UNIVERSITY OF CALIFORNIA SAN DIEGO

Contrast, Neutralization, and Systems of Invariance

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Linguistics and Cognitive Science

by

Anna Mai

Committee in charge:

Professor Eric Baković, Co-Chair
Professor Timothy Gentner, Co-Chair
Professor Leon Bergen
Professor Marc Garellek
Professor Brad Voytek

2022

The dissertation of Anna Mai is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

DEDICATION

For Tom

EPIGRAPH

The distinction to be made is not between exterior and interior, which are always relative, changing and reversible but between different types of multiplicities that coexist, interpenetrate, and change places—machines, cogs, motors and elements that are set in motion at a given moment, forming an assemblage of productive statements: "I love you" (or whatever).

— from *A Thousand Plateaus* by Gilles Deleuze and Félix Guattari

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

This work would not have been possible without the support of very many people, many of whom I may fail to mention by name here. Even for those I mention, please know that these words of thanks encompass barely an atom of everything I owe to you.

First and foremost, I thank my advisors Eric Baković and Tim Gentner. The kindness and generosity of all their advice and the rigor and inspiration they bring to our conversations have been mainstays of my graduate career.

I am grateful also for the incredible support I have received from the other members of my committee—for Brad Voytek welcoming me into his lab as I gained my footing in cognitive neuroscience; for Leon Bergen helping me think through projects from all angles with his thoughtful questions; and for Marc Garellek inspiring me with his teaching, inviting me onto projects, and providing detailed and considerate feedback on much of my written work.

Beyond the members of my committee, I could not imagine a better department in which to complete this dissertation. Thank you to Gaby Caballero for mentorship and support on my first published work; to Robert Kluender for early advice that shaped the trajectory of my graduate career and for consistent interest in and enthusiasm for my work; and to Sharon Rose for modeling excellent phonology pedagogy and providing teaching and feedback that has consistently elevated the level of my work.

I am deeply grateful for the existence of the Interdisciplinary Program in the Cognitive Sciences for its role in expanding my understanding of language and the brain, and I am grateful to Gary Cottrell, Bill Bechtel, and Arturs Semenuks in particular for their roles in my experience with the IDP.

Thank you also to my fellow graduate students in the Linguistics Department and to the members of Gentner Lab. I have learned so much from all of you over these years and am deeply grateful for your friendship, your support, and your patience with me.

Thank you to my coauthors, Andrés Aguilar and Gaby Caballero for teaching me how to navigate the peer review process; Adam McCollum, Eric Baković, and

Eric Meinhardt for going where the research takes us, even all the way to Reykjavík; Mai Moua and Marc Garellek for teaching me about voice quality in the Mong folk song tradition; Tim Sainburg and Tim Gentner for all our conversations on evolution, information theory, and comparative biology; and Jerry Shih, Sharona Ben-Haim, and Stephanie Riès for introducing me to working with intracranial patients.

Thank you to Ruth Levy, the Meltzer family, and all the other members of Ohr Shalom for welcoming me into your hearts and encouraging my work. Thank you also to Bill Morris specifically for the vivid and terrifying reminder that every day I don't write is a day added to the length of my degree.

Thank you to Brett Hyde and Brett Kessler without whom I would not have made it to San Diego, and thank you to my family without whom I would not have made it to St. Louis.

And thank you to Eric and Andrew and Mae for keeping me human.

Material from Chapters 2–4 of the dissertation has been submitted for publication as "The neural response to speech is language specific and irreducible to speech acoustics" with authors Anna Mai, Stephanie Riès, Sharona Ben-Haim, Jerry Shih, and Timothy Gentner. The dissertation author was the primary investigator and author of this paper.

VITA

2015	B. A. in Linguistics and Music *magna cum laude*, Washington University in St. Louis
2017	M. A. in Linguistics, University of California San Diego
2022	Ph. D. in Linguistics and Cognitive Science, University of California San Diego

PUBLICATIONS

A. Mai, S. Riès, S. Ben-Haim, J. Shih, & T. Gentner. under review. The neural response to speech is language specific and irreducible to speech acoustics.

E. Meinhardt, A. Mai, E. Baković, & A. McCollum. in press. Weak determinism and the computational consequences of interaction. *Natural Language and Linguistic Theory*.

T. Sainburg, A. Mai, & T. Gentner. 2021. Long-range sequential dependencies are phylogenetically pervasive in behavior and precede complex syntactic production in language. *Proceedings of the Royal Society B*.

A. McCollum, E. Baković, A. Mai, & E. Meinhardt. 2020. Unbounded circumambient patterns in segmental phonology. *Phonology*.

A. Mai. Phonetic effects of onset complexity on the English syllable. 2020. *Journal of Laboratory Phonology*.

A. Mai, A. Aguilar, & G. Caballero. 2018. Ja'a Kumiai. *Journal of the IPA*.

ABSTRACT OF THE DISSERTATION

Contrast, Neutralization, and Systems of Invariance

by

Anna Mai
Doctor of Philosophy in Linguistics and Cognitive Science
University of California San Diego, 2022
Professor Eric Baković, Co-Chair
Professor Timothy Gentner, Co-Chair

Speech is a highly-structured auditory signal that carries necessary and sufficient information for language comprehension among neurotypical, hearing populations. Speech perception, accordingly, involves the integration of both bottom-up auditory processing and top-down language knowledge. As explored in Chapter 1, however, while much is known about the brain's sensitivity to various acoustic and phonological manipulations, mechanistic transformative links between auditory sensation and linguistic perception remain poorly understood.

This dissertation contributes three studies to the ongoing pursuit of links between auditory and linguistic processing of speech in the brain. Intracranial stereo electroencephalography (sEEG) was recorded while participants listened to conversational English speech, and for each study, different sets of analyses were performed to target different aspects of these links.

Chapter 2 introduces a novel comparative technique for the isolation of phonological and morphological identity from the auditory speech signal. Through structured comparisons of the neural response to sounds in allophonic relationships with

one another, sites that dissociate phonemic and morphological identity from acoustic similarity were identified. These results provide crucial evidence that the neural response to speech is sensitive to abstract sub-lexical structure.

Chapter 3 provides general models to account for the unique contributions of categorical phonemic information and spectrographic information to the neural response. In lower frequency bands, phonemic category information explained a greater proportion of the neural response variance than spectrographic information. These results indicate that the explanatory value of phonemic category features in neural encoding models is not reducible to the surface spectrotemporal features that correlate with those categories.

Chapter 4 proposes a model for the relationship between the speech stimulus and neural response that integrates both phonemic and spectrographic information. Inclusion of the stimulus covariance structure increased model accuracy only when categorical phonemic information was available, indicating that phoneme-related neural activity is likely mediated through the acoustic covariance structure of speech. Moreover, phonemic label information conferred no benefit to model fit when participants listened to speech in an unfamiliar language (Catalan).

Together these studies show that neural responses associated with categorical phonemic information are language specific and irreducible to speech acoustics, and they scout a path towards mechanistic, transformative links between auditory sensation and language cognition.

Chapter 1

Phonology and the Brain

1.1 Introduction

Phonology is the structure of the smallest contrastive units of language and the ways in which they pattern within a language, whether signed or spoken. The basic unit of analysis in phonology is either the distinctive feature or the phoneme, and the study of phonology encompasses the study of possible phonemic sequences within a language (phonotactics); language-specific predictable variation in the production of phonemes conditioned by their phonological context (alternations); and the cross-linguistic properties of phonemic inventories and alternations (typology). Within the Chomsky Hierarchy (Chomsky, 1956), human language phonology is attested to be at most regular, and the vast majority of phonological patterns are subregular, due to the ubiquity of patterns that require only local context to compute. This review provides an overview of the major phenomena of study in phonology and our current understanding of how these phenomena are instantiated in the brain.

1.1.1 What is phonology?

Phonology encompasses linguistic phenomena that are specific to particular languages and most aptly described using discrete units that are devoid of semantic meaning. Core to this definition of phonology is the assumption that a finite set

of discrete analyzable units underpins the sound system of any language, and that set of units can be defined through the structured patterns of contrast and neutralization exhibited in that language.

When two sounds *contrast*, their presence is unpredictable from the surrounding phonological environment (Kenstowicz and Kisseberth, 2014). For example, based on phonological context alone—that is, without consideration of higher order forms of context, such as lexical frequency—the sound sequence [tæ], is equally as likely to be continued by [k] as [p] (i.e., [tæp] *tap* and [tæk] *tack*). Given this observation, [k] and [p] are recognized as discrete contrasting units in American English. On the other hand, if the distribution of two sounds is predictable from their phonological environment, they do not contrast with one another in that environment. For example, in American English, alveolar nasals regressively assimilate place. In a word like *phonebook*, the alveolar nasal [n] assimilates to the labial place of the following consonant such that [foʊn]+[bʊk] becomes [foʊmbʊk] in all but the most careful speech. This distribution of nasals is predictable from phonological context because we do not see [n] preceding labial consonants within a single word, but we do see [n] in other phonological contexts (i.e., word-finally in *phone*). Similarly predictable place assimilation occurs when [n] precedes alveolar consonants in words like *phone line* [foʊnlaɪn] or *phone tag* [foʊntæg]; velar consonants in words like *phone call* [foʊŋkɑl]; and even when [n] precedes labiodental consonants in phrases like *phone fee* [foʊɱfi].

In each of these collocations, the meaning of the word *phone* is constant despite the variability in the pronunciation of the final sound. The sounds are related to one another through their ability to *alternate* with one another in predictable phonological contexts. However, while the sounds [m, ŋ, ɱ, n] occur predictably pre-consonantally, some contrast with one another in other phonological environments. For example, if we consider the minimal pair *phone* [foʊn] and *foam* [foʊm], we observe that [n] and [m] do contrast word-finally. However, preceding a labial consonant, the contrast between [n] and [m] is neutralized due to the process of nasal place assimilation, whereby both *phonebook* and *foam book* surface as [foʊmbʊk].

Through such patterns of contrast and alternation, we begin to see the basis for discrete units in phonology: the [m] in *foam book* participates in a different set of contrasts and alternations than the [m] in the word *phonebook*. To account for these differences, linguists posit the existence of a type of abstract unit that unifies each distinct set of contrasts and alternations, called the phoneme. If one accepts this level of abstraction as reasonable, then alternations can be reframed as the selection or derivation of a pronounced sound, called a *surface form*, from a particular abstract unit, called an *underlying form*, given knowledge of the surrounding phonological context. The various surface forms of a phoneme are called its *allophones*, which comprise phonological features, and the underlying form is typically said to be the set of phonological features that corresponds to the most predictable allophone.[1] Sets of features for both surface forms and underlying forms are then represented in shorthand by a symbol. In the example given above, [m, ŋ, ɱ, n] are all the symbolic shorthands for the allophones of the phoneme called /n/.

In some ways, the study of alternations is like tossing a ball into a dark, empty space to probe its dimensions, using what one knows about predictable physical forces to reinterpret the sound of the ball rebounding off a surface as a distance. The structure of alternations provide a similar kind of evidence for the static primitives of a language's phonology. As discussed above, patterns of alternation provide evidence for the (synchronically invariant) content of a language's phonemic inventory, and in similar ways they also provide evidence for a language's phonotactics (permissible sound sequences) and syllable structure inventory.

For example, syllables with complex codas (e.g., CVCC)[2] or complex onsets (e.g., CCV) do not occur word-internally in Levantine Arabic (Broselow 1992, Watson 2007, Broselow 2018). When an underlying form like /kalb/ 'dog' is suffixed with a vowel, as in /kalb-i/ 'my dog', it can be syllabified as [kal.bi], but when it is suffixed with a CV syllable, as in /kalb-na/ 'our dog', syllabification alone is

[1] See Kiparsky (1968) and Chapter 3 of Kenstowicz and Kisseberth (2014) for more thorough discussion of this issue and the issue of abstraction in phonology more generally.

[2] The sequence CVCC represents any syllable that begins with a consonant, followed by a single vowel, followed by two more consonants. In English, the word [ɹɛst.ɹumz] 'restrooms' is two CVCC syllables. The syllable has a complex coda because it has more than one consonant following the vowel.

insufficient to prevent the occurrence of a word-internal CCV or CVCC syllable. Instead of [kalbna], the observed surface form of 'our dog' is [ka.lib.na]. The insertion of [i] (predictable alternation of ∅ with [i]) at this juncture, where lack of insertion would otherwise result in a syllable of type CCV or CVCC, provides evidence for the syllable structure constraints of the language.

1.1.2 Why is phonology an interesting problem for neuroscience?

One of the central goals of cognitive neuroscience is to understand how sensory input contributes to behavior. Human language phonology offers a handhold on this problem, in part because it presents a restricted "binding problem" in which infinitely variable sensory input is bound to finitely many behaviorally relevant (phonological) categories. From this perspective, aspects of phonological processing, such as phoneme recognition, appear similar to other psychophysical tasks that require categorization. For example, many of the hallmarks of categorical perception are shared with phonological processing: both are early, automatic perceptual phenomena that are variable across sensory space and modulated by both context and experience (Holt and Lotto, 2010).

Upon recognizing this similarity, it is typical to conclude that the psychophysical perception literature ought to be brought to bear on the study of phonological processing. However, the converse conclusion, though less common, is equally worthy of pursuit. That is, results from the study of phonological processing also have the potential to inform our understanding of psychophysics. With its restricted binding problem, phonological processing offers a manageable setting in which to investigate categorization without sacrificing ethological validity. Furthermore, depending on whether a signed or spoken language is chosen for study, categorization in both visual and auditory domains can be investigated. Within a single language, study can easily be further restricted to the visual or acoustic correlates of particular categories or to the categorical correlates of particular continuous visual or acoustic features. And different languages instantiate different categories, allowing for naturalistic study of the influence of category inventory on perception

as well. Thus, for those neuroscientists who wish to understand how sensory input contributes to cognition and behavior, phonological processing is a productive field in which to test and scale-up theories of perception.

Additionally, beyond its relation to psychophysics, phonology offers a tractable handhold on the interface between physical sensory stimuli and the abstract cognitive system of language. To date, the overwhelming majority of research on language and the brain focuses on how the brain processes the structures of linear sequences of words (syntax) and imputes meanings from them (semantics). Syntax and semantics are very difficult to study in the brain, in large part due to their levels of abstraction. A primary assumption in syntactic theory is that linear orders of words are underpinned by fundamentally non-linear structures, typically hierarchical structures in the generative grammar tradition (Chomsky, 2009). Studying syntax in the brain therefore requires asserting abstract structures for linear sequences and correlating sequences that share putative abstract structures with neural activity. This process is further complicated by the fact that (in principle) infinite linear sequences can have the same abstract structure, raising difficulties with stimulus sampling, and that identical linear sequences may permit multiple abstract structures.[3] Studying semantics in the brain is similarly difficult given not only its inherently abstract nature but also its interaction with syntax and lack of well-established primitives (cf., syntactic categories for syntax, features and phonemes for phonology).

Despite these difficulties, syntax and semantics are often privileged in the study of the neurobiology of language for what they represent to those who study them. The recursive property of syntax is often referred to as the core of the human language faculty and has been credited as the primary feature of human language that distinguishes it from non-human communicative systems (Hauser et al., 2002). Similarly, use of linguistic meaning has historically been argued to follow directly

[3]When identical sequences permit multiple abstract structures, sometimes the meaning of the sequence easily distinguishes which unique abstract structure is intended. For example, given the sentence 'I saw the pirate with the telescope,' only one abstract structure is tenable depending on who has the telescope, me or the pirate. However, in other situations, determining which structure a listener would arrive at is more difficult. For example, given a 'hot vegetable dinner' the difference between the vegetable being hot versus the vegetable dinner being hot is not easily determined, since the vegetable constitutes the dinner.

from distinctly human capacities for abstract thought (Fodor, 1975). In these ways, the studies of syntax and semantics stand as proxies for the study of what it means to be human. However, syntax and semantics are only aspects of the full system of human language, and while their draw is unmistakable, other aspects of language offer equally well-suited points from which to launch excavations of the neurobiological architecture of language.

Key among these, for this thesis, is the language-specificity of phonology and its webs of contrast and neutralization. Like syntax, phonology is governed by grammar in the computational sense (Chomsky, 1956). That is, whether a language permits a sequence (of words or sounds, respectively) is determined by a set of rules. However, whereas syntactic grammars typically require a high level of computational complexity that includes recursivity (Hauser et al., 2002), phonological grammars appear maximally regular (Kaplan and Kay, 1994). Similarly, while syntactic units of analysis are, in general, independent from the physical form they take whether signed, spoken, or written, units of phonological analysis interface directly with the physical signed or spoken linguistic signal. These are key advantages of phonology in the study of the neurobiology of language. Since all human languages must share some grammatical properties in common,[4] it thus follows that, in determining the phonological grammars of individual languages, core properties of all human languages are also determined. And in this way, aspects of phonology that are particular to specific languages, such as phonological alternations, shed light on core properties of the language system.

The remainder of this chapter provides an overview of what is currently known about how different aspects of phonology are instantiated in the brain. Section 1.2 considers phonological features and their relationship to the sensorimotor periphery. Section 1.3 discusses defining properties of segmental phonemes and evidence for the brain's sensitivity to those properties. Section 1.4 discusses suprasegmental units and lexical stress and tone in particular. Section 1.5 reviews how the neural response to speech is sensitive to different contextual domains, and Section 1.6 provides background for what is currently known about mechanisms for allophone

[4] in order to be learnable by, in principle, any infant

selection in the brain. Finally, Section 1.7 describes how phonological processing integrates with higher order grammatical structure through morphology, and Section 1.8 concludes.

1.2 Phonological features

In phonological theory since the late 19th century, phonemes of speech have been analysed as compositions of more basic units, called features, that are articulatory and acoustic in nature. Each phoneme in a language is defined by a set of typically binarily specified[5] (i.e., +/−) features that distinguishes it from all other phonemes in the language, and shared feature values define *natural classes* that are used to capture the observation that phonetically natural sets of sounds pattern together. For example, all and only the sounds [p], [t], and [k] are [−voice, −continuant] in English, and all and only these sounds are consistently aspirated when they occur in the initial position of stressed syllables. If English speakers suddenly adopted the use of the voiceless uvular stop [q], a sound that is also [−voice, −continuant], we would predict that it would also undergo aspiration in this context because it shares the features that motivate the pattern. In this way, the use of phonetically-motivated features makes strong predictions about the role of phonetic naturalness in phonology.

Because of the simplicity, predictive power, and broad empirical coverage afforded by the use of natural classes, feature-based representations of phonemes are generally uncontested within phonological theory (though see Silverman 2006, 2012). However, determining the neural mechanisms responsible for the explanatory power of feature-based representations in theoretical phonology remains an open area of investigation. Receptive fields estimated from intracranial recordings indicate that the spectrotemporal tuning of superior temporal gyrus (STG) is spatially organized such that sounds with a higher degree of temporal variation preferentially activate sites in posterior STG, while sounds with finer spectral structure and slower-varying temporal dynamics preferentially activate more anterior

[5]See Hall (2007) for discussion of alternatives and exceptions.

sites (Hullett et al., 2016). Accordingly, sounds that share common spectrotemporal features activate sites that are selective for those features (Mesgarani et al., 2014). Together, these results show that the brain tracks speech acoustics in a way that supports the behavior of sounds that share acoustic features behaving as a class. However, further work remains necessary to whether this activity drives phonological phenomena that are known to rely on natural class relatedness.

