times and the map neighborhood similarity index is recomputed on each permutation to generate a distribution of random neighborhood similarity indices. The true neighborhood similarity index is compared to the random neighborhood similarity index distribution to assess the level of significance. For the STRF map, the site neighborhood similarity metric was computed as the average of the correlation coefficients between an STRF and its nearest neighbors no more than one site away. For F1 frequency and latency maps, the site neighborhood similarity index was quantified as the absolute value of the difference between a site and its neighbors rather than the correlation coefficient.

References

Aitkin, L., and Webster, W. (1972). Medial geniculate body of the cat : organization and responses to tonal stimuli of neurons in ventral division. *Journal of Neurophysiology* 35, 365–380.

Bizley, J.K., Nodal, F.R., Nelken, I., and King, A.J. (2005a). Functional organization of ferret auditory cortex. *Cerebral Cortex* (New York, N.Y. : 1991) 15, 1637–1653.

Bizley, J.K., Nodal, F.R., Nelken, I., and King, A.J. (2005b). Functional organization of ferret auditory cortex. *Cerebral Cortex* (New York, N.Y. : 1991) *15*, 1637–1653.

Brauer, J., Neumann, J., and Friederici, A.D. (2008). Temporal dynamics of perisylvian activation during language processing in children and adults. *NeuroImage* *41*, 1484–1492.

Chichilnisky, E.J. (2001). A simple white noise analysis of neuronal light responses. *Network* (Bristol, England) *12*, 199–213.

Christianson, G.B., Sahani, M., and Linden, J.F. (2008). The consequences of response nonlinearities for interpretation of spectrotemporal receptive fields. *The Journal of Neuroscience : the Official Journal of the Society for Neuroscience* *28*, 446–455.

Crone, N.E., Boatman, D., Gordon, B., and Hao, L. (2001). Induced electrocorticographic gamma activity during auditory perception. *Clinical Neurophysiology* *112*, 565–582.

Dale, a M., and Halgren, E. (2001). Spatiotemporal mapping of brain activity by integration of multiple imaging modalities. *Current Opinion in Neurobiology* *11*, 202–208.

Dehaene-Lambertz, G., Hertz-Pannier, L., Dubois, J., Mériaux, S., Roche, A., Sigman, M., and Dehaene, S. (2006). Functional organization of perisylvian activation during presentation of sentences in preverbal infants. *Proceedings of the National Academy of Sciences of the United States of America* 103, 14240–14245.

Formisano, E., Kim, D.S., Di Salle, F., van de Moortele, P.F., Ugurbil, K., and Goebel, R. (2003). Mirror-symmetric tonotopic maps in human primary auditory cortex. *Neuron* 40, 859–869.

Garofolo, J.S., Lamel, L., Fisher, W., JG, F., and Pallet, D. (1993). Acoustic-phonetic continuous speech corpus. (Linguistic Data Consortium, Philadelphia).

George, J., Aine, C., and Mosher, J. (1995). Mapping function in the human brain with magnetoencephalography, anatomical magnetic resonance imaging, and functional magnetic resonance imaging. *Journal of Clinical Neurophysiology* 12, 406–431.

Guo, W., Chambers, A.R., Darrow, K.N., Hancock, K.E., Shinn-Cunningham, B.G., and Polley, D.B. (2012). Robustness of cortical topography across fields, laminae, anesthetic states, and neurophysiological signal types. *The Journal of Neuroscience : the Official Journal of the Society for Neuroscience* 32, 9159–9172.

Humphries, C., Liebenthal, E., and Binder, J.R. (2010). Tonotopic organization of human auditory cortex. *NeuroImage* 50, 1202–1211.

Kim, S., Richter, W., and Ugurbil, K. (1997). Limitations of temporal resolution in functional MRI. *Magnetic Resonance in Medicine* 37, 631–638.

Klein, D.J., Depireux, D. a, Simon, J.Z., and Shamma, S. a (2000). Robust spectrotemporal reverse correlation for the auditory system: optimizing stimulus design. *Journal of Computational Neuroscience* 9, 85–111.

Mendelson, J.R., Schreiner, C.E., and Sutter, M.L. (1997). Functional topography of cat primary auditory cortex: response latencies. *Journal of Comparative Physiology. A, Sensory, Neural, and Behavioral Physiology* 181, 615–633.

Merzenich, M., and Reid, M. (1974). representation of the cochlea within the inferior colliculus of the. *Brain Research* 77, 397–415.

Merzenich, M., and Roth, G.L. (1975). Representation of cochlea within primary auditory cortex in the cat. *Journal of Neurophysiology* 38, 231–249.

Moerel, M., De Martino, F., and Formisano, E. (2012). Processing of Natural Sounds in Human Auditory Cortex: Tonotopy, Spectral Tuning, and Relation to Voice Sensitivity. *Journal of Neuroscience* 32, 14205–14216.

Nourski, K.V., Steinschneider, M., Oya, H., Kawasaki, H., Jones, R.D., and Howard, M. a (2012). Spectral Organization of the Human Lateral Superior Temporal Gyrus Revealed by Intracranial Recordings. *Cerebral Cortex*.

Petkov, C.I., Kayser, C., Augath, M., and Logothetis, N.K. (2006). Functional imaging reveals numerous fields in the monkey auditory cortex. *PLoS Biology* 4, e215.

Reale, R. a, and Imig, T.J. (1980). Tonotopic organization in auditory cortex of the cat. *The Journal of Comparative Neurology* 192, 265–291.

Schreiner, C., and Mendelson, J. (1990). Functional topography of cat primary auditory cortex : distribution of integrated excitation. *Journal of Neurophysiology* 64, 1442–1459.

Sharpee, T., Rust, N.C., and Bialek, W. (2004). Analyzing neural responses to natural signals: maximally informative dimensions. *Neural Computation* 16, 223–250.

Steinschneider, M., Fishman, Y.I., and Arezzo, J.C. (2008). Spectrotemporal analysis of evoked and induced electroencephalographic responses in primary auditory cortex (A1) of the awake monkey. *Cerebral Cortex* (New York, N.Y. : 1991) 18, 610–625.

Steinschneider, M., Nourski, K.V., Kawasaki, H., Oya, H., Brugge, J.F., and Howard, M. a (2011). Intracranial study of speech-elicited activity on the human posterolateral superior temporal gyrus. *Cerebral Cortex* (New York, N.Y. : 1991) *21*, 2332–2347.

Striem-Amit, E., Hertz, U., and Amedi, A. (2011). Extensive cochleotopic mapping of human auditory cortical fields obtained with phase-encoding fMRI. *PloS One* *6*, e17832.

Talavage, T.M., Sereno, M.I., Melcher, J.R., Ledden, P.J., Rosen, B.R., and Dale, A.M. (2004a). Tonotopic organization in human auditory cortex revealed by progressions of frequency sensitivity. *Journal of Neurophysiology* *91*, 1282–1296.

Talavage, T.M., Sereno, M.I., Melcher, J.R., Ledden, P.J., Rosen, B.R., and Dale, A.M. (2004b). Tonotopic organization in human auditory cortex revealed by progressions of frequency sensitivity. *Journal of Neurophysiology* *91*, 1282–1296.

Theunissen, F.E., David, S.V., Singh, N.C., Hsu, a., Vinje, W.E., and Gallant, J.L. (2001). Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *Network: Computation in Neural Systems* *12*, 289–316.

Theunissen, F.E., Sen, K., and Doupe, a J. (2000). Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. *The Journal*

of Neuroscience : the Official Journal of the Society for Neuroscience 20, 2315–2331.

Woods, D.L., Stecker, G.C., Rinne, T., Herron, T.J., Cate, A.D., Yund, E.W., Liao, I., and Kang, X. (2009). Functional maps of human auditory cortex: effects of acoustic features and attention. *PloS One* 4, e5183.

Yang, X., Wang, K., and Shamma, S. a. (1992). Auditory representations of acoustic signals. *IEEE Transactions on Information Theory* 38, 824–839.

