ABSTRACT

Title of Dissertation: SPEECH SEGREGATION AND REPRESENTATION IN THE FERRET AUDITORY AND FRONTAL CORTICES

Neha Hemant Joshi
Doctor of Philosophy, 2022

Dissertation Directed by: Professor Shihab Shamma
Department of Electrical Engineering

The problem of separating overlapping streams of sound, and selectively attending to a sound of interest is ubiquitous in humans and animals alike as a means for survival and communication. This problem, known as the cocktail party problem, is the focus of this thesis, where we explore the neural correlates of two-speaker segregation in the auditory and frontal cortex, using the ferret as an animal model.

While speech segregation has been studied extensively in humans using various non-invasive imaging as well as some restricted invasive techniques, these do not provide a way to obtain neural data at the single-unit level. In animal models, streaming studies have been limited to simple stimuli like tone streams, or sound in noise. In this thesis, we extend this work to understand how complex auditory stimuli such as human speech is encoded at the single-unit and population level in both the auditory cortex, as well as the frontal cortex of the ferret.

In the first part of the thesis, we explore current literature in auditory streaming and design

a behavioral task using the ferret as an animal model to perform stream segregation. We train ferrets to selectively listen to one speaker over another, and perform a task to indicate detection of the attended speaker. We show the validity of this behavioral task, and the reliability with which the animal performs this task of two speaker stream segregation. In the second part, we collect neurophysiological data which is post-processed to obtain data from single units in both the auditory cortex (the primary auditory cortex, and the secondary region which includes the dorsal posterior ectosylvian gyrus) as well as the dorsolateral aspect of the frontal cortex of the ferret. We analyse the data and present findings of how the auditory and frontal cortices encode the information required to reliably segregate the speaker of relevance from the mixture of two speakers, and the insights provided into stream segregation mechanisms and the cocktail party solved by animals using neural decoding approaches.

We finally demonstrate that stream segregation has already begun at the level of the primary auditory cortex. In agreement with previous attention-modulated neural studies in the auditory cortex, we show that this stream segregation is more pronounced in the secondary cortex, where we see clear enhancement of the attended speaker, and suppression of the unattended speaker. We explore the contribution of various areas within the primary and secondary regions, and how it relates to speaker selectivity of individual neuronal units. We also study the neural encoding of top-down attention modulation in the ferret frontal cortex. Finally, we discuss the conclusions from these results in the broader context of their relevance to the field, and what future directions it may hold for the field.

SPEECH SEGREGATION AND REPRESENTATION IN THE FERRET AUDITORY AND FRONTAL CORTICES

by

Neha Hemant Joshi

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2022

Advisory Committee:
Professor Shihab Shamma, Chair/Advisor
Professor Jonathan Z. Simon
Professor Carol Espy-Wilson
Professor Behtash Babadi
Professor Naomi Feldman (Dean's Representative)

Dedication

Dedicated to my grandparents - it cannot be understated how much you've sacrificed to let me be where I am today, without ever making me feel guilty about the time I could have spent with you more.

Acknowledgments

It takes a village to complete a thesis, and I truly felt the support of my village. I owe my gratitude to all the people who have made this thesis possible. First of all, I would like to thank my advisor, Dr. Shihab Shamma who has been one of the kindest advisors I could ever ask for. Meeting him in India got me interested in this extraordinarily complex field of speech neuroscience, and I am glad it did. He has been there for me throughout my PhD, making sure I had everything I needed and was doing okay, meeting with me whenever I asked for it (even weekends!) and always encouraging me to learn more, and take on more academically challenging assignments.

I would also like to thank all my committee members, who were instrumental to my graduate school experience and kindly agreed to serve on my dissertation committee. Dr. Babadi taught me my first class at UMD, and has been a friendly face ever since. Dr. Espy-Wilson and Dr. Feldman co-taught a seminar that I learnt so much from, and which also introduced me to the large speech and language community on campus that has been instrumental in my extra-curricular development. Dr. Simon has been patiently answering all my questions about auditory science, graduate school, thesis proposal, and his mentoring was crucial to having a more rounded graduate experience. It was a pleasure working with him and Joshua on a project together.

Everyone at the Neural Systems Lab has been a pleasure to work with. I have learnt everything I know about animal physiology from Dani, Diego, Jonathan and Pingbo - they've

all been extremely patient and encouraged me to learn so much. Dani has been an incredibly wonderful mentor and friend throughout. Wing, Ali, Vasudha, Jaya, Rupesh, Daniel, Vinay, Shoutik, Rahil and many more have been such a dependable presence during their time at NSL. Special kudos to Kelsey and Mohsen for their invaluable friendship, I cannot put in words how much that meant to me.

There are so many more who were instrumental to my time at College Park - the ISR and ECE business office, especially Regina, the Language Science Center (thanks Shevaun and Colin!), Board and the Brew, Parkrun running community and so many others I must have inadvertently forgotten - thank you everyone! You've given me a lifetime of memories to cherish.

Friends in graduate school made College Park a home away from home for the longest time. Janhavi, Karthik, Manny, Nadee, Garrett and everyone else who I haven't listed here - thank you! I owe a big hug to Ranjani, Somie, Shruti and all friends and family all over the world who have had my back throughout these seven years.

I owe my deepest thanks to my family - my parents patiently listened to me, consoled me, encouraged me and supported me in every way they possibly could. They've respected me, my choices and boundaries and made me feel so loved, always. My brother was there for me in a way no one else has been, the backup sound board and co-foodie.

Last but not the least, I could not have done this without my husband, Yogarshi and my meowling, Shadow. They've given their everything to me during this period, standing by me through the good and the ugly. The thesis just would not have happened without you, Yogarshi, thank you for everything.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

A1	Primary Auditory Cortex
ACx	Auditory Cortex
ASG	Anterior Sigmoid Gyrus
dPEG	dorsal aspect of Posterior Ectosylvian Gyrus (secondary Auditory Cortex)
HG	Heschl's Gyrus
STG	Superior Temporal Gyrus
vPR	ventral Posterior Auditory Field
ECoG	Electrocorticography
EEG	Electroencephalography
MEG	Mangeto-Encephalography
PSTH	Peristimulus Time Histogram
SNR	Signal to Noise ratio
SSI	Speaker Selectivity Index
STRF	Spectro-temporal Receptive Field
TORC	Temporally Orthogonal Ripple Combinations

Chapter 1: Introduction

We perceive and navigate through the world with our senses. Hearing and audition form a very integral part of this sensory system, allowing us to perceive distance, enable communication, assist survival and much more. Yet very little is known about how humans and animals achieve these remarkably fast computations to navigate the world through a single stream of sound reaching each of their ears.

1.1 The Cocktail Party Problem

The remarkable ability of humans and other animals and birds to selectively tune into speech from a specific source in an otherwise noisy multi-speaker environment is more commonly known as the cocktail party problem. The cocktail party paradigm was first introduced formally by Cherry in 1953 [3]. This problem falls among the broad category of problems in auditory scene analysis [4] where a listener tries to navigate through an auditory environment, selectively listening and identifying components of it. This problem is fairly ill-posed mathematically due to the myriad of factors and variants within the problem, and extremely nuanced. Broadly speaking, there are two key challenges that a subject of the cocktail party needs to resolve.

The first challenge to solving the cocktail party is that of stream formation and segregation, as individuals are more often than not interested in one source rather than the whole environment.

They thus need to form auditory streams comprising of auditory objects, that possibly originate from different sources. There are a few widely accepted theories of how auditory streams are formed, and commonalities within a stream [5, 6, 7] that facilitate stream segregation, discussed further in the next chapter.

The second challenge an individual faces is that of directed attention. While auditory object and stream formation may or may not need attention, the processing of one particular segregated stream and directing resources to this task rather than processing the entire auditory scene is a strongly attention-driven process [5]. Cognitive aspects of attention are crucial to the solution of the cocktail party problem. However the underlying physiological mechanisms that encode selective auditory attention in the brain remain unclear.

In the context of a gathering of people, one person may wish to carry on a conversation with their partner. They need to separate out the partner's voice from that of the vehicles passing, the vendor in the backdrop, the sound of the keys dropped by the person passing by, conversations of the passers by and many more. In addition to separating all these sources, they need to direct attention to the auditory object of interest—their partner's voice. They can do it with great ease, yet these fast and complex computations within the brain are not yet completely understood. In this thesis, we study the encoding of such a cocktail party scenario at the level of single neuronal units in the auditory and frontal cortices in the brain to understand the mechanisms underlying attention driven stream segregation and directed attention during the cocktail party.

1.2 Animal models for Studying Neural Correlates during Auditory Streaming

Perceptual correlates of auditory streaming in animals have long been a subject of interest due to the validity of the cocktail party for animals, just as much as humans. Pre-attentive stream segregation can be studied using neurophysiological recordings in the absence of behavior. Features within Van Noorden streams [8] comprising of alternating tones have been studied extensively in humans and animals alike to understand what features affect stream formation in perceptual stream segregation. Early studies in the songbird primary cortical fields [9, 10] have shown difference in neural responses to tones during two-tone sequence presentation, validating the effects of features such as frequency difference, that induce varying levels of stream segregation. Neurophysiological evidence from awake, non-behaving bats to pulse-echo sequences demonstrate preattentive segregation of pulses and echo in population responses in the primary auditory cortex [11]. Elhilali et. al. [12] have shown the effect of temporal coherence of two streams on stream segregation.

While the above studies demonstrated the neural effects of differential responses in two-tone sequences pre-attentively without behavior, Carlyon et. al. [5] and others show that behavioral attention is indeed required to segregate acoustically interleaved tone sequences. To understand the neural correlates of attention driven auditory streaming, it is crucial to obtain neurophysiological data during behavior, as stream segregation is occurring. Stream segregation involves top-down processing relying on the ability of animals to pay attention and perform streaming. Evidence of streaming using tones is seen in primary cortical fields in terms of stimulus statistics and well segregated streams of tones in birds [13] and monkeys [14]. Itatani et. al. [13] show that while segregated responses are seen in the starlings' A1, the decision process

through behavior was not reflected in the responses. Using alternating tone sequences, it has been shown that in the macaques' A1, we can see evidence of build up of perceptual stream formation [15]. Ferrets can stream alternating tones and detect deviant frequencies in the stream, as well as detect a repeating tone in a random tone-cloud [16].

Most previous studies used sequences of simple stimuli such as tones and tone complexes in the primary auditory cortex. Repetition, as a cue for stream segregation, has also been demonstrated through detection of repeated noise segments (foreground) in a random noise segment (background) and the authors have shown that there is increased stream specific gain leading to enhancement of the foreground repeating noise in the ACx (primary and secondary cortices), and overall reduction of global gain [17]. In this work, we extend the study of streaming to more complex stimuli by using speech streaming in animals in both the primary and secondary auditory cortices, as well as the frontal cortex.

1.3 Ferret as an Animal Model for Studying the Cocktail Party Problem

For the purpose of behavior and neurophysiology, we use the ferret as an animal model to study auditory phenomena at a single cell level in the auditory and frontal cortices of the brain. Ferrets are an ideal model for auditory research since their hearing range is similar to that of the human hearing range [1] (Figure 1.1, adapted from [1]). A previous study has demonstrated that neuronal responses in the ferret primary auditory cortex are sufficiently diverse to encode and discriminate all English phoneme classes [18]. The functional connectivity of the ferret auditory cortex has also been studied, and well established [19], thus providing a framework for neurophysiological data collection.

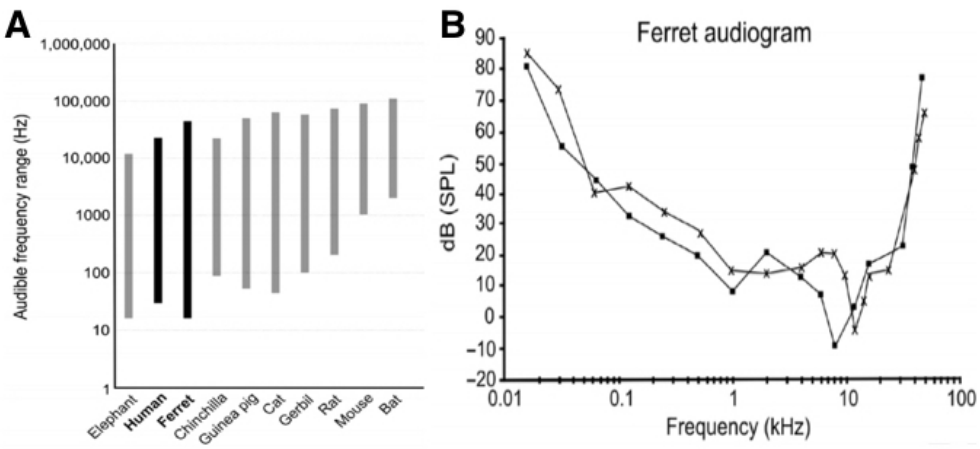


Figure 1.1: (A) The hearing range in the ferret is most comparable to that of humans, encompassing similar frequencies. (B). The ferret audiogram indicates (for two different ferrets) that they have high hearing acuity in the 20 Hz to 20 kHz frequency range (Adapted from Kelly et. al., Hearing Research, 1986 [1]).

Ferrets associate meaning to critical words in the form of cue words like danger, reward, punishment, food, mating vocalizations etc. and they can be trained on aversive and appetitive tasks to differentiate "target" words from various alternative "reference" words, by associating meaning to these words [20, 21]. Specifically, Bizley et. al. [22] showed that ferrets can be behaviorally trained using two alternative forced choice selection to discriminate between vowels of different timbre, and also to generalize this discrimination ability across various pitches. Ferrets can be trained to discriminate relative pitch movement (up vs down) in a sequence of tone-pairs [23]. Ferrets could be partially trained on pitch change detection on complex stimuli such as naturalistic speech, but the task is more difficult than a simple frequency change detection [24]. These studies in ferrets were all comprised a single auditory stream to test other auditory phenomena.

Streaming using tones has been studied behaviorally in ferrets to detect shifts in frequencies within alternating tone sequences, and regular repeating tones in otherwise varying tone clouds

[16]. This study demonstrated that some aspects of streaming in humans [15] are also observed in ferrets behaviorally. Ferrets only weakly distinguish harmonic tone complexes from in-harmonic complexes, often confusing pure tones with harmonic complexes [25]. These studies provide the literature based off which we design the experiment for two-speaker segregation to study this phenomena behaviorally, like in the above studies, but also neurophysiologically, for which current literature in ferrets is very limited.

1.4 Thesis Outline

In this thesis, we leverage the combination of behavioral experimentation in simple stimuli and neurophysiological extracellular recordings in the ferret to understand the neural underpinnings of the cocktail party with more complex stimuli like speech, in the auditory and frontal cortex. In this thesis, we discern what happens at the single unit and neuronal population level during the cocktail party scenario. While this has been studied in animals using simple stimuli such as tones, the more realistic scenario of speech has not yet been used to understand attention-driven solutions of the cocktail party. Prior studies except Saderi et. al. [17], especially in the ferret, focused on encoding in the primary auditory cortex during streaming, which we extend here to include the dPEG in the auditory cortex. Previous studies in human subjects try to dissociate what happens at the population level using a variety of techniques such as fMRI, EEG, MEG and ECoG, however access to single unit data is limited due to the invasive nature of single-unit electrophysiology. This thesis enables us to understand how the auditory and frontal regions of the brain encode speech during a cocktail party, especially at the single unit level and sets a framework for cross-species comparison of attention-driven speech stream segregation.

This thesis has been organized into six subsequent chapters. Following this introduction, we will start with a brief overview in Chapter 2 of the auditory system in mammals, focusing on our particular animal model and an in-depth literature survey of neural representations of speech streaming. We will focus on the pathway of auditory signals from the ear onward, and what role the auditory cortex has in this signal pipeline. We will then look at several hypotheses and theories of stream segregation and attention in the auditory system, reviewing their merits and demerits. In Chapter 3, we discuss the stimuli and neurophysiological recording methods that are the backbone of all data acquisition and experimental runs for the work in this thesis.

Chapter 4 presents a behavioral training paradigm for the ferrets to navigate a simulated cocktail party problem with two speakers. The two speakers are presented with equal perceptual loudness, and the ferret is taught to perform a task pertaining to one of the two speakers, while ignoring the other, interfering speaker. After training in such a setting, we demonstrate compelling evidence that the ferrets are successfully trained on this behavioral task. We also describe behavioral experiments and their results on adjacent tasks, which we have studied behaviorally for this thesis.

Once animals are trained on the task, we recorded extracellular activity in the primary and secondary regions of the ferret auditory cortex. In Chapter 5, we discuss the evidence from neurophysiological data from the ferret auditory cortex — its relevance and importance and what insights these data provide for solving the cocktail party problem, what the different cortical regions within the ACx contribute and what single unit characteristics are important to stream segregation. We demonstrate that the foreground speaker is enhanced in the primary region (A1) for a pop-out effect at the level of population responses, accompanied by suppression of the background speaker in the secondary area (dPEG). We then shift focus in Chapter 6

to the dorsal Frontal Cortices, and their role in encoding single and multi-speech streams in a stream segregation problem. We discuss literature to support our work on understanding neural correlates of frontal projections to the auditory cortex, and discuss evidence from our recordings to support this.

Finally, we summarize our findings in Chapter 7 and discuss the conclusions of this thesis, as well as future implications this thesis would have on understanding the neural mechanisms in the auditory system to support stream segregation, and more broadly, the auditory scene analysis problem.

Chapter 2: Overview of Auditory Streaming

Hearing has always been an integral part of communication, and survival for humans as well as animals. The auditory system takes in a mere pressure wave at both ears, and converts it into coherent electrical signals for us to use with language to provide meaning to the world. We are capable of using our auditory senses to discern the 'what' and 'where' of particular sounds in a variety of acoustic environments, and only realize the extent of reliance on the system once it fails to perform to our desired standards.

Traditionally, the auditory system has been thought of as a spectrum analyzer, with the ear playing the role of converting the pressure signal to a Fourier decomposition at the level of the cochlea that is further analyzed in the higher processing areas. This view however, discounts a lot of the more interesting intricacies of audio processing - what happens when those sounds occur adjacently? what happens when some sounds are more important? what happens when those sounds provide context or meaning? These are just a few of the questions within the realm of auditory processing that are obliterated in the simplistic frequency analyzer approach.

In the next section, we will outline how the auditory system functions, with specific attention to the ferret auditory system. We will visit literature understanding the behavioral encoding in the ferret auditory cortex, and how this propagates through various stages of the auditory system. In the following section, we understand the perceptual, behavioral and neural

underpinnings of auditory streaming, and specifically the cocktail party scenario.

2.1 The Auditory System

The peripheral and central auditory system, specifically the auditory cortex in the central auditory system, of the ferret has been well studied and shown to be consistent with that of various other mammalian species [19, 26]. In the ferret, the sound falls upon the cochlea through the ears. The hair cells in the cochlea convert the mechanical pressure signal into electro-chemical activity that is transmitted further. The cochlear signal, via the Auditory Nerve (AN) then passes on into the cochlear nucleus in the auditory brainstem [27]. Auditory Brainstem Responses (ABR) can be used to assess the hearing capacity of the ferret. These in turn send projections to the Superior Olivary Complex (SOC) [28] from both ears, which is responsible for integration of auditory signals from both the ears. These auditory signals are further processed in the Inferior Colliculus (IC) [29] in the midbrain, which are then projected into the Thalamus and Medial Geniculate Body (MGB) [30]. Finally, these signals are then sent to the auditory cortex located in the Ectosylvian gyrus via the thalamus and medial geniculate body.

Tonotopical organization in the ferret auditory cortex is studied using electrophysiology in the primary auditory cortex [1], Anterior AF [31], as well using optical imaging in the PEG [6], which also has some degree of frequency selectivity. These response properties were further confirmed with multi-electrode recordings in the ferret [19] throughout the MEG, AEG and PEG. The primary auditory cortex displays short latencies, and clear tonotopy from high frequency responses in the dorsal end, to lower frequency-specific responses as we move ventrally. There is a clear reversal line beyond which the PEG lies, where the tonotopy is less sharp, but still

prevalent, and latencies are longer. Further along the hierarchy, the vPR [32] shows even less clear tonotopy, and is instead modulated by attention and task-driven effects.

In the primary auditory cortex, the neurophysiological responses to tones reveal a change in specific spectrotemporal response fields (STRFs) of units during task performance [33]. These changes are rapid, reversible and are typically induced by perceptual learning of a particular task [34]. Multi-tone target sounds in the presence of background TORCs (Temporally orthogonal ripple combinations [35]) reveal enhancement for component tone frequency responses, and suppression of inter-tone frequencies [36]. These changes, presumably to enhance behavioral selectivity [37], are also seen albeit an even larger scale, in the PEG [2], and further amplified in the VPr [32], which is similar to tertiary areas in other species. Going higher up the hierarchy, the dorsolateral pre-frontal cortex encodes an even more abstract version on response changes [21] that is a stronger reflection of the behavioral meaning of sound, as a top-down attention-modulated signal, than physical attributes of the sound itself, as seen in STRFs in the A1 [21, 33].

Ferret recordings have shown evidence of multiplexing for parallel processing, which enable the ferret ACx to encode a more dynamic representation of incoming speech sounds [38]. For this work, we focus on encoding of speech in the auditory cortex, specifically the primary auditory cortex (A1) which displays short latencies to neuronal response post sound-onset, and has a sharp frequency response area, thus responding most selectively to certain frequencies, as well as the dorsal Peri-Ectosylvian Gyrus (dPEG), which is a secondary area encompassing responses to a wider set of stimuli [2]. We also record neuronal data from the Anterior Sigmoid Gyrus (ASG) and the dorsal aspect of the Orbital Gyrus (OG), together comprising the ferret frontal cortex.

2.2 Cocktail Party and Auditory Scene Analysis

The auditory system can segregate and isolate sounds from various different sources from a large range of acoustic environments remarkably well. We can separate sounds very well not only for sources that are spatially separated, but also in scenarios where the sounds originate from the same source. This problem of segregating sound sources is called the problem of 'Acoustic Scene Analysis' [4, 39], or more colloquially, the 'Cocktail Party Problem' [3]. Selection of appropriate features of sound and binding them into individual perceptual streams leads to formation of perceptual auditory objects, which typically contain sounds that are perceptually segregated and may or may not be emanating from different sources. Machine-based algorithms have emerged to attempt replication of biological systems for scene analysis, but current technology still severely under-performs compared to humans when allowed limited data for training. We will discuss the prevalent theories for formation and segregation of auditory objects in the following section.