Additional work exploring the brain basis of phonological features makes use of the Mismatch Negativity (MMN) response. The MMN is a pre-attentive, auditory event-related potential, elicited by violations of acoustic regularity, most commonly in passive oddball experimental paradigms. In these experiments, participants listen to a monotonous series of sounds of one type, called the standard, that is occasionally interrupted by a sound of a different type, called the deviant. The difference in evoked response between the standard and deviant stimuli appears as a negative waveform broadly distributed over the top of the head that peaks 100–200ms after stimulus presentation. Its magnitude is sensitive but not reducible to the acoustic difference between the standard and deviant stimulus (Näätänen and Alho, 1997; Näätänen et al., 2007; Winkler, 2007) and, it can be modulated by stimulus lexicality (Pulvermüller et al., 2001, 2004; Pettigrew et al., 2004; Shtyrov et al., 2005), frequency (Alexandrov et al., 2011), context (Jacobsen et al., 2004, 2005), and familiarity (Pulvermüller et al., 2001; Dehaene-Lambertz et al., 2000; Näätänen et al., 1997; Peltola et al., 2003; Sharma and Dorman, 2000; Winkler et al., 1999; Bonte et al., 2005).

For these reasons, the MMN has been used to investigate various questions about phonological processing in the brain (Näätänen, 2001). Early studies of the MMN indicate that it is dependent on specific language experience, reflecting knowledge of language-specific phonological categories (Näätänen et al., 1997; Dehaene-Lambertz, 1997; Cheour et al., 1998; Winkler et al., 1999) and phonotactics (Dehaene-Lambertz et al., 2000). These responses are robust to subphonemic acoustic variation (Rhodes et al., 2019; Shestakova et al., 2002) and reflect phonological learning, even in adults (Barrios et al., 2016).

Given these properties of the MMN, many researchers have used the MMN to

investigate the brain's sensitivity to various language-specific featural contrasts. These studies have observed asymmetries in the elicitation of the MMN that suggest that not all speech sounds are processed equally (Cummings et al., 2017; Scharinger et al., 2012a; Friedrich et al., 2008; Scharinger et al., 2012b; Eulitz and Lahiri, 2004; Hestvik and Durvasula, 2016; Scharinger et al., 2016). For example, some studies appear to provide neurological evidence for the theoretical claim that coronal sounds are unspecified for place (Avery and Rice, 1989). These studies show that when the standard is a non-coronal sound like [p, b, k,] or [g], and the deviant is a coronal sound like [t] or [d], the deviant elicits a large mismatch response. However, when the standard is coronal and the deviant is non-coronal, no MMN occurs (Cummings et al., 2017; Scharinger et al., 2012a; Friedrich et al., 2008). Thus, in theoretical terms, when the standard and deviant differ by one phonological feature, and the standard is unspecified for that feature, then the deviant does not elicit a mismatch response. Conversely, when the standard and deviant differ by one feature, and the deviant is unspecified for that feature, the deviant elicits a mismatch response. By this reasoning, the MMN has been used to argue for underspecification of voicing (Hestvik and Durvasula, 2016), vowel height (Scharinger et al., 2012b), and rounding (Eulitz and Lahiri, 2004), in addition to coronality (though c.f., Scharinger et al. 2011; Rivera-Gaxiola et al. 2000a,b).

1.3 Phonemes

Other work focuses at the level of the phoneme. Representationally, phonemes are sets of features that describe a single member of a language's sound inventory. For example, while [−voice, −continuant] describes the natural class of phonemes /p, t, k/, the feature set [−voice, −continuant, LABIAL] describes all and only the sound [p] in English. In this way, representationally, a phoneme is a natural class with a cardinality of one. Conceptually, however, three inter-related observations establish phonemes as a distinct level of linguistic analysis, worthy of study in the brain independently from features: 1) phonemic contrast is not sensitive to the details of featural differences, 2) the content of phonemes is fundamentally

abstract, and 3) phonemic contrasts are specific to particular languages.

First, phonemic contrast is such that two phonemes contrast with one another regardless of the precise featural differences between them. For example, 'clash' [klæʃ] and 'class' [klæs] are equally distinct words as 'clash' [klæʃ] and 'clap' [klæp]. From a synchronic perspective, the type or number of featural differences does not impact whether that difference can support a semantic contrast: either there is a featural difference between two sounds and a contrast in meaning is possible, or there is no featural difference, and differences in acoustic realization cannot support a difference in meaning. Though /s/ and /ʃ/ differ in only one feature ([±anterior]) while /ʃ/ and /p/ differ in both place and manner features, 'clash' [klæʃ] and 'class' [klæs] are no less different words than 'clash' [klæʃ] and 'clap' [klæp]. In this sense, though type and number of featural differences can impact aspects of language use, such as perceptual confusability, likelihood of sound change, or probability of speech errors, they do not impact synchronic contrast itself. In this way, the phonemic level exists alongside the featural level as an additional form of nonlinearity, buttressing linguistic categoricity against the infinite variability of speech acoustics.

Much work on speech sound representation in the brain has focused on this characteristic of phonemes. Often ignoring the precise featural differences between sound categories, these studies examine how neural activity distinguishes each sound category from all others. By regressing neural data on categorical speech sound labels, such work has taken steps towards decoding phoneme-level representations from the neural response to speech. Taking advantage of advances in machine learning techniques, contrasts between vowels and consonants (Pei et al., 2011) and English phonemes (Blakely et al., 2008; Brumberg et al., 2011; Mugler et al., 2014) have been decoded from intracranial neural activity. While this kind of work has proven particularly useful for clinical, brain-machine interface (BMI) applications (see Wolpaw et al. 2002 and Martin et al. 2018 for reviews), it generally does not address the second and third properties of phonemes that distinguish them from other levels of linguistic analysis. Namely, phonemes are fundamentally abstract representations that are specific to particular languages.

These two properties can be illustrated by a brief comparison of Yiddish and German.[6] Yiddish and German are closely related descendants of Middle High German, a West Germanic Indo-European language spoken in the High Middle Ages (c. 1050-1500) Their phonemic inventories both include paired series of voiced and voiceless stop consonants. In Yiddish, as shown in (1), both voiced and voiceless sounds occur at the ends of words.

(1) Yiddish (via Katz 1987, pgs. 29-31)

a.	[kalb]	'calf'	[kelbər]	'calves'
b.	[klap]	'strike'	[klapən]	'to strike'
c.	[rod]	'wheel'	[redən]	'wheels'
d.	[rot]	'council'	[rotən]	'councils'
e.	[folg]	'follow'	[folgən]	'to follow'
f.	[folk]	'people'	[felkər]	'peoples'

In contrast, as shown in (2), only voiceless obstruents occur word finally in German.[7] Voiced obstruents occur only in non-final positions.

(2) German

a.	[kalp]	'calf-NOM'	[kalbəs]	'calf-GEN'
b.	[klap]	'fold up-2S.IMP'	[klapə]	'fold up-1S'
c.	[ʁat]	'wheel-NOM'	[ʁadəs]	'wheel-GEN'
d.	[ʁat]	'council-NOM'	[ʁatəs]	'council-GEN'
e.	[folk]	'follow-2S.IMP'	[folgə]	'follow-1S'
f.	[folk]	'people-NOM'	[folkəs]	'people-GEN'

A general constraint prohibiting voiced obstruents of German from appearing in word-final position could explain the restricted distribution of voiced obstruents. However, if one posits only this general constraint, a puzzle remains: Consider the pairs of words given in (2c) and (2d). Generative approaches assume that

[6]This section describes Standard Yiddish, as presented in Lombardi (1999) and Katz (1987), and Standard colloquial German (p.c., Robert Kluender and Eric Meinhardt).

[7]Pre-obstruent vowels are produced 8.6ms (SE=2.04ms) longer before voiceless obstruents that alternate with voiced obstruents than before those that do not (Roettger et al., 2014). However, it remains controversial whether this acoustic cue is used during speech perception. This issue is discussed in more depth in Section 2.2.3.

the meaning 'wheel' has a single form and that nominative and genitive forms of that noun are derived through combination of the form meaning 'wheel' with forms meaning 'nominative' or 'genitive', respectively. From the word 'council', the nominative case appears to be unmarked (i.e., the meaning 'nominative' has no phonological form), and the form corresponding to the genitive appears to be [əs]. This leaves [ʁat] to be the form meaning 'council'. However, when the same reasoning is applied to 'wheel', two forms are left to mean 'wheel'— [ʁat] and [ʁad].

To decide which of these forms is the single, underlying form of 'wheel', we note that the nominative forms of 'wheel' and 'council' are homophonous,[8] but their genitive forms are not. Thus, while the nominative forms can be predicted from the genitive, the converse is not true. For this reason, we select /ʁad/ as the underlying form of 'wheel', and with the support of further data such as that shown in (2b) and (2e), we posit that the form of the nominative is derived by the devoicing of word-final obstruents. In this way, we conclude that the lack of word-final obstruents in German is not a purely static distributional requirement, but is maintained through the active application of a phonological process.

From these data, we also conclude—at a theoretical level—that an abstract, phonemic level of representation underlies the surface, acoustic forms that speech sounds take. As shown in (2c) and (2d), the phoneme /d/ is realized as [d] in word-medial contexts and [t] in word-final contexts, while the phoneme /t/ is realized as [t] in both word-medial and word-final contexts. Although the acoustic contrast between /d/ and /t/ is neutralized at the ends of words in German, it is maintained word-medially. In this sense, phonemes are defined through the system of acoustic forms (=allophones) that they take on in different phonological contexts.

Furthermore, the fact that this pattern of devoicing occurs in German but not in closely-related Yiddish suggests that phonemes are specific to particular languages and not artifactual of universal constraints on articulation or perception. Whereas phonological features such as [±voice] operate identically in both Yiddish and

[8]Although small consistent acoustic differences exist between these two words, they are too small to reliably distinguish between words during perception Roettger et al. (2014). Accordingly, (2c) and (2d) are have been transcribed identically. See also previous footnote.

German, with [+voice] attributed to all voiced sounds and [−voice] attributed to all voiceless sounds, voiced *phonemes* operate differently in Yiddish and German: in German, they devoice word finally, and in Yiddish, the do not. Thus, phonemes conceptually enable description and analysis of the intricate networks of sound patterns that underpin spoken languages in a way that phonological features alone do not.

A number of studies have assessed the extent to which the language-specific organization of allophones into phonemes impacts speech sound processing in the brain. While changes in sound sequences typically evoke a robust mismatch response, sound changes that are expected due to language-specific phonological context—such as the Russian consonant cluster simplification of [astna] to [asna]—fail to elicit a reliable MMN (Kharlamov et al., 2011). Conversely, when speakers of a language in which two sounds are allophones of the same phoneme (Spanish: [d], [ð]) are proficient in a second language in which those two sounds are allophones of different phonemes (English), they have a more reliable mismatch response to those two sounds than monolingual speakers of their first language (Barrios et al., 2016). A similar result has also obtained for Hungarian learners of Finnish with the vowels [æ] and [e], which contrast in Finnish, but not Hungarian (Winkler et al., 1999).

Such results concord with the core properties of phonemes enumerated in this section. Despite variability in the number and types of phonological features that distinguish the sounds involved, MMN responses are unreliably evoked when the deviant is an allophone of the same phoneme as the standard (e.g., [d] vs. [ð] (Spanish), and [æ] and [e] (Hungarian)), yet reliably evoked when the deviant is an allophone of a phoneme different from that of the standard (e.g., [d] vs. [ð] (English), and [æ] and [e] (Finnish)). Moreover, these results are specific to participants' experience with particular languages, with monolingual speakers of different languages exhibiting qualitatively different responses than one another. Together, these findings support the idea that neural responses to speech sounds are organized not only at an acoustic level but also a phonemic one.

The processes underpinning these results are likely complex and coordinated

across multiple regions. At a broad scale, both inferior frontal and superior temporal regions are implicated in the maintenance of phonemic contrasts during perception. Inferior frontal regions have been most strongly associated with properties of phonological processing that could be shared in common with other aspects of language and cognition, such as category formation and maintenance (Hutchison et al., 2008; Lee et al., 2012; Myers, 2007; Myers et al., 2009). In contrast, activity in superior temporal regions appear to support the processing of a range of complex auditory properties, some of which may contribute to the language-specificity and abstractness of phonemes, and some of which may be language general (Tremblay et al., 2013; Hullett et al., 2016; Mesgarani et al., 2014). Altogether, the precise role each lobe plays remains an issue of investigation and debate.

Activity in superior temporal regions is sensitive to familiarity and carries information that can be used to decode category membership. In particular, Liebenthal et al. (2005) show that bilateral anterior and middle superior temporal sulcus (STS) demonstrate stronger blood oxygenation level dependent (BOLD) response to categorization along acoustic continua between familiar phonemic speech sounds than for unfamiliar speech sound continua. Furthermore, once participants have trained on an unfamiliar continuum, categorization along that continuum also elicits greater BOLD response in STS (Liebenthal et al., 2010), further demonstrating that this region is sensitive to familiarity. Therefore, this area may play a role in phonological processing that requires experience with particular languages (=specific language knowledge).

Superior temporal gyrus (STG) and the supratemporal plane also carry complex acoustic information that could support phoneme-like activity. Several areas of the supratemporal plane demonstrate preference for speech over non-speech sounds (Tremblay et al., 2013) and sensitivity to linguistically relevant acoustic features of speech, such as the formant frequencies of vowels (Formisano et al., 2008; Obleser et al., 2003, 2006) and the spectral composition of consonants (Obleser et al., 2007; Raizada and Poldrack, 2007). Some areas of the supratemporal plane are even sensitive to degree of sequential statistical structure (i.e., transitional probabilities, Tremblay et al. 2013), a property that could be critical for the con-

textual conditioning of allophones.

Similarly, STG exhibits sensitivity to a variety of complex characteristics of acoustic stimuli. These characteristics include preference for intelligible speech (Binder et al., 2000; Price, 2012), entrainment to syllable rate (Giraud and Poeppel, 2012; Di Liberto et al., 2015), and sensitivity to phonological ambiguity (Gwilliams et al., 2018; Leonard et al., 2016). Moreover, both fMRI and intracranial data from this region carry information capable of decoding sound category identity (Chang et al., 2010; Arsenault and Buchsbaum, 2015). Altogether, these characteristics of superior temporal regions appear highly relevant to various aspects of phonological processing in ways that could support the emergence of the categoricity and language specificity that are hallmarks of phonemes.

Inferior frontal regions also appear sensitive to speech sound categoricity and display behavior consistent with category formation and maintenance. Presenting participants with open syllables varying in voice onset time (VOT), Myers et al. (2009) show that cross-category stimuli (/da/ vs. /ta/) disrupt adaptation effects in inferior frontal regions, whereas effects in superior temporal regions are gradient across acoustic continua. Similarly, when participants are presented with open syllables varying in the central frequency of their formant transitions, BOLD activation in IFG exhibits distinct patterns for stimuli classified as /ba/ and those classified as /da/ (Lee et al., 2012). Such results suggest that IFG participates in speech sound categorization, a role which may also support the maintenance of phonemic contrasts. Furthermore, Liebenthal et al. (2010) also found that IFG exhibit stronger BOLD response after training participants in categorization along an unfamiliar speech sound continuum, suggesting that activity in this area may also support category formation.

1.4 Suprasegmentals

Many aspects of phonology as it is studied in the brain rely on the assumption that phonological units (i.e., features, phonemes) are temporally segregable. That is, most neurolinguistic work on phonology concerns itself with *segmental*

phonemes. However, not all phonemes are segmental and temporally segregable from other phonemes. Phonemic elements whose primary acoustic cues include changes in f0, amplitude, and duration—such as lexical stress, phrasal stress, lexical tone, and intonational phrases—cannot be temporally segregated from consonants and vowels. Such phonemic units are considered *suprasegmental* because their acoustic realization occurs simultaneously with (i.e., *supra*, "on top of") that of segmental units such as vowels and consonants. This section focuses on lexical suprasegmentals.

Although lexical suprasegmentals, such as stress and tone, are acoustically realized simultaneously with vowels or consonants, they are nevertheless independent of segmental units. For example, the difference in quality, length, and loudness of the first vowel in the English words *atom* and *atomic* is best explained by the relative presence or absence of stress, rather than by a fundamental difference in the segments that constitute the lexeme *atom*. Similarly, in Mandarin Chinese, differences in pitch and voice quality on the word 你 /niˬ˩˧/ 'you' in phrases like 你喊 [niˬ˩˧ hanˬ˩˧] 'you shout' versus phrases like 你笑 [niˬ˩˧ ɕiao˧˥] 'you laugh' are best explained by an (allophonic) change in tone identity rather than segment identity.

Furthermore, just as segmental phonemes are defined through the system of acoustic forms they take on in different phonological contexts, suprasegmental phonemes are also defined through the systems of contrasts they participate in. For example, returning to English lexical stress, the location of stress within a word is sufficient to support a difference in meaning: *Missouri* /mɪˈzəɹi/ and *misery* /ˈmɪzəɹi/ differ only in the placement of stress yet bear clearly distinct meanings. In this way, differences in stress placement are contrastive in the same way that differences in segments like /t/ and /d/ are contrastive in English words like *taught* and *dot*. Moreover, just as the acoustic realizations of segments like /t/ or /d/ are sensitive to phonological context (recall word-final devoicing in German), the location of stress on a lexeme may also shift depending on its phonological context, as illustrated by the difference in stress placement on words like *atom* and *atomic*, *meteor* and *meteoric*, etc.

The same properties are true of lexical tone. Returning to Mandarin, tone

identity is sufficient to support a difference in meaning: 蓝 /lan˥/ 'blue' and 懒 /lan˨˩/ 'lazy' differ only in the tone they carry yet have clearly distinct meanings. In this sense, differences in tone are contrastive just as differences in segments are. Furthermore, tones are also subject to contextually-conditioned alternations. In a pattern called T3 sandhi, the contrast between the falling-rising tone (T3) of 懒 /lan˨˩/ 'lazy' and the rising tone (T2) of 蓝 /lan˧˥/ 'blue' is neutralized when T3 is followed by another T3. That is, when a word like 懒 /lan˨˩/ 'lazy' combines with a word like 鬼 /gui˨˩/ 'ghost', it produces 懒鬼 [lan˧˥ gui˨˩] 'slacker' (lit. 'lazy ghost'), which is homophonous with 蓝鬼 [lan˧˥ gui˨˩] 'blue ghost'.

These examples showcase how suprasegmentals behave independently of the segmental level and are subject to their own structured patterns of alternation and neutralization. In these ways, suprasegmentals are phonemic, yet the fact that they are acoustically realized simultaneously with segmental phonemes both raises fundamental questions about how they are processed during comprehension and can make them challenging to study in the brain. The following sections review what is currently known about the processing of lexical stress and tone.

1.4.1 Stress

Some studies address how the acoustic correlates of stress are processed as stress, while others address speaker knowledge about how stress patterns within words of the language.

In an AX task asking German speakers to determine whether pairs of bisyllabic pseudowords were the same or different, pairs differing only in stress elicited an increase in broad language network activation over the activation elicited by pairs of words that differed in vowel quality (Klein et al., 2011). Similarly broad activation was observed in Dutch speakers who were asked to determine whether bisyllabic words they heard were 'weak-initial' or 'strong-initial' (Aleman et al., 2005). These results may reflect that stress is the product of broadly distributed language network activity, or they may reflect increased effort, since subjects in both studies performed less accurately on stress perception tasks than control tasks.