Chapter3

A Systematic Approach to Investigating Neural Solutions to the Cocktail Party Problem

Abstract

The brain's ability to solve the cocktail party problem is not well understood and current engineering solutions are suboptimal. Knowledge of the brain's solution could be used to generate noise-robust automatic speech recognition algorithms and hearing aids. Here we demonstrate a systematic approach to identify those neural solutions. We create animals that are specialized at solving the cocktail party problem by rearing them in noise. We describe a method to identify neurons that can encode vocalization information in the face of noise and find that neurons in specialized animals encode significantly higher rates of vocalization information in the presence of noise. Further analysis of their receptive field properties will give insight on the underlying mechanism of this increased ability.

Introduction

The challenge to extract and process signals of interest embedded in a mixture of sound is commonly referred to as the cocktail party problem. (Cherry, 1953; McDermott, 2009). Understanding how the brain perceives signals of interest in the face of noise is an important problem with many real world applications including better automatic speech recognition software, more advanced hearing aids, and more broadly, any application which requires the extraction of a signal from noise. To understand how the brain solves a computational task of interest, it is useful to work with organisms specialized for solving that task because differences in neural processing between specialized

and non-specialized animals can give insight on the neural mechanisms that underlie that specialization. To investigate mechanisms by which the auditory system recognizes behaviorally relevant signals in the face of noise, we created rats that are specialized at solving cocktail party problem and compare them to non-specialized controls.

To create specialized animals we reared rats in naturalistic noise throughout the course of development and into adulthood. This serves as a natural training paradigm in which rats are forced to continuously perceive signals of interest in the face of noise during a time when the brain is most plastic and can robustly adapt to the acoustic environment (Insanally, Köver, Kim, & Bao, 2009; Nakahara, Zhang, & Merzenich, 2004; Zhang, Bao, & Merzenich, 2001a; Xiaoming Zhou & Merzenich, 2008a; de Villers-Sidani, Chang, Bao, & Merzenich, 2007; de Villers-Sidani, Simpson, Lu, Lin, & Merzenich, 2008). We determined if noise-rearing facilitates the development of specialized abilities to solve the cocktail party problem, by testing animals using a vocalization recognition task and show noise-reared animals are significantly better at recognizing vocalizations in the face of noise. To identify neural correlates of this increased ability, we describe a method that allows for identification of neurons that can extract vocalization information in the presence of noise. Using this method we identify populations of these neurons in noise-reared and naïve animals. Although the prevalence of these neurons is the same in both groups, neurons in noise-reared animals extract vocalization information in the face of noise at significantly higher rates. Further receptive field analysis of these

neurons will give insight to the underlying mechanism of this ability. This process demonstrates a systematic approach that can be iteratively performed using different types of noise to refine models of processing and unravel the process by which neural systems solve the cocktail party problem.

Results

Synthesis of Naturalistic Noise

Although some definitions of the cocktail party problem specify noise as competing vocalizations, we use a broader definition which defines noise as any competing environmental sound. To identify noise sources in a lab rat's natural environment, we recorded the acoustic scene of 6 female rat siblings starting post-natal day 30 and ending post-natal day 40. Visual inspection of these recordings identified vocalizations in the 20-30 kHz range and less prevalent vocalizations in the 50-80 kHz range. A typical rat vocalization consisted of a bout of narrowband whistles in which each whistle could be modulated up or down in frequency to generate up-sweeps, down-sweeps, U-shaped whistles, or even more complex vocalizations. This diversity of spectrotemporal structure in individual whistles and the frequency ranges in which they were identified are consistent with previous characterizations of rat vocalization repertoire (S M Brudzynski, Bihari, Ociepa, & Fu, 1993; S M Brudzynski, Kehoe, & Callahan, 1999; Stefan M Brudzynski, 2005; Kaltwasser, 1990; Portfors, 2007). In addition to vocalizations, the other dominant sound found in the recordings was associated with quick movement in the sawdust bedding during play or any movement

related behavior and contained broadband frequency content coherently modulated in time (Figure 1A). We estimated the statistics of this sound and generated a synthetic version that was characteristically similar by modulating the sound pressure level of broadband white noise to give a similar amplitude distribution and temporal modulation spectrum, which we refer to as “naturalistic noise” (Figure 1B).

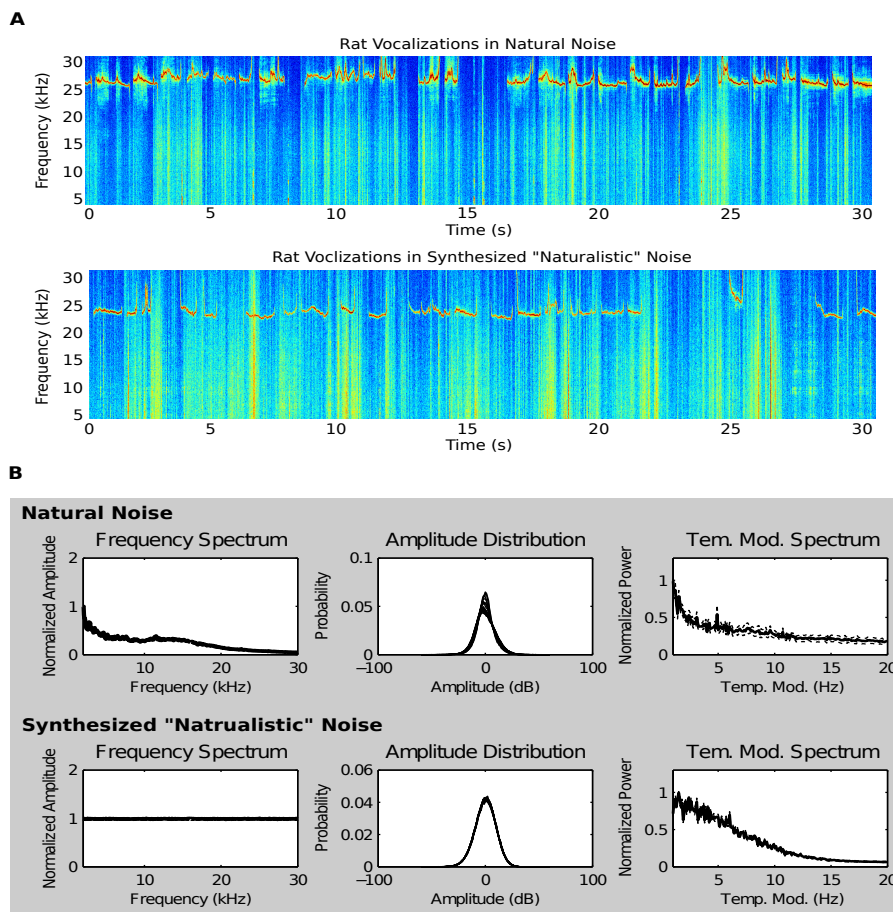


Figure 1. Natural and synthesized “naturalistic” noise. (A) Rat vocalizations in natural noise (top) and rat vocalization in the synthetic “naturalistic” noise. (B) Statistics between the movement related noise and the synthetic version. Compared to the natural version, the synthetic noise has a flatter frequency spectrum to increase noise power in the 20-30 kHz vocalization range. The frequency spectrum of synthetic noise also extends to higher frequencies to prevent animals from shifting their vocalization frequency range to compensate for background noise.

Noise-rearing

To create animals, which are specialists at solving the cocktail party problem, we exposed rats to naturalistic noise throughout the course of their life, starting at p6, prior to the opening of the ear canal. During the exposure rats were housed in acoustic attenuation chambers to restrict their auditory environment to self-generated sound and naturalistic noise. The average sound pressure level the natural movement related noise was 51.1 dB. To avoid pathologic sound pressure levels (SPLs), but still create a challenging listening conditions, we exposed animals to naturalistic noise with an average SPL of 60 dB (Cappaert, Klis, & Muijser, 2000; Ward, Cushing, & Burns, 2000). Although the mean SPL was not significantly louder than natural conditions, noise was played continuously without interruption for the life of the animal. Noise rearing lasted for a minimum of 39 days but extended up to 99 days for a subgroup of animals that went through two rounds of behavioral testing. Animals were checked daily to ensure no interruptions in the noise exposure and no noticeable differences were observed in the overall behavior or temperament of the exposed animals compared to naïve animals raised in normal laboratory conditions.