Auditory scene analysis is supported by two phenomena - formation of streams to detect various auditory objects, and selection of relevant streams to successfully segregate sounds. Both of these have been extensively studied in the last few years. It has been demonstrated that the auditory signal entering the ears is decomposed into various features along the auditory pathway, starting with frequency decomposition at the cochlea [40] itself, gradually propagating to the ACx. Selection of features that are bound into perceptually segregated streams leads to the formation of an 'auditory object' which typically (but not always) originate from different sound sources. Humans perceive these sounds as coming from a single stream or object by binding them along typically low level features such as onset times, location of sound, prosodic cues [41], intonation and pitch differences [42]. This streaming occurs despite local changes in features

such as loudness [43], pitch [44]. Higher level processes such as linguistic cues, recognition of a known sound and attention driven mechanisms also contribute to auditory stream selection. There are numerous psychoacoustic studies showing that attention can be directed to certain features such as frequency regions [45], pitch [46] and gender [47] to coherently detect streams of sounds.

2.2.1 Theories of Pre-Attentive Auditory Streaming

Studies done in the past few years have led to numerous models and hypotheses explaining the psycho-acoustic and neural foundations of perceptual streaming. One prominent hypothesis suggests separation of streams based on well separated regions in the brain that are selective to specific attributes of the sound stimuli [14, 15, 48]. As frequency distribution along the tonotopic axis is one of the key spatial separation paradigms, this model explains tone sequence perception well but does not account for the temporal binding across tones separated in frequency. For example, tones that are an octave or more apart should be heard as two distinct streams according to this hypothesis. However, psycho-acoustic and neurophysiological studies have shown that synchronous tones more than an octave apart are heard as a single perceptual stream [12]. The hypothesis also doesn't probe the question of two streams containing common frequencies such as speech from two speakers sharing a higher order harmonic, but having other distinct features such as timbre and pitch - something omnipresent in a multi-talker scenario.

Early human imaging studies used tone pips to show an enhanced response to bottom-up attended streams. Perceptual studies using Van Noorden streams [8] showed using various techniques such as EEG [49], intracranial EEG [50], MEG [51] that there is an increased response

based on binding of tones to form auditory objects, wherein the increase is observed only when the tone pips are bound into an attended object. Gutschalk et. al. [52] showed that these enhancements are observed using MEG only when a pop-out effect of the tone is detected in an otherwise rhythmic tone pip. These effects are specifically seen in non-primary areas of the human auditory cortex such as the Intra-Parietal Sulcus. All the aforementioned studies explain the formation of auditory perceptual objects and streams through bottom-up driven processing, without the influence of attention.

2.2.2 Role of Attention in Auditory Streaming

The formation of auditory objects raises the important question of what influences their formation and segregation. One school of thought suggests that streams are formed pre-attentively based on auditory objects that originate from the same sound source [53, 54]. That is, the sound mixture is decomposed into various streams possible, followed by selective attention and enhancement of a particular attended stream [55]. In this pre-attentive streaming model, attention plays a critical role in stream selection, but not in the formation of individual streams [56]. This would, however, require a lot of overhead storage, and appears to be a sub-optimal process and hence less likely the mechanism the brain is using for complete stream segregation.

Another school of thought suggests that attention is a crucial modulator to not just select streams but also to form distinct auditory objects. This has been shown in visual science where features like color, edges were integrated into visual objects when attention was focused on the features [57]. Multi-modal studies have also demonstrated how diversion of attention through another modality can disrupt stream formation [58], supporting that attention is indeed required

in stream formation during a cocktail party. Top-down influences of directed attention have been studied in detail using tones and tone-clouds over the last decade to understand their role in auditory object formation and perceptual segregation. Using MEG, Elhilali et. al. [59] showed that there is a strong response at the target frequency during a target deviance detection task in a tone-cloud, where participants paid attention to the stream to specifically detect appearance of a deviant tone. On the other hand, the response at the target frequency was reduced when the participants paid attention to the tone-cloud instead of the target frequency tone pip. These studies were also replicated using MEG [60] to study the build-up of the stream representation, where the authors showed that stream formation occurs over around 3 seconds. While these deviants were strictly in the spectral domain, temporal deviance detection also showed a similar effect of enhanced response to the tracked sound using MEG [61] and using ECoG [62]. The theory of temporal coherence explains what computational role attention may play in auditory stream formation.

To address the question of whether attention is a crucial element of stream formation, Elhilali et. al. [12] have proposed an alternate temporal coherence based stream formation model, wherein attention plays a role in the stream formation. They propose that after spectrotemporal feature extraction of the auditory signals from the scene, there is an additional coherence analysis phase. This is based on slow-varying stimulus features, the cortical phase-locking of which determines which sounds occur as a coherent ensemble. These co-occurrences are most likely to have originated from a single source or auditory object, which can then be perceptually enhanced with selective attention. That is, when temporally coherent elements occur in a sound mixture, these elements are bound together and pop-out as a single perceptual object from the rest of the sound mixture. This has been applied computationally to tone sequences [12] as well as speech

stream mixtures [63] successfully.

At a neural level, attention can influence stream segregation in two different ways. Firstly, it can enhance neuronal responses to attended features of a particular auditory object. This has been shown in awake behaving animals, where the attended repeated noise is better segregated and elicits stronger responses as compared to the background random noise [17]. Other studies [64] have also shown that attention can rapidly modulate neural responses to adapt for task behavior. For example, in the ferret primary auditory cortex, Yin et. al. [65] have shown that neural responses to attended stream are segregated better during behavior in a stream segregation behavioral paradigm. Secondly, attention influences temporal coherence and timing of neural responses by enhancing timing co-occurrence within distinct populations of neurons, thus binding them within a single stream [66]. This is indicative of a feature-based theory of stream formation. If a feature is selectively attended to, it anchors to accumulate a perceptual stream of sounds within the auditory scene that are temporally coherent with the anchoring feature.

2.2.3 Encoding and Decoding of Attention in Speech Stream Segregation

One of the most studied forms of the realistic auditory scene analysis problem is that of the cocktail party, where two speakers talk concurrently and listeners pay attention to one or the other. Various encoding and decoding techniques have been used in recent years to dissociate different aspects of the cocktail party encoding in the brain. The target of attention can be decoded using linear decoders in various scenarios to show how temporal coherence binds streams across various attended features [67]. Linear spectrotemporal receptive fields have been trained on MEG responses [68] to predict responses to the cocktail party scenario. Reconstructing the speech

envelope from MEG responses [69] showed that these speech envelopes closely matched the speech envelopes of the attended speaker in the original stimulus. The relative loudness of the two speakers did not affect the envelope reconstruction fidelity. Using EEG, it was shown that temporal response functions showed segregation only at the the M100 level, but not before, indicating that segregation was a reflection of a perceptual process, and not a strictly auditory stimulus-driven process. Mesgarani et. al. [70] showed that high gamma activity from ECoG could be used to reconstruct not just the temporal envelope, but the whole spectrogram of the attended speaker in correct trials of a two-speaker cocktail party. All these attention effects were exclusively seen in the Planum temporale (secondary region) but not in the primary region of the ACx in humans.

Various neuroscience techniques have also been used to decode attention selectivity from a cocktail party-like setting. Magnetoencephalography (MEG) and EEG techniques yield high decoding accuracy for attention in a limited setting of two speakers through a state-space model representation[71]. EEG [72], MEG [69] and ECoG studies show that the attentional subject can be decoded from as little neural data as a single trial (60s). Mesgarani et al. [70] have shown with ECOG techniques, similar attention selectivity in a multi-speaker environment where the responses were strongly correlated with spectrotemporal features of the attended speaker in a mixture. A classifier was built to decode attended words and speaker identity through reconstruction, thus revealing the features at the perceptual level used in encoding multi-talker speech. While the focus of the solution to the cocktail party problem has been on the attended stream, these decoding-based approaches also provide a framework to examine the processing of the unattended stream. MEG and EEG data from these experiments offer new evidence suggestive of significant processing and representation of unattended speech in the cortex [73]. The precise

nature and amount of the representation of multiple signals in a selectively attended environment in the brain is still unclear, and this thesis is an attempt to understand what happens at the single cell level in the auditory cortex of animals during two speaker segregation.

Chapter 3: Methods and Stimuli

In this chapter, we outline the behavioral and neurophysiological methods used for the thesis. In addition to that, we also describe the stimuli used, however, we defer a detailed explanation of the task design and structure for future chapters. All animal experiments were approved by the University of Maryland Animal Care and Use Committee (IACUC) and National Institute for Health (NIH).

3.1 Participants

We used ferrets (Mustela Putorius, female) as an animal model, obtained from Marshall Farms, New Jersey. For the purpose of behavioral experiments, we trained 4 ferrets (Samba, Savanna, Taiga, Arabica) and for the neurophysiological experiments, we used two of these trained ferrets from the set of behavioral animals (Samba, Arabica).

3.2 Animal Surgery and Transition to head-fixed setup

Animals are initially trained using conditioned avoidance in a free-moving setup, where the animals learn to lick from a spout at the front of a Faraday cage, but otherwise free to not perform the task by being away from the spout in the cage. Negative reinforcement is provided in the form of a mild shock on the paws. Water flows continuously through the spout at a rate

between 0.3ml/min to 4 ml/min. Training occurs with intermediate steps to confirm performance at easier versions of the task before training on the final version of the task design. Once they perform consistently above chance level at the final version of the task design, where consistency is defined as Accuracy (Hit Rate) $\geq$ 75%, Consistent Licking before stimuli presentation (Safe Rate) $\geq$ 50%, and Discrimination Rate (Hit rate controlled for inconsistent licking, chance = 25%) $\geq$ 40%, we consider the animal ready for implantation. Details about these metrics can be found in the next chapter.

Implantation surgery is done to add a stainless steel post to the animal's skull for head fixation. During this surgery, the Auditory Cortex is also identified and wells made over the auditory and frontal cortices for future craniotomies. Ferrets are anesthetized using ketamine and dexmedetomidine, and then maintained under anesthesia with 1%-2% iso-flurane in oxygen throughout the surgery. The heart rate, respiration, core body temperature and oxygen levels of the ferret are monitored throughout the surgery by a trained anesthetist, and emergency drugs administered with vet approval if required. The stainless steel head-post is secured in the skull using titanium screws and embedded in Charisma (up to 2018) or Dental cement with Gentamycin (2019 onwards). The area around the auditory and frontal cortices are covered in a single layer of cement, whereas surrounding areas are built up to about 5 -7 mm thickness to cover the exposed skull.

Animals are allowed recovery for three weeks post surgery, or as needed after that, and pain reliever drugs are administered as considered necessary by the attending veterinarian doctors. Once fully recovered, the animals are gradually habituated to the head-fixed setup and water acquisition through the spout over a period of two weeks. Then, ferrets are re-trained in the head-fixed setup to regain performance to consistent levels, as determined earlier (Hit Rate $\geq$ 75%,

Safe Rate $\geq$ 50%, Discrimination Rate $\geq$ 40%) before any neurophysiological recordings are performed. Retraining is required due to modified position of negative reinforcement, since the shock pin is placed on the tail in the head-fixed setup for comfort and lower noise interference with recordings, where as reinforcement happens through the animal's paws in the free moving setup.

3.3 Neurophysiological Recordings and Data Acquisition

All neurophysiological recordings are conducted in a sound attenuating chamber. In a typical experiment, four tungsten microelectrodes (FHC) with high impedance (2–7MΩ) are used for extra-cellular recordings of action potentials of neuronal units. A small 1–2 mm craniotomy is made in the skull, exposing a part of the auditory cortex for recording. The electrodes are independently advanced through the craniotomy in the auditory cortex using an EPS motor-driver system until good spike isolation is observed on the channels.

All stimuli are played from a speaker located one metre away from the center of the ferret's nose at 65 dB, unless otherwise noted. Sound stimuli presentation, behavioral data acquisition, online analysis and record triggering were done through a custom MATLAB-based open source software called Behavioral Auditory PHYsiology *(BAPHY)*. Data acquisition was performed by AlphaLab Data Acquisition Systems (Alpha-Omega), signals recorded at a sampling frequency of 25 kHz and amplified 15000x. Single units were isolated using Principal Component Analysis (PCA), K-Means clustering and subsequent template matching using custom MATLAB-based software called *Meska_PCA* to determine isolated units of activity by removing noise as well as labelling multiple neuronal units separately if there was clearly distinguishable activity from

more than only unit at a single electrode recording site. Only units with greater than 70% isolation and a typical refractory period were saved as single units.

3.4 Localization of Recording Sites

Each recording location in each ACx hemisphere is characterized by its location along the dorso-ventral and rostro-caudal axes with a fixed angle of viewing (30 degrees) such that the electrodes penetrate the surface of the brain perpendicular to the auditory cortex. An artificial mark, which is constant across experiments, made by drilling a depression in the headcap is used for this purpose. Furthermore, tones of various intensities and frequencies (duration = 250 ms) are played and neural responses to them analysed. These measurements help determine the best frequency and latency of each individual recording site. These are compared to a standard tonotopy-latency map of the ferret auditory cortex to determine recording locations [19, 31].

Each recording in the auditory cortex is characterized as primary auditory cortex (A1) and secondary auditory cortex — also known as the Posterior-Ectosylvian Gyrus (PEG). Each frontal cortex recording is characterized by the location in co-ordinate space along the dorso-ventral and rostro-caudal axis with reference to the marks made in the cortex well. Further anatomical analysis of units in frontal cortex was not done, as anatomically, that would be beyond the scope of this thesis.

We determine the Best Frequency at an electrode location in the auditory cortex using the average of neural responses to various tones in the range of 125 - 32000 Hz, spaced apart with intensities varying from 0 to -50 dB relative to SPL in steps of -10 dB. These responses are used to construct the Frequency Response Area (FRA) per site (Figure 3.1). The Best Frequency was

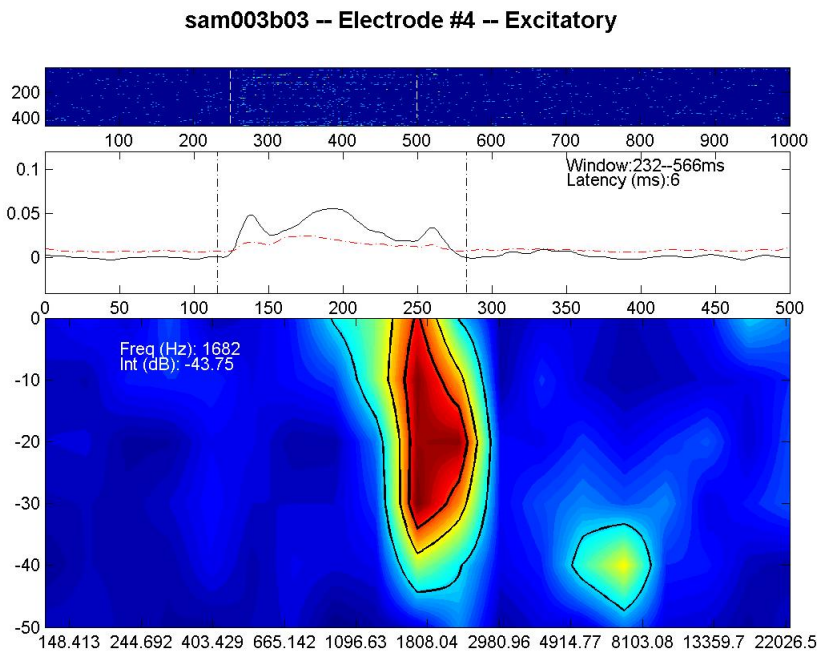


Figure 3.1: Responses to tones at different intensities help determine the Frequency Response Area (FRA) of a neuron, which can be used to infer the best frequency the neuron responses to, and the latency post which it responds. In this figure, the neuron responds with a latency of 6 ms, and best frequency of 1682 Hz.

determined as the frequency that responded the most to the lowest sound level, i.e. the lowest trough of the FRA figure per unit. The latency of a neuron is computed as the time delay for the spiking response during tone presentation to exceed a threshold of 3 times the standard deviation of the spontaneous activity of that unit. We confirm that collected recordings are from desired areas within the Auditory Cortex by comparing the tonotopy obtained through computing best frequency and neuron latency responses to pure tones and Torcs to those in systematic reviews of the Auditory Cortex in the ferret [19]. We only use these data for our analysis of responses in the Auditory Cortex.

We also record neurophysiological data from the frontal cortex. Since the frontal cortex has not demonstrated consistent responses to tones in earlier studies in the ferret, or simple auditory

stimuli nor demonstrated any evidence for tonotopy, we only note down the location of recording, but do no further anatomical analysis of these positions.

3.5 Stimuli

Speech is generated using the Straight synthesizer for speech [74] for earlier animals (Samba, Taiga, Savanna), and the inbuilt Apple synthesizer for later animals (Arabica). The fundamental frequency (F0) of the male speaker is 107 Hz, and that of the female speaker is 220 Hz for all electrophysiological experiments on Animal 1 (Samba), and is 100Hz (male) and 180 Hz (female) for animal 2 (Arabica). Each syllable is an English syllable of the form 'CV', comprising of a consonant, followed by a vowel, and lasts for 180 – 220 ms. The duration is maintained constant to control for effect of syllable duration on responses, with each syllable zero-padded to 250 ms. Words were comprised of three to four of these syllables (E.g. Fa-Be-Ku), and these words included utterances by male and female speakers.

For the purpose of our research, we used the auditory spectrograms of these syllables. Auditory spectrograms are a time-frequency representation of the sound mimicking the early stages of the auditory pathway. These stages include (i) using a constant-Q filter bank along the spectral axis as the cochlear output, (ii) a hair cell transduction stage with low pass filtering to account for loss of phase-locking on the auditory nerve and (iii) a lateral inhibitory network to mimic the mid-brain. We then filter these data for a specific scale (in cycles/ octave) and half-wave rectify them to represent an output similar to that postulated at the primary auditory cortex level. In particular we use the same set scale parameter for all our data and while our results through the thesis hold for all values of scales we used (0.5 cycles/octave - 2 cycles/octave),

all figures use a scale parameter of 0.5 cycles/ octave. More details about how the auditory spectrogram are generated and its interpretation can be found in Chi et. al. [75]. We do not use different scales for different neuronal units in this thesis, and leave that for future exploration.

For figures that indicate spectral components of either stimuli or responses, for example, in the reconstructed spectrograms in Chapter 5, we use a single vertical column comprised of the mean auditory stimulus collapsed across time to visualize the dominant frequencies present in the two speakers individually. This column figure is called the 'acoustic spectrum indicator', when used in the thesis.

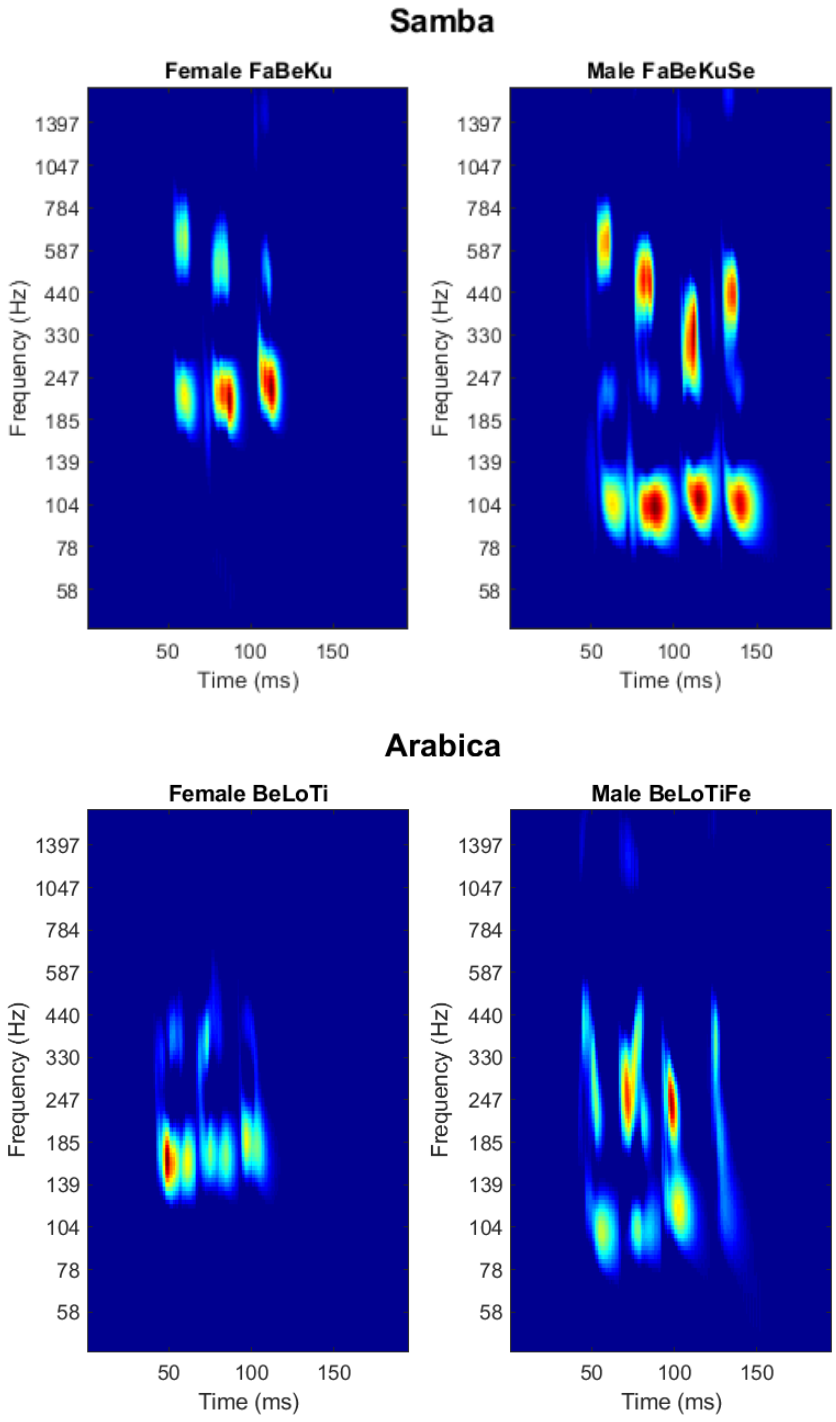


Figure 3.2: Auditory Spectrograms of the Stimuli saying target word by the female speaker (Fa-Be-Ku for Samba, Be-Lo-Ti for Arabica), and the word containing the target subword presented by the male voice (Fa-Be-Ku-Se for Samba, Be-Lo-Ti-Fe for Arabica)

Chapter 4: Auditory Two Stream Segregation: Behavior

The ferret has been used as an animal model for behavioral studies using appetitive and aversive conditioning using a variety of stimuli such as tone pips, TORCs, chords and harmonic complexes [21, 24, 36, 40, 64, 76]. The hearing range of ferrets, at 16 Hz - 44 kHz [1] is similar to that of humans (20 HZ - 20 kHz), making it possible to systematically study the processing of the same audio cross-species between humans and ferrets. Specifically, speech stimuli (we use English syllables as described in the previous chapter) can be used during a ferret behavioral task.

Here, we first explain the design of a behavioral cocktail party task that enables ferrets to learn to respond to a target word by a single speaker, by itself and in the presence of a competing speaker, taking into account the balance between task complexity and ferret capabilities. We then present the evidence through behavioral data that the animals are trained to successfully complete the task on one speaker in the presence of a competing speaker, as in a classical two-speaker stream segregation paradigm. Finally, we discuss the implications of these behavioral data, and what these mean in the context of the task.