Stress perception, especially for words with regular stress, is often difficult

for speakers to consciously monitor, and this difficulty appears to be reflected in patterns of electrophysiological responses to stress violations as well. For example, as studied in Domahs et al. (2012), trisyllabic words in Polish have regular, penultimate stress with relatively few exceptions, and speakers perform poorly at perceiving stress shifts away from the correct position, achieving only $56 - 68\%$ accuracy at identifying stress violations. Speakers had the greatest accuracy at determining when words were incorrectly stressed on the initial syllable. Regardless of whether the correct and expected stress for these words would have been regular penultimate stress or irregular antepenultimate stress, forms with initial stress evoked a positivity 210-560ms after word onset. On the other hand, presentation of stress violations where regular penultimate stress had been shifted to the antepenult evoked a reliable early negativity (170-430ms) followed by a late positivity (620-1050ms), and stress violations where irregular antepenultimate stress has been shifted to the penult evoked only an early negativity (Domahs et al., 2012).

Similarly, as studied in Domahs et al. (2013a), Turkish has regular final stress with the exception of a relatively small, semantically restricted set of words that have penultimate stress. Regardless of whether the correct and expected stress for words would have been regular final stress or irregular penultimate stress, forms with incorrect initial stress elicited a positivity 400-700ms post word onset. Presentation of stress violations where regular final stress had been shifted to the penult evoked a positivity 850-1100ms after word onset, and violations where exceptional penultimate stress had been shifted to the final syllable produced a negativity 500-750ms post word onset.

These patterns seem to suggest that deviations *from* the regular are processed differently than deviations *to* the regular, and that probable deviations—such as stress on the penult or antepenult for Polish and stress on the final or penult for Turkish—are processed differently than improbable deviations, such as initial stress in either Polish or Turkish. Results from Dutch seem to corroborate this interpretation: Böcker et al. (1999) found that when Dutch speakers were asked to discriminate between 'weak-initial' and 'strong-initial' words, 'weak-initial' words

elicited a negativity roughly 325ms after the onset of the word. Although task materials were controlled for word frequency, the asymmetric response to iambic, 'weak-initial' words may have been the result of a global imbalance of 'weak-initial' and 'strong-initial' words in the Dutch lexicon, since roughly 90% of the Dutch lexicon is 'strong-initial', based on counts from CELEX (Vroomen and De Gelder, 1995). Similar, although more complicated, results are found in Cairene Arabic as well (Domahs et al., 2014).

1.4.2 Tone

Studies of lexical tone have established that the brain responds differently to manipulations of f0 that cue lexical tone identity than to other linguistic manipulations of f0 (e.g., phrasal intonation, pitch declination). Additionally, neural response to lexical tone information also varies according to the linguistic function of tone in a particular language.

In a crosslinguistic positron emission tomography (PET) study with Thai, Mandarin, and English participants, Gandour et al. (2000) show that Thai-like tone patterns elicit activation over baseline in left IFG for Thai speakers only. Despite familiarity with lexical tone from Mandarin, Mandarin speakers did not exhibit increased activation in this region relative to a non-linguistic pitch baseline, nor did English speakers. From these results, the authors conclude that linguistic pitch patterns are processed in a top-down manner that is sensitive to particular language knowledge. Additionally, for Thai speakers, these results demonstrate that linguistic manipulations of f0 are processed differently than non-linguistic f0 manipulations. Similar results were also observed in fMRI (Gandour et al., 2002).

Neural responses are also sensitive to the language-specific systems of contrast and neutralization that lexical tones participate in. As introduced above, in Mandarin T3 sandhi, falling-rising tones (T3) are realized as rising tones (T2) when they occur before another falling-rising tone (e.g., 懒 /lan↓↗/ + 鬼 /gui↓↗/ ⟶ 懒鬼 [lan↗ gui↓↗]). When Mandarin speakers listen to a string of T3 syllables interrupted by a T2 syllable, the MMN is attenuated relative to when they listen to a string of T3 syllables interrupted by a syllable with a tone that does not participate

in sandhi with the falling-rising T3 tone (Chandrasekaran et al., 2007a,b; Cheng et al., 2013; Hsu et al., 2014; Li et al., 2016; Li and Chen, 2015; Chang et al., 2019). This effect is explained as a product of T3 sandhi reducing the perceptual dissimilarity between rising and falling-rising tones. That is, while the rising tone alternates with the falling-rising tone in Mandarin, other tones do not alternate with the falling-rising tone. Thus, by virtue of their alternation during T3 sandhi, rising and falling-rising tones are phonologically more similar to one another.

Similar attenuation of the MMN response has been observed for a sandhi pattern in Southern Min. In Southern Min, tones participate in a complex sandhi pattern where each tone alternates with a different tone when in occurs in a non-final position within a word. For example, T7 is a mid-level tone that surfaces as mid-falling tone T3 when it occurs in a non-final position. When a series of T7 syllables is interrupted by a T3 syllable, the MMN is attenuated relative to when a series of T7 syllables is interrupted by a syllable with a tone that it does not alternate with (Chang et al., 2019). Moreover, Chang et al. (2019) observe this attenuation only in skilled speakers of Southern Min, suggesting that knowledge of Southern Min drives the effect, rather than acoustic features of the tones themselves.

When listeners are presented with sentences containing a semantic error caused by either a tonal change or a segmental change, the neural response is similar. That is, when hearing a Cantonese sentence such as "...he sees the doctor to treat his ___", if the gap is filled with /beŋ˧˥/ 'biscuit' or /bou˨/ 'step' instead of /beŋ˨/ 'illness', the N400 response elicited by the segmental error is similar in amplitude and latency to that elicited by the tonal error (Schirmer et al., 2005). From this result, Schirmer et al. (2005) concluded that both segmental and suprasegmental phonemic information are accessed at roughly the same time during word recognition. In a similar manipulation with Mandarin, tonal errors were found to elicit an earlier and more left-lateralized negativity than segmental errors (Malins and Joanisse, 2012). However, segmental errors were more likely to phased later than tonal errors in this study (i.e., changes to the the second vowel targets of diphthongs or changes to coda consonants), so the difference in latency between these

two conditions may be a consequence of this difference. At the very least, as Malins and Joanisse (2012) concluded, this result suggests that tone information is not a weaker cue to word recognition than segmental information.

Altogether, electrophysiological studies indicate that the brain is sensitive to lexical tone in ways comparable to its treatment of other phonemic information. To-date, studies exploring the neural basis of lexical tone have been heavily biased towards Sinitic languages and Mandarin in particular. While there are clear experimental advantages to studying well-documented languages with large speaker populations, conclusions about the neural treatment of lexical tone would be strengthened by study of a broader sample of languages in which differences in f0 are phonemic. For example, are lexical and morphological tone processed similarly in a language like Akan? How are tone melodies processed in a language like Mende? Are the processing consequences of sandhi patterns in Sinitic languages similar to those of Otomanguean languages? Answering these and other fundamentally crosslinguistic questions will undoubtedly strengthen our understanding of how the brain integrates f0 information during spoken language processing.

1.5 Phonological domains

Many phonological processes are constrained or conditioned by factors other than their immediate phonological surroundings. Such processes provide evidence for the existence of abstract, language-specific domains above the level of individual sounds or articulatory gestures. For example, the adaptation of English loanwords into Hawaiian provides evidence for the existence and structure of the syllable. In English, the sequence of sounds [ˈmɛɹi ˈkɹɪsməs] is used as a greeting in late December: 'Merry Christmas'. However, several phonological properties of this phrase do not occur in Hawaiian, so when the greeting 'Merry Christmas' was borrowed into Hawaiian, it was adopted as the sequence [mele kalikimaka].

In Hawaiian, [ˈmɛɹi ˈkɹɪsməs] becomes [mele kalikimaka] for three reasons. First, the sound [ɹ] does not occur in Hawaiian; it is replaced with [l]. Second, the sound [s] does not occur in Hawaiian; it is replaced with [k]. And third, Hawaiian does

not permit consonants that are not immediately followed by a vowel, so a vowel is inserted between the clusters [kɹ] (or [kl]) and [sm] (or [km]) and after the final consonant of 'Christmas'. The generalization that Hawaiian does not allow closed syllables or complex onset clusters follows from this third generalization. In this sense, the phonotactic constraints of Hawaiian are sensitive to the word's syllable structure: while vowels are permitted both word-initially and word-finally, consonants are not permitted word-finally because syllables that end with a consonant are disallowed in Hawaiian. Appealing to the syllable in this way also allows the analyst to explain why sequences of consonants are also not permitted in the language: they would result in impermissible syllable structures.

In addition to phonotactics, other phonological phenomena that are sensitive to the syllable domain include lexical stress, Arabic emphasis spread, reduplication, and language games like Pig Latin in English or Verlans in French. Of these, lexical stress and its interaction with higher prosodic domains, such as the foot, has been studied most extensively in the brain. For example, German words are composed of trochaic feet constructed from the right word edge (Féry, 1998; Giegerich, 1985; Vennemann, 1990). Regardless of footing, listening to incorrectly stressed words resulted in increased activation in bilateral superior temporal gyrus (STG), while correctly stressed words evoked greater activation in left posterior angular gyrus and right retrosplenial cortex. For incorrectly stressed words, this STG activation was primarily driven by stress violations that would require refooting, such as a word like [biˈkini] being stressed as [bikiˈni], whereas incorrectly stressed words that would not require refooting engaged bilateral inferior and middle frontal gyrus (including Broca's area) and supplementary motor area (Domahs et al., 2013b). Using similar task materials in EEG, Domahs et al. (2008) found that incorrectly stressed words evoked a late positivity, interpreted as a P300 effect, and that stress violations that would require refooting elicited positivities of greater magnitude than those that would not.

Other phonological phenomena take place over longer timescale domains, such as the word, phrase, or utterance. Some of these processes, such as pitch declination, are tempting to think of as physiologically deterministic and outside of

speaker control. After all, as air pressure in the lungs lowers over the course of an utterance, the frequency of vocal fold vibration and consequent vocal pitch have a tendency to fall. However, not only is the rate of pitch declination under speaker control, the way it is controlled varies crosslinguistically (Yuan and Liberman 2014; Cui and Kuang 2020, among others). Although the neural correlates of pitch declination specifically have not yet been explored, correlates of voluntary pitch control have been observed in English speakers (Dichter et al., 2018)., Often, the beginnings and endings of these domains are subject to distinctive phonological processes, such as phrase-initial strengthening or word-final devoicing. Some work suggests these domain edges have special, prominent representation in the brain (Hamilton et al., 2018).

1.6 Phonological processes

In generative phonological theory, these kinds of patterns are typically considered to be the result of a process whose inputs are abstract phonological units, such as feature bundles, and whose outputs are equivalent to attested spoken language at a certain level of generality.[9] Phonological theories can then be divided into two major types based on the form they propose that these processes take. Rule-based theories conceive of the mapping from phonological to phonetic space as the application of a series of maximally regular rules, and constraint-based theories conceive of this mapping as a constraint optimization problem where the optimal output form for a given input form is selected from a set of possible candidates based on the way in which it satisfies a set of linguistically-informed constraints.

A number of papers have attempted to provide evidence that the brain organizes its response to spoken language in a way that would be consistent with either rule- or constraint-based processes. For the French obstruent voicing assimilation pattern, Sun et al. (2015) observe a MMN response in left central electrodes 140-

[9]That is, the outputs of phonological processes are not in general specific enough to, say, generate a fully natural sounding speech signal, but are intended to be specific enough that a natural sounding speech signal could be derived from them through the application of some additional deterministic process that makes more explicit reference to articulators and their acoustics.

240ms after the deviant consonant followed by a late positive component (LPC) across parietal electrodes 360-520ms after the deviant consonant. A similar MMN-LPC complex has been observed in response to violations of Finnish vowel harmony (Ylinen et al., 2016). In a study on German stress shift, however, only an early negativity was evoked by phrases with ungrammatical stress clash, and only an LPC was evoked by phrases where the stress shift process had applied (Bohn et al., 2013). These studies are framed in terms of phonological rules, and deal with patterns that would otherwise require the interaction of more than one constraint, if they were to be framed from the perspective of a constraint-based theory. Instead, constraint-based neurolinguistic work tends to focus on the significance of patterns that showcase the relevance of single, often phonotactic, constraints. In studies focusing on phonotactic constraint violations, novel forms evoke an N400, and phonotactically illicit forms additionally elicit an LPC response. This pattern of responses has been observed for German OCP violations (Domahs et al., 2009), Finnish vowel co-occurrence restrictions (Aaltonen et al., 2008), and the violation of laboratory-learned artificial phonotactic constraints (Moore-Cantwell et al., 2018).

While conclusions about whether rules or constraints best characterize how the brain organizes its response to natural language phonology appear to be a ways off, this literature provides converging evidence for two important generalizations. First, neural responses are consistent with mapping from an abstract underlying form to a physical acoustic form, validating the way of thinking about phonology as a kind of mapping problem. And second, some of the neural components evoked by phonological manipulations—such as the N400 and LPC—are commonly evoked by syntactic and semantic manipulations as well. This observation suggests that some of the neural mechanisms supporting higher order linguistic organization (i.e., syntax, semantics) may support abstract language organization at the phonological level as well.

1.7 Morphophonology

One of the largest questions animating research in morphological processing concerns how the brain performs composition and decomposition of morphological structure during production and comprehension.

Work in this field has identified left lateralized cortical areas linked to compositional morphology through the study of both lesioned and healthy brains. Patients with left-hemisphere damage that includes inferior frontal regions demonstrate selective impairment in auditory immediate repetition priming for decompositional inflectional morphology (i.e., pairs like *join, joined*) but not irregular inflectional morphology (i.e., pairs like *find, found*) (Marslen-Wilson and Tyler, 1997; Tyler et al., 2002). Similarly, in healthy listeners asked to perform a semantic match-to-sample (AX) task, related pairs with regular, decompositional inflectional morphology elicit greater BOLD response in left IFG, left anterior cingulate, and bilateral STG compared to related pairs with irregular inflectional morphology (Marslen-Wilson and Tyler, 2007). Moreover, when participants are asked to produce regular past tense forms out loud from visually-presented present tense forms, left STG in particular generates a greater BOLD response for decomposable regular forms over non-decomposable irregular forms (Desai et al., 2006).

From these and other results, compositional processing for inflectional morphology is well established. However, whether the processing of derivational morphology involves composition remains a topic of some contention. Derived forms do not appear to be processed compositionally in fMRI studies of Polish (Bozic et al., 2013a) and English (Bozic et al., 2013b). But in EEG, using an ERP design, Havas et al. (2012) provide evidence from Spanish nominalization to suggest that recognition of derived word forms engages both word-level and decompositional processes. Similar engagement of both word-level and decompositional processes has also been observed in Finnish (Leminen et al., 2013).

To date, morphophonological studies generally subserve these larger questions around morphological structure and compositional processing. For example, several studies have looked at how morphological structure influences the predictability of subsequent phonemes. Ettinger et al. (2014) found that phoneme sequences

like [kɪl] predicted morphologically related forms like *killed* and *killer* more strongly than sequences like [bɪb] predicted words like *bourbon* and *burble*. Similarly, in languages with nonlinear (non-sequential) morphological structure, phoneme-level processing also seems to be influenced by morphological constituency. For example, in Modern Standard Arabic, the neural response to the /t/ in a word like /nabata/ 'grow' is better predicted by the preceding root consonants /nb/ than by the full preceding context /naba/ (Gwilliams and Marantz, 2015).

Furthermore, Marslen-Wilson and Tyler (2007) find that both regular pairs of present and past tense verbs (i.e., *join, joined*) and pseudo-regular pairs (i.e., *crey,creyed*) evoke greater BOLD response in left IFG and bilateral posterior STG than phonologically similar pairs that could not be interpreted as regular English morphology (i.e., *clay, claim* or *blay, blane*), indicating that these areas are sensitive to morphophonological cues to potential inflectional affixes. However, when compared against pairs of pseudo-words that could not be interpreted as past tense forms (i.e., *blay, blane*), past-like pseudo forms (i.e., *crey,creyed*) evoked greater BOLD response only in left IFG. This result suggests not only that IFG is sensitive to possible compositionality, but also that lexicality is critical to the differential response in STG observed for the comparison that included both real and pseudo-words (i.e., *play, played* and *crey, creyed* vs. *clay, claim* and *blay, blane*).

Additionally, misapplications of regular morphological rules have been shown to elicit an early left lateralized negativity in Catalan (Rodriguez-Fornells et al., 2001), German (Weyerts et al., 1997; Lück et al., 2006), English (Newman et al., 1999), and Italian (Gross et al., 1998), which has been hypothesized to index morphological decomposition. In morphological domains where L2 learners are highly proficient, this early negativity is also observed, but in domains where they are less proficient, L2 learners lack this negativity, suggesting that sensitivity to compositional morphological structure may be a sign of more efficient, expert language use in these generally agglutinative languages (Hahne et al., 2006).

1.8 Conclusion

This brief survey skims the surface of both phonology and the neuroscience of language. In doing so it aims to provide the reader with a sense of the vast and inchoate nature of the the intersection between these two areas of study. Broad treatment of phonological phenomena was prioritized, including discussion of phonological features, phonemes, suprasegmentals, phonological domains, phonological processes, and the interaction of phonology of with morphology. Although the total number of studies on phonology and the brain is large, coverage across these phenomena is nevertheless relatively thin, highlighting the vastness of the territory that phonology encompasses.

Moreover, from phenomenon to phenomenon many of the same brain areas and neural signatures recur frequently. In fMRI, phonological phenomena from phonemic categorization to morphophonological processing implicate both STG and IFG, and in EEG, the MMN response is sensitive to an equally broad range of phonological manipulations. These observations are both frustratingly and tantalizingly broad, providing ample evidence for "sensitivity" to numerous kinds of structures whose existences are motivated by phonological theory, but relatively little understanding of the neural mechanisms that could support instantiation of those structures.

This thesis leans into the tantalization of the literature. In some places, it provides further evidence for neural sensitivity to certain kinds of phonological and morphophonological structure (Chapter 2), in others it provides an overview of phonological sensitivity across a range of neural features (Chapter 3), and in yet others it provides evidence for the structure of the bridge between sensory and linguistic processing of speech sounds (Chapter 4). In doing so, this work contributes one chapter at a time to answering the questions posed in Section 1.1 of this chapter ("What is phonology?" And, "Why is it an interesting problem for neuroscience?") and hopefully models ways for future work to approach answering those questions too.

Bibliography

Aaltonen, O., Hellström, Å., Peltola, M. S., Savela, J., Tamminen, H., and Lehtola, H. (2008). Brain responses reveal hardwired detection of native-language rule violations. *Neuroscience letters*, 444(1):56–59.

Aleman, A., Formisano, E., Koppenhagen, H., Hagoort, P., De Haan, E. H., and Kahn, R. S. (2005). The functional neuroanatomy of metrical stress evaluation of perceived and imagined spoken words. *Cerebral Cortex*, 15(2):221–228.

Alexandrov, A. A., Boricheva, D. O., Pulvermüller, F., and Shtyrov, Y. (2011). Strength of word-specific neural memory traces assessed electrophysiologically. *PLoS one*, 6(8):e22999.

Arsenault, J. S. and Buchsbaum, B. R. (2015). Distributed neural representations of phonological features during speech perception. *Journal of Neuroscience*, 35(2):634–642.

Avery, P. and Rice, K. (1989). Segment structure and coronal underspecification. *Phonology*, 6(2):179–200.

Barrios, S. L., Namyst, A. M., Lau, E. F., Feldman, N. H., and Idsardi, W. J. (2016). Establishing new mappings between familiar phones: Neural and behavioral evidence for early automatic processing of nonnative contrasts. *Frontiers in Psychology*, page 995.

Binder, J. R., Frost, J. A., Hammeke, T. A., Bellgowan, P. S., Springer, J. A., Kaufman, J. N., and Possing, E. T. (2000). Human temporal lobe activation by speech and nonspeech sounds. *Cerebral cortex*, 10(5):512–528.