Behavioral Testing

To determine if noise-rearing increases an animal's ability to solve the cocktail party problem, we trained noise-reared ($n = 3$) and naïve animals ($n = 4$) on a vocalization recognition task. After reaching peak performance under quiet

conditions, naturalistic noise was added to the task during a short testing phase and the performance of each group's ability to recognize vocalizations in the face of noise was quantified. Outside of training and testing sessions, noise-reared animals were placed back into the noise exposure. The vocalization recognition task consisted of a go/no-go paradigm in which animals were trained to respond to a target vocalization (breaking a sensor beam in a circular port with their nose) for a food reward and withhold a response if a non-target vocalization was played. The vocalization stimuli consisted of one target vocalization (vocalization-1) and three non-target vocalizations plus (Figure 2).

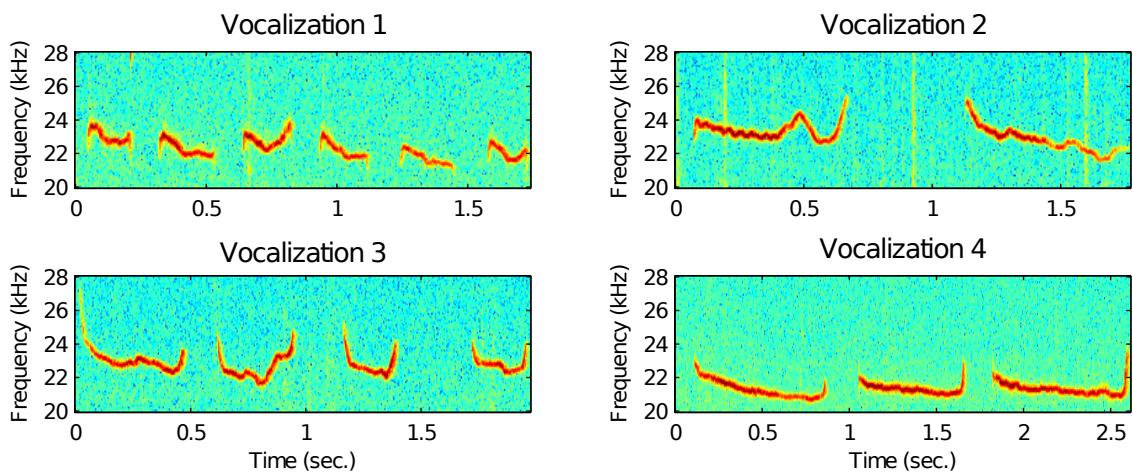


Figure 2. Vocalization stimuli used for behavioral training. All vocalizations are within the 20-30 kHz range and have similar duration. Although there are many differences between the vocalizations, the most apparent are the number of whistles and duration of individual whistles within a single vocalization bout. Vocalization 1 served as the target for the first experiment.

The target vocalization remained the same throughout the training and testing phase. Vocalizations for the task were taken from rat acoustic recordings. The

vocalizations used had frequency content in the 20-30 kHz range and were similar in total duration, but differed in number of whistles and the temporal

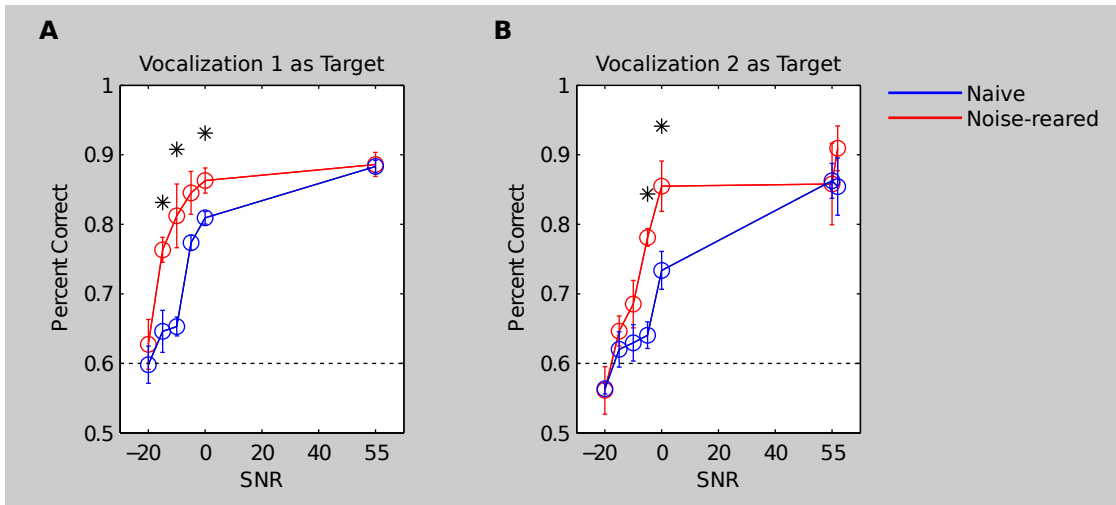


Figure 3. Vocalization recognition performance in naturalistic noise. Both groups were trained to recognize vocalizations in the absence of noise and then tested on their ability to recognize vocalizations in naturalistic noise during a short two-day testing phase. (A) Vocalization recognition performance using vocalization 1 as the target. Both groups have similar performance in zero noise conditions (+ 55 dB SNR), but noise-reared animals significantly outperform naïve animals at SNRs of 0 dB, -10 dB, and -15 dB (pvalue < 0.05, Wilcoxon rank sum test; -5 dB SNR pvalue = 0.08, Wilcoxon rank sum test). (B) Vocalization recognition performance using vocalization 2 as the target. Both groups have the same performance level in zero noise conditions (+ 55 dB SNR), but noise-reared animals significantly outperform naïve animals at SNRs of 0 dB, and -5 dB (pvalue < 0.05, Wilcoxon rank sum test). In B, an additional zero-noise block at the end of the low SNR-blocks was added to determine if performance in zero-noise conditions remained the same before and after the SNR blocks.

duration of individual whistles within the vocalization. Rats began training at p45 and trained on the task twice daily (1.5 hours each session) for 28 to 39

consecutive days until each rat reached a plateau in performance. After reaching peak performance, each rat was tested for the ability to perform vocalization recognition in the face of naturalistic noise during a two day testing phase. Each testing day consisted of 30 warm-up trials (without background noise) followed by 6 testing blocks of 45 trials per block. Each block had a different signal to noise ratio ranging from +55 dB (zero noise) to -20 dB SNR. Figure 3A shows vocalization recognition performance at various SNRs for rats trained using vocalization 1 as the target. Both groups of rats had similar performance levels in zero noise conditions (+55 dB SNR). As naturalistic noise was added the performance in both groups decreased. However, noise-reared animals exhibited a less dramatic decrease and had significantly higher performance scores at three of four signal-to-noise ratios in which performance was above chance. To ensure this trend was not specific to the particular target vocalization chosen, we switched the target to vocalization 2 and re-trained the same animals (vocalization-1 became a non-target). Again, the noise-reared animals significantly outperformed the naïve animals in conditions with naturalistic noise added to the background (Figure 3B). This data suggests that noise rearing facilitates the development of animals that are able to better recognize vocalizations in the face of noise compared to naïve animals.