4.1 Task Design

We use a conditioned avoidance (Go/ No-Go) task for training ferrets for the tasks in this thesis. The ferrets are water-restricted for a few hours before a behavioral session, and receive water during the session. Animal hydration is monitored throughout, and behavioral training is restricted to no more than five days a week as per the IACUC protocol. The ferret learns to continuously lick water from a spout located 2-7 mm in front of it, initially in the free moving setup and later in the head fixed setup. Water flow is varied between 0.3 ml/min to 4 ml/min during the session such that the animal stays engaged during the behavioral session and licks from the spout continuously. If water consumption during training falls short, additional water is provided after the training to the animal.

The ferrets are taught to continuously lick the spout when reference words are being played (Go Phase). However, on learning the task, in addition to licking during the reference word presentation, the ferrets learn to refrain from licking when the speaker of interest utters the target word (No-Go Phase). During initial training, a loudness difference of 30 dB is introduced for the target relative to the reference word, which is gradually reduced till 0 dB as the animal learns the task. For this work, all words are essentially pseudo-words comprising of three syllables, each lasting 250ms in duration, thus forming a 'word' of 750 ms. After the completion of the word, we provide a 400ms buffer window where the animal can make the decision to stop licking, and then execute the required movements for it. If the ferret licks during the target word presentation in the reinforcement period following the word (400 ms - 800 ms after word ends), it is presented with a mild shock on its paws (free-moving) or tail (head-fixed) to reinforce that the target word is a No-Go word.

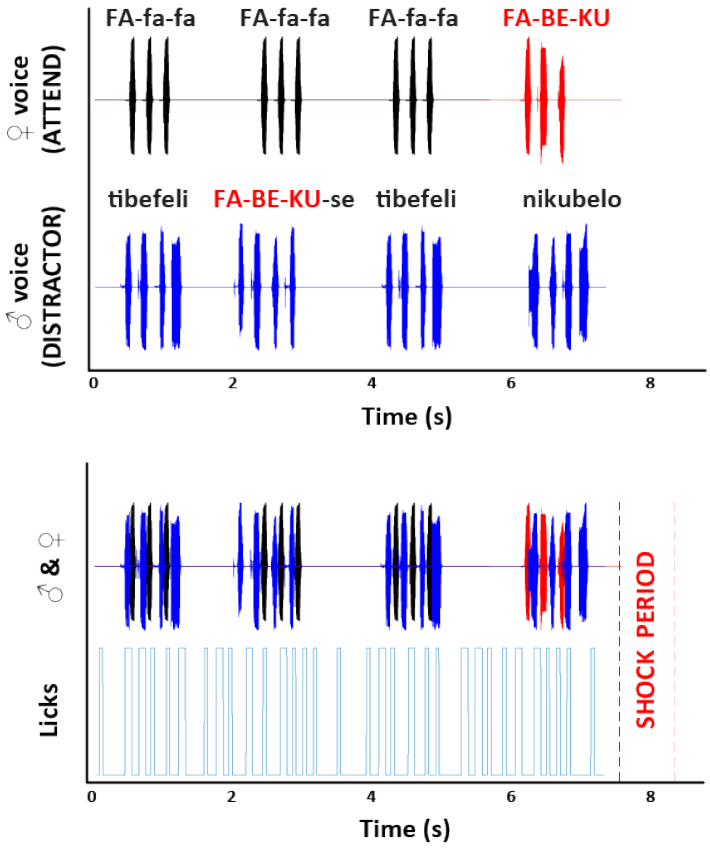


Figure 4.1: Example of cocktail party trial showing the speech signal of the female attended speaker (top panel) and male distractor (second panel), with their combined voices in the third panel. Target words are shown in red, and reference words are shown in black. (Bottom panel) Ideal licking behavior expected of the ferret for the voice mixture.

Ferret training begins on a single speaker only. Once the ferret can perform the target word detection task reliably on a single speaker, we add a second background speaker. We simulate the cocktail party paradigm by using two distinct speakers, one male and one female, for ease of distinction between speakers. The male background speaker is added at a lower loudness (-20dB relative to female), increasing in two steps of 10 dB to finally attain same loudness for both the speakers of the cocktail party. In all cases, the animal is taught to pay attention to one of two speakers. In all cases but one, the animal was trained to attend to the female speaker, and ignore the male speaker. As only one animal was successfully trained on the 'Attend Male' condition, and all neurophysiological data is from animals trained on the 'Attend Female' condition, we present the results pertaining to this animal in a separately dedicated sub-section. Outside of that, we will refer to the female speaker as the foreground speaker, and male speaker as the background speaker throughout the thesis.

During training, the animal is taught to detect the word ABC, where A, B and C refer to three distinct syllables (Eg. Fa-Be-Ku for Samba, Be-Lo-Ti for Arabica, Lo-Be-Ti for Taiga), from a series of reference words made of either AAA (Eg. Fa-Fa-Fa for Samba and Lo-Lo-Lo for Taiga), or AXY/ AQR (Be-Ka-Gu and Be-Ku-Za for Arabica). These reference words occur between 1 and 6 times before the target word presentation (Figure 4.1). The background comprises of the male speaker saying one of seven four syllabic words (DEFG) picked randomly, where D, E, F and G may or may not be the same as A, B, C. One of the seven words explicitly contains the target sequence ABC, to check control analyses of behavior to the unattended speaker (Figure 4.1, second row).

To simulate partial overlap and temporal incoherence of speakers, which is typically also seen in real situations where two people rarely speak at the exact start time, but rather with at

least a small delay, we introduce a relative delay between the background and the foreground. Each of the background words are shifted by starting them 400 ms before, 80 ms before or 200 ms after the target word chosen randomly at every word instance. The manner in which a word of the foreground and background co-occur, including their relative time onset delays are shown in Figure 4.2. The three cases of relative delay, 400 ms before, 80 ms before and 200 ms after are shown in the three rows of the figure. Each of these word combinations, with a total duration of 1950s are combined to form a trial comprising of 1-6 words with reference foregrounds, followed by a target foreground word. Each trial duration is thus between 3.8s - 13.65s. Together, they provide a perceptual illusion of listening to two continuous streams by two speakers uttering a variety of sequences with multi-syllabic words (Figure 4.1, third row). During this period, the animal is taught to lick from the spout during reference word utterances by the attended speaker (female), and refrain from licking during the 400 ms window that occurs 400 ms after the target word presentation.

This task design allows us to test that the animal is listening to the attended female speaker by performing the target word detection task on this speaker, while ignoring the background speaker who may or may not be saying the same target word.

4.2 Analysis of Behavioral Performance

To reliably determine whether the animal performed the intended two stream segregation task, we define some metrics to track the animal's behavioral progress on the task. An animal is considered to be trained if she performs above criterion on the following metrics for at least three consecutive training sessions:

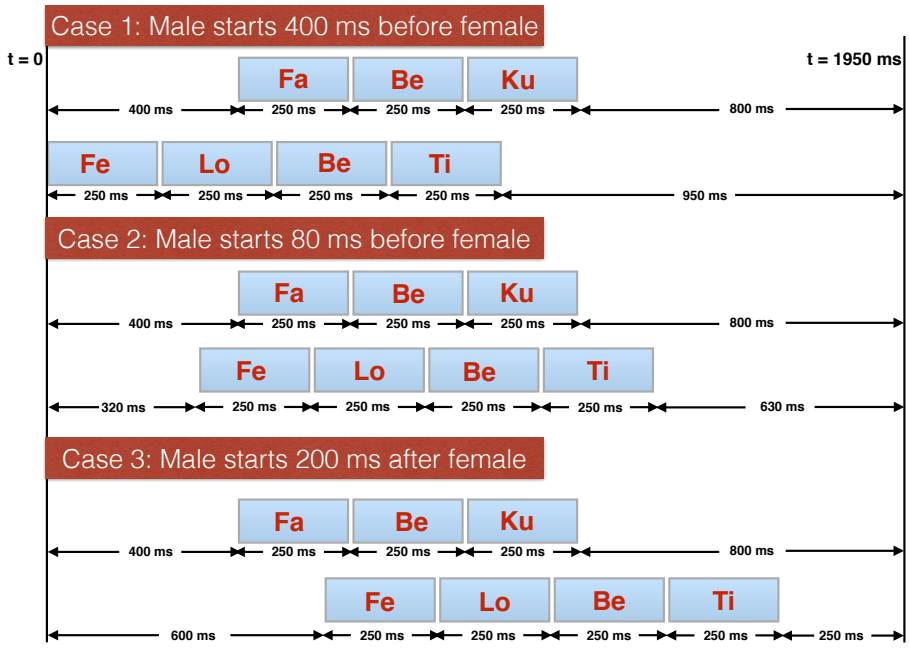


Figure 4.2: The three possible cases for delays in the male background word where the background word starts (1) 400 ms before the female word onset, (2) 80 ms before the female word onset and (3) 200 ms after the female word onset. Seven different male words are used in the stimuli, however, for the purpose of illustration, we use a placeholder word 'Fe-Lo-Be-Ti'.

1. Hit Rate (HR) = Accuracy of getting target, criterion = 75% accuracy

$$HR = \frac{\text{Number of trials with No-Go at Target}}{\text{Total number of trials with Target}} \tag{4.1}$$

2. Lick Rate or Safe Rate (SR) = Amount of time animal licked before word presentation

   In the case of trials where the animal was not licking prior to target word presentation, we label the trials as 'snooze' trials and do not count these in accurate trials. The Safe Rate is used to determine whether the animal actually stopped licking after the target presentation due to target detection, or whether the animal was not licking during the target word presentation prior to the target, in which case there is no 'detection' of target involved.

3. Discrimination rate (DR) = Accuracy modified by licking rate, criterion = 40%

$$DR = HR * SR \tag{4.2}$$

Thus, chance performance in such a scenario is determined to be that in which accuracy (HR) = 50%, and SR is also 50%, this DR = 0.5 x 0.5 = 25%. For criterion performance, we define the minimum performance DR $\geq$ 40%, and HR $\geq$ 75% for at least three consecutive training sessions. In table 4.1 below, we report the Hit Rate (HR) and Discrimination Rate (DR) means and standard deviation of all animals.

Secondly, we use the D-Prime, which indicates the signal separation, to determine how different two signals are, in terms of their z-score. Here, we use them to determine the separation

Table 4.1: Mean and Standard Deviation of Hit Rate and Discrimination rates in single and multi-speaker conditions

Animal Name	Female Alone		Mixture (Attend Female)	
	Hit Rate	Discrimination Rate	Hit Rate	Discrimination Rate
Samba	74.9 ± 13.5	42.2 ± 12	69.9 ± 19.6	38.7 ± 14
Arabica	61.1 ± 24.7	31.1 ± 11	62.8 ± 25	35.6 ± 14
Taiga	69.9 ± 22.6	42.3 ± 19.3	71.82 ± 11.69	41.86 ± 8

of licking profiles as a function of time between hits and false alarms. D-prime = 0 is chance.

$$\text{D-prime} = z(\text{False Alarm}) - z(\text{Hits}) \tag{4.3}$$

We computed D-Prime of all trials in the head-fixed setup, post surgery. In the case of each recorded animal, as reported in Table 4.2, the animal performed consistently well (D-Prime above chance) in the case of single speaker as well as multi-speaker scenarios.

4.2.1 Lick Behavior

We compare licking for the target (both, hits and misses) and reference words separately, to compare the licking profiles. In Figure 4.3, y axis denotes the rate of licking, x axis denotes time.

Table 4.2: D-Prime of Animals during Female only and Attend Female Conditions

Animal Name	Female Alone	Mixture (Attend Female)
Samba	2.33	2.15
Arabica	1.71	1.99
Taiga	1.92	1.93

We separate target word presentations into correct (hit) and incorrect (miss) trials. In all panels, the solid red line indicates target word licking (hits), dashed red indicates missed targets, black line(s) indicate reference word licking and in the case of mixture conditions on the right, the blue word indicates the licking target word embedded in the unattended speaker. For these figures, the first word in a trial is omitted during these lick analysis since a trial is triggered by the first lick, causing initial lick locking at the beginning of the trial due to spout position. Arabica, in Row 2 of the Figure, has two black traces corresponding to the two reference words used. The other two animals had one reference word each.

In the case of a single speaker presentation, the black line(s) indicating licking rate to the reference word shows a near horizontal trajectory. Contrasted against that, the target word (hits) licking profile, shown here in solid red, indicates that the animal stops licking during the third syllable of the word presentation. However, when the animal misses the target word, the profile of which is shown in dashed red, there is indication of refraining from licking in the case of two animals (Samba, Taiga in rows 1 and 3), which was restarted earlier than the end of the No-Go period.

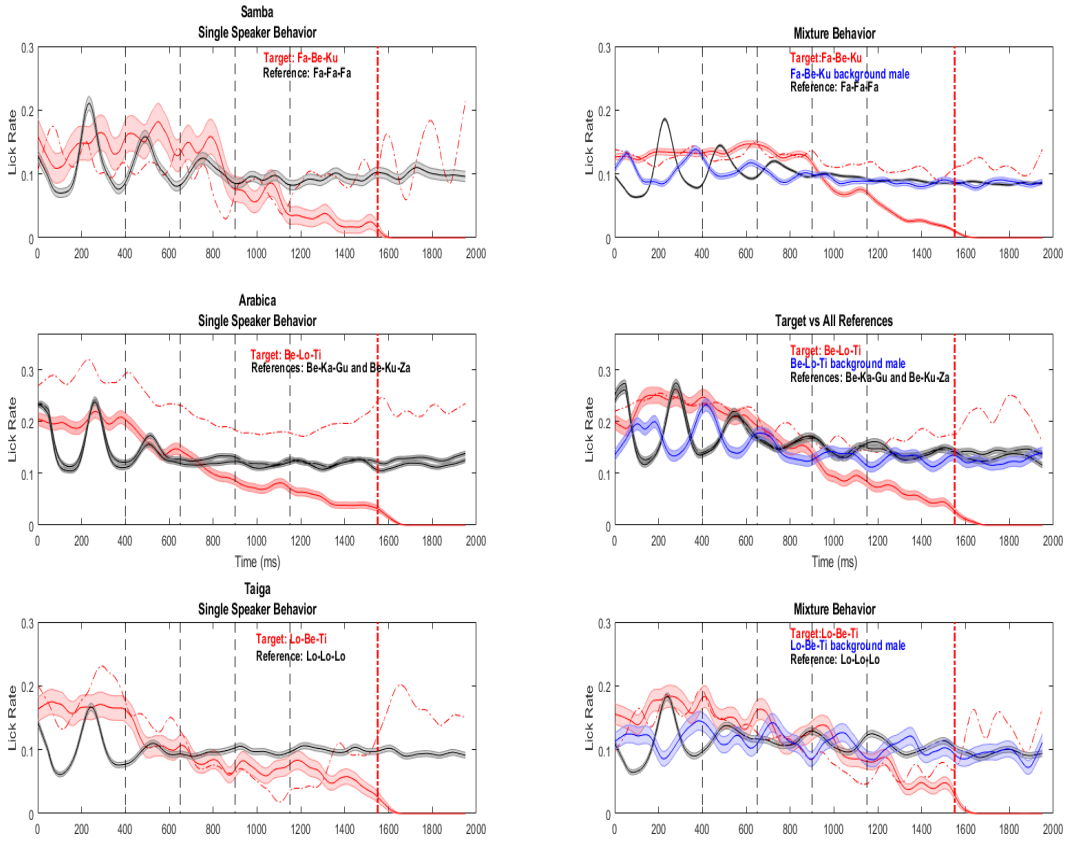


Figure 4.3: The three rows indicate three animals used in the behavioral aspect of this study - (top to bottom) Samba, Arabica, Taiga. Each panel shows the rate of licking on the y axis, and time on the x axis. In all figures, target licking is shown in red, reference(s) in black, male control word during the mixture in blue and dashed lines for missed targets. Each licking profile also shows a shaded region indicating one standard deviation. For each of the three animals, on the left is the licking behavior for the single speaker, female alone condition and on the right is the mixture condition where the animal attends to the female target word.

During two speaker stream presentation, shown in 4.3 in the right panels, we note similar patterns of licking at a near-constant rate to the reference word(s) in black, and pause in licking during the target word (in bold red), towards the end of the word presentation (in red, solid). The target word presentation in which the animal missed the No-Go period, shown in dashed red here, indicates a profile similar to the reference word presentation in black, suggesting that the animal missed the occurrence of the target word. Incorrect, missed trials occur about 30% of the time during training. (For the more complex mixture condition, for one animal, Hit Rate Avg = 69.98, Std Deviation = 19.61 during training. All animals reported in detail in the previous section).

As a control sequence, we also compare the licking of the animal when the background speaker utters the foreground target word. Shown in blue here, we see that the animal licks through this presentation similar to licking during a reference word, demonstrating no indication of attention to the target word by the background speaker.

Two tailed t-test between licking profiles are used to determine significant difference in licking between the target and reference word presentations to determine points of divergence of licking profiles, based on significance (** $p \leq 0.05$). This is done using the ttest2 function in MATLAB. The results from computing all these metrics, and what they mean in the context of two speaker stream segregation are presented below.

We determine the time of divergence as the the point beyond which the word presentation maintains significant difference between the reference and target, as determined by a two-tailed ttest with $p \leq 0.05$. In case multiple references were used for an animal, all such references were concatenated as there were no behavioral differences between the two conditions.

Arabica, Single Speaker - 782 ms

Arabica, Multi-Speaker - 893 ms

Samba, Single Speaker - 1226 ms (With secondary divergence at 794 ms)

Samba, Multi-Speaker - 955 ms

We discuss what these divergence points mean in the Discussion session ahead.

4.2.2 Behavior against Words in Trial

In this behavioral paradigm, as mentioned in the section above, we used trials containing between 1 and 6 reference words. For trained animals, the performance was comparable across all these conditions as seen below in fig 4.4. i.e. segregation was built up even by the end of one reference word to provide enough information to the animal to perform the task successfully.

Additionally, we used three different delays for the background word: -400 ms, -80 ms and + 200 ms, as compared to the onset of the foreground. These delays induced different kinds of challenges to the task such as, in the case of late start, there is sound stimuli playing during the reinforcement period as well, where as sound has stopped before the reinforcement period in the other two cases. One question we seek to ask behaviorally is whether the animal would obtain performance benefits if the stream to be ignored started before the foreground stream, or after the foreground stream. To answer this, we compared the hit rates across the target word presentation with the three separate delays in the background. we noticed comparable performance across the three delay conditions, indicating no clear evidence of perceptual benefit to unattended stream starting before or after the foreground word onset.

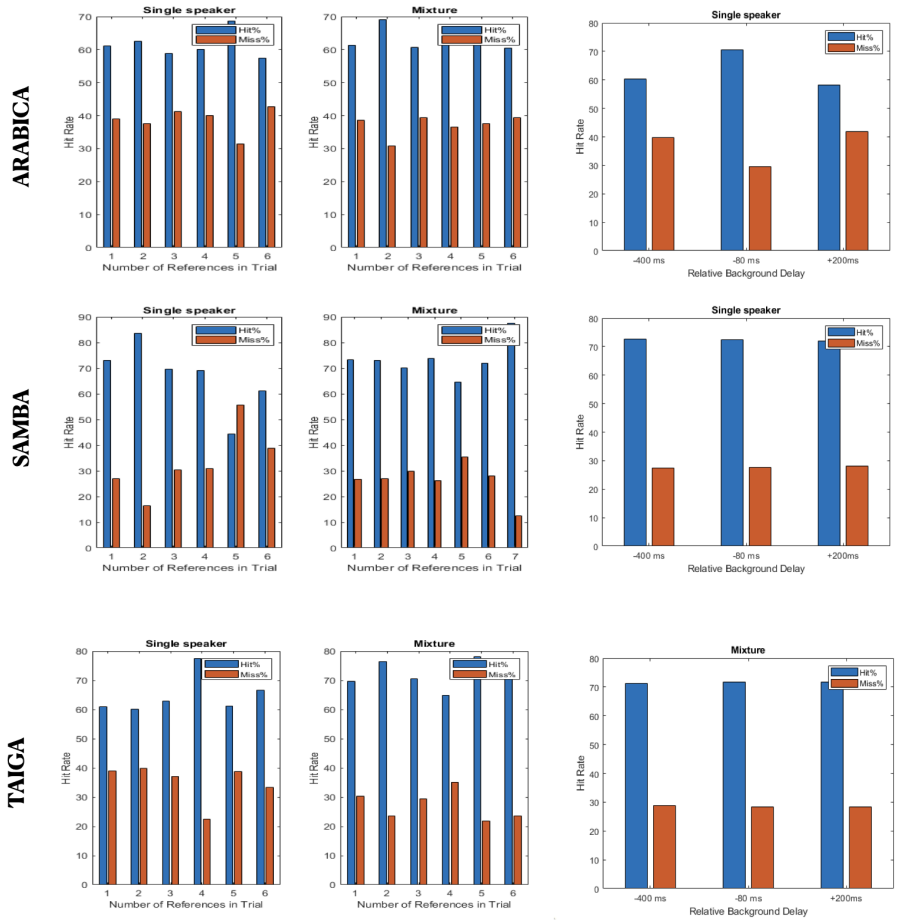


Figure 4.4: Behavioral performance as a function of number of words in trial, and relative background onset delay

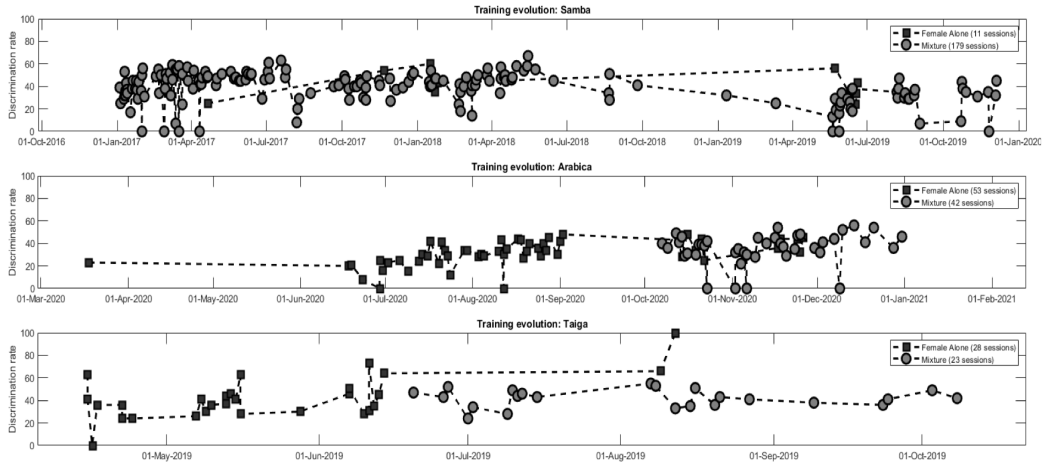


Figure 4.5: Evolution of learning through Discrimination Rate for animals attending to female speaker in a mixture of two speakers. The the x axis is time (date) and y axis shows the discrimination rate. Single speaker conditions are indicated by a square, and multi-speaker conditions by a filled circle.