Blakely, T., Miller, K. J., Rao, R. P., Holmes, M. D., and Ojemann, J. G. (2008). Localization and classification of phonemes using high spatial resolution electrocorticography (ECoG) grids. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4964–4967. IEEE.

Böcker, K. B., Bastiaansen, M. C., Vroomen, J., Brunia, C. H., and De Gelder, B. (1999). An ERP correlate of metrical stress in spoken word recognition. *Psychophysiology*, 36(6):706–720.

Bohn, K., Knaus, J., Wiese, R., and Domahs, U. (2013). The influence of rhythmic (ir)regularities on speech processing: Evidence from an ERP study on German phrases. *Neuropsychologia*, 51(4):760–771.

Bonte, M. L., Mitterer, H., Zellagui, N., Poelmans, H., and Blomert, L. (2005). Auditory cortical tuning to statistical regularities in phonology. *Clinical Neurophysiology*, 116(12):2765–2774.

Bozic, M., Szlachta, Z., and Marslen-Wilson, W. D. (2013a). Cross-linguistic parallels in processing derivational morphology: Evidence from Polish. *Brain and language*, 127(3):533–538.

Bozic, M., Tyler, L. K., Su, L., Wingfield, C., and Marslen-Wilson, W. D. (2013b). Neurobiological systems for lexical representation and analysis in English. *Journal of Cognitive Neuroscience*, 25(10):1678–1691.

Brumberg, J. S., Wright, E. J., Andreasen, D. S., Guenther, F. H., and Kennedy, P. R. (2011). Classification of intended phoneme production from chronic intracortical microelectrode recordings in speech motor cortex. *Frontiers in neuroscience*, 5:65.

Chandrasekaran, B., Gandour, J. T., and Krishnan, A. (2007a). Neuroplasticity in the processing of pitch dimensions: A multidimensional scaling analysis of the mismatch negativity. *Restorative neurology and neuroscience*, 25(3-4):195–210.

Chandrasekaran, B., Krishnan, A., and Gandour, J. T. (2007b). Mismatch negativity to pitch contours is influenced by language experience. *Brain research*, 1128:148–156.

Chang, C. H., Lin, T.-H., and Kuo, W.-J. (2019). Does phonological rule of tone substitution modulate mismatch negativity? *Journal of Neurolinguistics*, 51:63–75.

Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., and Knight, R. T. (2010). Categorical speech representation in human superior temporal gyrus. *Nature neuroscience*, 13(11):1428–1432.

Cheng, Y.-Y., Wu, H.-C., Tzeng, Y.-L., Yang, M.-T., Zhao, L.-L., and Lee, C.-Y. (2013). The development of mismatch responses to Mandarin lexical tones in early infancy. *Developmental neuropsychology*, 38(5):281–300.

Cheour, M., Ceponiene, R., Lehtokoski, A., Luuk, A., Allik, J., Alho, K., and Näätänen, R. (1998). Development of language-specific phoneme representations in the infant brain. *Nature neuroscience*, 1(5):351–353.

Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on information theory*, 2(3):113–124.

Chomsky, N. (2009). Syntactic structures. In *Syntactic Structures*. De Gruyter Mouton.

Cui, P. and Kuang, J. (2020). Pitch declination and final lowering in Northeastern Mandarin. In *INTERSPEECH*, pages 1928–1932.

Cummings, A., Madden, J., and Hefta, K. (2017). Converging evidence for [coronal] underspecification in English-speaking adults. *Journal of neurolinguistics*, 44:147–162.

Dehaene-Lambertz, G. (1997). Electrophysiological correlates of categorical phoneme perception in adults. *NeuroReport*, 8(4):919–924.

Dehaene-Lambertz, G., Dupoux, E., and Gout, A. (2000). Electrophysiological correlates of phonological processing: a cross-linguistic study. *Journal of cognitive neuroscience*, 12(4):635–647.

Desai, R., Conant, L. L., Waldron, E., and Binder, J. R. (2006). fMRI of past tense processing: the effects of phonological complexity and task difficulty. *Journal of Cognitive Neuroscience*, 18(2):278–297.

Di Liberto, G. M., O' Sullivan, J. A., and Lalor, E. C. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current Biology*, 25(19):2457–2465.

Dichter, B. K., Breshears, J. D., Leonard, M. K., and Chang, E. F. (2018). The control of vocal pitch in human laryngeal motor cortex. *Cell*, 174(1):21–31.

Domahs, U., Genc, S., Knaus, J., Wiese, R., and Kabak, B. (2013a). Processing (un-)predictable word stress: ERP evidence from Turkish. *Language and Cognitive Processes*, 28(3):335–354.

Domahs, U., Kehrein, W., Knaus, J., Wiese, R., and Schlesewsky, M. (2009). Event-related potentials reflecting the processing of phonological constraint violations. *Language and Speech*, 52(4):415–435.

Domahs, U., Klein, E., Huber, W., and Domahs, F. (2013b). Good, bad and ugly word stress–fMRI evidence for foot structure driven processing of prosodic violations. *Brain and language*, 125(3):272–282.

Domahs, U., Knaus, J., Orzechowska, P., and Wiese, R. (2012). Stress "deafness" in a language with fixed word stress: An ERP study on Polish. *Frontiers in Psychology*, 3:439.

Domahs, U., Knaus, J. A., El Shanawany, H., and Wiese, R. (2014). The role of predictability and structure in word stress processing: an ERP study on Cairene Arabic and a cross-linguistic comparison. *Frontiers in Psychology*, 5:1151.

Domahs, U., Wiese, R., Bornkessel-Schlesewsky, I., and Schlesewsky, M. (2008). The processing of German word stress: Evidence for the prosodic hierarchy. *Phonology*, 25(1):1–36.

Ettinger, A., Linzen, T., and Marantz, A. (2014). The role of morphology in phoneme prediction: Evidence from MEG. *Brain and language*, 129:14–23.

Eulitz, C. and Lahiri, A. (2004). Neurobiological evidence for abstract phonological representations in the mental lexicon during speech recognition. *Journal of cognitive neuroscience*, 16(4):577–583.

Féry, C. (1998). German word stress in Optimality Theory. *Journal of Comparative Germanic Linguistics*, 2(2):101–142.

Fodor, J. A. (1975). *The language of thought*, volume 5. Harvard University press.

Formisano, E., De Martino, F., Bonte, M., and Goebel, R. (2008). "Who" is saying "what"? brain-based decoding of human voice and speech. *Science*, 322(5903):970–973.

Friedrich, C. K., Lahiri, A., and Eulitz, C. (2008). Neurophysiological evidence for underspecified lexical representations: asymmetries with word initial variations. *Journal of Experimental Psychology: Human Perception and Performance*, 34(6):1545.

Gandour, J., Wong, D., Hsieh, L., Weinzapfel, B., Lancker, D. V., and Hutchins, G. D. (2000). A crosslinguistic PET study of tone perception. *Journal of cognitive neuroscience*, 12(1):207–222.

Gandour, J., Wong, D., Lowe, M., Dzemidzic, M., Satthamnuwong, N., Tong, Y., and Li, X. (2002). A cross-linguistic fMRI study of spectral and temporal cues underlying phonological processing. *Journal of cognitive neuroscience*, 14(7):1076–1087.

Giegerich, H. J. (1985). Metrical phonology and phonological structure. German and English. *Cambridge Studies in Linguistics London*, (43):1–301.

Giraud, A.-L. and Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. *Nature neuroscience*, 15(4):511–517.

Gross, M., Say, T., Kleingers, M., Clahsen, H., and Münte, T. F. (1998). Human brain potentials to violations in morphologically complex Italian words. *Neuroscience Letters*, 241(2-3):83–86.

Gwilliams, L., Linzen, T., Poeppel, D., and Marantz, A. (2018). In spoken word recognition, the future predicts the past. *Journal of Neuroscience*, 38(35):7585–7599.

Gwilliams, L. and Marantz, A. (2015). Non-linear processing of a linear speech stream: The influence of morphological structure on the recognition of spoken Arabic words. *Brain and language*, 147:1–13.

Hahne, A., Mueller, J. L., and Clahsen, H. (2006). Morphological processing in a second language: Behavioral and event-related brain potential evidence for storage and decomposition. *Journal of cognitive neuroscience*, 18(1):121–134.

Hall, T. A. (2007). Segmental features. *The Cambridge handbook of phonology*, pages 311–334.

Hamilton, L. S., Edwards, E., and Chang, E. F. (2018). A spatial map of onset and sustained responses to speech in the human superior temporal gyrus. *Current Biology*, 28(12):1860–1871.

Hauser, M. D., Chomsky, N., and Fitch, W. T. (2002). The faculty of language: what is it, who has it, and how did it evolve? *science*, 298(5598):1569–1579.

Havas, V., Rodríguez-Fornells, A., and Clahsen, H. (2012). Brain potentials for derivational morphology: An ERP study of deadjectival nominalizations in Spanish. *Brain and Language*, 120(3):332–344.

Hestvik, A. and Durvasula, K. (2016). Neurobiological evidence for voicing underspecification in English. *Brain and Language*, 152:28–43.

Holt, L. L. and Lotto, A. J. (2010). Speech perception as categorization. *Attention, Perception, & Psychophysics*, 72(5):1218–1227.

Hsu, C.-H., Lin, S.-K., Hsu, Y.-Y., and Lee, C.-Y. (2014). The neural generators of the mismatch responses to Mandarin lexical tones: An MEG study. *brain research*, 1582:154–166.

Hullett, P. W., Hamilton, L. S., Mesgarani, N., Schreiner, C. E., and Chang, E. F. (2016). Human superior temporal gyrus organization of spectrotemporal modulation tuning derived from speech stimuli. *Journal of Neuroscience*, 36(6):2014–2026.

Hutchison, E. R., Blumstein, S. E., and Myers, E. B. (2008). An event-related fMRI investigation of voice-onset time discrimination. *Neuroimage*, 40(1):342–352.

Jacobsen, T., Horváth, J., Schröger, E., Lattner, S., Widmann, A., and Winkler, I. (2004). Pre-attentive auditory processing of lexicality. *Brain and Language*, 88(1):54–67.

Jacobsen, T., Schröger, E., Winkler, I., and Horváth, J. (2005). Familiarity affects the processing of task-irrelevant auditory deviance. *Journal of Cognitive Neuroscience*, 17(11):1704–1713.

Kaplan, R. M. and Kay, M. (1994). Regular models of phonological rule systems. *Computational linguistics*, 20(3):331–378.

Katz, D. (1987). *Grammar of the Yiddish language*. Bloomsbury Academic.

Kenstowicz, M. and Kisseberth, C. (2014). *Generative Phonology: Description and Theory*. Academic Press.

Kharlamov, V., Campbell, K., and Kazanina, N. (2011). Behavioral and electrophysiological evidence for early and automatic detection of phonological equivalence in variable speech inputs. *Journal of Cognitive Neuroscience*, 23(11):3331–3342.

Kiparsky, P. (1968). *How abstract is phonology?* Indiana University Linguistics Club.

Klein, E., Domahs, U., Grande, M., and Domahs, F. (2011). Neuro-cognitive foundations of word stress processing-evidence from fMRI. *Behavioral and Brain Functions*, 7(1):1–17.

Lee, Y.-S., Turkeltaub, P., Granger, R., and Raizada, R. D. (2012). Categorical speech processing in Broca's area: An fMRI study using multivariate pattern-based analysis. *Journal of Neuroscience*, 32(11):3942–3948.

Leminen, A., Leminen, M., Kujala, T., and Shtyrov, Y. (2013). Neural dynamics of inflectional and derivational morphology processing in the human brain. *Cortex*, 49(10):2758–2771.

Leonard, M. K., Baud, M. O., Sjerps, M. J., and Chang, E. F. (2016). Perceptual restoration of masked speech in human cortex. *Nature communications*, 7(1):1–9.

Li, X. and Chen, Y. (2015). Representation and processing of lexical tone and tonal variants: evidence from the mismatch negativity. *PloS one*, 10(12):e0143097.

Li, X., Kager, R., and Gu, W. (2016). Surface vs. underlying listening strategies for cross-language listeners in the perception of sandhied tones in the Nanjing dialect. In *The 5th International Symposium on Tone Aspects of Languages*, pages 33–37.

Liebenthal, E., Binder, J. R., Spitzer, S. M., Possing, E. T., and Medler, D. A. (2005). Neural substrates of phonemic perception. *Cerebral cortex*, 15(10):1621–1631.

Liebenthal, E., Desai, R., Ellingson, M. M., Ramachandran, B., Desai, A., and Binder, J. R. (2010). Specialization along the left superior temporal sulcus for auditory categorization. *Cerebral Cortex*, 20(12):2958–2970.

Lombardi, L. (1999). Positional faithfulness and voicing assimilation in Optimality Theory. *Natural Language & Linguistic Theory*, 17(2):267–302.

Lück, M., Hahne, A., and Clahsen, H. (2006). Brain potentials to morphologically complex words during listening. *Brain research*, 1077(1):144–152.

Malins, J. G. and Joanisse, M. F. (2012). Setting the tone: An ERP investigation of the influences of phonological similarity on spoken word recognition in Mandarin Chinese. *Neuropsychologia*, 50(8):2032–2043.

Marslen-Wilson, W. D. and Tyler, L. K. (1997). Dissociating types of mental computation. *Nature*, 387(6633):592–594.

Marslen-Wilson, W. D. and Tyler, L. K. (2007). Morphology, language and the brain: The decompositional substrate for language comprehension. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1481):823–836.

Martin, S., Iturrate, I., Millán, J. d. R., Knight, R. T., and Pasley, B. N. (2018). Decoding inner speech using electrocorticography: Progress and challenges toward a speech prosthesis. *Frontiers in neuroscience*, 12:422.

Mesgarani, N., Cheung, C., Johnson, K., and Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*, 343(6174):1006–1010.

Moore-Cantwell, C., Pater, J., Staubs, R., Zobel, B., and Sanders, L. (2018). Event-related potential evidence of abstract phonological learning in the laboratory.

Mugler, E. M., Patton, J. L., Flint, R. D., Wright, Z. A., Schuele, S. U., Rosenow, J., Shih, J. J., Krusienski, D. J., and Slutzky, M. W. (2014). Direct classification of all American English phonemes using signals from functional speech motor cortex. *Journal of neural engineering*, 11(3):035015.

Myers, E. B. (2007). Dissociable effects of phonetic competition and category typicality in a phonetic categorization task: An fMRI investigation. *Neuropsychologia*, 45(7):1463–1473.

Myers, E. B., Blumstein, S. E., Walsh, E., and Eliassen, J. (2009). Inferior frontal regions underlie the perception of phonetic category invariance. *Psychological Science*, 20(7):895–903.

Näätänen, R. (2001). The perception of speech sounds by the human brain as reflected by the mismatch negativity (MMN) and its magnetic equivalent (MMNm). *Psychophysiology*, 38(1):1–21.

Näätänen, R. and Alho, K. (1997). Mismatch negativity–the measure for central sound representation accuracy. *Audiology and Neurotology*, 2(5):341–353.

Näätänen, R., Lehtokoski, A., Lennes, M., Cheour, M., Huotilainen, M., Iivonen, A., Vainio, M., Alku, P., Ilmoniemi, R. J., Luuk, A., et al. (1997). Language-specific phoneme representations revealed by electric and magnetic brain responses. *Nature*, 385(6615):432–434.

Näätänen, R., Paavilainen, P., Rinne, T., and Alho, K. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: a review. *Clinical neurophysiology*, 118(12):2544–2590.

Newman, A., Izvorski, R., Davis, L., Neville, H., and Ullman, M. (1999). Distinct electrophysiological patterns in the processing of regular and irregular verbs. *Journal of Cognitive Neuroscience, Supplement*, 47.

Obleser, J., Boecker, H., Drzezga, A., Haslinger, B., Hennenlotter, A., Roettinger, M., Eulitz, C., and Rauschecker, J. P. (2006). Vowel sound extraction in anterior superior temporal cortex. *Human brain mapping*, 27(7):562–571.

Obleser, J., Elbert, T., Lahiri, A., and Eulitz, C. (2003). Cortical representation of vowels reflects acoustic dissimilarity determined by formant frequencies. *Cognitive Brain Research*, 15(3):207–213.

Obleser, J., Zimmermann, J., Van Meter, J., and Rauschecker, J. P. (2007). Multiple stages of auditory speech perception reflected in event-related fMRI. *Cerebral Cortex*, 17(10):2251–2257.

Pei, X., Barbour, D. L., Leuthardt, E. C., and Schalk, G. (2011). Decoding vowels and consonants in spoken and imagined words using electrocorticographic signals in humans. *Journal of neural engineering*, 8(4):046028.

Peltola, M. S., Kujala, T., Tuomainen, J., Ek, M., Aaltonen, O., and Näätänen, R. (2003). Native and foreign vowel discrimination as indexed by the mismatch negativity (MMN) response. *Neuroscience letters*, 352(1):25–28.

Pettigrew, C. M., Murdoch, B. E., Ponton, C. W., Finnigan, S., Alku, P., Kei, J., Sockalingam, R., and Chenery, H. J. (2004). Automatic auditory processing of

English words as indexed by the mismatch negativity, using a multiple deviant paradigm. *Ear and hearing*, 25(3):284–301.

Price, C. J. (2012). A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. *Neuroimage*, 62(2):816–847.

Pulvermüller, F., Kujala, T., Shtyrov, Y., Simola, J., Tiitinen, H., Alku, P., Alho, K., Martinkauppi, S., Ilmoniemi, R. J., and Näätänen, R. (2001). Memory traces for words as revealed by the mismatch negativity. *Neuroimage*, 14(3):607–616.

Pulvermüller, F., Shtyrov, Y., Kujala, T., and Näätänen, R. (2004). Word-specific cortical activity as revealed by the mismatch negativity. *Psychophysiology*, 41(1):106–112.

Raizada, R. D. and Poldrack, R. A. (2007). Selective amplification of stimulus differences during categorical processing of speech. *Neuron*, 56(4):726–740.

Rhodes, R., Han, C., and Hestvik, A. (2019). Phonological memory traces do not contain phonetic information. *Attention, Perception, & Psychophysics*, 81(4):897–911.

Rivera-Gaxiola, M., Csibra, G., Johnson, M. H., and Karmiloff-Smith, A. (2000a). Electrophysiological correlates of cross-linguistic speech perception in native English speakers. *Behavioural Brain Research*, 111(1-2):13–23.

Rivera-Gaxiola, M., Johnson, M. H., Csibra, G., and Karmiloff-Smith, A. (2000b). Electrophysiological correlates of category goodness. *Behavioural Brain Research*, 112(1-2):1–11.

Rodriguez-Fornells, A., Clahsen, H., Lleo, C., Zaake, W., and Münte, T. F. (2001). Event-related brain responses to morphological violations in Catalan. *Cognitive Brain Research*, 11(1):47–58.

Roettger, T. B., Winter, B., Grawunder, S., Kirby, J., and Grice, M. (2014). Assessing incomplete neutralization of final devoicing in German. *Journal of Phonetics*, 43:11–25.

Scharinger, M., Bendixen, A., Trujillo-Barreto, N. J., and Obleser, J. (2012a). A sparse neural code for some speech sounds but not for others. *PloS one*, 7(7):e40953.

Scharinger, M., Merickel, J., Riley, J., and Idsardi, W. J. (2011). Neuromagnetic evidence for a featural distinction of English consonants: Sensor-and source-space data. *Brain and language*, 116(2):71–82.

Scharinger, M., Monahan, P. J., and Idsardi, W. J. (2012b). Asymmetries in the processing of vowel height.