Although increased performance levels demonstrate noise-reared animals are better at recognizing vocalization in the face of noise, this effect in behavioral performance could be due to non-auditory specific factors. For example, the addition of noise could distract or startle the naïve animals, which have not been

reared in the noise environment. This could cause a decrease in performance relative to noise-reared animals due to decreased attention to the behavioral task. To rule out non-auditory specific differences that could account for increased performance by noise-reared animals we trained a new group of animals on a visual stimulus recognition task and tested their performance on this task in the

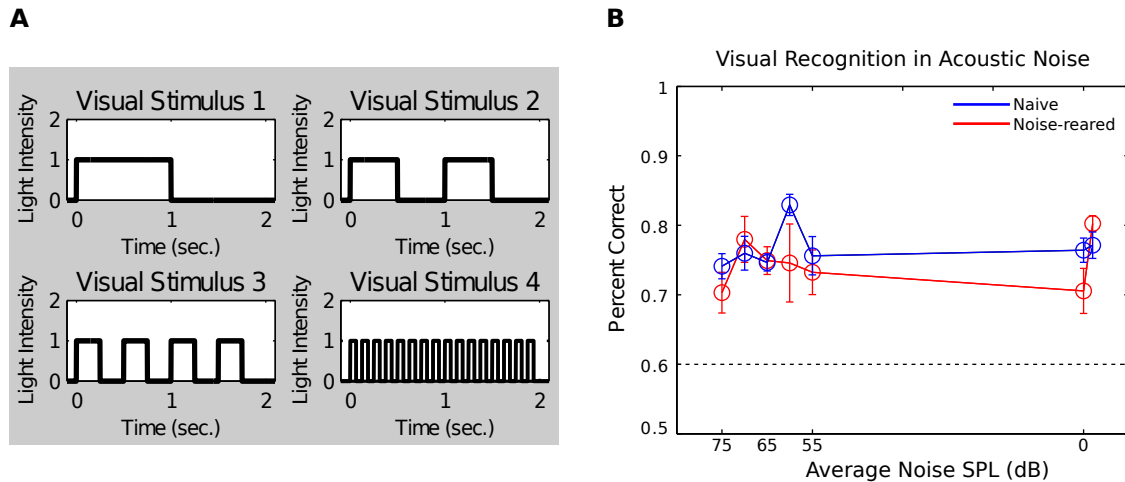


Figure 4. Visual stimulus recognition performance in naturalistic noise. (A) Visual stimuli used for the visual recognition. (B) Visual recognition performance in the presence of naturalistic acoustic noise used for rearing. Noise conditions have the same noise SPLs as the low SNR blocks of the previous vocalization recognition task. Noise sound pressure level is plotted in reverse order for easy comparison with figure 3. There is no significant difference between groups at any noise SPL, and no decrement in performance as noise SPL is increased. This indicates the difference in vocalization recognition performance (Figure 3) is not due to non-auditory specific differences between groups such decreased attention to the task by naïve animals who are not familiar with the noise.

presence of the acoustic noise (noise-reared, $n = 3$; naïve, $n = 3$). If the difference leading to increased performance in noise-reared rats is not auditory specific (i.e. they are habituated to the noise and therefore less distracted by it) the noise reared animals should again outperform naïve animals when naturalistic noise is added. The visual recognition task is identical to the vocalization recognition task, with the exception of visual stimuli rather than vocalization stimuli. Visual

stimuli were generated by pulsing the light within the nose-poke response port at different rates to generate four 2-second visual stimuli (Figure 4A). After reaching peak performance, the ability to recognize visual stimuli in acoustic noise was measured at the same noise sound pressure levels used to create low SNR conditions in the previous experiment. Figure 4b shows that there is no difference between the groups at any noise SPL level indicating that the results from the vocalization recognition task cannot be explained by nonspecific effects such as habituation to the noise environment. This suggest increased performance by noise-reared animals at solving the cocktail party problem is due to an auditory specific adaptation that allows for better recognition of behaviorally relevant signals in the face of noise. Differences in auditory processing between these specialized animals and naïve animals can be used to identify mechanisms by which neural systems recognizes signals of interest in the face of noise.

Identification of Cocktail Party Neurons

To identify neural correlates of noise-reared animal's increased performance on the cocktail party task, we recorded neural responses in primary auditory cortex of anesthetized animals to various vocalization stimuli, and vocalization stimuli embedded in noise. To specifically look at the effects of noise rearing, we recorded single unit activity from untrained noise-reared animals and age matched animals raised in normal laboratory conditions (noise-reared: n = 4, 1055 units; naïve: n = 4, 756 units). Recordings were localized to the 20-30

kHz region of A1 based on a tonotopic map that was determined for each animal. To identify neural correlates of the increased behavioral performance we played two sets of stimuli.

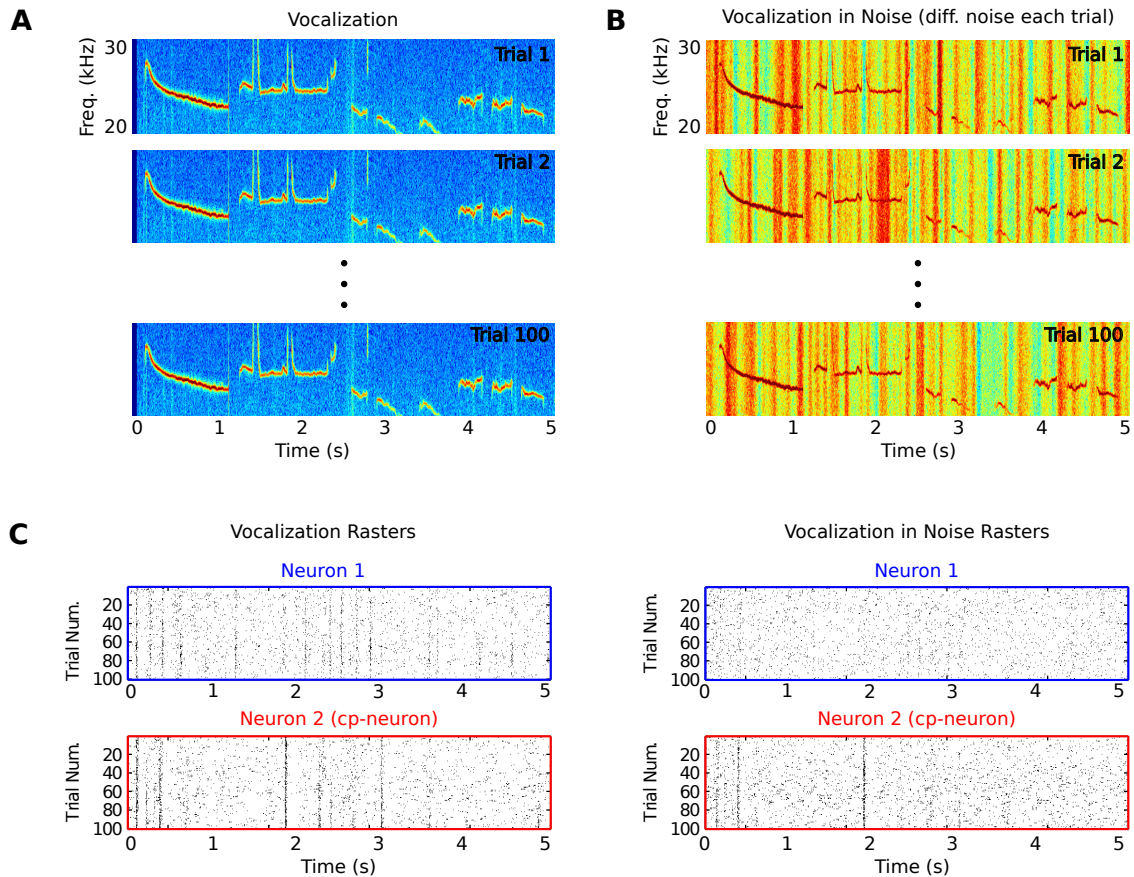


Figure 5. Identification of cocktail party neurons (cp-neurons). (A) One hundred trials of a 5 second rat vocalization bout used to identify neurons sensitive to vocalizations in the 20-30 kHz range. (B) One hundred trials of the vocalization bout with a different instance of naturalistic noise on each trial (0 SNR). (C) Example response rasters from a vocalization sensitive neuron (Neuron 1, blue) and a cocktail party neuron (cp-neuron, Neuron 2, red). The vocalization sensitive neuron (blue) has structure in the vocalization response raster, corresponding to significant information, but no structure when noise is added. The cp-neuron (red) has structure in the vocalization response raster and retains structure despite the addition a unique instance of naturalistic noise on each trial.

The first stimulus set consisted of a 5 second bout of vocalizations in the 20-30 kHz range repeated 100 times and was used to identify neurons that were responsive to 20-30 kHz vocalizations (Figure 5A). The second stimulus set was meant to identify neurons that can extract vocalization information in the face of noise and consisted of the same bout of vocalizations, but with naturalistic noise in the background at 0-dB SNR (Figure 5B). A 0-dB SNR has resulted in significant performance differences between naive and noise-rear animals in the behavioral task (see Fig. 3). A key feature of the second stimulus set is that the specific instance of background noise on each trial is different, thus, any consistent response across trials, reflects encoded information about the vocalization and not the noise. This makes it possible to identify neurons that encode vocalization information in the face of noise by identifying response rasters that have consistent responses across trials. To identify these rasters, we compute information values in each raster to determine if each raster contains a significant amount of stimulus information. Neurons with significant information values for the vocalization-in-noise stimulus, encode a significant amount of vocalization specific information in the face of noise and are classified as cocktail party neurons (cp-neurons). We use this method to identify cp-neurons in both groups of animals, and look for differences that could underlie the increased behavioral performance in the noise-reared group.