4.3 Evolution of Training

While detailed studies about neurophysiological correlates of evolution of learning during a stream segregation task are beyond the scope of this thesis and may require chronically implanted electrodes, it is interesting to note the trajectory of evolution of learning, shown below with the Discrimination Rate as an indicator of learning.

The above figure shows the relearning trajectory for each of three animals after the head-post implantation surgery. The x axis indicates the date of training, while the y axis shows the Discrimination Rate. Both single speaker (square) and mixture (circle) conditions are shown in the evolution plots in Fig 4.5. While the animals are trained prior to surgery in a free-moving setup, the learning gap during surgery and subsequent recovery, as well as the slightly modified

apparatus for head-fixed training with a tail shock does require some training sessions before the animal starts performing consistently well on the task.

4.4 Attention Switch: What happens if the animal learns to attend to both speakers?

For one animal, we attempted to change the focus of attention. Once the animal was successfully performing the attend female in the mixture task, the focus of attention was switched to attend male in a mixture. The goal was two fold - (1) to gauge if the animal could learn to attend to more than one speaker, with training and (2) to determine if switching attention to both speakers was possible in the realm of a day for neuronal recordings.

We note that the animal was able to successfully switch attention from attend female to attend male, as evidenced by the Hit rate and Discrimination rate in each of these time periods. However, this switch was not a one-day switch as initially hoped, and instead happened over the course of 8-10 training sessions. This provides evidence of the capability of ferrets to perform a switch, however, more intricate experimental design may be needed before experiments are designed to obtain switched-speaker attention data from the same neuronal units within sessions on the same day.

4.5 Discussion

Ling Ma et al. [16] showed initial evidence of streaming by ferrets, however, the stimuli were limited to tones. By designing a task such that the animal would have to stream a single speaker amongst two equally loud speakers with speech streams to successfully perform the task,

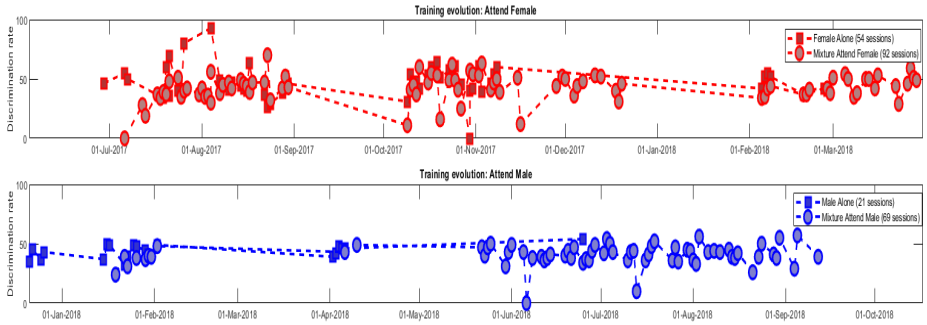


Figure 4.6: Evolution of learning through Discrimination Rate for animals attending to female speaker in a mixture of two speakers. Attend female is shown in the top in red and Attend male in the bottom panel in blue. Squares represent single speaker condition, which were used in the first training session of a switch, and circles denote two-speaker scenarios.

we are able to demonstrate that the ferret is capable of such a complex behavioral study. This paves the way for future studies to study streaming in further detail in awake, behaving animals using the ferret as an animal model. Comparing cross species, the behavioral aspect of this thesis helps establish that animal models smaller than monkeys and humans may be used to study the phenomena of acoustic scene analysis, especially at the single unit level. There are limitations on task design due to medium of instruction to ferrets being restricted to rewards or punishments, however.

Ferrets can perform this two-speaker stream segregation task very reliably, with consistent performance well above chance. This behavior, even in animals trained prior to surgery, steadily improves till it reaches the animal's plateaued capabilities. The behavioral licking profiles, as seen in Figure 4.3, and the times of divergences provide insight into how these tasks fared for the animals. Mistakes seemed to be of two kinds – for Samba and Arabica (single speaker) and Taiga (single speaker and mixture), licking dipped during the target but recovered right after. This indicates that some mistakes were mistakes of restarting licking, rather than that of detection itself. Mistakes during the mixture case, for Samba and Arabica, however show no

indication of detection whatsoever and instead closely tracked the reference word. By obtaining the exact points of divergences in these profiles, we note that the animals respond through licking differently during (in case of female alone, Arabica) the occurrence of second syllable (650 - 900 ms, differing syllable), or after the second syllable (third syllable: 900 - 1150 ms). In cases of both animals, the lick divergence point was earlier for the single speaker condition (except the late lick in case of Samba) than the multi-speaker condition, possibly due to additional cognitive load involved in listening to two speakers instead of one. While the one animal trained on switching attention could reliably perform it, it took several training sessions to achieve criterion performance. These observations provide insight into what may or may not be feasible during future designs of behavioral experiments to study streaming when using the ferret as an animal model.

In the future, additional behavioral studies may be conducted using various pitches for the two speakers, keeping the gender fixed, as well as ascertain robustness of this behavioral task by changes to the background speaker within a trial. Due to the limited trial capacity of ferrets during a single behavioral session due to thirst limitations, additional experimental design would be required to compare across these various conditions.

Table 4.3: Mean and Standard Deviation of Hit Rate and Discrimination rates in single and multi-speaker conditions across four periods of switched attention

Period 1: Attend Female 47 sessions of the single speaker condition, and 68 sessions of the mixture condition		
Hit rate	70 ± 19.3	62.7 ± 14.8
Discrimination Rate	47.9 ± 14.4	42.3 ± 7.5
Period 2: Attend Male 15 sessions of the single speaker condition, and 8 sessions of the mixture condition		
Hit rate	65.8 ± 9.3	60.3 ± 6.9
Discrimination Rate	42.4 ± 5.8	37 ± 7
Period 3: Attend Female 7 sessions of the single speaker condition, and 24 sessions of the mixture condition		
Hit rate	77.5 ± 12.7	66.3 ± 11.8
Discrimination Rate	45.79 ± 7	43.7 ± 7.5
Period 4: Attend Male 6 sessions of the single speaker condition, and 61 sessions of the mixture condition		
Hit rate	69 ± 5.9	69.4 ± 11.5
Discrimination Rate	45.8 ± 5.1	41 ± 9.3

Chapter 5: Auditory Two Stream Segregation: Encoding in Auditory Cortex

Stream segregation has been studied extensively in humans using simple streams of tones, as well as more complex speech stimuli. In animal models, studies have focused predominantly on complex streaming in passive, sometimes anesthetized settings exclusively. With advancements to record extracellular activity in awake, task-engaged animals most studies have been limited to tones, harmonic tone complexes, or white noise based repetition paradigms to understand streaming. By using human speech to study the stream segregation problem in behaving animals, we can obtain high temporal and spatial resolution as well as enhanced control of coverage of the Auditory Cortex to study the neural correlates during attention-driven speech streaming.

In the ferret, we are able to collect extra-cellular neurophysiological activity from an ensemble of single units in the Auditory and Frontal Cortex. These data are binned into either 1 ms bins for responses to tones and TORCs, or 10ms bins for responses to speech sounds, providing a high degree of temporal resolution. Since we use single micro tungsten electrodes for these recordings, they also provide a finer spatial resolution, limited only by recording capacity and sampling of the surface and depth of the auditory cortex. We use these high resolution data to understand how behavioral gating affects the encoding of stream segregation in a mixture of two speakers. For electrophysiological recordings from animal models, we access the auditory cortex

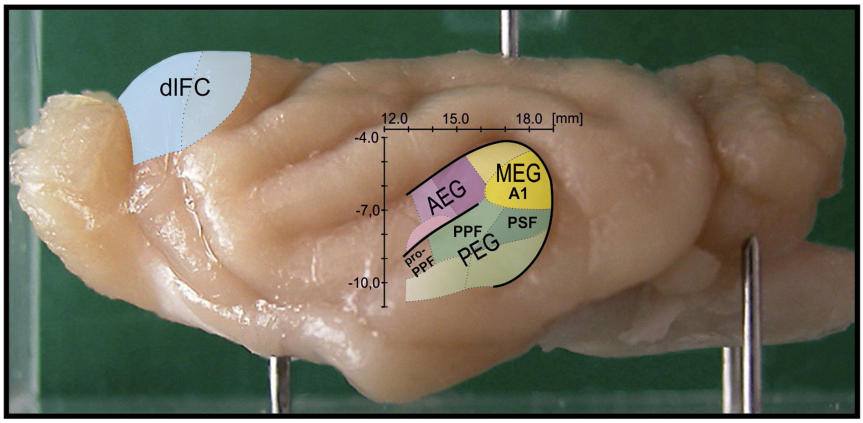


Figure 5.1: Schematic of the ferret brain. Highlighted areas indicate the auditory cortex (in yellow, pink and green) and dorso-lateral frontal cortex (in blue). Adapted from Atiani et. al., 2011. [2]

through a small 1-2 mm cranial opening made in the skull prior to electrode placement. This prevents us from obtaining a direct location of electrode placement, however, we use previous laboratory expertise recording in the ACx as well as established tonotopical features of the ferret auditory fields [19], tonotopic organization fields or selectivity shown in previous studies in A1 [40], AAF [31] and A/PEG [6, 19] to validate our position of electrodes.

5.1 Neurophysiological Responses: Tonotopy

The two main regions we record neurophysiological data from the ferret auditory cortex are the primary auditory cortex, known as A1 (shown in yellow, in Figure 5.1) and the dorsal aspect of the Posterior Ectosylvian Gyrus (dPEG, shown in green, in Figure 5.1) located just anterior of the sylvian fissure. More details about the ferret auditory system are found in Chapter 2, and the neurophysiological recording methods in Chapter 3.

The tonal response characteristics of each single unit are obtained as outlined in Chapter 3. The responses to tones of varying frequencies and intensities are used to determine the best

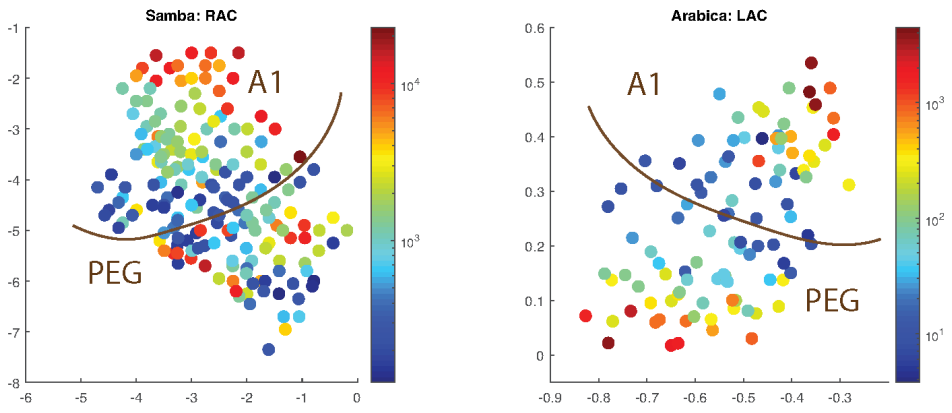


Figure 5.2: The Best Frequencies for each recording site are obtained, and their tonotopies compared to standard tonotopies in the Auditory Cortex to determine area of recording. Each mark on the maps indicates a recording site, colored according to its best frequency. Red indicates high frequency selectivity, and blue indicates low frequency selectivity of the recording units at that location. Samba's Right ACx tonotopy is shown on the left here and Arabica's tonotopy of the Left ACx shown on the right.

frequency of a unit. Using these best Frequency (BF) data, along with recording coordinates, BF maps obtained are compared to those studied in extensive detail previously [19], and the primary and secondary regions marked, as can be seen in Figure 5.2 below. In this figure, each point on the scatter plot corresponds to one experimental recording electrode in coordinate space. In case there were multiple recordings at the same penetration location, the latest site was considered for computation of the best frequency. Higher frequency selective units are marked in red, and lowest frequencies in blue, on a blue-to-red scale. The A1/ PEG reversal line is determined by where the low frequencies start tending high again as we move from top to bottom (moving ventrally).

5.2 Peri-Stimulus Time Histogram Responses to Stream Segregation

In the temporal domain, we bin the neural responses into 10 ms, to extract spikes. Further, we average these windowed spike data across trials and stimulus presentation to obtain peri-

stimulus time histograms (PSTHs) of the various stimuli that are presented to the animal during an experimental run.

When individual syllables are presented to the animal, we obtain different PSTHs averaged across neurons, one for each syllable in the stimulus presentation. Some representative average PSTHs are shown in the schematic for stimulus reconstruction in Figure 5.6. About 94.5% of AC neurons evoked responses to syllables that were at-least 2 standard deviations away from the baseline neuronal firing of the unit. Dividing between areas and between animals – For animal 1 (Samba): 126/130 of A1 and 92/101 of PEG neurons showed neural response to syllables at least two standard deviations more than baseline activity and for animal 2 (Arabica): 78/86 of A1 and 65/65 of PEG neurons showed a similar significant response to syllables.

In the case of PSTHs of the individual speakers in both animals in Figure 5.3, the baseline firing of the units has been removed from the response. The green dashed lines indicate syllabic boundaries and the blue bar at the bottom indicates the whole word duration. The PSTHs themselves are depicted with bold red lines, with shaded regions encompassing one standard deviation above and below the mean PSTHs. We note that there are clear, syllabic responses to the female target word (Fa-Be-Ku for Samba, Be-Lo-Ti for Arabica), and male control word containing the target utterance (Fa-Be-Ku-Se for Samba, Be-Lo-Ti-Ka for Arabica), in the ACx of both animals. Such clear syllabic responses are also observed for other words presented during the experiment. This is in agreement with studies showing that Auditory Cortex in ferrets neurally responds well to English syllabic sounds [18]. It has been shown that these responses to syllables in ferret primary cortex are robust enough that responses can lead to good decoding of the individual syllables[18].

We next compare the PSTH responses in the ACx in the passive condition to those during

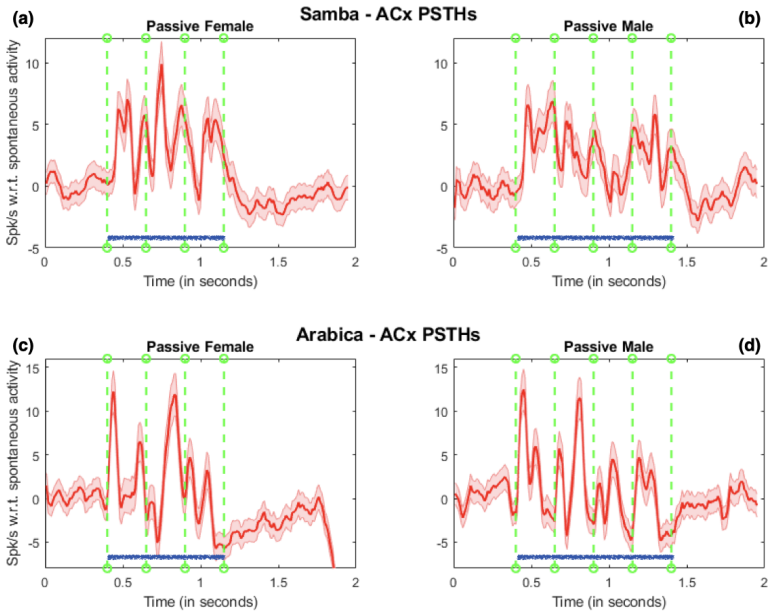


Figure 5.3: PSTH Responses in the Auditory Cortex of Samba to (a) Female Fa-Be-Ku (b) Male Fa-Be-Ku-Se and in the Auditory Cortex of Arabica to (c) Female Be-Lo-Ti and (d) Male Be-Lo-Ti-Ka during passive word presentation. The blue bar on the bottom indicates the sound 'on' segment of the word presentation, where syllable boundaries are indicated by the green dashed lines.

task-engagement. We look at the PSTHs for the mixture data from two different lenses - one where the mixture data are aligned to the female target words, and the other where we consolidate information from the three delays of the background male word, and present data for the presentation of mixture, aligned to the male control word i.e. the female foreground has three relative delays. This was done to check for any averaging effects that may occur in one condition over another. While PSTHs do lose their syllabic peaks slightly when aligned to the male word (right side) during behavior, the two conditions do not reveal any clear, gross differences in terms of responses indicating segregation of the two temporally overlapping streams. This prompted us to look into analyses that convert these temporal responses to a feature-space better suited to study temporally overlapping streams. We describe this method of stimulus reconstruction, and

consequent analyses in the following section.

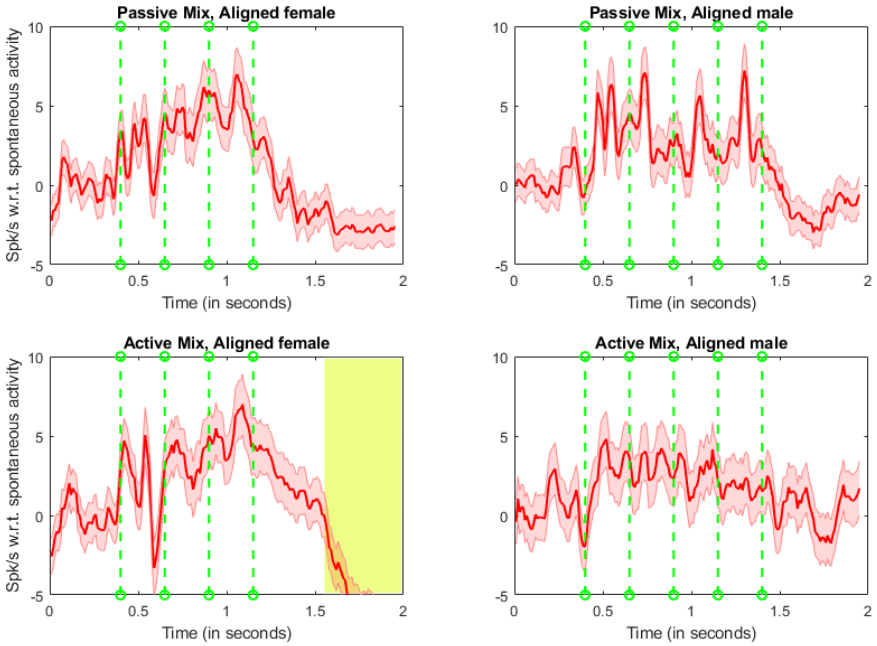


Figure 5.4: PSTH Responses in Samba ACx to Mixture Passive (top) and Active (bottom) aligned to Female on the left, and Mixture Passive (top) and Active (bottom) aligned to the Male speaker on the right. Shaded region shows one standard deviation from the mean PSTH. The green vertical lines indicate syllable boundaries, and the yellow window during the active condition is the reinforcement period, responses during which are disregarded due to noise contamination

5.3 Frequency Domain Analysis: Decoding via Stimulus Reconstruction

The response PSTHs we discussed in the previous section were limited to the temporal domain of analysis. Here, we introduce the frequency domain of the stimuli since human speech in general spans across multiple frequency bands in parallel, and hence can possibly capture the difference in the two speakers by the difference in the response modulation of the frequency bands they comprise of. The various pitches and details of the male and female speakers used for this task have been outlined in Chapter 3. These speakers have relatively stationary pitches

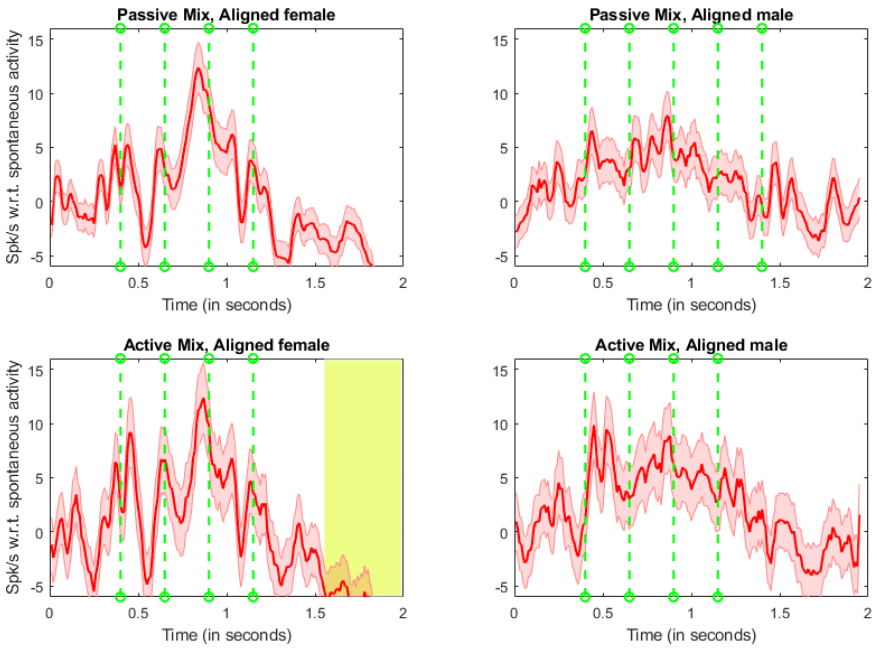


Figure 5.5: PSTH Responses in Arabica ACx to Mixture Passive (top) and Active (bottom) aligned to Female on the left, and Mixture Passive (top) and Active (bottom) aligned to the Male speaker on the right. Shaded region shows one standard deviation from the mean PSTH. The green vertical lines indicate syllable boundaries, and the yellow window during the active condition is the reinforcement period, responses during which are disregarded due to noise contamination

due to the short length of the utterance and for one animal, the synthetic nature of the speaker. During every neurophysiological recording, we obtain responses to individual syllables that are present throughout the stimuli set, words uttered by the male unattended speaker (in passive), and attended female single-speaker and female/ male multi-speaker conditions in passive, then during behavior with active task-engagement, followed again by a passive set of recordings. In the case of single speaker and mixture conditions with an active and passive component, the trial structure follows as outlined in the previous chapter. Here, passive refers to the animal not being engaged in a task, whereas during behavior, the animal is actively engaged in the task of detecting the target word (also referred to as the active condition).

To study the effect of frequency bands in the stimuli, we use the method of stimulus reconstruction [77, 78]. A schematic representation of the linear filter-based reconstruction is shown below in Figure 5.6. The idea behind this method is that we obtain individual syllable PSTHs for each neuronal unit, and along with the auditory spectrograms of the syllables, we trained a linear filter using syllabic and word responses alone presented to the animal in the passive condition, by minimizing the squared error between the training auditory spectrograms and neural responses. Analytically, this can be obtained by reverse correlating syllabic responses with the stimuli auditory spectrogram. This linear filter is a representation of the linear mapping between stimuli and their neural responses. We then use this trained filter to estimate the stimuli from the different conditions such as active engagement and passive listening using the neural responses to the stimuli in the two-speaker mixture presentations. In this fashion, the reconstructed stimuli represent the encoded neural responses rather than the presented stimuli itself, thus reflecting the decoded stimuli from task-specific encoding in the auditory cortex.