Scharinger, M., Monahan, P. J., and Idsardi, W. J. (2016). Linguistic category structure influences early auditory processing: Converging evidence from mismatch responses and cortical oscillations. *NeuroImage*, 128:293–301.

Schirmer, A., Tang, S.-L., Penney, T. B., Gunter, T. C., and Chen, H.-C. (2005). Brain responses to segmentally and tonally induced semantic violations in cantonese. *Journal of Cognitive Neuroscience*, 17(1):1–12.

Sharma, A. and Dorman, M. F. (2000). Neurophysiologic correlates of cross-language phonetic perception. *The Journal of the Acoustical Society of America*, 107(5):2697–2703.

Shestakova, A., Brattico, E., Huotilainen, M., Galunov, V., Soloviev, A., Sams, M., Ilmoniemi, R. J., and Näätänen, R. (2002). Abstract phoneme representations in the left temporal cortex: magnetic mismatch negativity study. *Neuroreport*, 13(14):1813–1816.

Shtyrov, Y., Pihko, E., and Pulvermüller, F. (2005). Determinants of dominance: is language laterality explained by physical or linguistic features of speech? *Neuroimage*, 27(1):37–47.

Silverman, D. (2006). *A critical introduction to phonology: Of sound, mind, and body*. A&C Black.

Silverman, D. (2012). *Neutralization*. Cambridge University Press.

Sun, Y., Giavazzi, M., Adda-Decker, M., Barbosa, L. S., Kouider, S., Bachoud-Lévi, A.-C., Jacquemot, C., and Peperkamp, S. (2015). Complex linguistic rules modulate early auditory brain responses. *Brain and Language*, 149:55–65.

Tremblay, P., Baroni, M., and Hasson, U. (2013). Processing of speech and non-speech sounds in the supratemporal plane: Auditory input preference does not predict sensitivity to statistical structure. *Neuroimage*, 66:318–332.

Tyler, L. K., deMornay Davies, P., Anokhina, R., Longworth, C., Randall, B., and Marslen-Wilson, W. D. (2002). Dissociations in processing past tense morphology: Neuropathology and behavioral studies. *Journal of Cognitive Neuroscience*, 14(1):79–94.

Vennemann, T. (1990). Syllable structure and simplex accent in Modern Standard German. *CLS*, 26(2):399–412.

Vroomen, J. and De Gelder, B. (1995). Metrical segmentation and lexical inhibition in spoken word recognition. *Journal of Experimental Psychology: Human perception and performance*, 21(1):98.

Weyerts, H., Penke, M., Dohrn, U., Clahsen, H., and Münte, T. F. (1997). Brain potentials indicate differences between regular and irregular German plurals. *NeuroReport*, 8(4):957–962.

Winkler, I. (2007). Interpreting the mismatch negativity. *Journal of Psychophysiology*, 21(3-4):147–163.

Winkler, I., Kujala, T., Tiitinen, H., Sivonen, P., Alku, P., Lehtokoski, A., Czigler, I., Csépe, V., Ilmoniemi, R. J., and Näätänen, R. (1999). Brain responses reveal the learning of foreign language phonemes. *Psychophysiology*, 36(5):638–642.

Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G., and Vaughan, T. M. (2002). Brain–computer interfaces for communication and control. *Clinical neurophysiology*, 113(6):767–791.

Ylinen, S., Huuskonen, M., Mikkola, K., Saure, E., Sinkkonen, T., and Paavilainen, P. (2016). Predictive coding of phonological rules in auditory cortex: A mismatch negativity study. *Brain and language*, 162:72–80.

Yuan, J. and Liberman, M. (2014). F0 declination in English and Mandarin broadcast news speech. *Speech Communication*, 65:67–74.

Chapter 2

Phonological and Morphological Abstraction in Speech Processing

2.1 Introduction

2.1.1 Context-dependent speech sound processing

The high spatiotemporal resolution of intracranial recordings has enabled fine-grained investigation of the neural basis of speech sound processing. The millimeter-scale spatial resolution afforded by high-density grid recordings has allowed estimation of receptive fields across superior temporal gyrus (STG) (Holdgraf et al., 2016; Hullett et al., 2016) and provided evidence of a richly structured mosaic of functionally distinct sub-centimeter regions (Flinker et al., 2011; Chan et al., 2014). The millisecond-scale temporal resolution has allowed observation of fluctuations in population activity that track the rapid acoustic transitions of natural speech (Berezutskaya et al., 2017). Altogether, intracranial recordings have provided the degree of resolution necessary to observe the selective encoding of speech sounds relevant for phonological processing.

In addition to detailed representations of sub-phonemic acoustic features, intracranial recordings have also provided evidence for higher order auditory processing in STG, including the encoding of categorical perception (Chang et al., 2010) and representations modulated by context. Contexts which have been shown to

modulate responses to speech sounds include the degree of intelligibility of spectrally degraded speech (Nourski et al., 2019), lexical sequence probability (Leonard et al., 2016), phonological neighborhood density (Cibelli et al., 2015), and selective attention (Mesgarani and Chang, 2012). Contexts like unintelligible, spectrally degraded speech have been shown to decrease the SNR of the response relative to intelligible speech, while selective attention contexts and high lexical sequence probability contexts increase response SNR. These results demonstrate the extent to which neural activity, particularly in STG, reflects information beyond the acoustics of the sensory signal.

In light of the context-sensitive activity already observed in STG, this chapter investigates the extent to which phonological and morphological context influence sensory processing. Previous work has provided evidence for computations that would support language-specific phonological processing: namely, the encoding of both subphonemic detail and phonemic category (Chang et al., 2010; Toscano et al., 2010) and the representation of speech sounds as combinations of acoustic features relevant for phoneme distinction (Mesgarani et al., 2014), (cf. distinctive feature theory: Gussenhoven and Jacobs 2017; Chomsky and Halle 1968; Jakobson et al. 1963). However, these studies examine features and phonemic contrasts that—while focusing exclusively on English speakers—fail to illustrate the extent to which their results reflect structured knowledge of the English phonological system specifically or properties of human speech perception more generally. For this reason, the study described here leverages language-specific phonological and morphophonological alternations to adequately assess the degree to which language-specific linguistic knowledge shapes sublexical speech processing.

2.1.2 Target phones and environments

The primary phonological opposition considered is that of [d] and [t] and their contextual neutralization to [ɾ]; and the morphophonological processes considered are the formation of the regular past tense and regular plural. These phenomena were chosen to guide the investigation because each involves alternation and neutralization of contrast and occurs with relative frequency in typical English speech.

In the case of the the phonological neutralization of coronal stops to tap, the neutralization context provides a window on the one-to-many mapping of acoustic forms to phonological constructs, while the morphophonological alternations provide a window on a slightly more abstract many-to-one mapping from phonological forms to morphological exponence.

The phonemes /t/ and /d/ participate in several alternation processes in which their contrast is either preserved or neutralized, including word-final devoicing, initial aspiration, and intervocalic tapping. From their participation in these and other phonological patterns, we observe that these two sounds form a minimally contrastive pair. They contrast in neither place nor manner of articulation, but primarily in the timing of their release gesture relative to the onset of a following vowel (VOT: voice onset time). While /t/ may be found with a large positive VOT (aspiration) at the beginning of stressed syllables and prevocalically, /d/ has a small, but typically still positive VOT. In traditional analyses, the phonetic basis for the contrast between /t/ and /d/ has been associated with the binary feature [±voice], such that /d/ is referred to as a voiced stop, and /t/ is referred to as a voiceless stop. However, from an articulatory standpoint, the laryngeal gestures associated with these sounds are similar; thus, the contrast between /t/ and /d/ would be more accurately referred to as one of aspiration rather than voicing. For clarity and efficiency of prose, we continue to refer to the opposition between /t/ and /d/ as one of "voicing," but the reader should keep the basic phonetic facts of the opposition in mind, especially as we consider the relationship of phonetic form to phonological representation.

While /t/ is typically characterized as a voiceless aspirated coronal stop, and /d/ is typically characterized as a voiceless unaspirated coronal stop, their realization varies based on phonological context. The canonical aspirated [t^h] is typical of syllable-initial, prevocalic, stressed contexts, and we refer to the aspiration of /t/ in this environment as initial aspiration. In word-final contexts however, stops are frequently unreleased, unaspirated. Without its definitive aspiration, the phonetic realization of /t/ is similar to that of /d/, and the partial neutralization of this voicing contrast in word-final contexts is referred to as word-final devoicing.

However, neutralization of the voicing contrast is not the only possible realization of /t/ in this environment. Realization of /t/ as [ʔ] is also common word-finally or when /t/ is followed by a syllabic nasal, as in *button* or *mountain*. Another contextual neutralization of /t/ and /d/ occurs intervocalically, preceding an unstressed syllable, where both /t/ and /d/ surface as a coronal tap [ɾ], as in *writer* and *rider*. Taking into consideration the substantial, systematic acoustic and articulatory variation underpinning the phonological constructs we refer to as /t/ and /d/, each context of interest is coded both for its putative underlying form (i.e., /t/, /d/) as well as for its phonetic realization (e.g., [t], [ʔ], [ɾ]).

The two morphophonological contexts of interest also exhibit substantial variation in their realization depending on the phonological context. The regular past tense takes one of three forms: a voiceless coronal stop [t], a voiced coronal stop [d], or a syllabic voiced coronal stop [əd]. The voiceless form surfaces following voiceless consonants other than [t], the voiced form surfaces following voiced segments other than [d], and the syllabic form surfaces elsewhere (i.e., following [t] and [d]). Similarly, the regular plural manifests in one of three forms: a voiceless coronal sibilant [s], a voiced coronal sibilant [z], or a syllabic voiced coronal sibilant [əz]. The contexts in which each of these forms surfaces also mirrors that of the past tense. The voiceless form surfaces following voiceless consonants other than sibilants [s, tʃ, ʃ], the voiced form surfaces following voiced consonants other than sibilants [z, dʒ, ʒ], and the syllabic form manifests elsewhere (i.e., following sibilants [s, z, tʃ, dʒ, ʃ, ʒ]).

Natural speech stimuli for the passive listening task were designed to exploit the (morpho)phonological patterns described above to dissociate phonetic, phonological, and morphological representations from one another. All stimuli were phonetically segmented and labeled with the surface form of each phoneme, and each brief monologue contained at least ten instances of each surface form relevant to the predetermined (morpho)phonological contexts of interest: t, ɾ, d, s, z. These surface forms are allophones of the phonemes /t/, /d/, /s/, and /z/, and participate in intervocalic tapping, the formation of the regular plural, or the formation of the regular past tense.

Given that auditory processing primarily proceeds in a feedforward manner from spectrotemporal features of the sensory input, it is anticipated that many speech selective sites will demonstrate an acoustic 'surface response', where the responses for all taps are more similar to one another than to the voiceless coronal stop allophone of /t/. However, if it is the case that phonological context is used to compute phonemic identity during language processing, even when the acoustic contrast between two phonemes is neutralized, then there also should exist sites demonstrating a phonemic 'underlying response', where the neural response to underlyingly /t/ taps (**tap_t**; i.e., *writing*) is more similar to other allophones of /t/ (**t**; i.e., voiceless alveolar stops) than to underlyingly /d/ taps (**tap_d**; i.e., *riding*).

Similarly, for the two morphophonological comparisons, it is anticipated that some speech selective sites will demonstrate a 'surface response', where responses for [z] forms of the plural and [d] forms of the past tense are more similar to word-final non-plural [z] and non-past [d], respectively, than to the voiceless forms of the plural and past tense, respectively. Additionally, if it is the case that morphological identity is abstracted from phonological context, then there should also exist sites demonstrating a morphological 'underlying response', where the neural responses to both voiced and voiceless forms of the plural and past tense pattern together to the exclusion of word final non-plural [z] and non-past [d], respectively.

In this way, observing the distribution of surface and underlying sites for the coronal tap, plural, and past tense comparisons has the potential to provide critical evidence for the mental reality of phonemes and morphemes.

2.2 Methods

2.2.1 Participants

Study participants were ten patients at UC San Diego Health who underwent subdural electrode implantation as part of treatment for refractory epilepsy. All were native English speakers who had no prior experience with Catalan, and all reported normal hearing, performed within the acceptable range on a battery of

Table 2.1: Summary of basic patient information. *While patient SD013 did not experience a clear, prolonged speech arrest with either injection, he demonstrated greater language deficits following left hemisphere injection.

Patient	Age	Gender	Handedness	Wada	Coverage	Language Experience
SD010	28	M	R	–	sEEG: LH,RH	English
SD011	32	M	R	LH	sEEG: LH,RH	English, Creole, French
SD012	25	F	R	LH	sEEG: LH,RH; Grid,strips: LH	English
SD013	43	M	L	LH*	sEEG: LH,RH	English, Spanish
SD015	55	F		–	sEEG: LH,RH	English
SD017	21	M	R	–	sEEG: RH	English, Spanish
SD018	42	F	L	RH	sEEG: LH,RH	English, Spanish
SD019	21	M	R	–	sEEG: LH,RH	English
SD021	33	F	R	–	sEEG: LH,RH	English
SD022	23	M	R	–	sEEG,Grid: RH	English

neuropsychological language tests, and exhibited left hemisphere language dominance as assessed by clinicians using the Wada test. The research protocol was approved by the UC San Diego Institutional Review Board, and all subjects gave written informed consent prior to surgery.

2.2.2 Stimuli

Participants passively listened to short excerpts of conversational American English speech taken from the Buckeye Corpus (Pitt et al., 2007). Interspersed with the passages from the Buckeye Corpus were short passages of conversational Catalan taken from the Corpus de Català Contemporani de la Universitat de Barcelona (Alturo et al., 2002). To assess participant attention to the task, participants responded orally to a two-alternative question about the content of each English passage, and for Catalan passages, participants were instructed to listen for English words embedded in the recording and press the space bar when they heard an English word.

English passages were 25-76s long (mean 38s) taken from 27 (12 women; 15 men) native speakers of Midwestern American English living in Columbus, OH between October 1999 and May 2000. Per corpus documentation (Pitt et al., 2007), excerpts were recorded monophonically on a head-mounted microphone (Crown CM-311A) and fed to a DAT recorder (Tascam DA-30 MKII) at a 48kHz sampling

rate through an amplifier (Yamaha MV 802). Following each excerpt, participants heard a two-alternative choice question regarding the content of the passage they heard, and they were asked to respond orally to the question to ensure that they were alert and paying attention to the passages that they heard. These questions were recorded with a Blue Yeti USB microphone sampling at a rate of 48kHz by a speaker of American English. The amplitude of all English passages and questions was normalized to -20dBFS prior to padding with 500ms silence.

Catalan passages were 44-55s long (mean 49s) taken from 2 (1 woman; 1 man) native speakers of Catalan—- one woman from Benabarre, Huesca and one man from Tamarit, Catalonia. Three passages were excerpted from each speaker for a total of six Catalan excerpts, and Catalan excerpts were interspersed into the passive listening task such that one Catalan excerpt was included in each block of the task. Since patients are not expected to understand Catalan speech, two-alternative questions about the content of the Catalan passages were not suitable to assess participant attention. Instead, English nouns were digitally spliced into the Catalan excerpts, and participants were asked to press the space bar when they heard an English word. For each Catalan passage, three English nouns were excised from the Buckeye corpus digitally spliced into the Catalan recording. These English nouns were taken from one female speaker and one male speaker from portions of the Buckeye Corpus not otherwise heard in the passive listening task. Nouns were excised from their original contexts and inserted into the Catalan speech at zero crossings to minimize acoustic artifacts. Inserted nouns were gender-matched with their Catalan carrier passages, but no further controls were taken to match for voice similarity. Prior to insertion into Catalan speech, the English nouns were normalized to -20dBFS, and after insertion into Catalan speech, the entire mixed-language passage was normalized to -20dBFS. This process of amplitude normalization resulted in English nouns which were clearly audible and segregable from the Catalan speech stream. Nouns were inserted into the Catalan speech during natural pauses at roughly equal but not regularly-spaced intervals to increase uncertainty and encourage attention to the unintelligible speech stream.

All auditory stimuli, both English and Catalan were orthographically and pho-

netically transcribed, segmented, and labeled. Transcription, segmentation, and labeling procedures for passages from the Buckeye Corpus are described in Pitt et al. (2007). Phonetic and orthographic transcriptions for Catalan passages were archived alongside audio recordings in the Dipòsit Digital de la Universitat de Barcelona (http://diposit.ub.edu/dspace/handle/2445/10413), and the passages used in this study were subsequently segmented and labeled in Praat (Boersma and Weenink, 2017) by a phonetically-trained researcher at UC San Diego. Given the similarity in the phonetic inventories of Catalan and English, the segmentation criteria used for the Buckeye corpus (reported in Pitt et al. 2007) were also used to segment the Catalan passages used in this study.

2.2.3 Coronal tap acoustics

The central logic of the study requires coronal taps to be acoustically distinct from coronal stops, and requires there to be no significant acoustic difference between coronal taps based on their phonemic identity. This section demonstrates that this is indeed the case for the coronal stops and taps present in the Buckeye stimuli.

The primary acoustic feature that distinguishes coronal stops from coronal taps is their duration (Zue and Laferriere, 1979; Braver, 2013; Derrick and Schultz, 2013). While voiced and voiceless American English stops are respectively roughly 70-90ms and 98-130ms in duration (Hillenbrand et al., 1984; Hogan and Rozsypal, 1980; Revoile et al., 1982), coronal taps last a mere 10-40ms (Zue and Laferriere, 1979). Previous investigation into the neutralization of /t/ and /d/ in intervocalic environments has found that the variability in coronal tap duration is primarily dependent on the quality of the preceding vowel and not on the phonemic identity of the tap itself (Zue and Laferriere, 1979). In fact, numerous studies have found that /d/ and /t/ taps do not differ significantly in their closure duration (Charles-Luce, 1997; Zue and Laferriere, 1979; Fox and Terbeek, 1977; Sharf, 1962). This finding is replicated in the stimuli used in this study. As shown in Figure 2.1A, voiceless coronal stops are significantly longer than voiced stops, which are in turn significantly longer than coronal taps, but /d/ and /t/ taps do not differ

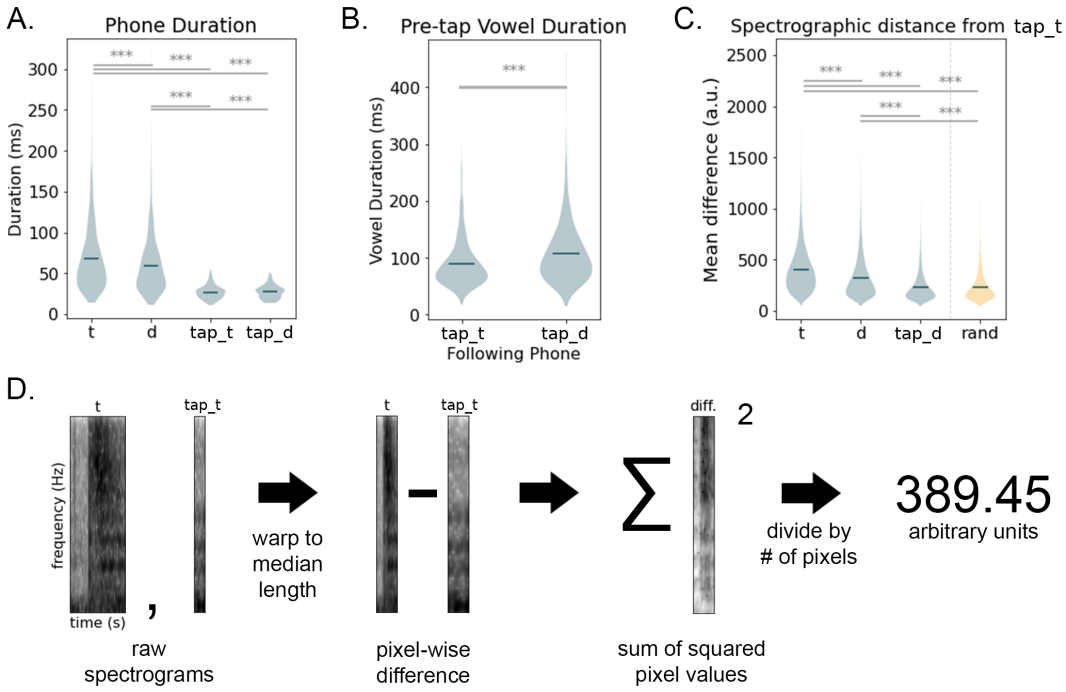

Figure 2.1: Acoustics of coronal stops and taps. (A) Durations of coronal stops and taps in the Buckeye Corpus. (B) Durations of vowels immediately preceding coronal taps in the Buckeye Corpus. (C) Spectrographic distance of t, d, and tap_d sounds from tap_t (green); Spectrographic distance of a random split of all taps from the remainder of taps (yellow). (D) Schematic representation for the calculation of spectrographic distance. Distribution means are shown by the thick dark green lines in plots A-C, and significance levels are indicated on the following scale: * $\leq$ 0.05, ** $\leq$ 0.01, *** $\leq$ 0.001.

significantly from one another in duration. On the basis of their duration, coronal taps are distinct from coronal stops yet not distinct from one another on the basis of their phonemic identity, satisfying the study's foundational assumptions.