With this method to identify cp-neurons we asked two questions which could underlie noise-reared animals increased ability to solve the cocktail party problem. Do noise reared animals have more cp-neurons? And, are cp-neurons in noise-reared animals better at encoding vocalization information in the face of noise? To address the first question we classified neurons based on which stimuli they encoded significant information for by computing information values from the vocalization and vocalization-in-noise response rasters (Figure 6A). Based on the two stimuli, neurons could encode significant information for both stimuli,

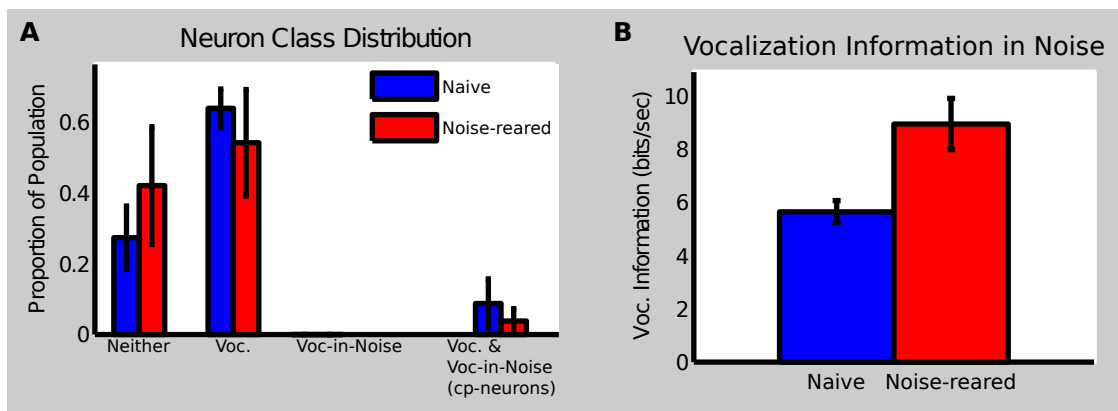


Figure 6. Cp-neuron prevalence and information transmission rates. (A) Distribution of neurons types. Neurons were classified according to the presence of significant information in their vocalization and vocalization in noise response rasters. Four neuron classes are possible based on their responsiveness to two stimuli. cp-neurons have significant vocalization information for vocalization and vocalization in noise stimuli and make up ~ 5 -10% of the recorded neural population. There were no significant differences in the proportion of cp-neurons or any other neuron class between naïve and noise-reared animals. (B) Cp-neuron vocalization specific information. Because the noise on each trial of the vocalization in noise stimulus is different, information values reflect the magnitude of vocalization specific information present in the response. Cp-neurons in noise reared-animals represent significantly higher amounts of vocalization specific information in the presence of noise compared to naïve animals (9.0 ± 1.0 bits/sec vs 5.7 ± 0.4 bits/sec, mean \pm s.e.m, p value < 0.005 Wilcoxon rank sum test).

the vocalization-in-noise stimulus only, the vocalization stimulus only, or neither stimulus. Neurons that encode significant information for both stimuli or the

vocalization-in-noise stimulus only would be classified as cp-neurons. However, we observed no “stochastic resonant” neurons that encoded significant information for the vocalization-in-noise only. Approximately 5-10% of recorded neurons encoded significant information for both vocalization and vocalization-in-noise stimuli and make up the cp-neuron population. As shown, there is no statistically significant difference in the proportion of cp-neurons between naïve and noise-reared animals. Furthermore there is no significant difference in the proportion of any neuron type between the two groups. These data suggest that an increased proportion of cp-neurons in the noise-reared animals is not responsible for increased performance in recognizing vocalizations in the face of noise. In addition, the overall distribution of neuron types in A1, as classified by this method, does not change with noise-rearing.

To determine if individual cp-neurons from noise-reared animals are better at encoding vocalization information in the face of noise we compared the distribution of information values between noise-reared and naïve animals for the vocalization-in-noise response rasters (Figure 6B). Cp-neurons from noise-reared animals encode significantly more vocalization specific information per second in the face of naturalistic noise at 0-dB SNR (9.0 ± 1.0 bits/sec vs 5.7 ± 0.4 bits/sec, mean \pm s.e.m, p-value < 0.005, Wilcoxon rank sum test). This finding shows that although there is no difference in the overall proportion of cp-neurons between the two groups, individual cp-neurons in noise-reared animals encode more vocalization information in the face of noise compared to naïve animals. This increased encoding ability by cp-neurons in noise reared animals

could underlie their increased behavioral performance on the cocktail party task and further analysis of the receptive field properties of these neurons will give insight to the mechanisms by which neurons encode vocalization information in the presence of noise.

Discussion

The aim of this study was to create animals that are specialists at solving the cocktail party problem, and then identify auditory processing differences between specialized and non-specialized animals to investigate the mechanisms neural systems use to process signals of interest in the presence of background noise. To create specialized rats we reared animals in synthetic noise with characteristics of natural noise from a lab rat's acoustic environment. Through behavioral testing we have shown noise-rearing increases behavioral performance on a cocktail party problem task, and this increased performance is due to an auditory specific adaptation rather than non-auditory specific effects like familiarity with noise. Using a method designed to identify cocktail party neurons, we found a subpopulation of cp-neurons in both naïve and noise-reared animals with the ability to extract vocalization information in the presence of noise. Although there was no difference in the relative proportion of cocktail party neurons between groups, noise-reared cocktail party neurons encoded vocalization information at significantly higher rates. This increased ability to extract vocalization information in the presence of noise could translate to larger amounts of encoded vocalization information in the awake behaving animal and

could account for the increased behavioral performance shown by noise-reared rats.

To create specialized rats we reared animals in naturalistic noise, however, any type of noise could be used, and mechanisms by which neural systems deal with noise may be different depending on the particular type of noise. In addition, within a particular class of noise, models of the neural mechanisms which facilitate signal perception can be iteratively refined by varying the characteristics of noise used for rearing. For example, it is possible that shifts in temporal tuning are important for cp-neuron function in our naturalistic noise rearing conditions. This hypothesis could be tested by rearing animals in different versions of naturalistic noise with varying temporal modulation content.

An alternative strategy to create specialized animals is to directly train them to perform an auditory task in the face of noise and characterize changes in processing that correlate with increased performance (Whitton, Hancock, Polley, 2012). This is potentially a more direct approach, however, it is possible the adaptive changes induced by training may be very specific to the signal and noise used, and may be dominated by highly specific shifts in tuning for the particular signal stimulus rather than general adaptive changes that would allow the system to better process any potential signal in the face of noise. By contrast, because animals in a noise-rearing environment have many signals of interest that are only defined by what is behaviorally relevant at the moment, adaptive

changes are likely to highlight general mechanisms a neural system uses to process any signal of interest in the particular class of noise used for rearing. In fact, one can think of noise rearing as a “*natural*” training paradigm in which the training occurs continuously, the behavior-reward contingencies are natural (i.e. a pup may be the last to start feeding if it doesn’t hear the mother’s vocalization), and the training begins at infancy and lasts through development- a time in which the brain is most plastic and can adapt to challenges most efficiently (Zhang, Bao, & Merzenich, 2001b; Xiaoming Zhou & Merzenich, 2008b; de Villers-Sidani et al., 2007).