We represent the original sound stimulus as its auditory spectrogram representation S(t, f), where t = 1, .. , T and f = 1, .. , 128 channels, logarithmically spaced from 50 Hz to 4000 Hz. The auditory spectrogram is a time frequency representation that models the early auditory processing stages of the auditory pathway, as outlined in [75], as well as in Chapter 3. For a population of N neurons, we represent the response to stimuli of neuron n ($1 \leq n \leq N$) at time t = 1 .. T as R(t, n). These are the same as the PSTHs in the previous section. These PSTHs have been processed to remove neuronal baseline firing activity. The inverse function g(t, f, n) is a function mapping the responses R(t,n) to stimulus S(t, f) as follows:

$$\hat{S}(t, f) = g(f, t, n) * R(t, n) \tag{5.1}$$

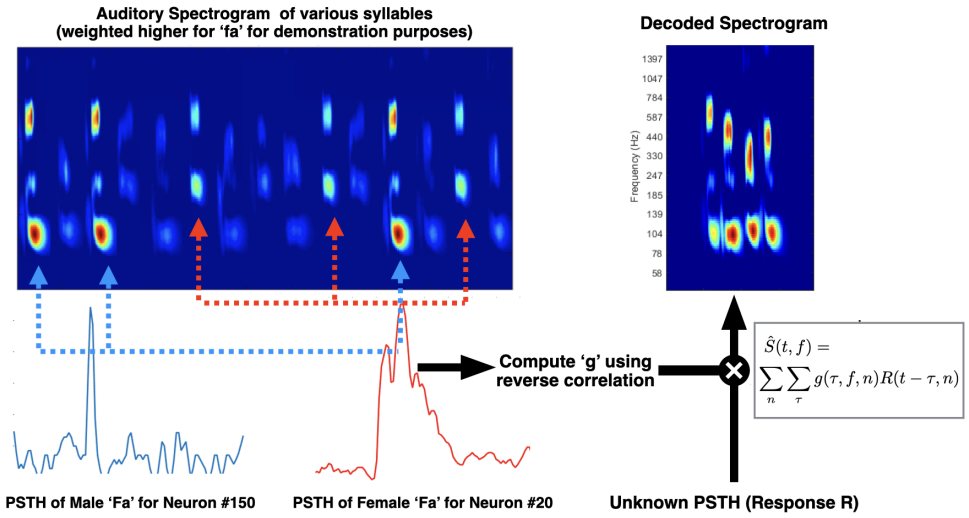


Figure 5.6: Stimulus Reconstruction - Syllabic Responses are collected per neuron, along with their auditory spectrogram. 'g' filters are obtained through reverse correlation, to finally get decoded spectrograms from unknown PSTHs

$$\hat{S}(t,f) = \sum_n \sum_\tau g(\tau,f,n)R(t-\tau,n) \tag{5.2}$$

Since the reconstruction at a particular frequency f, is independent of other frequency channels, we use a single channel of frequency f for our further analyses. Let g_f be the channel inversion filter. g_f is estimated by minimizing the mean-squared error between the actual and reconstructed spectrogram for that particular frequency channel.

$$\min e_f = \sum_t [S_f(t) - \hat{S}_f(t)]^2 \tag{5.3}$$

Solving this analytically yields a close form solution using normalized inverse correlation, given by

$$g_f = C_{RR}^{-1} C_{RS_f} \tag{5.4}$$

where C_{RR} and C_{RS_f} are the auto-correlation of neural responses and cross-correlation between

neural responses and stimulus at different lags. We experimented with different sets of causal and non-causal lags during which no difference in the results for our task was observed. The data presented in the thesis use lags from 0 to 200 ms.

$$C_{RR} = RR^T C_{RS_f} = RS_f^T \tag{5.5}$$

and R (assuming causality) and S_f are defined as:

$$\begin{bmatrix} r_1(0) & r_1(0) & ... & r_1(0) & ... & r_1(0) \\ \vdots & \vdots & & \vdots & & \vdots \\ 0 & 0 & ... & r_1(0) & ... & r_1(T-\tau_{max}) \\ \vdots & \vdots & & \vdots & & \vdots \\ r_n(0) & r_1(0) & ... & r_1(0) & ... & r_1(0) \\ \vdots & \vdots & & \vdots & & \vdots \\ 0 & 0 & ... & r_1(0) & ... & r_n(T-\tau_{max}) \end{bmatrix} \tag{5.6}$$

and

$$S_f = [S(0,f)..S(T,f)] \tag{5.7}$$

The matrix C_{RR} is full rank because of the stochastic and causal nature of the neural responses, and hence easily invertible. For this project, time lags up to τ_{max} = 200 ms were used in steps of 10 ms. The reconstruction filter for the various independent frequency channels were composed together to obtain

$$G = \{g_1, g_2 \dots g_F\} \tag{5.8}$$

Since the spectrograms of the two speakers differed significantly, we present both results separately throughout the thesis.

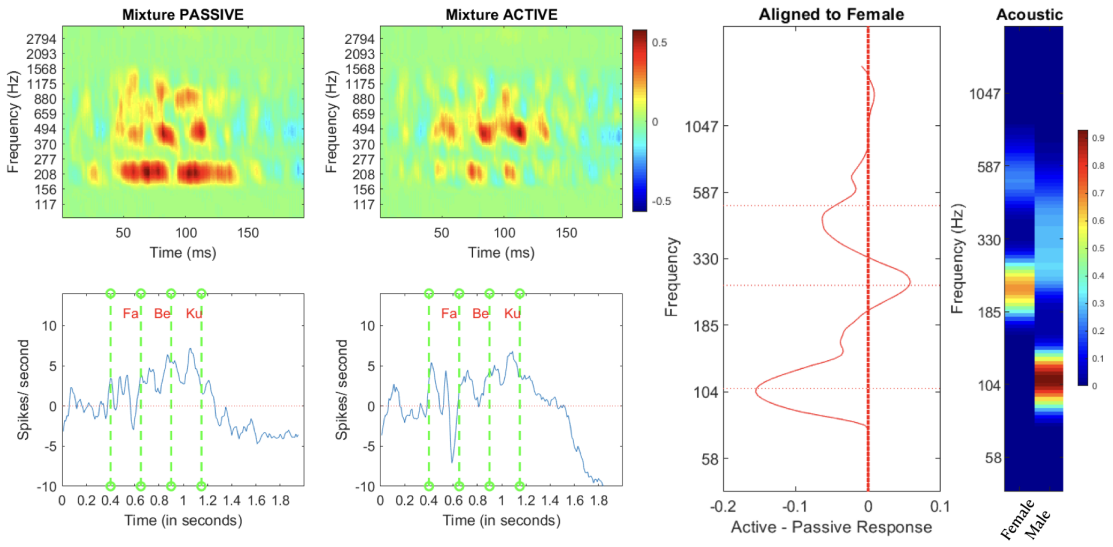


Figure 5.7: Samba: All Auditory Cortex. (Top) Reconstruction of the passive on the left, and active condition on the right. (Bottom) PSTHs corresponding to the reconstructions. (Right) Difference in the reconstructed spectrograms of active and passive target word presentations, along with an acoustic indicator showing the averaged spectrum of the two speakers. Positive values indicate enhancement, and negative values indicate relative suppression.

In the whole Auditory Cortex of both animals, the reconstructed auditory spectrogram of the mixture target word in the active condition correlates more with the attended female speakers' spectrogram than that of the male, unattended speaker. For Samba, correlation of the reconstructed spectrogram of the target word in the active condition with the female speaker is 0.60 (95% confidence interval: 0.59-0.60), while with the male speaker is 0.23 (95% confidence interval: 0.22 - 0.24). For the second animal, the correlation during active engagement with female spectrogram is 0.29 (95% confidence interval: 0.28 - 0.3) and with the male spectrogram is 0.12 (95% confidence interval: 0.11 - 0.13).

Within the Auditory Cortex as a whole, as seen in Figures 5.7 and 5.8, there is a overall

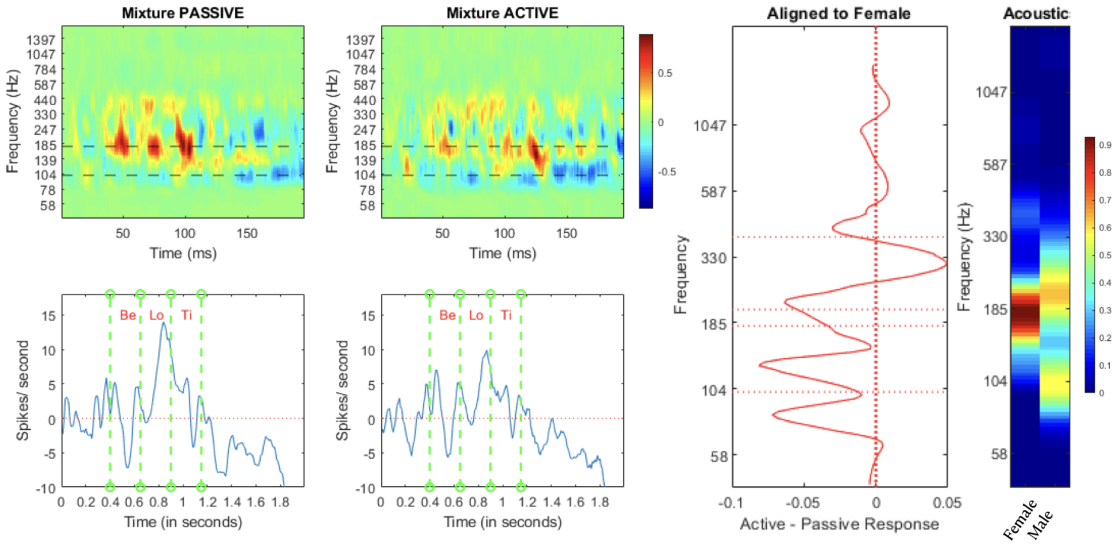


Figure 5.8: Arabica: All Auditory Cortex. (Top) Reconstruction of the passive on the left, and active condition on the right. (Bottom) PSTHs corresponding to the reconstructions. (Right) Difference in the reconstructed spectrograms of active and passive target word presentations, along with an acoustic indicator showing the averaged spectrum of the two speakers. Positive values indicate enhancement, and negative values indicate relative suppression.

suppression during the behaving, active condition as compared to the passive condition of the target word. In these figures, the reconstruction of the target word spectrogram (Fa-Be-Ku for Samba in Figure 5.7 and Be-Lo-Ti for Arabica in Figure 5.8 are shown on the top with Passive on the left, and active condition on the right side. Right below them are their respective PSTHs. The acoustic indicator is used to the right of these reconstructions, as well as to the right of the average difference at each frequency between the active and passive reconstruction, shown to the right in each figure. These figures show a clear suppression aligning with the male fundamental frequency (right side of acoustic indicator, for reference). Both animals, to varying degrees, also show enhancement of higher frequency regions which have a predominantly female but also partially male presence. This is seen for the target word, as in Figures 5.7 and 5.8 and for the reference word, as seen in Figure 5.9. At the population level, in ACx, we thus show that our

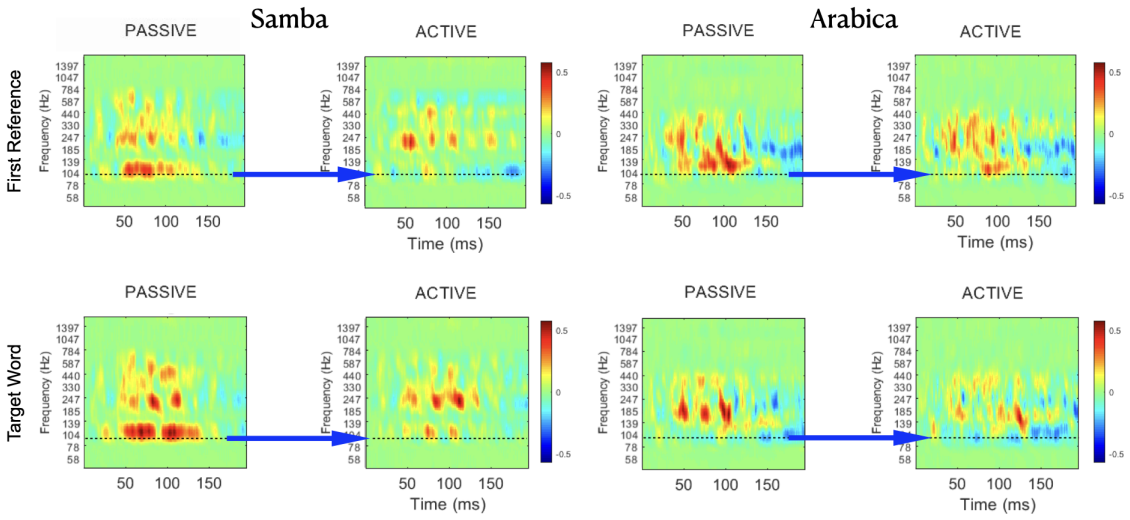


Figure 5.9: Reconstruction of the reference and target words from data in all Auditory Cortex. (Top) Reconstruction of the reference word and (Bottom) Reconstruction of the target word. From left to right, Samba: passive condition, Samba: Active Condition, Arabica: passive condition and Arabica: active condition. Blue arrows indicate position of the male fundamental frequency. Positive, red values indicate enhancement, and negative, blue values indicate relative suppression.

neural analysis using data in the ferret auditory cortex agrees broadly with human literature that attended target enhancement and suppression of unattended speakers occurs neurally during a stream segregation problem.

5.3.1 Area-wise Attention-modulation to Stream Segregation

An important aspect of the primary and secondary auditory cortices is the tonotopy that spans the frequency range. In this section, we specifically look at how individual regions within the ACx respond in a two speaker stream segregation task, starting with the well-defined and separated primary (A1) and secondary (dPEG) regions.

In the PSTHs of the responses to the target word Fa-Be-Ku (Samba) or Be-Lo-Ti (Arabica)

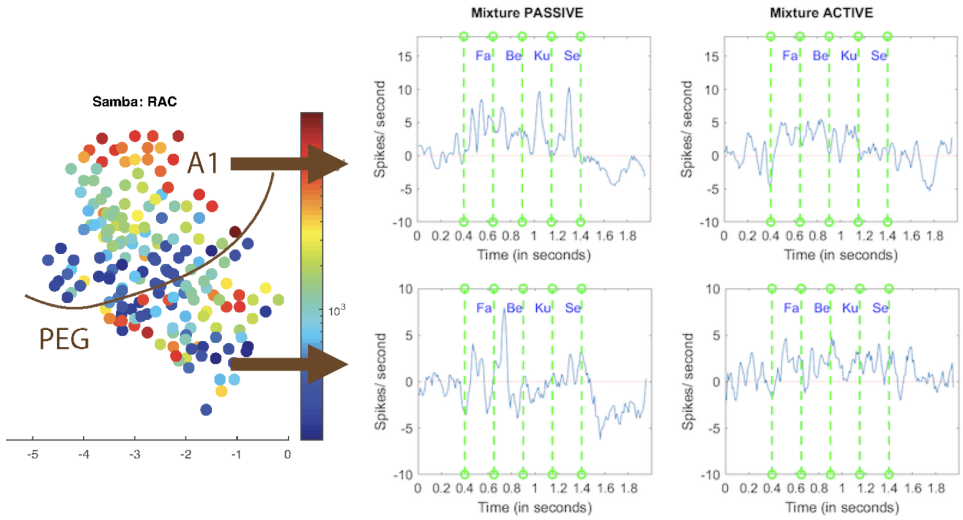


Figure 5.10: (Left) Best Frequency map of Samba in the Right ACx, showing the approximate boundary between A1 and PEG. (Right) PSTH Responses in Samba in A1 (top row) and PEG (bottom row) to (a) Female Alone, Mixture Passive (left) and Active (right)

in Figures 5.10 and 5.11 we notice that overall, response amplitudes to words in the PEG (second row) had lower peak amplitudes than those in A1 (top row). When compared to the passive responses (on the left), there is very slight suppression of the response amplitude in both the A1 and the PEG during the active task-engaged condition (on the right). This is seen in both animals. These differences during behavior however were not significant, or provided meaningful insight into how the data were represented in the two areas. Hence we employ the stimulus reconstruction method, detailed in the previous section, to understand what differential encoding occurs in these two areas.

In Figures 5.12 and 5.13, on the left, we show the reconstructions of A1 (top row) and PEG (bottom row). In both the areas, the passive mixture reconstruction is shown on the left, and the reconstruction of decoded stimuli during task-engagement in the active condition is shown on the right. Regions around the fundamental frequency of both speakers are highlighted using dashed

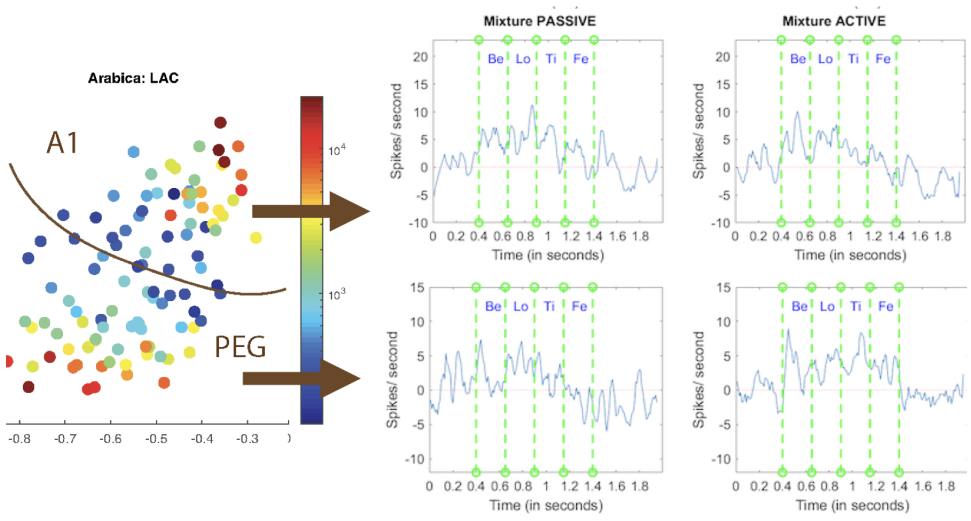


Figure 5.11: (Left) Best Frequency map of Arabica in the Left ACx, showing the approximate boundary between A1 and PEG. (Right) PSTH Responses in Arabica in A1 (top row) and PEG (bottom row) to (a) Female Alone, Mixture Passive (left) and Active (right)

horizontal lines. An acoustic spectrum indicator is shown on the right of these figures to indicate a summary of spectrograms of the two individual speakers. Details on how these indicators were obtained can be seen in Chapter 3. On the right side, we present the difference in reconstructions of the active and passive conditions, collapsed across time. The vertical line indicates a 'zero' difference, but should be interpreted in a relative sense due to the overall response suppression that occurs in the active condition, especially during A1.

In the A1 of both animals, the process of segregation has clearly commenced, as seen in the difference of active and passive spectral reconstructions. There is slight suppression at the fundamental frequency of the unattended speaker in A1, however this suppression is significantly more pronounced in the PEG. In addition to that, in the PEG, we note a clear enhancement of the attended female speaker frequencies - this enhancement of the attended speaker agrees with human literature that also demonstrates that in the secondary auditory cortices of humans, namely,

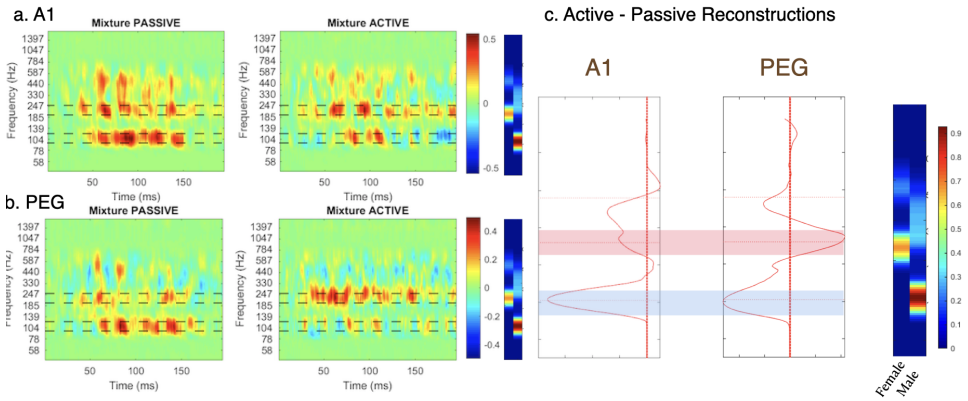


Figure 5.12: Reconstruction of responses for Samba (Left) in A1 (top row) and PEG (bottom row). On the left is mixture passive condition, and on the right, mixture during task engagement. Dashed lines correspond to region around the fundamental frequencies of the individual speakers, shown with a vertical spectrum indicator panel. (Right) The difference in responses of active and passive conditions in both the areas with the red rectangle highlighting the female fundamental frequency region, and blue indicating the male fundamental frequency region

the Superior Temporal Gyrus (STG), there is enhancement of the attended speaker that may lead to better task performance.

These results further prompt the question of what populations of neurons are most critical to the differential enhancement and suppression of the attended and unattended speakers respectively. To further test these questions, we lassoed portions of the Auditory Cortex, and obtained all data contained within the lassoed region. These regions were made approximately parallel to the tonotopic gradation in best frequencies. In Figure 5.14 - Figure 5.16, we demonstrate on the left the areas lassoed for each animal on their Best-Frequency maps where high frequencies are shown in red and low frequency tuning in blue, and on the right the reconstructed spectrograms in the passive (left) and active (right) condition. Similar to previous figures, we provide an acoustic spectrum indicator to the right of each row to compare the individual speakers' spectral characteristics to the reconstructed spectrograms.