Nevertheless, the stimuli used in this study were taken from natural speech and undoubtedly vary along more acoustic dimensions than duration. As a means of further ensuring that the acoustic differences between coronal stops is greater than the difference between coronal taps, a measure of spectrographic distance was calculated for coronal stops and taps. First, wide-band spectrograms were created for all stimuli using Praat's default settings, and the spectrographic segments corresponding to coronal stops and taps were excised using the segmentation and

labelling given in the Buckeye Corpus. The median duration of these segments was calculated (in bins) and all spectrographic segments were resized to this duration using the resize function from the transform module of the Python library scikit-image (van der Walt et al. 2014). This function performs bi-linear interpolation to resize the image. Prior to down-scaling any images, it applies a Gaussian filter with a kernel size of (s-1)/2, where s is the down-scaling factor, to prevent any anti-aliasing artifacts. For each pair of resized spectrograms, the sum of the squared pixel-wise difference was calculated, and divided by the size of the spectrogram in pixels. This process is shown schematically in Figure 2.1D. Figure 2.1C plots the distances of t, d, and tap_d spectrograms from tap_t spectrograms (green), as well as the distances for a random split of all coronal taps (yellow). A one-way ANOVA indicated that there was a significant effect of phonemic category on spectrographic distance from tap_t [F(3; 339,842)=9124.47, p 0.001]. Post hoc comparisons using the Tukey HSD test indicated that the distance between tap_t and tap_d is significantly less than the distance between tap_t and either of the coronal stops at the p 0.001 level. However, when mean distance is calculated for a random split of taps not based on phonemic identity, it is not significantly different from the mean distance calculated based on the phonemic identity of the taps ($p = 0.44$). Once again, this shows that the phonemic identity of coronal taps cannot be determined from the acoustic properties of taps themselves.

Although taps derived from /d/ and /t/ are acoustically indistinguishable from one another, the length of the vowel that precedes a tap systematically varies in duration based on the phonemic identity of the tap. When preceding a tap derived from /d/, vowels are approximately 10% longer than those preceding a tap derived from /t/ (Sharf, 1962; Fischer and Hirsch, 1976; Zue and Laferriere, 1979; Herd et al., 2010; Braver, 2013). This pattern is also observed in the Buckeye Corpus, where vowels preceding taps derived from /d/ were on average 18.31ms (SD=[13.23, 23.38]) longer than vowels preceding taps-derived from /t/ [t(1309)=50.16, p 0.001] (Figure 2.1B). Thus, there does exist an acoustic cue to the phonemic identity of taps. On this basis, one could argue that any observed difference in the neural response to medial /d/and /t/ is due to the duration of

the preceding vowel, seemingly undermining the foundational assumption of the study. For this reason, this study focuses on the 500ms following the onset of coronal stops and taps. By timelocking the analyzed response to the beginning of closure and using the preceding 100ms to baseline the signal, the acoustic impact of the preceding vowel is effectively neutralized. Though it remains possible that the duration of the preceding vowel cues the listener to the phonemic identity of the following tap, since the mean signal in the 100ms preceding the tap is subtracted from the analyzed response, and /d/ and /t/ taps do not differ acoustically from one another, differences observed between /d/ and /t/ taps must be based on their categorization, and not the acoustics of the preceding vowel itself. In other words, though the duration of the vowel may cue the phonemic identity of the tap, by timelocking the response to the beginning of the tap, any difference in response cannot be reducible to the preceding acoustic context.

Furthermore, the duration of the preceding vowel does not appear to be used reliably as an acoustic cue to phonemic identity during speech perception. Malécot and Lloyd (1968) found that participants performed nearly at chance (56.6% accuracy) when asked to select the written form of the word they heard from a pair of minimal pairs with medial /d/ and /t/. While subsequent studies have variously found that listeners are able to discriminate such minimal pairs with up to 80% accuracy, (Fischer and Hirsch, 1976; Charles-Luce, 1997; Charles-Luce and Dressler, 1999; Charles-Luce et al., 1999), when experimental materials are well-controlled for word frequency and the bias to perceive /d/, accuracy falls to 52% (Herd et al., 2010). Although the apparent lack of attention to the difference in vowel duration during perception does not negate the fact of the difference in duration itself, it suggests that this particular acoustic feature may not be encoded as faithfully as other features.

2.2.4 Data acquisition and processing

Experimental instructions and stimuli were presented to participants in their hospital rooms on a Windows 10 desktop PC (Dell XPS 8910) using PsychoPy for Python 2.7 (Peirce, 2007, 2008). The task was conducted in six blocks, each

of which contained eight English trials and one Catalan trial. The average block duration was 6 minutes 30 seconds. All participants completed at least two of the six blocks, and the average participant completed four blocks, listening to 26:45 minutes of speech totaling 15,070 phones. Each English trial consisted of a short, spoken passage followed by a content question and an oral response, and each Catalan trial consisted of a reminder to press the space bar for English words followed by a short spoken passage.

Intracranial EEG signals were amplified using a multi-channel amplifier system (Natus Quantum) and recorded using Natus NeuroWorks software. Auditory stimuli and oral responses were recorded simultaneously with the EEG data by feeding the output of a Zoom H2n microphone as an additional input channel to the Natus Quantum amplifier.

After recording, neural data were deidentified and exported from the clinical NeuroWorks system in .edf (European Data Format) format for pre-processing using the Python package MNE Python (Gramfort et al. 2013). Channels displaying excessive artifacts or line noise were removed (45 channels total). Remaining channels common average referenced, notch filtered at 60Hz and its harmonics, bandpass filtered 0.1-170Hz, and downsampled to three times the lowpass cutoff (510Hz). Independent Component Analysis (ICA) was used to remove stationary artifacts from the filtered data, and time intervals containing remaining artifacts were visually identified and discarded. Channels selected for analysis were those which exhibited reliable evoked response to speech stimuli, determined by a sliding window t-test between responses to randomly-selected time frames during the passive listening task and in silence ($p < 0.05$).

2.2.5 Band power

The frequency bands used in this study were delta (1–4Hz), theta (4–7Hz), alpha (8–12Hz), beta (13–30Hz), gamma (31–50Hz), and high gamma (70–150Hz) bands. To compute the power for each band, the analytic amplitude from eight Gaussian band-pass filters with logarithmically increasing center frequency (Crone et al., 1998; Bouchard et al., 2013) was averaged. For the ERP analysis, this high-

gamma signal was then segmented into peri-target epochs with 100 ms pre-target and 500 ms post-target, and each epoch was z-scored relative to the mean and standard deviation of its 100ms pre-target baseline. The target is the onset of a phone of interest.

2.2.6 Electrode localization

Stereo EEG electrodes where localized by registering each patient's pre-op T1-weighted MR volume to an interoperative CT in 3dSlicer (Fedorov et al., 2012; Johnson et al., 2007) and manually marking the contact CT artifacts.

2.3 Results

2.3.1 Significant electrodes

Broadband neural activity was recorded from a total of $1,355$ valid electrodes[1] across the ten subjects, and each electrode was assessed for speech selectivity using a sliding-window one-way t-test, where 500ms silent epochs were compared against an equivalent number of randomly sampled speech epochs. Within each group, epochs were z-scored against a 100ms baseline and t-tests were calculated for nine 100ms windows with 50ms overlap across the 500ms non-baseline portion of the epoch. Within each window, values for each token were averaged over time. Channels were considered speech selective if the t-test for at least one window was significant with $p < 0.05$.

Speech-selective electrodes were defined independently for each band, and across the ten subjects, $1,091$ electrodes were found to be speech selective for at least one band, with an average of 339 electrodes found to be speech selective for each band. While some overlap was observed in the sites categorized as speech-selective across bands, the majority of speech-selective sites were speech-selective for only one band, as shown in Figure 2.2. Phonological and morphophonological comparisons were then performed only for speech selective electrodes.

[1]That is, electrodes not discarded due to the presence of excessive line noise or artifacts.

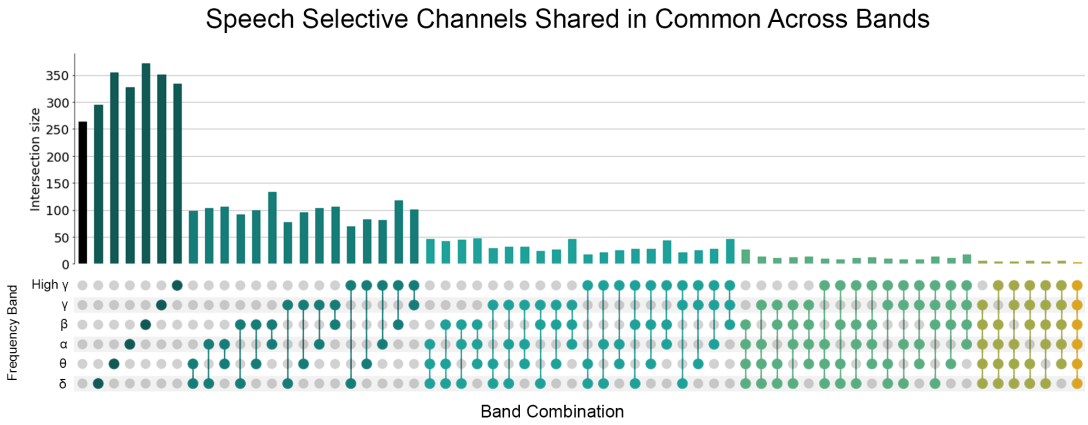

Figure 2.2: Degree of overlap across bands for speech selective channels. Lower dot matrix plot indicates the combination of bands being considered, and the upper bar plot shows the number of speech-selective channels shared in common for that combination of bands. The left most column (black bar) shows the number of electrodes that were not selective for speech for any band, and color indicates the number of bands in each combination.

Phonological Comparisons

Acoustic sites and phonemic sites were defined using a sliding-window one-way ANOVA. For a site to be considered an acoustic site, there must have existed at least one time window with a significant ANOVA for which a Tukey's post hoc test indicated that there was a significant difference ($\alpha = 0.05$) between **t** and **tap_t** tokens and between **t** and **tap_d** tokens but no significant difference between **tap_t** and **tap_d** tokens. Alternatively, for a site to be considered a phonemic site, there must have existed at least one time window with a significant ANOVA for which a Tukey's post hoc test indicates that there was a significant difference between **tap_d** and **tap_t** tokens and between **tap_d** and **t** tokens but no significant difference between **t** and **tap_t** tokens.

Of the roughly 339 speech selective electrodes for each band, an average of 5 acoustic sites (SD±4.7) and 30 phonemic sites (SD±6.3) were observed for the coronal stop–tap alternation across all bands. The number of acoustic and phonemic sites observed across bands is summarized in Table 2.2, and examples of the response observed at acoustic and phonemic sites are shown in Figure 2.3. Only one site was categorized as both an acoustic and phonemic site. This site, RA11,

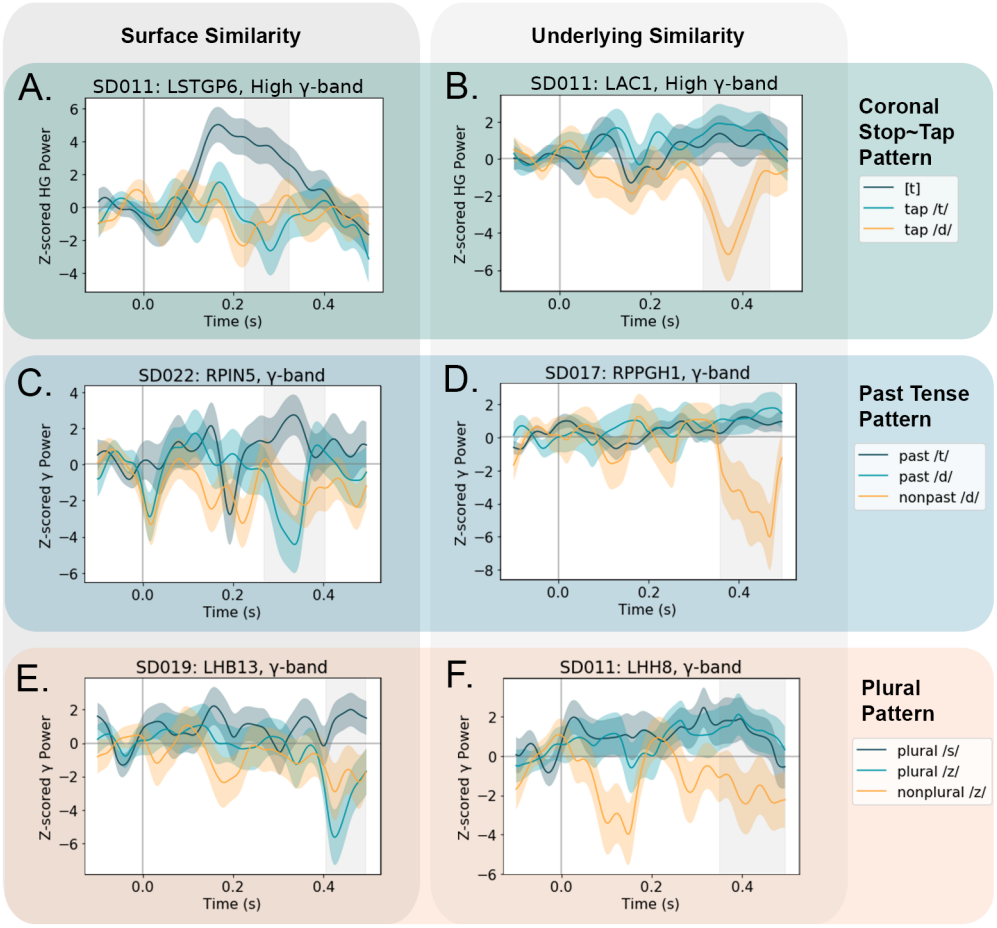

Figure 2.3: Examples of sites exhibiting acoustic and phonemic patterns of activity identified by the coronal tap comparison. Upper plots of (A) and (B) show the time course of high gamma power z-scored relative to baseline (-100ms–0ms). The evoked response to tokens of [t] are shown in charcoal; the evoked response to /t/-taps is shown in teal; and the evoked response to /d/-taps is shown in gold. Shading indicates ±SEM. Lower plots of (A) and (B) show the anatomical location of the channels whose activity are plotted above.

was recorded in white matter beneath right superior temporal sulcus in patient SD022. The phonemic response at this site occurred at 50–150ms after target onset, preceding the acoustic response at 350–450ms, and the effect was observed only in delta band power.

To assess the likelihood of observing these numbers of acoustic and phonemic response sites by chance, an expected null distribution was generated for each frequency band by performing the statistical analysis described above using $1,000$

arbitrary pairs of phones (i.e., A, B) with an arbitrary split of one phone (i.e., A, B_x, B_y). For each each arbitrary set of phones, phony-acoustic sites were identified as those with at least one time window in which there was a significant difference between the evoked response to A phones and the evoked response to B phones, but no significant difference in the evoked response to B_x and B_y phones. Phony-phonemic sites were identified as those with at least one time window with no significant difference between A and B_x phones but a significant difference between those phones and B_y phones. In this way, the null distribution is generated from the real, recorded data. Spatially correlated activity is thus preserved in the null distribution, accounting for the possibility that such correlated activity could inflate the number of observed acoustic and phonemic sites.

Compared against these null distribution, the numbers of observed sites exhibiting acoustic and phonemic patterns of activity were greater than would be expected by chance for each band (Figure 2.4). From the generated distribution, the probability of observing at least 39 phonemic sites and at least 21 acoustic sites was <1%. Figure 2.4 shows the null distribution of phonemic and acoustic sites, with a dotted white line marking the bound for 95% of the distribution's mass and a single white point indicating the observed number of phonemic and acoustic sites.

Morphophonological Comparisons

For the two morphophonological comparisons, rather than categorize sites as either acoustic or phonemic, sites were classified as *surface* or *morphological* using a sliding-window one-way ANOVA assessment similar to that described for the purely phonological coronal stop alternation. For the regular past tense, surface sites were considered to be those where the sliding-window ANOVA was significant for at least one time window and for which a Tukey's post hoc test indicated that there was a significant difference ($\alpha = 0.05$) between past tense t and past tense d tokens and between past tense t and word final non-past d tokens but no significant difference between past tense d and word final non-past d tokens. Similarly, for the regular plural, surface sites were defined as those with at least one time window

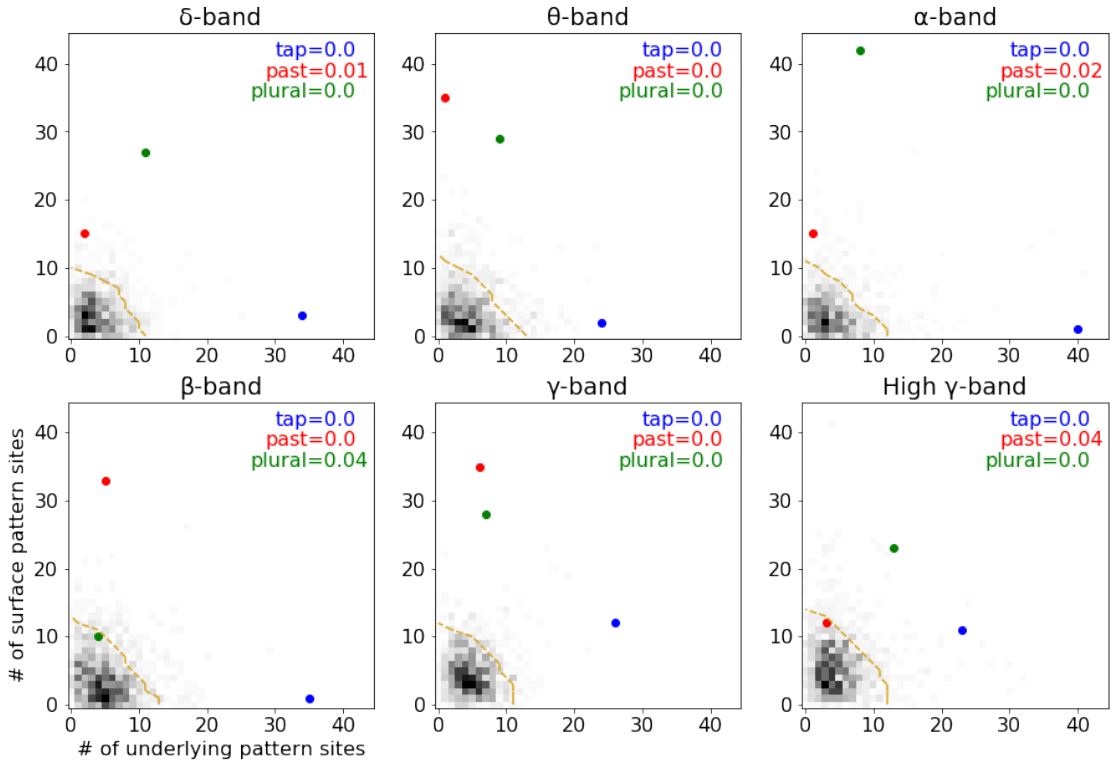

Figure 2.4: Number of significant sites observed for each linguistic comparison for each neural response band relative to the generated null distribution for that band. Vertical axes indicate the number of sites selective for surface identity (i.e., 'acoustic' sites for the tap comparison and 'surface' sites for the plural and past tense comparisons) observed for each comparison, while horizontal axes indicate the number of sites selective for underlying identity (i.e., 'phonemic' sites for the tap comparison and 'morphological' sites for the plural and past tense comparisons) observed for each comparison. The proportion of the null distribution that contains at least as many surface and underlying sites as were observed for each comparison is indicated in the top right corner of each plot. Values for the tap comparison are green; values for the past tense comparison are blue; and values for the plural comparison are teal.

for which a Tukey's post hoc test indicated that there was a significant difference ($\alpha = 0.05$) between plural s and plural z tokens and between plural s and word final non-plural z tokens but no significant difference between plural z and word final non-plural z tokens.