Although we have demonstrated a useful approach to studying the cocktail party problem in which we create specialized animals and then identify the neural correlates of that specialization, there is still considerable work to be done to understand the processes by which neural systems extract information about signals of interest in the presence of noise. One limitation as well as benefit of the study is the use of anesthetized animals for neural recordings. In general, recordings of single units in anesthetized animals are more stable than recordings in awake, behaving animals, which allows for longer recording times. Information estimates, which were essential to our method of identifying cp-neurons and the study as a whole, are data intensive (Panzeri, Senatore, Montemurro, & Petersen, 2007a; Strong, Koberle, de Ruyter van Steveninck, & Bialek, 1998). Therefore, we used an anesthetized preparation to increase the likelihood of obtaining enough data from single units to obtain the necessary information estimates. However, it is unclear how well characterizations of

processing in the anesthetized state reflect processing in awake animals that are performing the task of interest. Future recordings in awake, behaving animals are likely to be necessary for a complete understanding of the problem. A second limitation of the study is the restricted region in which we made recordings. The auditory system, and brain in general, is a highly integrated system with many levels of processing. We recorded neural activity in a single level of this system—the primary auditory cortex. It is possible that synchronous measurements at each level of the system will be necessary to formulate a deep understanding of how the brain solves the cocktail party problem. Lastly, we identified a population of neurons in noise-reared animals that extracts higher amounts of vocalization information in the face of noise, but we gave no account of a mechanism to explain this. It should be possible to gain insight on this mechanism by characterizing the spectrotemporal receptive field properties of these neurons. Recently several methods have been developed to specifically characterize neural receptive fields using natural signals (David, Mesgarani, & Shamma, 2007; Paninski, 2004; Sharpee, Rust, & Bialek, 2004; Theunissen et al., 2001). These methods will undoubtedly be useful in characterizing receptive field properties and identifying potential mechanisms of cp-neuron function.

Since the description of a critical period in the primary auditory cortex of rats, there have been many studies characterizing the effects of rearing animals in different exposure stimuli (Chang & Merzenich, 2003a; Insanally et al., 2009; Nakahara et al., 2004; Zhang et al., 2001b; X Zhou, Nagarajan, Mossop, & Merzenich, 2008; Xiaoming Zhou & Merzenich, 2008b; de Villers-Sidani et al.,

2008). Rearing animals in the presence of pulsed, pure-tone stimuli causes tonotopic map expansion at the exposure frequency (Zhang et al., 2001b). Because training adult animals on a frequency discrimination task also induces tonotopic map expansion it is possible that tone-reared animals would show better frequency discrimination abilities near the exposure frequency (Polley, Steinberg, & Merzenich, 2006; Recanzone, Schreiner, & Merzenich, 1993). However, tone-reared animals show decreased discrimination performance near the exposure frequency contrary to training induced map expansion (Han, Köver, Insanally, Semerdjian, & Bao, 2007). This example shows the necessity of testing animals behaviorally to determine the perceptual effects of stimulus-rearing induced changes in processing. A particularly relevant study reared animals in continuous unmodulated white noise and showed abnormal development of frequency selectivity and degraded tonotopic organization (Chang & Merzenich, 2003). This was interpreted as an overall degradation in auditory processing and white noise was implicated as a potential risk factor for abnormal child development. However, in the absence of perceptual testing to quantify the effects of unmodulated white noise exposure it is unclear exactly what the perceptual effects were. Although some perceptual tasks, like pure tone frequency discrimination, may be degraded, other tasks, like vocalization recognition in the face of unmodulated white noise, may be improved. It is likely that the combined effects of changes in processing induced by stimulus-rearing experiments impart a net benefit to the animal which allows the organism to better survive in the environment it was raised in. Through carefully chosen rearing conditions, we can look for correlations between adaptive changes in processing and increased perceptual performance to gain insight on mechanisms

the brain uses to extract information from the environment to maximize an organisms chances of survival.

Methods

Acoustic Recordings and Noise Synthesis

Acoustic recording of 6 female siblings housed in a sound attenuation chamber were made continuously over a ten day period (p30 – p40). The sound attenuation chamber restricted the acoustic environment to self generated sound. All acoustic recordings and noise rearing were performed in sound attenuation chambers to maintain control and consistency over the acoustic environment. Recordings were made using a microphone suspended 10 cm above the cage. Recordings were collected sampled at 192 kHz, in 10 minute increments, and saved to disk. A two second 4 kHz tone of known SPL was played at the beginning of each recording for calibration purposes. Daily checks were performed to ensure no interruption in the recording process. Spectrogram representations of acoustic recordings were generated with a short-time Fourier transform and visually inspected to extract rat vocalizations and non-vocalization sound. The dominant non-vocalization sound, which we classify as noise, was associated with quick movement related behavior and was characterized by broadband frequency content coherently modulated in time. The synthetic version consists of broadband white noise whose sound pressure level is modulated in time by a modulation envelope $[m_{dB}(t)]$ to give the desired variation in SPL. To generate the modulation envelope, a Gaussian white signal

was filtered in the frequency domain to give a modulation envelope with the desired temporal modulation content. After filtering, the modulation envelope was normalized to have zero mean and unit variance. To generate the linear modulation envelope, $m_{db}(t)$ was exponentiated according to :

$$m_{lin}(t) = 1e^{-5} \times 10^{[\mu + \sigma \times m_{db}(t)]/20} \quad (1)$$

Broadband white noise was then multiplied by the linear modulation envelope ($m_{lin}(t)$) to give the desired synthetic noise signal with sound pressure level as a function of time given by

$$SPL_{dB}(t) = \mu + \sigma \times m_{db}(t) \quad (2)$$

Where the mean (μ) and standard deviation (σ) were 60 dB and 8 dB respectively. The mean noise SPL was chosen to create a challenging listening conditions, but avoid pathologic sound pressure levels (Cappaert, Klis, & Muijser, 2000; Ward, Cushing, & Burns, 2000).

Noise Rearing

Sprague Dawely female rats were housed in groups of 6 in sound attenuation chambers. A speaker was suspended 30 centimeters above the cage. A 1-hour instance of naturalistic noise was played continuously at an average SPL of 60 dB starting at p6, prior to ear canal opening. Noise-reared animals were kept in the noise exposure until the day of electrophysiology recordings (average age p60). For behavioral testing, noised-reared animals were returned to the noise exposure outside of training and testing sessions. The average age of animals at the end of testing was p84 for the first vocalization recognition in noise experiment and the visual stimulus recognition in noise experiment. Animals from the first vocalization recognition in noise experiment were then trained and tested on a second target. The average age at the end of testing was p104. Animals were checked daily to ensure that the noise exposure was not interrupted and the average SPL remained at the desired level. No noticeable differences were observed in the behavior or temperament of noise-reared animals compared to naïve animals raised in normal laboratory conditions.

Behavioral Training and Testing

To determine if noise rearing confers an increased ability to recognize rat vocalizations embedded in noise, three Sprague Dawely rats raised in normal laboratory conditions (without noise exposure and outside sound attenuation chambers) and four noise reared rats were trained to discriminate rat vocalizations using an operant Go / No-Go training paradigm. All rats started training at p45. Rats were placed on food restriction to provide motivation

during training sessions. Both groups of rats were trained on the task in the absence of noise. After reaching peak performance, noise was added to the task during a short two session testing phase in which their ability to perform the task at various SNRs was quantified. For the vocalization recognition task, rats were trained in quiet conditions to respond with a nose-poke (“Go”) within 4 seconds of the onset of a target vocalization or withhold a response (“No-Go”) when presented with one of four non-targets (three vocalizations and one silent catch trail). A Correct “Go” response to the target was scored as a hit and was rewarded with a food pellet (45 mg; BioServe, Frenchtown, NJ). Failure to perform a “Go” response to the target was scored as miss and punished with an 10 second timeout period in which the house light was turned off. A correct “No-Go” response to a non-target was scored as a withhold and neither rewarded nor punished. An incorrect “Go” response to a non-target was scored as a false positive and punished with a 30 s timeout. The long timeout period for a false positive served to discourage a strong bias for this group of rats to respond to all stimuli and accrue high false positive rates. A single trial was defined as the length of time between the start of the stimulus and the end of the 4 second response window plus any time out period resulting from an incorrect response. Intertrial intervals were randomly selected between 1 – 4 seconds long and nose-poke responses during this time prevented stimulus presentation, were scored as a false alarm, and resulted in a 30 second time out. Following the timeout, the intertrial interval was reset and the next trial allowed to proceed.