In Figure 5.14, we divide the area into higher frequency regions of A1 and PEG, and

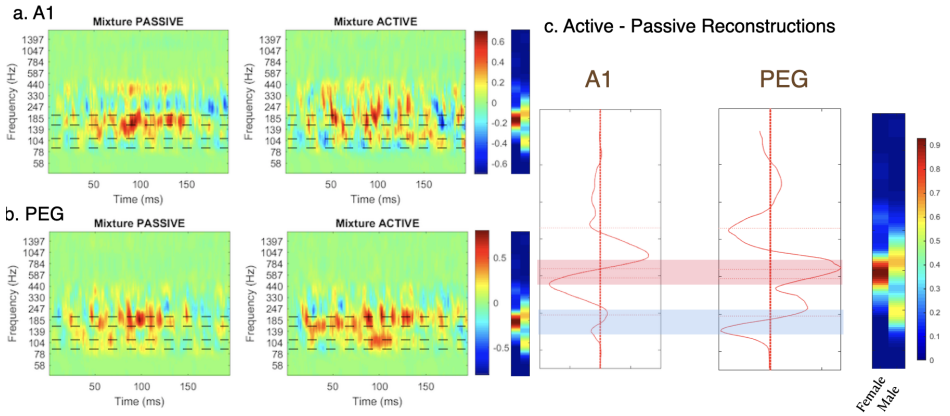


Figure 5.13: Reconstruction of responses for Arabica (Left) in A1 (top row) and PEG (bottom row). On the left is mixture passive condition, and on the right, mixture during task engagement. Dashed lines correspond to region around the fundamental frequencies of the individual speakers, shown with a vertical spectrum indicator panel. (Right) The difference in responses of active and passive conditions in both the areas with the red rectangle highlighting the female fundamental frequency region, and blue indicating the male fundamental frequency region

separate these two regions with a third, low frequency region. The physical border between the A1 and PEG is approximate without accompanying anatomical studies, and this lasso provides a way to consider the more ambiguous region as a separate entity. Another, more stronger motivation lies in the nature of human speech. Fundamental frequencies of all speakers in all cases do not exceed 300 Hz, leading to the lowest frequency tuned region being of utmost importance. True to our hypothesis, the middle region represented more faithfully both the speakers, even during the active task-engaged condition, as compared to the higher frequency PEG region.

Within A1 in Figure 5.15 and PEG in Figure 5.16 in both animals, the highest frequency regions, especially of the PEG (middle and lower panels) show the clearest suppression of the unattended speaker, while the region with the lowest best frequencies show better enhancement of the attended speaker. These area-wise analysis provide insight into what further considerations would have to be involved in encoding of segregation at the micro-population level.

5.3.2 Speaker Selectivity

In this section, we develop a speaker selectivity index (SSI) in a fashion similar to that in O'Sullivan et. al. [79], and attempt to understand what selective characteristics of a single unit contribute to the population level effects in two-speaker segregation. This speaker selectivity index is defined per neuron, and is based on the responses of each neuron to the male and female speakers, when presented passively in isolation. We obtain the variance in the response of the male and female alone word presentation without removing spontaneous activity, considering only the first three syllables to ensure similar comparisons. We obtain the difference in these variances, normalized by the sum, so as to obtain a value between -1 and 1 for each unit such that a higher positive value indicates male speaker selectivity, and a higher absolute negative value indicates female speaker selectivity.

$$SSI = \frac{var(\text{male PSTH response}) - var(\text{female PSTH response})}{var(\text{male PSTH response}) + var(\text{female PSTH response})} \tag{5.9}$$

Some representative neuronal units with their responses to the female and male speaker in isolation, and their SSI are shown in Figure 5.17. Some cells displayed visual selectivity to male or female speakers, by virtue of their PSTH response, like in rows 2 and 3 of the figure. There were however units that responded equally well to both speakers, thus having a small absolute SSI value. Note here that both speakers do have overlap in the frequency domain, as well as temporal domain, and hence, a majority of cells (251/382) indicate a smaller variance difference effect for speaker selectivity.

Our SSI computation indicated that 43/130 in A1 and 31/101 in PEG (Samba) and 25/86 in

A1 and 30/65 in PEG (Arabica) showed a absolute selectivity ≥ 0.2. These single units, however, were spread across the ACx and did not reveal any correlation with the tonotopy or position of these units in the Auditory Cortex. In these data, we do not see two significantly different SSI distributions in the A1 against the PEG (p=0.18 for Samba, p=0.26 for Arabica by running a ttest2 between A1 and PEG SS distributions).

While human ACx has shown a higher spread of selectivity in Heschl's Gyrus (primary), compared to the STG [79], we see comparable histogram spreads of SSI in both A1 and PEG. This held when SSI was computed as defined above, as well as in SSI computations that exactly followed O'Sullivan et. al. [79].

When units have low SSI (i.e. SSI $\leq$ -0.2) it indicated stronger female selectivity for that unit. The population of units with negative SSI, as seen in Figure 5.21 and 5.23, indicate that these units get the signal to enhance the response as a whole. This enhancement reflects as increase in the frequency bands of not only the attended female speaker, but also the unattended male speaker.

The neuronal population with positive SSI (SSI $\geq$ 0.2) which are more selective to the male speaker, as seen in Figure 5.20 and Figure 5.22 more faithfully represent the male speaker in the passive mixture reconstructions except in the case of PEG in Arabica (correlation of male with male-aligned passive mixture are 0.63 (A1,Samba), 0.65 (A1,Arabica), 0.66 (PEG, Samba) and 0.17 (PEG, Arabica) as compared to the correlation of the female alone spectrogram with female-aligned passive mixtures, which are 0.38 (A1, Samba), 0.68 (A1, Arabica, comparable), 0.37 (PEG, Samba) and 0.58 (PEG, Arabica).

As a population, the responses in this group have a reduced response amplitude, presenting as overall reduction of response in the spectrogram, as well as suppression of the auditory

reconstructed spectrogram - even at the attended, female speaker. While our current dataset could only provide very preliminary insight into how speaker selectivity could affect responses due to overlapping spectrograms with shared frequencies, future experiments with individual, synthetic syllables or tone complexes may be designed to further assess the contribution of cell selectivity to segregation mechanisms.

5.4 Discussion and Conclusions

Segregation is already happening in the ferret in the primary auditory cortex. While most studies in human imaging and ECoG [70] show segregation at the level of secondary and higher order cortices, we show representation of segregation in A1, in agreement with other animal studies [17]. The PSTHs do not reveal any significant gross differences during behavior, however, the method of stimulus reconstruction provides insight into what happens at the various frequency components of the speakers in the mixture. There is clear enhancement of features exclusive to the attended speaker, and suppression of features exclusive to the unattended speaker. This effect is clearer in the PEG, in agreement with various studies showing that the PEG responses show more attention-modulated effects during behavior across a variety of tasks. The active mixture reconstructed spectrograms are consistently more correlated with the attended than unattended speaker.

Different frequency regions through the primary and secondary auditory cortices reveal the nature of processing occurring across different regions. The low frequency regions (regions that have best frequencies approximately $\leq$ 1000 Hz) most faithfully represent both speakers. The highest frequency regions in the A1 remain relatively unmodulated, while the PEG neurons

lend their broader tuning to more suppression of unattended channels. We also tried to compare speaker selectivity studies cross-species by designing an SSI metric similar to a human ECoG cocktail party study [79] - while in humans there is a clearer preference for speaker in the Heschl's Gyrus than the STG, we see a more equal spread of selectivity between the primary and secondary auditory cortex. We hypothesize that cells that are more selective to the unattended speaker also contribute more to the overall suppression of the population's responses, while the more attended-speaker selective populations do not. Further studies using a combination of tone complexes may be more suitable to ascertain the exact nature of the effect of selectivity on stream segregation.

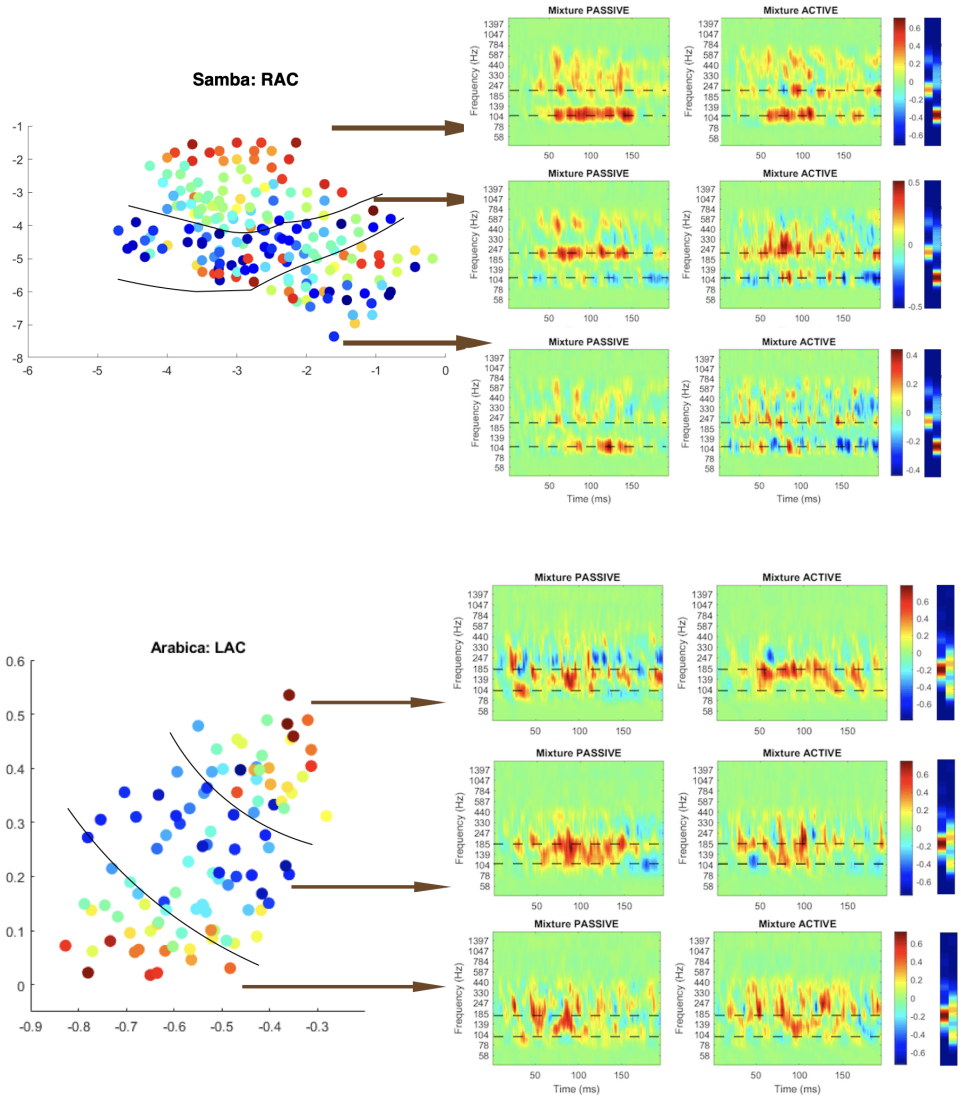


Figure 5.14: Reconstruction of responses in the Auditory Cortex within specific lassoed regions. (Left) Lassoed regions on the best frequency maps of Samba (top) and Arabica (bottom). (Right) Passive and Active Reconstructed spectrograms, with dashed lines indicating the fundamental frequencies of both speakers corresponding to the vertical acoustic spectrum indicator on the right of each panel

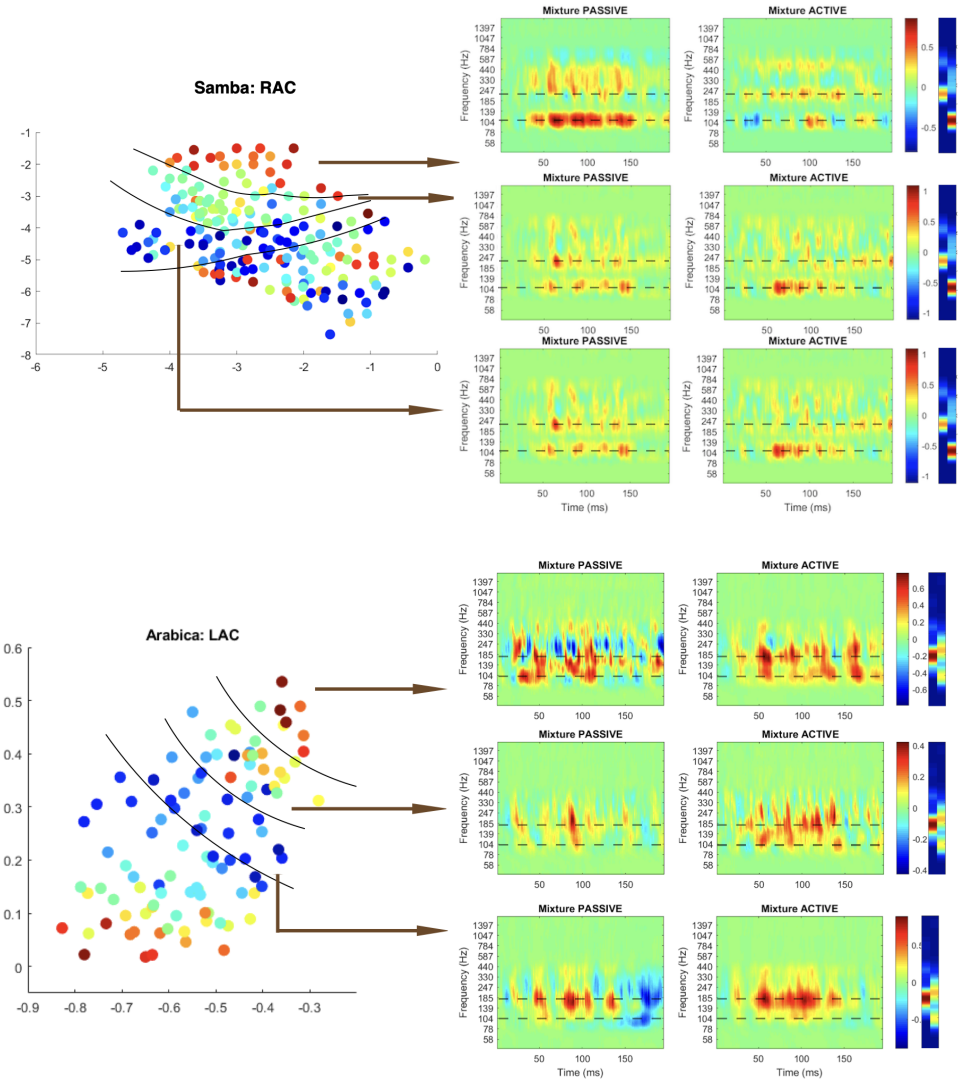


Figure 5.15: Reconstruction of responses in the Primary Auditory Cortex within specific lassoed regions. (Left) Lassoed regions on the best frequency maps of Samba (top) and Arabica (bottom). (Right) Passive and Active Reconstructed spectrograms, with dashed lines indicating the fundamental frequencies of both speakers corresponding to the vertical acoustic spectrum indicator on the right of each panel

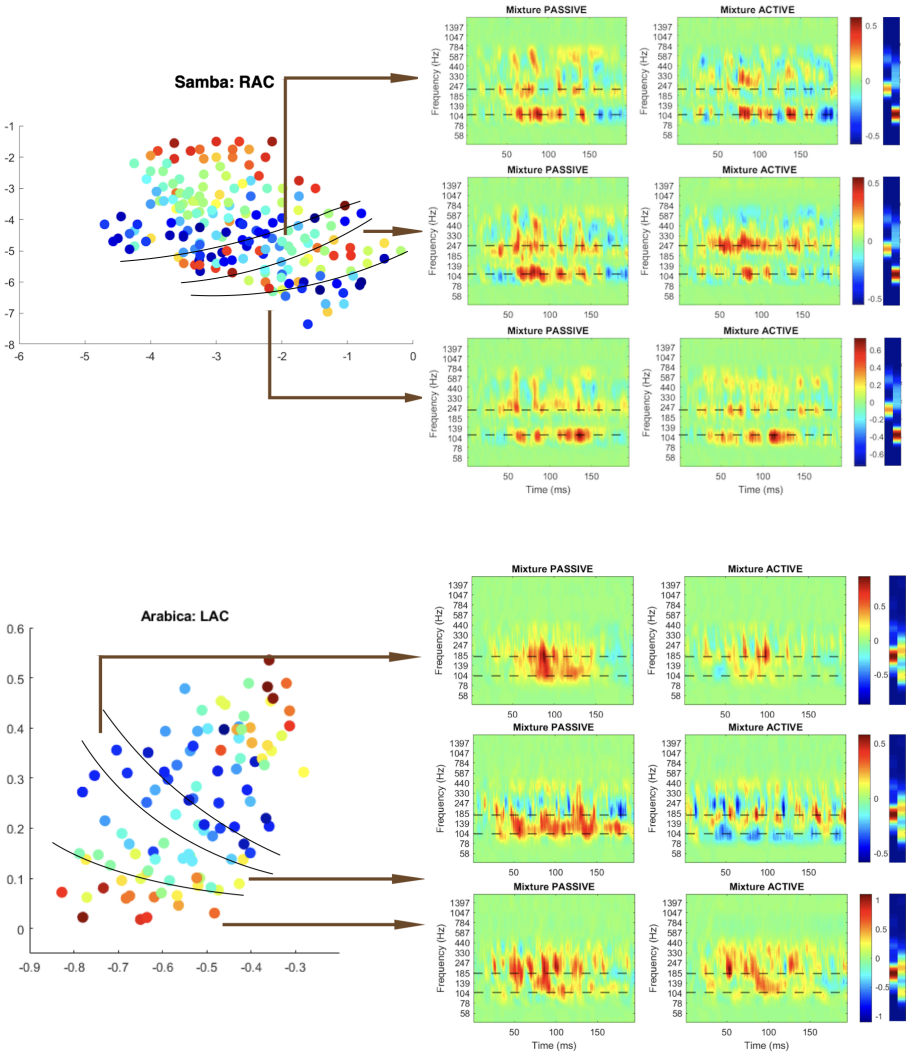


Figure 5.16: Reconstruction of responses in the Secondary Auditory Cortex within specific lassoed regions. (Left) Lassoed regions on the best frequency maps of Samba (top) and Arabica (bottom). (Right) Passive and Active Reconstructed spectrograms, with dashed lines indicating the fundamental frequencies of both speakers corresponding to the vertical acoustic spectrum indicator on the right of each panel

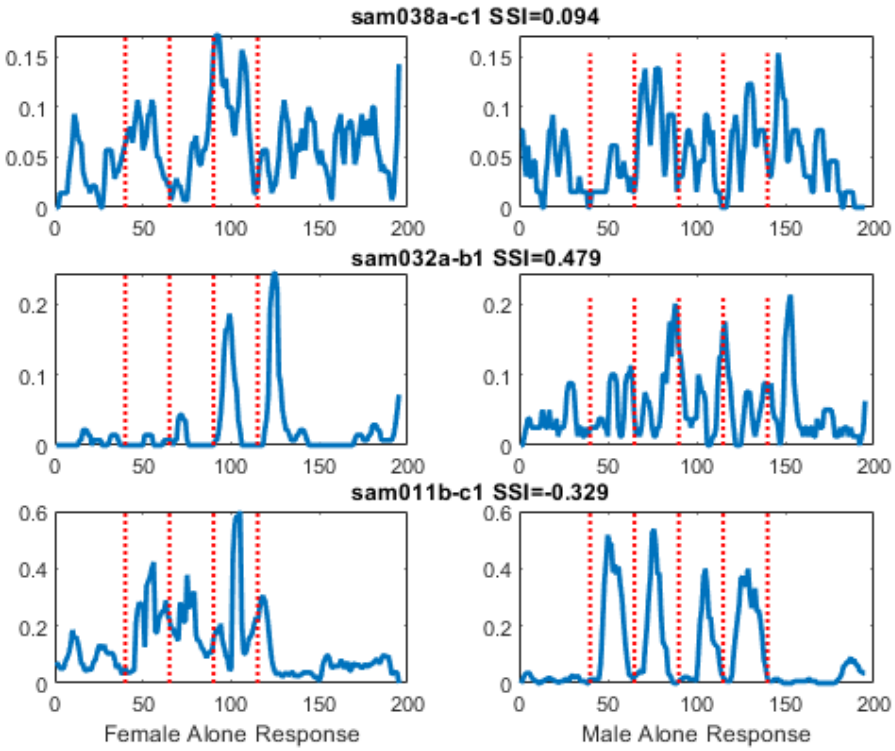
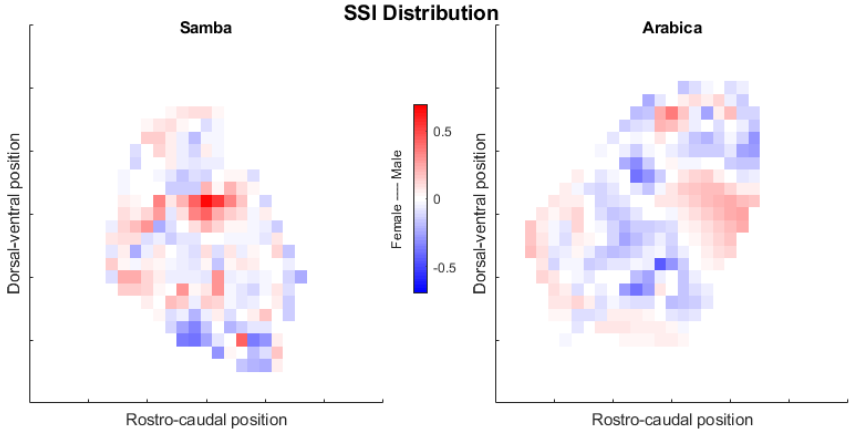


Figure 5.17: Three representative neuronal units with their female PSTH (left), male PSTH (right) when present without an interfering speaker, and SSI indicated above the unit.



a.

Figure 5.18: SSI Distribution in Samba (left) and Arabica (right) indicate a mixed distribution. Positive SSI indicates a preference for the male speaker, and a negative SSI indicates preference for a female speaker

a.

b.