These sites were considered surface sites rather than acoustic or phonemic sites because, given the nature of the comparisons, it is not possible to distinguish

between the two alternatives. For example, sites distinguishing plural and non-plural z from plural s could be doing so either on the basis of acoustic similarity or phonemic identity. To convey this ambiguity, we refer to these sites simply as 'surface sites' because whichever unit over which they are maintaining similarity is less abstract or lower in the linguistic hierarchy (=closer to the surface) than the unit tracked by morphological sites.

For the regular past tense alternation, morphological sites were considered to be those for which the comparison of evoked responses to past tense t, past tense d, and word-final non-past d resulted in at least one time window indicating a significant difference between word-final non-past **d** and past tense **t** tokens and between word-final non-past **d** and past tense **d** tokens but no significant difference between past tense **t** and past tense **d** tokens. Similarly, for the regular plural alternation, morphological sites were those for which there was at least one time window indicating a significant difference between word-final non-plural **z** and plural **s** tokens and between word-final non-plural **z** and plural **z** tokens but no significant difference between plural **z** and plural **s** tokens. In this way, morphological sites were those which maintained language-specific morphological similarity based on the meaning of similar word-final sounds rather than maintaining the comparably less abstract similarity of their acoustic or phonemic realization.

For the past tense alternation, across all bands an average of 24.2 (SD±10.2) were categorized as surface sites, 3 (SD±1.9) were categorized as morphological sites, and four sites were categorized as both surface and morphological sites. For the plural alternation, 26.5 (SD±9.4) sites were categorized as surface sites, 8.7 (SD±2.9) were categorized as morphological sites, and six sites were categorized as both surface and morphological sites. As for the purely phonological comparison, the numbers of sites exhibiting acoustic and morphophonological patterns of activity were greater than would be expected by chance for both the regular past tense and plural comparisons. These likelihoods were assessed against the same null distribution as was used to assess the significance of the phonological coronal stop alternation, a distribution generated by performing this statistical analysis for $1,000$ arbitrary pairs of phones (i.e., A, B) with an arbitrary split of one phone

Table 2.2: Count of sites demonstrating a surface similarity pattern or an underlying similarity pattern for each frequency band. The observation of more surface sites than underlying sites for the morphological comparison is likely due to the fact that surface similarity for the morphological comparisons is analogous to both surface and underlying similarity at the phonological level.

	Coronal Tap Comparison		*Plural Comparison*		*Past Tense Comparison*	
Band	Surface	Underlying	Surface	Underlying	Surface	Underlying
δ: (1–3Hz)	3	34	27	11	15	2
θ: (4–7Hz)	2	24	29	9	35	1
α: (8–12Hz)	1	40	42	8	15	1
β: (13–30Hz)	1	35	10	4	33	5
γ: (31–50Hz)	12	26	28	7	35	6
High-γ: (70–150Hz)	11	23	23	13	12	3

(i.e., A, B_x, B_y).

From this distribution, the probability of observing the past tense pattern of at least 12 acoustic sites and at least 3 morphophonological sites was 1.7%. The probability of observing the plural pattern of at least 23 acoustic sites and at least 13 morphophonological sites was less than 1%. Figure 2.4 shows the null distribution of phonemic and acoustic sites, with a dotted white line marking the bound for 95% of the distribution's mass and a single white point indicating the observed number of phonemic and acoustic sites.

2.3.2 Timing

The time windows with a significant acoustic or phonemic response varied across sites. For both the acoustic and phonemic response patterns, each time window was significant for at least one site, with the exception of the 0–100ms window. However, although the timing of significant windows was variable overall, the 200–300ms window was significant for nearly half the significant phonemic sites (19 of 39) and 85% of significant acoustic sites (18 of 21). The distribution of significant time windows for these response patterns is shown in Figure 2.5.

Similar to the coronal tap comparison, significant time windows for both the past tense and plural comparisons varied substantially across sites. For the past tense comparison, 5 of the 12 acoustic sites and 2 of the 3 morphophonological sites

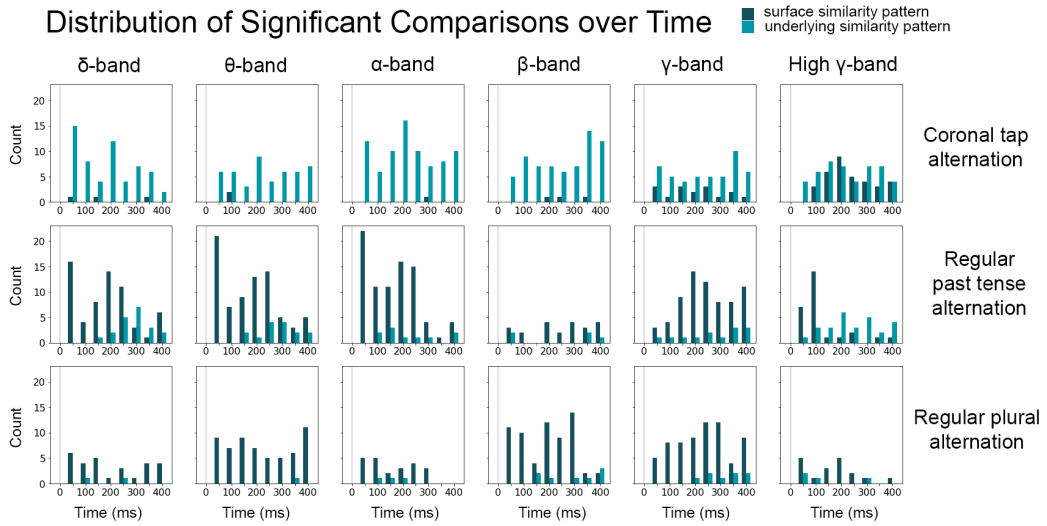

Figure 2.5: Distribution of significant channels over time. For each neural response band (columns) and each linguistic comparison (rows), plots show the number of channels exhibiting surface and underlying responses for each time bin of the sliding window ANOVA. Bins correspond to 100ms windows beginning with the value indicated on the plot's horizontal axis. For each time bin, two bars are present: the navy bar shown on the right of each bin counts the number of sites selective for surface identity (i.e., 'acoustic' sites for the tap comparison and 'surface' sites for the plural and past tense comparisons), while the teal bar on the left counts the number of sites selective for underlying identity (i.e., 'phonemic' sites for the tap comparison and 'morphological' sites for the plural and past tense comparisons)

had a significant window from 50–150 ms. Five acoustic sites also had a significant window from 200–300. The plural comparisons were somewhat similar. Fourteen of the 23 significant acoustic sites had significant windows from 100–200 ms, and 6 of the 13 morphophonological sites had significant windows from 200–300 ms. The full distributions of significant windows for past tense and plural comparisons are also shown in Figure 2.5.

2.3.3 Localization

Locations of significant sites are spread bilaterally throughout fronto-temporal areas, as shown in Figure 2.6. However, given the paucity of coverage over posterior cortices, the broad distribution of significant sites in traditional language areas

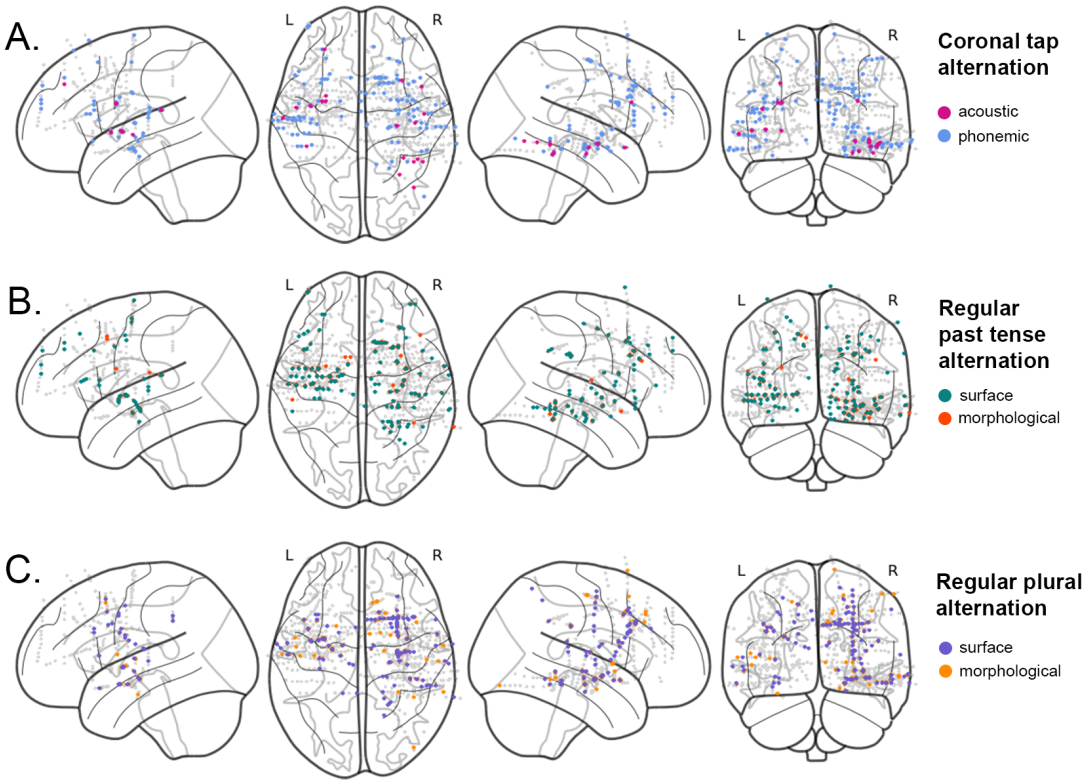

Figure 2.6: Distributions of significant sites for all neural response bands. (A) Locations of acoustic (pink) and phonemic (blue) sites identified by the coronal tap alternation. (B) Locations of surface (green) and morphological (orange) sites identified by the regular past tense alternation. (C) Locations of surface (purple) and morphological (gold) sites identified by the regular plural alternation.

should not be mistaken for a positive result; rather, without further corroboration from whole brain imaging studies, this distribution can be conservatively considered artifactual of commonalities among clinical presentations that require invasive neuromonitoring. Similarly, the larger number of significant sites in right hemisphere areas should not be taken as evidence for the lateralization of any effect, since more sites were recorded from right hemisphere than from the left. In general, due to the clinically motivated biases inherent in the distribution of channel locations, claims about the relative distributions of channels based on any number of factors (e.g. response band, linguistic comparison, etc.) should be avoided because they cannot be responsibly substantiated given the data available. For this

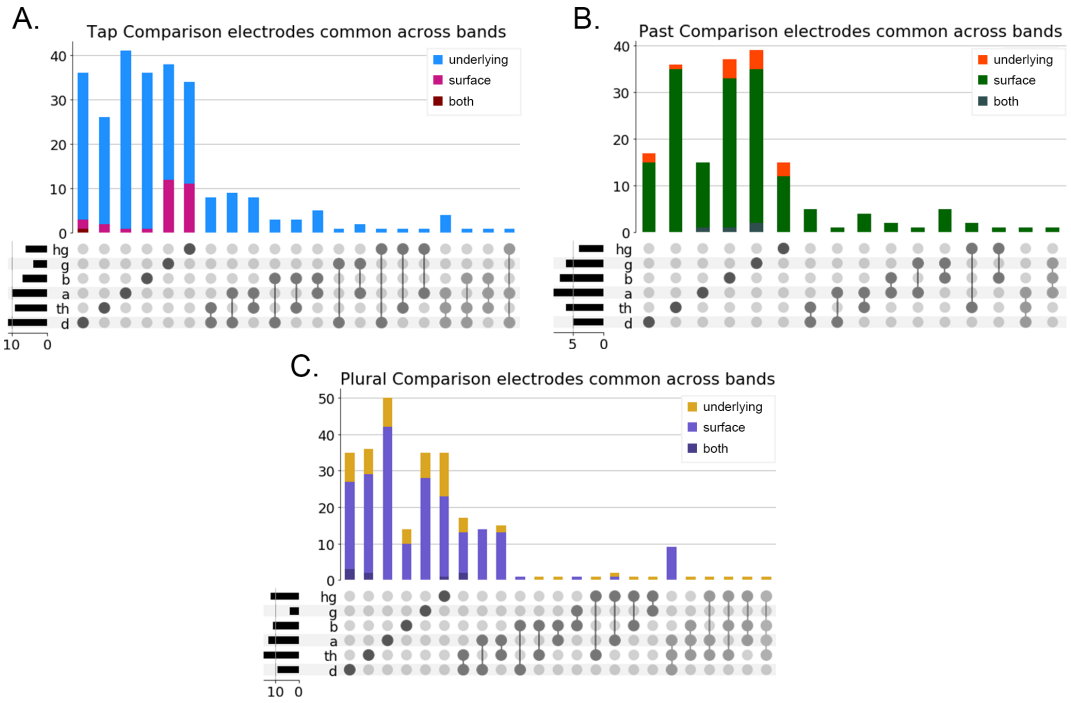

Figure 2.7: Degree of overlap across bands for channels identified as significant for each of the three (morpho)phonological comparisons. For each subfigure, lower dot matrix plots indicate the combination of bands being considered, and upper bar plots show the number of speech-selective channels shared in common for that combination of bands. Combinations of bands not indicated in the dot matrix plots shared no significant sites in common. (A) Number of sites identified as acoustic (pink), phonemic (blue) or both (red) by the coronal tap alternation. (B) Number of sites identified as surface (green), morphological (orange), or both (dark green) by the regular past tense alternation. (C) Number of sites identified as surface (purple), morphological (gold), or both (dark purple) by the regular plural alternation.

reason, information about the localization of various electrodes is presented purely descriptively, to aid in cross-referencing with effects found in future work.

For both surface and underlying patterns for all three linguistic comparisons, the majority of significant electrodes are located in white matter. For the coronal tap alternation, acoustic and phonemic sites are diffusely distributed, and do not cluster separately from one another, as shown in Figure 2.6A. Similarly, surface and morphological sites for the plural and past tense alternations do not cluster separately from one another and are broadly distributed across recorded sites, as

shown in Figure 2.6B,C. These general properties hold across bands, and over the full time course of the neural response. That is, it is not the case that clusters of sites identified as either surface or underlying emerge in particular bands or at particular times. While many aspects of language processing are characterized as distributed, for phonological phenomena, the degree of distribution observed here is uncharacteristic, and likely attributable to underdocumented properties of white matter in linguistic processing.

2.4 Discussion

2.4.1 Implications for phonology

The central finding of this study is that more sites sensitive to phonemic identity were observed than would be predicted by chance. While simple in its articulation, this result has far-reaching consequences for our conceptions of phonology and language representation in the brain. In particular, these findings support a reallocation of probability mass away from a number of common ideas about the nature of phonology and phonological processing.

First, implicit in many theories of speech production and processing is the assumption that language-specific grammar only occurs at the level of syntax. Phonological knowledge, including knowledge of language-specific phonotactics, is cast as epiphenomenal of lexical knowledge: speech perception is more-or-less a word recognition problem, and speech production proceeds in a universal and deterministic manner from whatever articulo-acoustic content is specified in the lexicon. This study provides evidence that these common assumptions are incomplete. Language-specific phonological grammar guides the neural response to speech, and it does so at a level of granularity commensurate with the phoneme.

To see how this conclusion follows from the observed data, consider what one would predict if phonemes were not a relevant unit of organization for language in the brain. To maintain true skepticism about the cognitive reality of phonemes, one would need to envision a state of affairs in which underlying representations of speech sounds were altogether unnecessary. That is, at no point in speech process-

ing would a listener assign the value /d/ to some taps and /t/ to others. Perhaps in lieu of underlying representations, the forms of words that the brain had access to would be exemplars, such as pronunciations the individual had heard before. Laying aside the fact that such a system would likely fail to explain how speakers produce forms they haven't heard before, if it were the case that words containing taps were specified as such in their lexical entries, it would be unexpected to observe any of the sites in this paper that have been called 'phonemic'. Instead, one would expect to see only acoustic sites.

Having observed phonemic sites, however, if one were to maintain that lexical representations reflect sounds as they are produced, the distinction between /d/-taps and /t/-taps observed in phonemic sites would need to be accounted for by some other means. The most readily available explanation would be orthographic influence. In most all cases, the orthographic forms of words containing taps reliably differentiates between words containing taps that are hypothesized to derive from /d/ and those that are hypothesized to derive from /t/. On this basis, one could argue that phonemic sites are indicative of the pervasiveness of orthographic influence on literate language users. There is some evidence to suggest that this level of orthographic influence is possible. For example, using the French minimal triplet *prix* pʁi, *tri* tʁi, *cri* kʁi, during a passive listening task, Pattamadilok et al. (2014) show that a larger MMN is elicited when the spelling of the deviant is incongruent with the spelling of the standard (i.e., *prix* vs. *cri*) than when the spelling of the deviant is congruent with the standard (i.e., *cri* vs. *tri*). However, as the authors themselves note, the task of passively listening to strings of repeated words is not particularly natural, and due to the inherent constraints of the MMN and the French lexicon, only these three words were used in the study.

However, in the context of the current study, the likelihood that orthographic influence alone drives the appearance of phonemic sites is low. If it were the case that orthographic influence drove the occurrence of phonemic sites for the tap comparison, we would not expect to see surface sites for the past tense comparison. The regular past tense is consistently spelled with an '-ed' and yet, to observe surface sites for the past tense comparison in this study, pasts produced as [d]

pattern with other word final [d] sounds, to the exclusion of pasts produced as [t]. If orthography was primarily responsible for the results of this study, we would not expect to see this structured split between past tense [d] and past tense [t], since orthography is consistent across all forms. A very similar argument can be made from the observation of surface sites for the regular plural comparison.