Exposed and naïve rats were trained to discriminate vocalizations the absence of background noise. After reaching satisfactory vocalization discrimination performance, the synthetic background noise was added and their ability to discriminate vocalizations in the face of noise was tested at various SNR levels. Outside the training sessions exposed rats were placed back in the noise exposure. Training rats to discriminate vocalization consisted of three phases. In phase one, rats were trained to nose-poke for a food reward. In phase two, rats were trained to nose-poke within 4 seconds of the onset of any vocalization and withhold a response on silent catch trials. Once responses were clearly under stimulus control and false positive rates were below 15% on silent catch trials, rats were advanced to phase three and trained to respond to a single target vocalization and withhold responses to all non-targets and silent catch trials. To keep the rats motivated despite long trial durations the probability of a target was 0.6 (Schnupp, Hall, Kokelaar, & Ahmed, 2006). The probability of the three non-targets and silent catch trials were 0.1. Once animals reached a plateau in performance defined by a standard deviation in percentage correct below 5% over a four day window, they were advanced to the testing stage in which discrimination ability in the face of noise was tested. Testing consisted of two sessions that took place on a consecutive two-day period. Prior to the beginning of a test session, 30 warm-up trials were given without noise. Within a test session there were 6 SNR blocks of 45 trials per block in which the blocks had the following order: 55 dB vocalization (zero noise), 0 dB SNR, -5 dB SNR, -10 dB SNR, -15 dB SNR, -20 dB SNR. This ordering was randomized during the testing phase of the second target. During each block the background noise was continuously played rather than starting and stopping at the beginning and end

of each trial to mimic the conditions of the actual noise exposure. After the testing phase for the first target, both groups were trained and tested on a second target. For the testing phase of the second target the order of low SNR blocks was randomized but was still preceded and followed by 55 dB (zero noise) SNR blocks. In addition a second 55 dB (zero noise) SNR block was added at the end to quantify performance in zero noise after progression through the noise blocks. Behavioral training and testing of visual stimuli recognition in acoustic noise was identical with the exception of visual rather than auditory target and non-target stimuli. In addition the dependence of house light illumination was reversed. During trials the house light was turned off to increase visibility of the visual stimulus. During time out periods the house light was turned on.

Stimuli

Three sets of stimuli were used to characterize auditory processing of rat vocalizations embedded in naturalistic noise. The first stimulus consists of a 5 second bout of rat vocalizations repeated 100 times. The vocalizations span the 20 – 30 kHz frequency range and were generated by concatenating 4 isolated vocalizations from the acoustic recordings. This stimulus is used to identify neurons sensitive to vocalizations in the 20-30 kHz frequency range in the absence of noise. The second stimulus consists of the same vocalization repeats with naturalistic noise added to each trial at 0 SNR, however, on each trial a different instance of naturalistic noise is used. This stimulus is used to quantify the amount of vocalization specific information present in responses to

vocalizations embedded in naturalistic noise. The third stimulus consist of 12 minutes of rat vocalizations in the 20 – 30 kHz range taken from the acoustic recordings and embedded in naturalistic noise at 0 SNR. This stimulus will be used to thoroughly probe the 20 – 30 kHz vocalization in noise stimulus space for the purpose of characterizing receptive field properties

Electrophysiology

Naïve (n = 4, average age = p60.8± 9.8) and noise-reared animals (n = 4, average age = p60.5 ± 9.1) were anesthetized with pentobarbital and a craniotomy was performed unilaterally over auditory cortex. Stimuli were delivered in the contralateral ear via a closed speaker system. Responses to pure tone stimuli were recorded with Tungsten electrodes (1-2 M Ω) to measure multiunit frequency-intensity response areas and generate a map of characteristic frequency. Primary auditory cortex was functionally identified as having short latency onset responses to tones (8-20 ms) and located as the most dorsal tonotopically organized region with a gradient running in the posterior to anterior direction ~ 10° relative to the transverse plane. After identification of primary auditory cortex, responses to the set of vocalization and vocalization in noise stimuli were recorded from the 20-30 kHz regions of the map. Recordings were made with two NeuroNexus A series 16 channel probes (2-3 M Ω impedance) in which each probe had 4 shanks oriented in a plane (125 μ m between shanks), and 4 channels oriented along the length of each shank (100 μ m

spacing). Neural traces were band-pass filtered between 600 and 6000 Hz and recorded to disk using a Nueralynx-Cheetah A/D system. Spikes were then sorted offline using a Bayesian spike sorting algorithm (M. Lewicki, 1998; M. S. Lewicki, 1994).

Data Analysis

Information values for the vocalization repeats and vocalization in noise repeats were computed using the direct method (Strong et al., 1998). Spike train data was binned at 4 ms resolution. Bias due to finite data set size was corrected using quadratic extrapolation, in which estimates of total and noise entropy were computed using subsets of the data (90%, 80%, 70%, 60%) and then extrapolated for infinite data set size (Panzeri, Senatore, Montemurro, & Petersen, 2007b; Strong et al., 1998). Each data fraction estimate was bootstrapped using different data fraction segments. After extrapolation for infinite data set size, total and noise entropy estimates were computed for word durations of 4, 8, 16, 32, 64, and 128, ms and extrapolated for infinite word length to obtain information estimates for vocalizations and vocalizations in noise at 0 SNR. Because the noise on each trial is different for the vocalization in noise stimulus, components of the response related to the noise are different on each trial, and as such, are quantified as noise (noise entropy) in the information estimates thus allowing for the isolation of vocalization specific information. For all information estimates, the first 750 ms of each trial was discarded to remove information related to the onset of each trial. This is particularly important for vocalization in noise information

estimates to isolate vocalization specific information. The onset of noise preceded the onset of vocalizations on each trial. Although the noise on each trial is different, its onset still generated a consistent response across trials, which would contaminate vocalization specific information estimates if not removed.

Significance values were used to classify neurons according to which stimuli they encode significant information for. Permutation tests were used to determine if responses contained significant information for each stimulus. To determine significance, each trial of a given raster was circularly permuted a random amount, leaving the statistics of the spike train unchanged but removing its dependence on the stimulus. Information estimates were computed for 1000 instances of the permuted raster to generate a distribution of information values for responses with the same spike train statistics but no information stimulus information. This distribution was compared to the information estimate for the actual unpermuted raster to determine significance. Information estimates using the direct method are time-intensive due bootstrapping, and extrapolations for infinite data set size and word length. To make permutation testing feasible, we used a less time intensive estimate of information, I_{spike} , which is correlated with total information and quantifies the information given by the arrival time of a single spike (Brenner, Strong, Koberle, Bialek, & de Ruyter van Steveninck, 2000).

References

- Brenner, N., Strong, S. P., Koberle, R., Bialek, W., & de Ruyter van Steveninck, R. R. (2000). Synergy in a neural code. *Neural computation*, *12*(7), 1531–52. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/22049416>
- Brudzynski, S M, Bihari, F., Ociepa, D., & Fu, X. W. (1993). Analysis of 22 kHz ultrasonic vocalization in laboratory rats: long and short calls. *Physiology & behavior*, *54*(2), 215–21.
- Brudzynski, S M, Kehoe, P., & Callahan, M. (1999). Sonographic structure of isolation-induced ultrasonic calls of rat pups. *Developmental psychobiology*, *34*(3), 195–204.
- Brudzynski, Stefan M. (2005). Principles of rat communication: quantitative parameters of ultrasonic calls in rats. *Behavior genetics*, *35*(1), 85–92. doi:10.1007/s10519-004-0858-3
- Buus, S. (1985). Release from masking caused by envelope fluctuations. *The Journal of the Acoustical Society of America*, *78*(6), 1958–65.
- Cappaert, N., Klis, S., & Muijser, H. (2000). Noise-induced hearing loss in rats. *Noise and Health*, *3*(9), 23–32.

Chambers C.R., Hancock K.E., Polley D.B., (2012). An adaptive stimulus search method for rapid characterization of multidimensional receptive fields in auditory cortex of awake animals. Poster presented at Advances and Perspectives in Auditory Neurophysiology. New Orleans, LA.

Chang, E. F., & Merzenich, M. M. (2003). Environmental noise retards auditory cortical development. *Science (New York, N.Y.)*, 300(5618), 498–502. doi:10.1126/science.1082163

Cherry, E. (1953). Some experiments on the recognition of speech, with one and two ears. *The Journal of the acoustical society of America*, 25, 975–979.