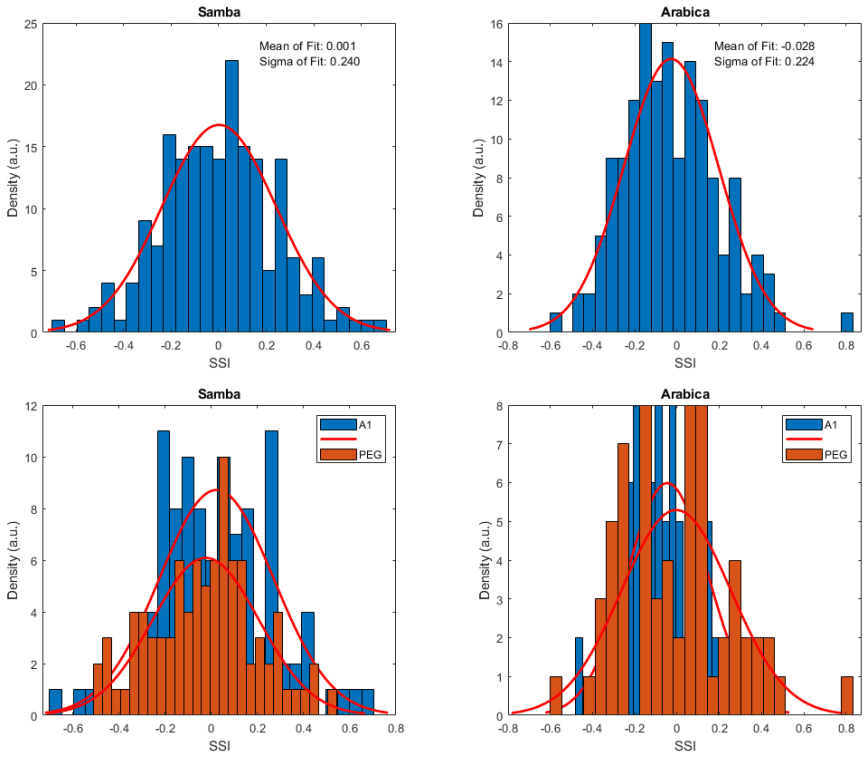


Figure 5.19: SSI Distribution in Samba (left) and Arabica (right) indicate a normal distribution centered around SSI = 0, that is, no specific selectivity. This lack of significant difference in distribution is also seen when the units are separated by area, into the A1 and PEG

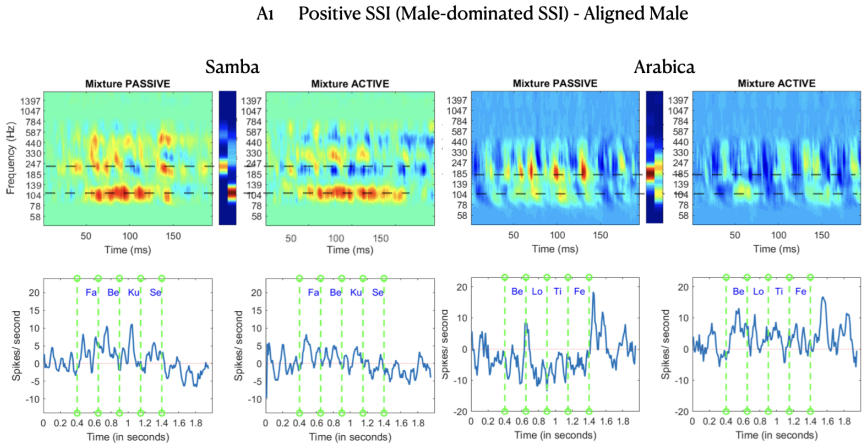


Figure 5.20: Positive SSI population in the A1 in Samba in (left) in the passive and active conditions, with an acoustic indicator separating them, and in Arabica (right) in the passive and active conditions, with an acoustic indicator separating them. Below them, in the bottom row, are the respective PSTHs.

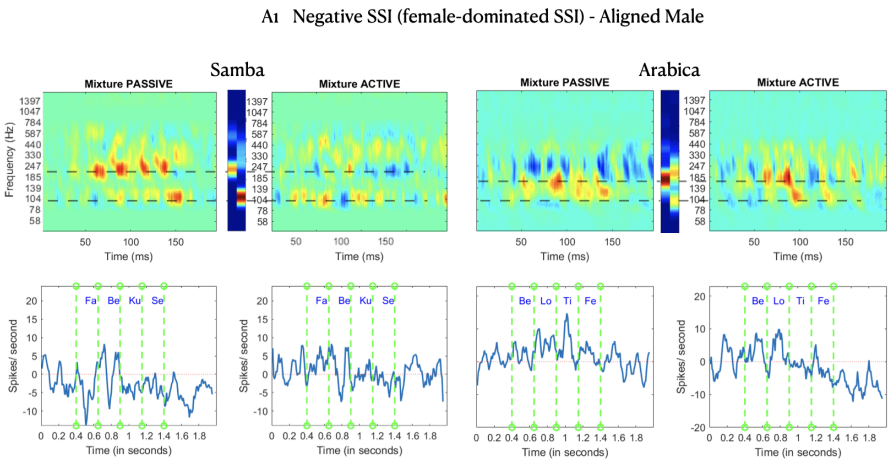


Figure 5.21: Negative SSI population in the A1 in Samba in (left) in the passive and active conditions, with an acoustic indicator separating them, and in Arabica (right) in the passive and active conditions, with an acoustic indicator separating them. Below them, in the bottom row, are the respective PSTHs.

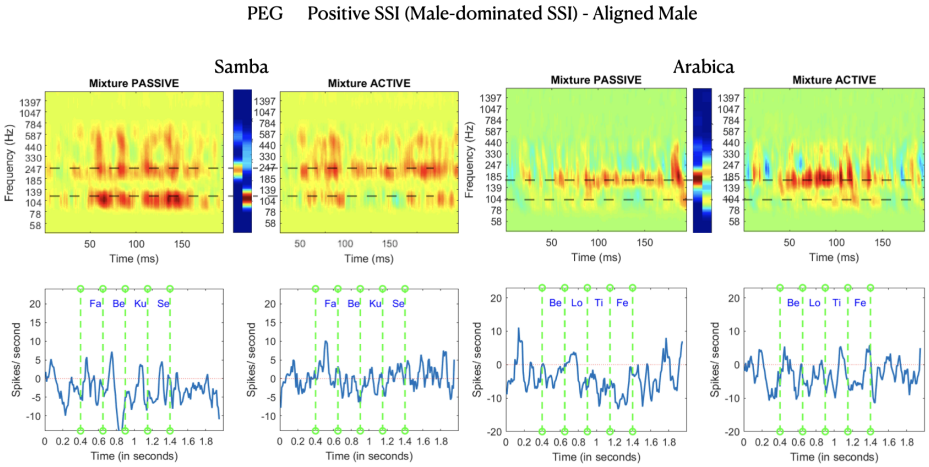


Figure 5.22: Positive SSI population in the PEG in Samba in (left) in the passive and active conditions, with an acoustic indicator separating them, and in Arabica (right) in the passive and active conditions, with an acoustic indicator separating them. Below them, in the bottom row, are the respective PSTHs.

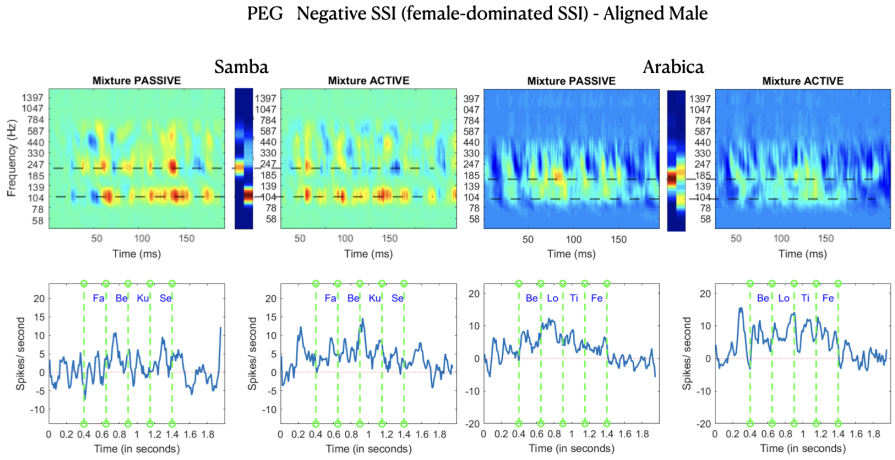


Figure 5.23: Negative SSI population in the PEG in Samba in (left) in the passive and active conditions, with an acoustic indicator separating them, and in Arabica (right) in the passive and active conditions, with an acoustic indicator separating them. Below them, in the bottom row, are the respective PSTHs.

Chapter 6: Auditory Two Stream Segregation: Frontal Cortex

The frontal cortices are well known for their executive function, decisions and working memory [80, 81] in humans as well as non-human primates. Frontal cortex activation by stimulation of the orbitofrontal cortex has shown to cause changes in the primary cortex of mice [82]. The prefrontal cortex is involved in maintaining a working memory to select and represent behaviorally relevant information [83], and is known to flexibly adapt to the needs of the task at hand [84]. Top-down selective attention has been implicated through numerous studies [5, 64] to play a very important role in stream formation and selection in a cocktail party scenario. Such top-down changes, likely from the frontal cortices, have been shown to initiate rapid, sustained changes in neural responses during task behavior in the primary auditory cortices [21].

In the previous chapter, we looked at manifestations of attention to a single stream in a two stream cocktail party as modulation of frequency components of responses in the auditory cortex. Here in this chapter, we present neural data from the right frontal cortex of the ferret brain from one animal (Samba), to understand what top-down changes from the frontal cortex could have developed this differential processing of attended speech stream.

6.1 Role of Frontal Cortices in Auditory Attention

The frontal cortex is divided into further sub-regions encompassing the orbital frontal lobe, dorso-lateral prefrontal cortex, and premotor area, amongst others. Within the auditory modality, lesion studies have provided evidence of projections between the auditory and prefrontal regions [85, 86, 87]. Romanski et. al. [87, 88] showed that the frontal cortex projections were more robust than those in the auditory cortical regions, suggesting a directionality from early to late stage auditory processing. In ferrets, attention-gated neural responses showed an increasing hierarchy of effects when moving along the primary [33] to secondary [2] to tertiary, belt like areas [32], which demonstrated responses very alike the dorsolateral prefrontal cortex [21] and suggesting that the connections between the frontal cortices for top-down modulation are strongest in the higher order areas. Similar gating driven by attention is also seen in the monkey PFC during a visual task [89].

Since most of this past work focused on frontal projections to the auditory cortex during behavioral tasks (tone detection, click discrimination) in a single stream, we focus on what happens in the frontal cortex of the ferret when detecting a word sequence by a single speaker, as well as neural correlates of stream segregation, modulated by attention and auditory memory. For the single speaker task, we also draw parallels from another task in the laboratory.

6.2 Analysis of Neural Responses

In this work, we record from approximately the anterior sigmoid gyrus (ASG) and the dorsal aspect of the orbital gyrus (OGS) as well as pre-motor cortex similar to neurophysiological

recordings done in Fritz et al. [21]. Our dataset comprises of neural activity from 54 single units. We refer to our recording region broadly as the frontal cortex. Future studies with anatomical expertise using injections in the site of recording post-mortem would be needed to determine the exact recording location and is beyond the current thesis. The neurophysiological data we acquired is sorted into single units and is binned into 10 ms bins, similar to data in the auditory cortex. For details, see Chapter 3, Methods. The peri-stimulus time histogram is obtained by averaging across trials.

6.2.1 Single Speaker Scenario

We present the single stream target detection task, where the animal discriminated the target word 'Fa-Be-Ku' from the reference word 'Fa-Fa-Fa' in Figure 6.1. In the passive condition, we see PSTH responses to all three syllables. This is unlike some previous studies, where neuronal responses to stimuli were absent in the passive. However, this passive syllabic response was observed during other (yet unpublished) frontal cortex experiments in the laboratory. We discuss the possible reasons for this in the Discussions section of this chapter.

As contrasted against the passive condition, in the active condition we note that the response to the second and third target syllables, which differ between the target and reference, are significantly enhanced as compared to the first syllable (in red). At the same time, the neural responses to the second and third syllables in the reference show a sharp decline (in blue), thus suggesting that the frontal cortex rapidly enhances the target word of concern, which can be used to make a decision about whether to continue licking the water at the spout or stop. This is in agreement with other studies [21] where there is clear, large target enhancement relative to the reference

sounds in the frontal cortex.

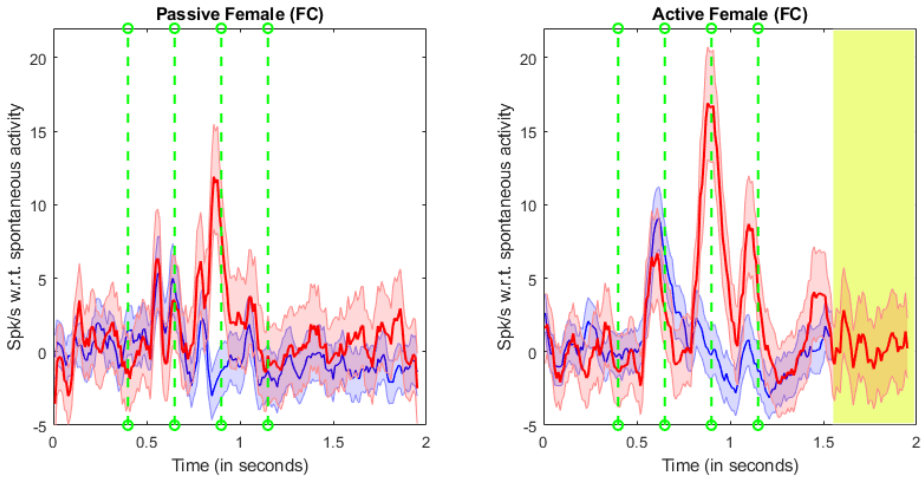


Figure 6.1: Neural responses to target word ABC (Fa-Be-Ku, Red) and reference word AAA (Fa-Fa-Fa, blue) during attention to single speaker when presented in isolation. The active condition (right) shows an enhanced target response as compared to the passive condition (left). Shaded region shows one standard deviation.

All these PSTHs are normalized such that each neuronal unit contributes equally to the PSTH, and baseline activity of the unit is removed from the response. Shaded regions depict one standard deviation of the responses. Enhancement of the target is clearly seen in the frontal cortex, in Figure 6.2 as compared to the auditory cortex.

The latency of responses in the frontal cortices reveal the temporal nature of processing in these areas. In Figure 6.2, we compare the auditory cortex responses of a single speaker to those in the frontal cortex. In the female alone condition, the peak responses for the syllables occur at 550 ms, 800 ms and 1050 ms for the passive condition (pre- and post-passive) and at 580 ms, 880 ms and 1090 ms in the active condition. The consistent delay of around 40 ms between active and passive is significant compared to the neural data in the auditory cortex, where the passive

syllables occur at 540 ms, 650 ms and 1030 ms, where as the active syllables occur at 540 ms, 640 ms and 1040 ms after the prestimulus onset (400 ms before sound starts), thus relaying no significant delay.

6.2.2 Two Speaker Stream Segregation

In the case of segregation with the two-stream mixture, while there are clear responses to the active condition, when aligned to female, the same data when aligned to the male speaker shows the lack of a response where the response is close to the spontaneous baseline firing of the neuron. This could be indicative of top-down modulation indicating a disinterest in the male speaker, and is seen even at the level of the PSTH.

6.3 Discussion and Future Work

Passive responses to sounds are not typically seen in all previous experiments in the laboratory in the frontal cortex. We do note these responses in our current animal, as well as in one other study in the lab. While a detailed anatomical study of recording areas would be required to understand passive encoding of sounds in trained animals, we hypothesize a possible reason these may occur. These responses, especially during the passive condition, could be a result of animal over-training, as the animal had to be trained and behavioral consistency maintained over a period of a few years during the recording sessions. In the future, naive animals' neural data may aid, in addition to post-recording tracer injections to ascertain exact recording areas, to facilitate a better understanding of these passive responses,

In this dataset of neural recordings from the frontal cortex, attention appears to behaviorally

gate responses during active task-behavior as compared to the passive condition. As contrasted against the comparable passive and active PSTHs in the Auditory Cortex, the frontal cortex showed a clear enhancement in the response at the target of the second syllable, which is also the differing syllable. This response was seen only for the target word, where as there was a suppression in the reference word PSTH. This effect existed slightly in the passive, but was really prominent through attention gating of the response. Through the enhancement of target and suppression of reference during behavior of a single speaker stream, even at the level of the PSTH, the frontal cortex neural responses also indicate that the responses encode not just the stimuli characteristics, but also meaning categories (as target or reference) of the words.

During the two-stream segregation task, the PSTH response in active compared to the passive condition is preserved when aligned to the male, but reduced to baseline activity level, when the same mixture data in active is aligned to the male. This indicates a top-down modulatory effect that suppresses responses during behavior that are aligned to the unattended speaker, even at the level of PSTH without foraying into the frequency feature space.

In the future, this work can be extended by conducting simultaneous recordings in the auditory and frontal cortices, to gain some insight through the correlation of responses, of the top-down influences exerted by the frontal cortex on auditory sensory processing of a two-stream segregation task.

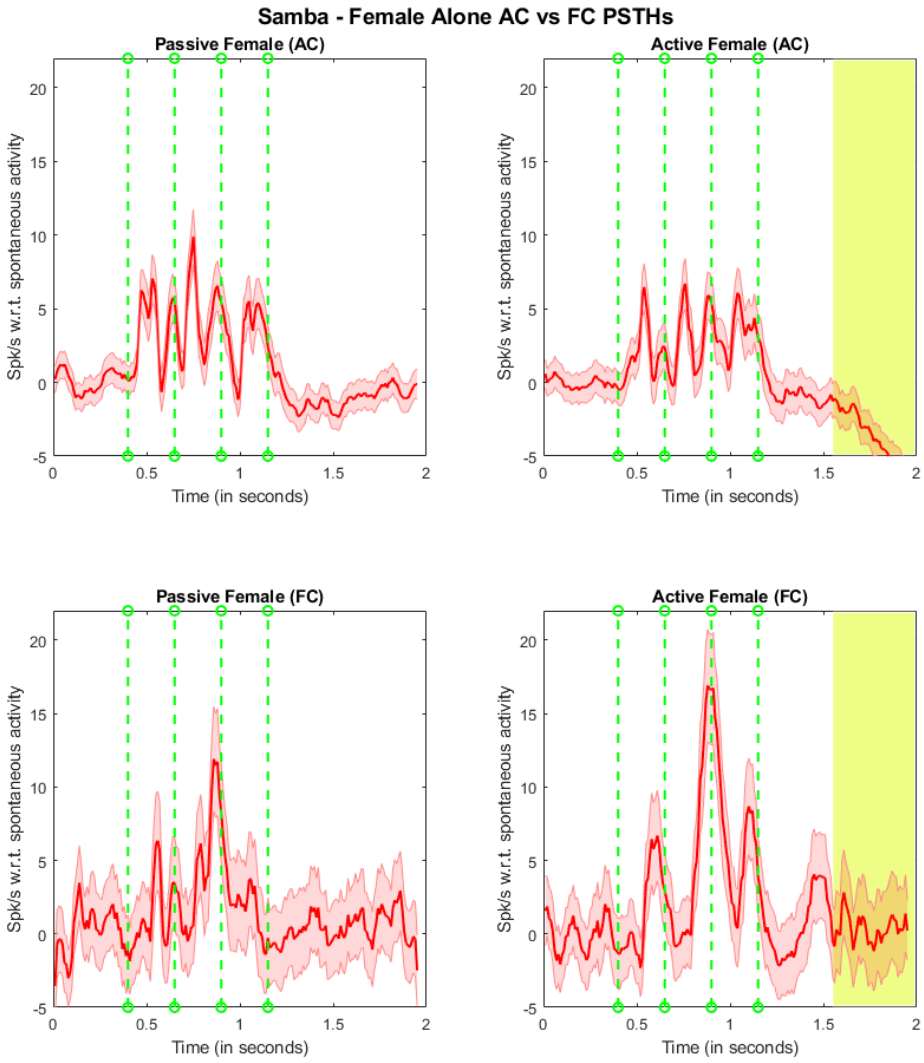


Figure 6.2: Neural responses in the Auditory Cortex (top row) and Frontal Cortex (bottom row) to target word ABC (Fa-Be-Ku, Red)in the passive condition (left) and during attention to a single speaker (right). Yellow indicates the shock period, during which noise contamination may occur in case of incorrect trials, and shaded region indicates one standard deviation.

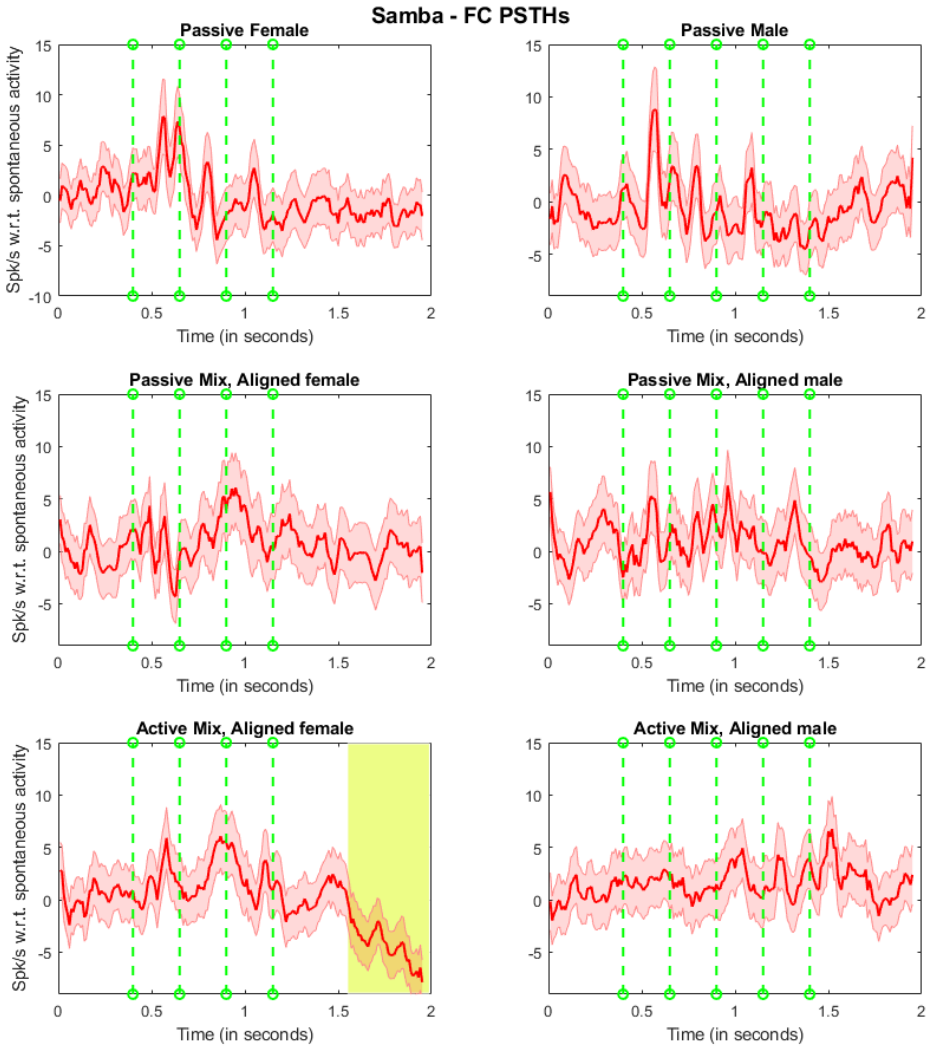


Figure 6.3: Neural responses to target word ABC (Fa-Be-Ku, Red) when presented individually (top row), in the passive condition (middle row) and active condition (bottom row). Figures on the left indicate alignment of single speaker of mixture responses to the female foreground speaker, and the panels on the right indicate male responses and mixture responses aligned to the male background word. Shaded region indicates one standard deviation.

Chapter 7: Conclusions

7.1 Summary

This thesis has provided an in-depth study of how the cocktail party and specifically, two speaker segregation, can be performed by animals behaviorally and how it is encoded in the cortex. Along with neural data during behavior, we demonstrated that the reconstructed spectrograms indicated similar representation of the cocktail party as in the human brain, and set the stage for further probing of speech streams in the ferret. Furthermore, we showed that this process of neural segregation has started at the primary cortex level, unlike in the human auditory cortex. There is clearer attentional modulation of attended and unattended speakers in the secondary areas in the form of enhancement of attended speaker features and suppression of features of the unattended speaker.

We demonstrated a novel behavioral paradigm to train ferrets to perform a two-speaker segregation task in a limited setting by showing that ferrets could reliably attend to one speaker and perform a target word detection task on that speaker even in the presence of an equally loud background speaker. Ferrets were able to consistently perform this task well above chance and behavioral capabilities evolved with training sessions. This behavioral ability in the ferrets did not depend on whether the background speaker started before or after the foreground word, nor did it depend on how long the trial was, as a function of number of words in the trial. This behavioral

task sets the stage to test future hypothesis on speech segregation and broadly, streaming, by showing that ferrets can perform streaming with complex acoustic stimuli such as speech.

The auditory cortex recordings revealed encoding of the two-speaker segregation paradigm in the ferret. We showed evidence for enhancement of the PSTH response during task engagement as compared to passive listening, that only occurred when the data are aligned to the female word. However, when the same two-speaker data are aligned to the male word occurrence, there was a suppression of overall response, as seen in the lower amplitudes of the PSTHs. To discern further what happened in the feature space of the stimuli, we employed the method of stimulus reconstruction by decoding auditory spectrograms from the neural responses. We showed that the segregation process has already started in the primary auditory cortex, but clearer suppression of frequencies belonging to the unattended speaker and enhancement of frequencies of the attended speaker is seen in the secondary cortex. We further subdivided each of these areas parallel to the tonotopic reversal line to explore the contributions of different sub-regions within the auditory cortex. Speaker selectivity indices have been used with human ECoG data to show that during a two-speaker segregation task, multi-unit like activity is modulated Here, we used a slightly modified version of speaker selectivity to demonstrate that populations which are more responsive to the male stimuli indicate overall suppression in the reconstructed spectrograms during behavior, whereas the female speaker selective units are more likely to either slightly enhance or bypass immediate modulation.

In the frontal cortex, there is evidence of modulation at the level of the PSTHs themselves. Responses in the frontal cortex showed a rapid enhancement of the target word and suppression of the passive word response even in the single speaker. This active enhancement of the target word was unique to the frontal cortex in that the responses to the same stimuli in the auditory cortex

population did not indicate such direct modulation. This could be indicative of a mechanism to provide target-enhanced responses to make a decision, in this case, to stop licking, during behavior. Longer latencies of responses also indicated that the auditory responses' processing in the frontal cortex is otherwise downstream to the stimuli representation. During two stream segregation, responses to the male words were suppressed in the temporal gaps, and responses to the female target word were enhanced during active engagement, but not in the passive condition. Further anatomical studies to accurately obtain locations of recordings, along with simultaneous recordings in the frontal and auditory cortex could reveal the interactions that may occur in the cocktail party scenario.

7.2 Future Work

During the behavioral experiments in this thesis, we noted that in our experience ferrets were not able to switch the target speaker of attention within sessions of a single day. This may require additional behavioral training techniques such as possibly priming with the attended speaker only or possibly, positively reinforcing a switch. While our current behavioral paradigm and recording technologies may be insufficient to understand evolution of training and the trajectory of learning for ferrets on this task, a combination of chronic arrays with early behavioral recording prior to excessive stimuli exposure could be a possible future behavioral direction for this project on two-speaker segregation by the ferret.

On the one hand we can utilize ferret data at the single unit and micro-population level to understand how the auditory cortex encodes speech in a two-speaker segregation task, but on the other hand, due to the complexity of the cortices, limitations of recording techniques and

restrictions on what the animal can reliably perform, there are a number of interesting questions that were beyond the scope of this project. Further studies could be designed in the future to probe the generalization of attention to multiple speakers, neuronal interactions during simultaneous recordings from populations of neurons, and modeling different scale/ rate spectrograms for different single units based on individual neuron characteristics to decode the stimuli from the auditory cortex. These questions would, together with the work in this thesis, move closer to understanding how the animal auditory cortices encode information for the cocktail party, and more broadly, auditory scene analysis.

Bibliography

[1] Jack B Kelly, Gerard L Kavanagh, and James CH Dalton. Hearing in the ferret (mustela putorius): thresholds for pure tone detection. *Hearing research*, 24(3):269–275, 1986.

[2] Serin Atiani, Mounya Elhilali, Stephen V David, Jonathan B Fritz, and Shihab A Shamma. Task difficulty and performance induce diverse adaptive patterns in gain and shape of primary auditory cortical receptive fields. *Neuron*, 61(3):467–480, 2009.

[3] E Colin Cherry. Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America*, 25(5):975–979, 1953.

[4] Albert S Bregman. *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.

[5] Robert P Carlyon. How the brain separates sounds. *Trends in cognitive sciences*, 8(10):465–471, 2004.

[6] Israel Nelken. Processing of complex stimuli and natural scenes in the auditory cortex. *Current opinion in neurobiology*, 14(4):474–480, 2004.

[7] Christophe Micheyl, Robert P Carlyon, Alexander Gutschalk, Jennifer R Melcher, Andrew J Oxenham, Josef P Rauschecker, Biao Tian, and E Courtenay Wilson. The role of auditory cortex in the formation of auditory streams. *Hearing research*, 229(1-2):116–131, 2007.

[8] LPAS Van Noorden. Temporal coherence and the perception of temporal position in tone sequences. *IPO Annual Progress Report*, 10:4–18, 1975.

[9] Mark A Bee and Georg M Klump. Primitive auditory stream segregation: a neurophysiological study in the songbird forebrain. *Journal of neurophysiology*, 92(2):1088–1104, 2004.

[10] Mark A Bee and Georg M Klump. Auditory stream segregation in the songbird forebrain: effects of time intervals on responses to interleaved tone sequences. *Brain, behavior and evolution*, 66(3):197–214, 2005.

[11] Jagmeet S Kanwal, Andrei V Medvedev, and Christophe Micheyl. Neurodynamics for auditory stream segregation: tracking sounds in the mustached bat's natural environment. *Network: Computation in Neural Systems*, 14(3):413, 2003.

[12] Mounya Elhilali, Ling Ma, Christophe Micheyl, Andrew J Oxenham, and Shihab A Shamma. Temporal coherence in the perceptual organization and cortical representation of auditory scenes. *Neuron*, 61(2):317–329, 2009.

[13] Naoya Itatani and Georg M Klump. Neural correlates of auditory streaming in an objective behavioral task. *Proceedings of the National Academy of Sciences*, 111(29):10738–10743, 2014.

[14] Yonatan I Fishman, Joseph C Arezzo, and Mitchell Steinschneider. Auditory stream segregation in monkey auditory cortex: effects of frequency separation, presentation rate, and tone duration. *The Journal of the Acoustical Society of America*, 116(3):1656–1670, 2004.

[15] Christophe Micheyl, Biao Tian, Robert P Carlyon, and Josef P Rauschecker. Perceptual organization of tone sequences in the auditory cortex of awake macaques. *Neuron*, 48(1):139–148, 2005.

[16] Ling Ma, Christophe Micheyl, Pingbo Yin, Andrew J Oxenham, and Shihab A Shamma. Behavioral measures of auditory streaming in ferrets (mustela putorius). *Journal of Comparative Psychology*, 124(3):317, 2010.

[17] Daniela Saderi, Brad N Buran, and Stephen V David. Streaming of repeated noise in primary and secondary fields of auditory cortex. *Journal of Neuroscience*, 40(19):3783–3798, 2020.

[18] Nima Mesgarani, Stephen V David, Jonathan B Fritz, and Shihab A Shamma. Phoneme representation and classification in primary auditory cortex. *The Journal of the Acoustical Society of America*, 123(2):899–909, 2008.

[19] Jennifer K Bizley, Fernando R Nodal, Israel Nelken, and Andrew J King. Functional organization of ferret auditory cortex. *Cerebral cortex*, 15(10):1637–1653, 2005.

[20] Henry E Heffner and Rickye S Heffner. Conditioned avoidance. In *Methods in comparative psychoacoustics*, pages 79–93. Springer, 1995.

[21] Jonathan B Fritz, Stephen V David, Susanne Radtke-Schuller, Pingbo Yin, and Shihab A Shamma. Adaptive, behaviorally gated, persistent encoding of task-relevant auditory information in ferret frontal cortex. *Nature neuroscience*, 13(8):1011–1019, 2010.

[22] Jennifer K Bizley, Kerry MM Walker, Andrew J King, and Jan WH Schnupp. Spectral timbre perception in ferrets: discrimination of artificial vowels under different listening conditions. *The Journal of the Acoustical Society of America*, 133(1):365–376, 2013.

[23] Pingbo Yin, Jonathan B Fritz, and Shihab A Shamma. Do ferrets perceive relative pitch? *The Journal of the Acoustical Society of America*, 127(3):1673–1680, 2010.

[24] Kerry MM Walker, Jan WH Schnupp, Sheelah MB Hart-Schnupp, Andrew J King, and Jennifer K Bizley. Pitch discrimination by ferrets for simple and complex sounds. *The Journal of the Acoustical Society of America*, 126(3):1321–1335, 2009.

[25] Sridhar Kalluri, Didier A Depireux, and Shihab A Shamma. Perception and cortical neural coding of harmonic fusion in ferrets. *The Journal of the Acoustical Society of America*, 123(5):2701–2716, 2008.

[26] Jennifer K Bizley, Victoria M Bajo, Fernando R Nodal, and Andrew J King. Cortico-cortical connectivity within ferret auditory cortex. *Journal of Comparative Neurology*, 523(15):2187–2210, 2015.

[27] David R Moore. Auditory brainstem of the ferret: sources of projections to the inferior colliculus. *Journal of Comparative Neurology*, 269(3):342–354, 1988.

[28] CK Henkel and Mark L Gabriele. Organization of the disynaptic pathway from the anteroventral cochlear nucleus to the lateral superior olivary nucleus in the ferret. *Anatomy and embryology*, 199(2):149–160, 1999.

[29] Victoria M Bajo, Fernando R Nodal, Jennifer K Bizley, David R Moore, and Andrew J King. The ferret auditory cortex: descending projections to the inferior colliculus. *Cerebral Cortex*, 17(2):475–491, 2007.

[30] Alessandra Angelucci, Francisco Clascá, and Mriganka Sur. Brainstem inputs to the ferret medial geniculate nucleus and the effect of early deafferentation on novel retinal projections to the auditory thalamus. *Journal of Comparative Neurology*, 400(3):417–439, 1998.

[31] Nina Kowalski, Huib Versnel, and Shihab A Shamma. Comparison of responses in the anterior and primary auditory fields of the ferret cortex. *Journal of Neurophysiology*, 73(4):1513–1523, 1995.

[32] Diego Elgueda, Daniel Duque, Susanne Radtke-Schuller, Pingbo Yin, Stephen V David, Shihab A Shamma, and Jonathan B Fritz. State-dependent encoding of sound and behavioral meaning in a tertiary region of the ferret auditory cortex. *Nature neuroscience*, 22(3):447–459, 2019.

[33] Jonathan Fritz, Shihab Shamma, Mounya Elhilali, and David Klein. Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nature neuroscience*, 6(11):1216–1223, 2003.

[34] Shaowen Bao, Edward F Chang, Jennifer Woods, and Michael M Merzenich. Temporal plasticity in the primary auditory cortex induced by operant perceptual learning. *Nature neuroscience*, 7(9):974–981, 2004.

[35] David J Klein, Didier A Depireux, Jonathan Z. Simon, and Shihab A. Shamma. Robust spectrotemporal reverse correlation for the auditory system: optimizing stimulus design. *Journal of computational neuroscience*, 9(1):85–111, 2000.

[36] Jonathan B Fritz, Mounya Elhilali, and Shihab A Shamma. Adaptive changes in cortical receptive fields induced by attention to complex sounds. *Journal of neurophysiology*, 98(4):2337–2346, 2007.

[37] Stephen V David, Jonathan B Fritz, and Shihab A Shamma. Task reward structure shapes rapid receptive field plasticity in auditory cortex. *Proceedings of the National Academy of Sciences*, 109(6):2144–2149, 2012.

[38] Kerry MM Walker, Jennifer K Bizley, Andrew J King, and Jan WH Schnupp. Multiplexed and robust representations of sound features in auditory cortex. *Journal of Neuroscience*, 31(41):14565–14576, 2011.

[39] Albert S Bregman, Christine Liao, and Robert Levitan. Auditory grouping based on fundamental frequency and formant peak frequency. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 44(3):400, 1990.

[40] Jack B Kelly, Peter W Judge, and Dennis P Phillips. Representation of the cochlea in primary auditory cortex of the ferret (mustela putorius). *Hearing research*, 24(2):111–115, 1986.

[41] Frederick Weber, Linda Manganaro, Barbara Peskin, and Elizabeth Shriberg. Using prosodic and lexical information for speaker identification. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–141–I–144, 2002.

[42] Johannus Petrus Leonardus Brokx and Sibout Govert Nooteboom. Intonation and the perceptual separation of simultaneous voices. *Journal of Phonetics*, 10(1):23–36, 1982.

[43] Ernst Terhardt. On the perception of periodic sound fluctuations (roughness). *Acta Acustica united with Acustica*, 30(4):201–213, 1974.

[44] Andrew J Oxenham. Pitch perception and auditory stream segregation: implications for hearing loss and cochlear implants. *Trends in amplification*, 12(4):316–331, 2008.

[45] Willard Larkin and Gordon Z Greenberg. Selective attention in uncertain frequency detection1. *Perception & Psychophysics*, 8(3):179–184, 1970.

[46] Ross K Maddox and Barbara G Shinn-Cunningham. Influence of task-relevant and task-irrelevant feature continuity on selective auditory attention. *Journal of the Association for Research in Otolaryngology*, 13(1):119–129, 2012.

[47] CJ Darwin and RW Hukin. Limits to the role of a common fundamental frequency in the fusion of two sounds with different spatial cues. *The Journal of the Acoustical Society of America*, 116(1):502–506, 2004.

[48] William Morris Hartmann and Douglas Johnson. Stream segregation and peripheral channeling. *Music perception*, 9(2):155–183, 1991.

[49] Joel S Snyder, Claude Alain, and Terence W Picton. Effects of attention on neuroelectric correlates of auditory stream segregation. *Journal of cognitive neuroscience*, 18(1):1–13, 2006.

[50] Andrew R Dykstra, Eric Halgren, Thomas Thesen, Chad E Carlson, Werner Doyle, Joseph R Madsen, Emad N Eskandar, and Sydney S Cash. Widespread brain areas engaged during a classical auditory streaming task revealed by intracranial eeg. *Frontiers in human neuroscience*, 5:74, 2011.

[51] Alexander Gutschalk, Christophe Micheyl, Jennifer R Melcher, André Rupp, Michael Scherg, and Andrew J Oxenham. Neuromagnetic correlates of streaming in human auditory cortex. *Journal of Neuroscience*, 25(22):5382–5388, 2005.

[52] Alexander Gutschalk, Christophe Micheyl, and Andrew J Oxenham. Neural correlates of auditory perceptual awareness under informational masking. *PLoS biology*, 6(6):e138, 2008.

[53] Daniel Pressnitzer, Mark Sayles, Christophe Micheyl, and Ian M Winter. Perceptual organization of sound begins in the auditory periphery. *Current Biology*, 18(15):1124–1128, 2008.

[54] Elyse S Sussman, János Horváth, István Winkler, and Mark Orr. The role of attention in the formation of auditory streams. *Perception & psychophysics*, 69(1):136–152, 2007.

[55] HT Tiitinen, J Sinkkonen, K Reinikainen, Kimmo Alho, J Lavikainen, and R Näätänen. Selective attention enhances the auditory 40-hz transient response in humans. *Nature*, 364(6432):59–60, 1993.

[56] Barbara G Shinn-Cunningham. Object-based auditory and visual attention. *Trends in cognitive sciences*, 12(5):182–186, 2008.

[57] Anne M Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980.

[58] Jennifer Adrienne Johnson and Robert J Zatorre. Neural substrates for dividing and focusing attention between simultaneous auditory and visual events. *Neuroimage*, 31(4):1673–1681, 2006.

[59] Mounya Elhilali, Juanjuan Xiang, Shihab A Shamma, and Jonathan Z Simon. Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene. *PLoS biology*, 7(6):e1000129, 2009.

[60] Sahar Akram, Bernhard Englitz, Mounya Elhilali, Jonathan Z Simon, and Shihab A Shamma. Investigating the neural correlates of a streaming percept in an informational-masking paradigm. *PloS one*, 9(12):e114427, 2014.

[61] Juanjuan Xiang, Jonathan Simon, and Mounya Elhilali. Competing streams at the cocktail party: exploring the mechanisms of attention and temporal integration. *Journal of Neuroscience*, 30(36):12084–12093, 2010.

[62] Aurélie Bidet-Caulet, Catherine Fischer, Francoise Bauchet, Pierre-Emmanuel Aguera, and Olivier Bertrand. Neural substrate of concurrent sound perception: direct electrophysiological recordings from human auditory cortex. *Frontiers in Human Neuroscience*, page 5, 2008.

[63] Mounya Elhilali and Shihab A Shamma. A cocktail party with a cortical twist: how cortical mechanisms contribute to sound segregation. *The Journal of the Acoustical Society of America*, 124(6):3751–3771, 2008.

[64] Jonathan B Fritz, Mounya Elhilali, Stephen V David, and Shihab A Shamma. Auditory attention—focusing the searchlight on sound. *Current opinion in neurobiology*, 17(4):437–455, 2007.

[65] Pingbo Yin, Ling Ma, Mounya Elhilali, Jonathan Fritz, and Shihab Shamma. Primary auditory cortical responses while attending to different streams. In *Hearing–From Sensory Processing to Perception*, pages 257–265. Springer, 2007.

[66] Ernst Niebur, Steven S Hsiao, and Kenneth O Johnson. Synchrony: a neuronal mechanism for attentional selection? *Current opinion in neurobiology*, 12(2):190–194, 2002.

[67] Mohsen Rezaeizadeh and Shihab Shamma. Binding the acoustic features of an auditory source through temporal coherence. *Cerebral cortex communications*, 2(4):tgab060, 2021.

[68] Nai Ding and Jonathan Z Simon. Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of neurophysiology*, 107(1):78–89, 2012.

[69] Nai Ding and Jonathan Z Simon. Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences*, 109(29):11854–11859, 2012.

[70] Nima Mesgarani and Edward F Chang. Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, 485(7397):233–236, 2012.

[71] Sahar Akram, Alessandro Presacco, Jonathan Z Simon, Shihab A Shamma, and Behtash Babadi. Robust decoding of selective auditory attention from meg in a competing-speaker environment via state-space modeling. *NeuroImage*, 124:906–917, 2016.

[72] James A O'Sullivan, Alan J Power, Nima Mesgarani, Siddharth Rajaram, John J Foxe, Barbara G Shinn-Cunningham, Malcolm Slaney, Shihab A Shamma, and Edmund C Lalor. Attentional selection in a cocktail party environment can be decoded from single-trial eeg. *Cerebral cortex*, 25(7):1697–1706, 2015.

[73] Lars Hausfeld, Lars Riecke, Giancarlo Valente, and Elia Formisano. Cortical tracking of multiple streams outside the focus of attention in naturalistic auditory scenes. *NeuroImage*, 181:617–626, 2018.

[74] Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain De Cheveigne. Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds. *Speech communication*, 27(3-4):187–207, 1999.

[75] Taishih Chi, Powen Ru, and Shihab A Shamma. Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, 118(2):887–906, 2005.

[76] Andrew J King, Victoria M Bajo, Jennifer K Bizley, Robert AA Campbell, Fernando R Nodal, Andreas L Schulz, and Jan WH Schnupp. Physiological and behavioral studies of spatial coding in the auditory cortex. *Hearing research*, 229(1-2):106–115, 2007.

[77] Nima Mesgarani, Stephen V David, Jonathan B Fritz, and Shihab A Shamma. Influence of context and behavior on stimulus reconstruction from neural activity in primary auditory cortex. *Journal of neurophysiology*, 102(6):3329–3339, 2009.

[78] Brian N Pasley, Stephen V David, Nima Mesgarani, Adeen Flinker, Shihab A Shamma, Nathan E Crone, Robert T Knight, and Edward F Chang. Reconstructing speech from human auditory cortex. *PLoS biology*, 10(1):e1001251, 2012.

[79] James O'Sullivan, Jose Herrero, Elliot Smith, Catherine Schevon, Guy M McKhann, Sameer A Sheth, Ashesh D Mehta, and Nima Mesgarani. Hierarchical encoding of attended auditory objects in multi-talker speech perception. *Neuron*, 104(6):1195–1209, 2019.

[80] J Rowe, L Hughes, D Eckstein, and AM Owen. Rule-selection and action-selection have a shared neuroanatomical basis in the human prefrontal and parietal cortex. *Cerebral cortex*, 18(10):2275–2285, 2008.

[81] Daniel Bor, Nick Cumming, Catherine EL Scott, and Adrian M Owen. Prefrontal cortical involvement in verbal encoding strategies. *European Journal of Neuroscience*, 19(12):3365–3370, 2004.

[82] Daniel E Winkowski, Sharba Bandyopadhyay, Shihab A Shamma, and Patrick O Kanold. Frontal cortex activation causes rapid plasticity of auditory cortical processing. *Journal of Neuroscience*, 33(46):18134–18148, 2013.

[83] Gregor Rainer, Wael F Asaad, and Earl K Miller. Selective representation of relevant information by neurons in the primate prefrontal cortex. *Nature*, 393(6685):577–579, 1998.

[84] Stefan Everling, Chris J Tinsley, David Gaffan, and John Duncan. Selective representation of task-relevant objects and locations in the monkey prefrontal cortex. *European Journal of Neuroscience*, 23(8):2197–2214, 2006.

[85] Michael Petrides and Deepak N Pandya. Association fiber pathways to the frontal cortex from the superior temporal region in the rhesus monkey. *Journal of Comparative Neurology*, 273(1):52–66, 1988.

[86] Deepak N Pandya and Henricus GJM Kuypers. Cortico-cortical connections in the rhesus monkey. *Brain research*, 13(1):13–36, 1969.

[87] Lizabeth M Romanski, Joseph F Bates, and Patricia S Goldman-Rakic. Auditory belt and parabelt projections to the prefrontal cortex in the rhesus monkey. *Journal of Comparative Neurology*, 403(2):141–157, 1999.

[88] Jon H Kaas and Troy A Hackett. 'what'and'where'processing in auditory cortex. *Nature neuroscience*, 2(12):1045–1047, 1999.

[89] David J Freedman, Maximilian Riesenhuber, Tomaso Poggio, and Earl K Miller. Categorical representation of visual stimuli in the primate prefrontal cortex. *Science*, 291(5502):312–316, 2001.