David, S. V., Mesgarani, N., & Shamma, S. a. (2007). Estimating sparse spectro-temporal receptive fields with natural stimuli. *Network (Bristol, England)*, 18(3), 191–212. doi:10.1080/09548980701609235

de Villers-Sidani, E., Chang, E. F., Bao, S., & Merzenich, M. M. (2007). Critical period window for spectral tuning defined in the primary auditory cortex (A1) in the rat. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 27(1), 180–9. doi:10.1523/JNEUROSCI.3227-06.2007

de Villers-Sidani, E., Simpson, K. L., Lu, Y.-F., Lin, R. C. S., & Merzenich, M. M. (2008). Manipulating critical period closure across different sectors of the

primary auditory cortex. *Nature neuroscience*, 11(8), 957–65.

doi:10.1038/nn.2144

Han, Y. K., Köver, H., Insanally, M. N., Semerdjian, J. H., & Bao, S. (2007).

Early experience impairs perceptual discrimination. *Nature neuroscience*,

10(9), 1191–7. doi:10.1038/nn1941

Insanally, M. N., Köver, H., Kim, H., & Bao, S. (2009). Feature-dependent

sensitive periods in the development of complex sound representation. *The*

Journal of neuroscience : the official journal of the Society for Neuroscience, 29(17),

5456–62. doi:10.1523/JNEUROSCI.5311-08.2009

Kaltwasser, M. T. (1990). Acoustic signaling in the black rat (*Rattus rattus*).

Journal of comparative psychology (Washington, D.C. : 1983), 104(3), 227–32.

Lewicki, M. (1998). A review of methods for spike sorting: the detection and

classification of neural action potentials. *Network: Computation in Neural*

Systems, 9(4), R53–R78. doi:10.1088/0954-898X/9/4/001

Lewicki, M. S. (1994). Bayesian Modeling and Classification of Neural

Signals. *Neural Computation*, 6(5), 1005–1030. doi:10.1162/neco.1994.6.5.1005

McDermott, J. H. (2009). The cocktail party problem. *Current biology : CB*,

19(22), R1024–7. doi:10.1016/j.cub.2009.09.005

Nakahara, H., Zhang, L. I., & Merzenich, M. M. (2004). Specialization of primary auditory cortex processing by sound exposure in the “critical period”. *Proceedings of the National Academy of Sciences of the United States of America*, 101(18), 7170–4. doi:10.1073/pnas.0401196101

Paninski, L. (2004). Maximum likelihood estimation of cascade point-process neural encoding models. *Network: Computation in Neural Systems*, 15(4), 243–262. doi:10.1088/0954-898X/15/4/002

Panzeri, S., Senatore, R., Montemurro, M. a, & Petersen, R. S. (2007). Correcting for the sampling bias problem in spike train information measures. *Journal of neurophysiology*, 98(3), 1064–72. doi:10.1152/jn.00559.2007

Polley, D. B., Steinberg, E. E., & Merzenich, M. M. (2006). Perceptual learning directs auditory cortical map reorganization through top-down influences. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 26(18), 4970–82. doi:10.1523/JNEUROSCI.3771-05.2006

Portfors, C. V. (2007). Types and functions of ultrasonic vocalizations in laboratory rats and mice. *Journal of the American Association for Laboratory Animal Science : JAALAS*, 46(1), 28–34.

Recanzone, G. H., Schreiner, C. E., & Merzenich, M. M. (1993). Plasticity in the frequency representation of primary auditory cortex following discrimination training in adult owl monkeys. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 13(1), 87–103.

Schnupp, J. W. H., Hall, T. M., Kokelaar, R. F., & Ahmed, B. (2006). Plasticity of temporal pattern codes for vocalization stimuli in primary auditory cortex. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 26(18), 4785–95. doi:10.1523/JNEUROSCI.4330-05.2006

Sharpee, T., Rust, N. C., & Bialek, W. (2004). Analyzing neural responses to natural signals: maximally informative dimensions. *Neural computation*, 16(2), 223–50. doi:10.1162/089976604322742010

Strong, S., Koberle, R., de Ruyter van Steveninck, R., & Bialek, W. (1998). Entropy and Information in Neural Spike Trains. *Physical Review Letters*, 80(1), 197–200. doi:10.1103/PhysRevLett.80.197

Theunissen, F. E., David, S. V., Singh, N. C., Hsu, a., Vinje, W. E., & Gallant, J. L. (2001). Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *Network: Computation in Neural Systems*, 12(3), 289–316. doi:10.1088/0954-898X/12/3/304

Verhey, J. L., Pressnitzer, D., & Winter, I. M. (2003). The psychophysics and physiology of comodulation masking release. *Experimental brain research. Experimentelle Hirnforschung. Expérimentation cérébrale*, 153(4), 405–17. doi:10.1007/s00221-003-1607-1

Ward, W. D., Cushing, E. M., & Burns, E. M. (2000). Effective quiet and moderate TTS: Implications for noise exposure standards., 59(1), 160–165.

Zhang, L. I., Bao, S., & Merzenich, M. M. (2001). Persistent and specific influences of early acoustic environments on primary auditory cortex. *Nature neuroscience*, 4(11), 1123–30. doi:10.1038/nn745

Zhou, X, Nagarajan, N., Mossop, B. J., & Merzenich, M. M. (2008). Influences of un-modulated acoustic inputs on functional maturation and critical-period plasticity of the primary auditory cortex. *Neuroscience*, 154(1), 390–6. doi:10.1016/j.neuroscience.2008.01.026

Zhou, Xiaoming, & Merzenich, M. M. (2008). Enduring effects of early structured noise exposure on temporal modulation in the primary auditory cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 105(11), 4423–8. doi:10.1073/pnas.0800009105

Concluding Remarks

This work was motivated by a broad interest to understand how the brain processes information to give rise to intelligence. This is a very broad question so we restrict our efforts on understanding neural processing of sensory inputs to the system and focus on two important questions pertaining to auditory processing: (1) What is the function organization of speech processing in human superior temporal gyrus (STG), and (2) how does the brain extract information about signals of interest from noise. In superior temporal gyrus we show, for the first time, a robust form of organization called modulotopy. Modulotopic organization is a higher order analog of tonotopic organization based on tuning for the Fourier components of the time-frequency (spectrogram) representation of sound. This form of functional organization provides evidence for a succinct model of auditory processing: the auditory system initially decomposes sound into its Fourier components (frequency) to generate a time-frequency spectrogram representation of the sound. This time-frequency representation of sound is then decomposed in terms of the Fourier components of a spectrogram known as spectrotemporal modulations. This can be conceptualized as a stacked Fourier decomposition of the signal (Chapter 1, Figure 6) in terms of frequency then in terms of spectrotemporal modulations.

With regard to understanding how neural systems solve the cocktail party problem, we have established a systematic approach in rats to study how neural

systems extract information about signals of interest embedded in noise. In this work we demonstrated a method to create rats that are specialized at solving the cocktail party problem based on rearing them in naturalistic noise. We tested animals behaviorally on a cocktail party problem task and show that noise-rearing facilitates the development of an increased ability to recognize vocalizations in the presence of the naturalistic noise. In addition, we demonstrated a method to identify individual that have the ability to extract vocalization information in the face of noise. Examination of the response properties of these neurons will illuminate neural mechanisms which allow the brain to extract signals of interest from noise. We believe this will be a powerful tool to probe the system with different signal and noise classes to investigate how the system solves the cocktail party problem in the range of different signal and noise conditions that organisms are confronted with during everyday life.

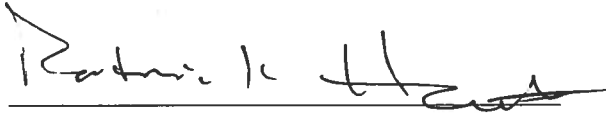
In the course of this work we have identified a highly informative description of the organization of speech processing in humans and demonstrated a systematic framework for studying the cocktail party problem in rats. Although we have not explicitly shown the computations the brain performs to give rise to our speech perceptual abilities or to extract signal from noise we believe the work presented here constitutes a significant contribution to understanding these computations and can be built upon to eventually understand the nature of intelligent processing in the auditory system.

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.



Author Signature

6/14/13
Date