Perhaps then one would argue that orthography influences perception only when there is very little acoustic difference between two sounds. The [d] and [t] forms of the past tense are sufficiently acoustically distinct from one another, while /t/-taps and /d/-taps are practically indistinguishable. If orthographic knowledge is only called upon to bootstrap lexical recognition when acoustic discriminability is poor, then one could argue that orthographic influence is responsible for the the so-called phonemic sites in the case of the tap comparison, but not for the surface sites of the regular plural and past tense comparisons. This very well may be the case, and is worthy of its own empirical investigation. However, any study which would attempt to show that orthographic influence is responsible for the appearance of phonemic sites in this study must also take care to control for the substantial correlations between orthography and underlying phonemics, for orthographic systems are not randomly structured and often reflect codification of phonological knowledge (see Chomsky 1970 for a discussion of English).

However, if one accepts that orthographic effects alone cannot account for the results observed, the question remains: How can it be that more phonemic sites are observed than would be expected by chance? Here, it is argued that the most parsimonious explanation for this result is language-specific phonological knowledge of the kind generally assumed by working phonologists. More precisely, the existence of phonemic sites suggests that the representations of speech sounds and words are not merely amalgamations of pronunciations that a listener has encountered before. Rather, there is a degree of abstraction between surface, acoustic forms and the prelexical forms that speech sounds are mapped onto in the brain. The nature of this abstraction is language-specific, reflecting not only differences in the phonemic inventories of different languages, but also the language-specific, contextually-sensitive sound alternations that comprise a specific language's phonological gram-

mar.

Corollary to this point, the often naive assumption that phonological structure is subservient to word recognition ought to be re-examined. This assumption has early origins in modern psycholinguistics, that cannot be explored exhaustively here. However, we can take Forster's (1979) seminal work "Levels of Processing and the Structure of the Language Processor" as a representative example, both given its age and influence. Forster's model admits three linguistic processing modules: the lexical processor, the syntactic processor, and the message processor. Of these three modules, the lexical processor is most closely related to phonological processing, the latter two modules being responsible for sentence and discourse comprehension, respectively. In Forster's (1979) model, "the lexical processor accepts input from peripheral perceptual systems in the form of feature lists and attempts to segment this input and to access the entries in the lexicon that correspond to the lexical elements in the input string" (p. 34).

In light of the results observed in this study, however, the assumption that discrete sublexical units merely support word recognition is inadequate. If this were truly the case, there would be no need for sensitivity to the kind of contrast exemplified by the phonemic sites in this study. Words could be segmented and identified on the basis of their acoustic form and semantic context alone, with no need to posit an abstract level of structure where sounds are processed differentially based on the surrounding acoustic context. However, the existence of phonemic sites in this study suggests that such an abstract level of structure does exist, since acoustic context, such as the presence of a preceding stressed vowel and subsequent unstressed vowel, *does* influence how otherwise identical sounds are processed. Certainly, an argument can be made that this kind of phonemic structure aids lexical recognition by providing increased predictability for certain subsequences of sounds (see Meinhardt et al. 2020). Thus, at the very least, theories of speech processing ought to include this level of structure as an intermediary between acoustic processing and lexical recognition.

In general, however, given the language-specificity of the tap, plural, and past tense alternations, it is most parsimonious to adopt a theory of the phonological

level that is analogous to one's theory of the syntactic level. That is, to see the relationship between phonology and syntax truly as a case of duality of patterning, phonological knowledge is language-specific grammatical knowledge of the same kind as syntactic knowledge, merely over a smaller scale domain. Consequent of this, phonological knowledge ought not be considered any more teleologically subservient to lexical recognition than syntactic knowledge is to semantic comprehension. Instead, as is generally assumed for syntax, phonological structure ought to be considered a design feature of language in its own right, enabling a range of processes in linguistic cognition, from speech comprehension in noisy and accented contexts to speech stream segmentation and, of course, lexical recognition.

Beyond evidence for the cognitive reality of the phoneme and the existence of language-specific phonological grammar, these findings also problematize theories of perception that posit syllables as a basic unit of linguistic processing. For example, the hierarchical state feedback control (HSFC) model of speech production postulates two levels of phonological processing for spoken languages, a syllable level and a segmental level. In this model, syllables constitute cycles of constriction in the oral tract, syllables are maximally CV in structure, and phonemes are defined by the abstract configuration of articulators at half-cycles of movement, roughly corresponding to the minima (C) and maxima (V) of jaw displacement (Hickok 2012, p. 138).[2]

Interestingly, the HSFC model admits neither auditory representations of segmental targets nor somatosensory representations of syllables. That is, the basic unit of acoustic representation for the HSFC model is the syllable, while the basic unit of articulatory representation is the segment. In this way, the HSFC model, although popular and influential, appears contrary to work indicating that higher auditory cortex, and even motor cortex, are densely organized by segment-level acoustic features (Mesgarani et al., 2014; Cheung et al., 2016; Hullett et al., 2016). Moreover, if the target of speech sound perception was the syllable, rather than the phoneme, we would expect to see neither significant numbers of acoustic sites nor phonemic sites, because in general, the identity of syllables containing target

[2]It is unclear how a word like *strengths* would be produced using this model.

sounds did not covary with either the phonetic or phonemic identity of the target sounds themselves.

Furthermore, phonological theories that assume the sole linguistic content of phonemes takes the form of distinctive features are also likely inadequate. If we assume, uncontroversially, that the difference between /t/ and /d/ and the difference between /s/ and /z/ is their value for the feature [±voice], then we would predict that surface sites for the past tense comparison, where tokens of /t/ are distinguished from /d/ regardless of their morphological content, would be the same as surface sites for the plural comparison, where tokens of /s/ and /z/ are distinguished. However, across all subjects and bands, less than 10% of surface sites are shared between the past tense and plural comparisons, as shown in Figure 2.8. Moreover, when phonemic sites identified from the tap comparison, in which tokens of /d/ and /t/ are also hypothetically distinguished by the single feature [±voice], are considered alongside plural and past tense surface sites, the number of sites shared in common drops to one. Thus, it is highly unlikely that these sites index the presence or absence of the feature [±voice], since we would not expect the addition of phonemic sites identified by the tap comparison to further exclude candidate [±voice] sites that were identified by the past tense and plural surface comparisons. In this way, the findings of this study compel careful reconsideration of distinctive feature theories in phonology.

For instance, Dresher (2011) lays out three ways that phonologists have conceptualized of the phoneme as an object of inquiry. One class of definitions characterizes the phoneme as a *physical reality*. Representative definitions of this genre describe the phoneme as a language-specific "family" of sounds that "count for practical purposes as if they were one and the same" (Jones 1967: 258, via Dresher 2011), perhaps because "the speaker has been trained to make sound-producing movements in such a way that the phoneme features will be present in the sound waves, and [the speaker] has been trained to respond only to these features" (Bloomfield 1933: 77-78). Psychological theories of language, such as Forster's (1979), are more or less consistent with this category of explanation, given their focus on direct mappings from physical sounds to lexical entries. The

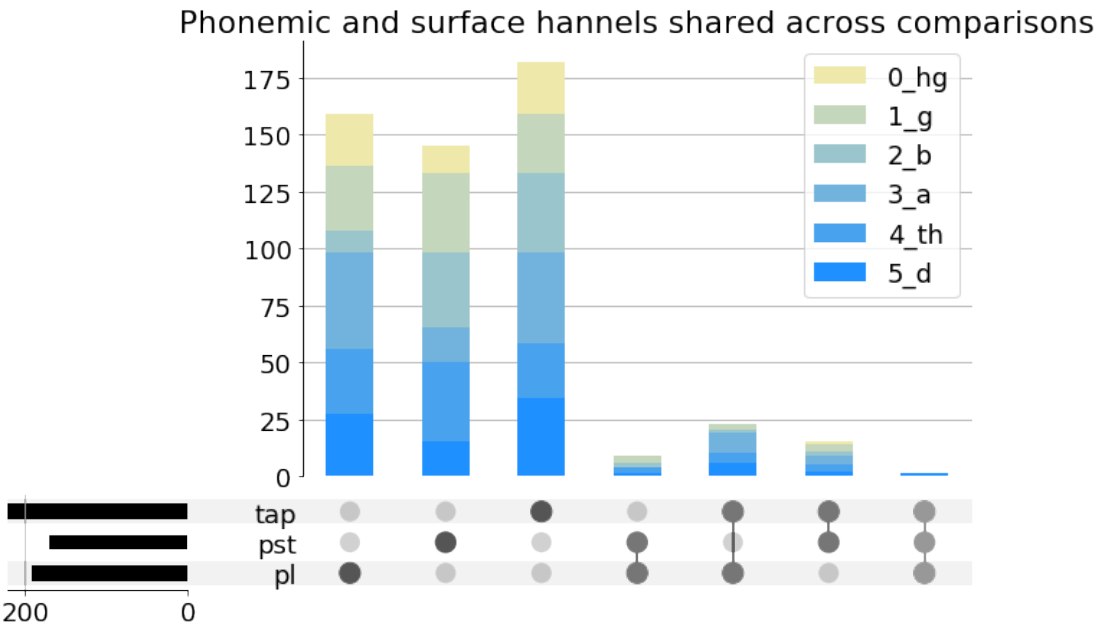

Figure 2.8: Phonemic and surface channels shared across comparisons, organized by response band. The lower dot matrix plot indicates the combination of comparisons being considered, and the upper bar plot shows the number of significant channels shared in common for that combination of bands. Colors within each bar indicate the distribution of significant channels across response bands.

second class of definitions views the phoneme as a *psychological concept*. This class of explanation is often presented as an alternative to the physical reality of the phoneme: if a physical constant coextensive with the phoneme does not exist, then perhaps a mental constant does instead. Finally, if both the physical and psychological reality of the phoneme are rejected, then Dresher (2011), echoing Twaddell (1935), concludes that the phoneme is a convenient *theoretical fiction*, without material basis in the mind, mouth, or middle ear.

In some sense, all three of these positions must be true, in different ways. There must exist some relationship between the physical properties of speech sounds and the roles they play in language, since the experience of spoken language comprehension has a principled relationship to physical speech. Additionally, there must exist some physical substrate in the brain to bridge the sensory periphery with the experience of language comprehension. In the abstract, this bridge may be considered to be composed of mental representations. However, ultimately, not all

phonological representations that have been theorized will have equal explanatory adequacy in the brain. Thus, some aspects of the phoneme will remain useful theoretical fictions.

The study presented in this paper addresses the second of these three positions, investigating whether the commonly held understanding of the phoneme as a unit of language-specific contrast has psychological reality or merely theoretical utility. In demonstrating the existence of sites sensitive to phonemic contrast in the absence of acoustic distinction, this study provides support for the classical understanding of the phoneme as a psychological entity and a core level of sublexical linguistic processing. At the same time, it suggests that some theoretical implementations of the phoneme as a unit characterized by minimal sets of distinctive features may have more theoretical utility than psychological reality. To the extent that linguistic and psychological theories of phonology intend to account for the same sets of behavioral phenomena, these results nevertheless support the continued use of phonemic units in phonological theory and reify the existence of such language-specific sublexical structure in language processing.

2.4.2 Implications for morphology

The morphological results of this study also engage with foundational principles of sublexical language processing. In observing more surface and morphological sites than would otherwise be expected by chance for both the plural and past tense comparisons, this study provides evidence for two interrelated processes in the neural basis of morphology.

It is generally accepted that morphological decomposition takes place during the processing of regularly inflected forms in English and related languages (Münte et al. 1999; Marslen-Wilson and Tyler 2007; Bozic and Marslen-Wilson 2010; Schiller 2020; though cf. Sereno and Jongman 1997). The presence of morphological sites in this study is generally consistent with these results. However, the paradigms typically employed to assess the compositionality of morphological units within words arguably gauge fairly abstract proxies for morphological structure. Many rely on priming, and assess compositionality through metrics such as

reaction time or expectation violation response. The evidence for morphological abstraction presented here has the advantage being gathered during a naturalistic listening task and straightforwardly comparing the difference in neural response to sounds that bear morphological exponence to those that do not carry any morphological meaning.

From this perspective, the difference in response between non-plural word-final [z] and plural [s] or [z] can be attributed straightforwardly to the morphological exponence of [s] or [z], without appeal to intermediary cognitive phenomena such as priming or expectation. That is, the meaning of 'plural' associated with the sounds [s] and [z] drives the response of the morphological sites for the plural comparison, and likewise, the meaning of 'past' associated with [t] and [d] drives the response of the morphological sites for the past tense comparison. This, in general, is a great strength of the paradigm employed in this study. However, these results themselves do not provide explicit evidence that the structure of regular plural and past tense forms is compositional, since it could be the case that, for instance, plural [s] and [z] pattern together at morphological sites through an analogical process. In such a case, plural [s] and plural [z] would still evoke similar neural responses because they index a common morphological exponent, but that commonality would be mediated through an analogical process via the lexicon rather than a compositional, syntax-like process within the word itself. Nevertheless, the results of this study provide tantalizing early evidence of the neural response to the presence of particular morphological exponents.

Moreover, these results support the idea that morphological identity is abstracted over phonologically distinct alternants in a structured, language specific manner. As was argued for the phonemic sites identified by the tap comparison, the structured relationships between speech sounds and their phonological contexts form a language-specific grammar of sounds. Given their interface with meaning, morphophonological alternations in particular have been of central importance to the development of linguistic theory (see Kenstowicz and Kisseberth 2014) and are often appealed to in classic texts as foundational evidence for the existence of the phoneme, since it is through the ways that sounds either change or fail to change

the meanings of words that phonemic contrast is most strikingly established. In this way, the existence of morphological sites identified by the plural and past tense comparisons demonstrates both that morphological identity is abstracted over distinct, phonologically conditioned alternants and that morphological exponence transcends phonemic particularities.

2.4.3 Localization and timing

Determining the nature and origin of the neural processes that support language processing is crucial for building a mechanistic understanding of linguistic cognition. In this study, band-limited power was examined for six frequency ranges. However, the interpretation of this measure, including its relation to other neural measures and its cellular origins, varies depending upon where in the brain it is recorded.

In intracranial recordings, LFP recorded from grey matter is credited primarily to local synaptically induced current flow. On the other hand, LFP recorded from white matter is understood to originate from a combination of cortical activity spread by volume conduction into adjacent white matter as well as current flows through the voltage-gated channels that mediate the travel of action potentials along axons (Mercier et al., 2017). Thus, while the dominant components of electrical activity recorded in grey matter result from the activity of a spatially restricted set of cells proximal to the electrode contact, activity recorded from white matter also contains signal that originates from a combination of cells, possibly separated by great distance, that happen to project axonal processes within proximity of the electrode contact. Additionally, the volume conduction properties of white and grey matter differ. Whereas LFP in grey matter (passively) propagates more-or-less equally in all directions—that is, both parallel and perpendicular to the cortical surface—propagation in white matter is directionally biased by the higher density of axonal myelin, such that passive spread occurs more readily along myelin sheaths rather than across them (Kajikawa and Schroeder, 2011, 2015). These key differences in the properties of grey and white matter impact the functional interpretation of electrophysiological measures collected from these

tissues.

Thus, while activity recorded from white matter is certainly physiologically valid, its interpretation differs from activity recorded in cortical tissue. At present, the electrophysiological properties of white matter are relatively under-explored, and consequently, activity recorded in white matter is considered difficult to interpret and likely under-reported when it is recorded. Accordingly, while this paper does not speculate on how any of the significant sites reported in Section **??** fit into larger research narratives about the language network, it nevertheless reports results that are primarily driven by activity recorded in white matter. In doing so, this paper contributes to an early understanding of what language-related electrophysiological activity looks like when recorded from these sites.

Relatedly, it is *a priori* unclear whether the time course of language-related activity recorded in white matter ought to follow the same time course as activity recorded from the scalp or grey matter. Some work suggests that the acoustic processing of phonetic detail takes place within 100–200ms following the onset of a speech sound (Chang et al., 2010; Toscano et al., 2010), activity related to speech sound categorization roughly 300–500ms after the sound's onset (Toscano et al., 2010), and morphological processing occurs even later (Leminen et al., 2019). However, other work suggests a messier temporal story, where both categorical and gradient speech sound information are represented simultaneously in the neural signal (Sarrett et al., 2020; Gwilliams, 2020), along with morphological information (Munding et al., 2016; Gwilliams, 2020). To the extent that it is possible to compare results observed in EEG, MEG, and ECoG data to those observed in SEEG contacts embedded in white matter, the results of this study appear to support the messier story.

For the coronal tap comparison in particular, the majority of sites that show an acoustic pattern occur in the gamma and high gamma bands. In these bands, the time windows where significant acoustic responses occurred correlate highly with the windows where significant phonemic responses also occurred (Figure 2.5). These results suggest that both phonemic category and acoustic detail are processed simultaneously and overwhelmingly separately, since relatively few sites

across comparisons exhibited both acoustic and phonemic patterns or surface and morphological patterns (Figure 2.7).

2.5 Conclusion

Using a novel experimental design based on the foundational concepts of phonological contrast and neutralization, this chapter investigates the representation of phonological units in the brain and the relationship between those units, auditory sensory input, and higher levels of language organization, namely morphology. Leveraging the phonological neutralization of coronal stops to tap in English, this study provides evidence of a dissociation between acoustic similarity and phonemic identity across sites recorded intracranially while participants listened to conversational speech. Moreover, leveraging morphophonological alternations of the regular plural and past tense, this study further demonstrates early (<500ms) evidence of dissociation between phonological form and morphological exponence. Together these results highlight the central nature of language-specific knowledge in sublexical language processing and improve our understanding of the ways language-specific knowledge structures and organizes speech perception in the brain.

Material from Chapter 2 has been submitted for publication as "The neural response to speech is language specific and irreducible to speech acoustics" with authors Anna Mai, Stephanie Riès, Sharona Ben-Haim, Jerry Shih, & Timothy Gentner. The dissertation author was the primary investigator and author of this paper.

Bibliography

Alturo, N., Boix, E., and Perea, M.-P. (2002). Corpus de Català contemporani de la Universitat de Barcelona (CUB). a general presentation. *dins C. PUTSCH*, pages 155–170.

Berezutskaya, J., Freudenburg, Z. V., Güçlü, U., van Gerven, M. A., and Ramsey, N. F. (2017). Neural tuning to low-level features of speech throughout the perisylvian cortex. *Journal of Neuroscience*, 37(33):7906–7920.

Bloomfield, L. (1933). *Language*. New York: Holt.

Boersma, P. and Weenink, D. (2017). Praat: doing phonetics by computer (version 6.0. 28)[software].

Bouchard, K. E., Mesgarani, N., Johnson, K., and Chang, E. F. (2013). Functional organization of human sensorimotor cortex for speech articulation. *Nature*, 495(7441):327–332.

Bozic, M. and Marslen-Wilson, W. (2010). Neurocognitive contexts for morphological complexity: Dissociating inflection and derivation. *Lang. Linguist. Compass*.

Braver, A. (2013). Incomplete neutralization in American English flapping: A production study. *Proc. Mtgs. Acoust.*, 19(1).

Chan, A. M., Dykstra, A. R., Jayaram, V., Leonard, M. K., Travis, K. E., Gygi, B., Baker, J. M., Eskandar, E., Hochberg, L. R., Halgren, E., et al. (2014). Speech-specific tuning of neurons in human superior temporal gyrus. *Cerebral Cortex*, 24(10):2679–2693.

Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., and Knight, R. T. (2010). Categorical speech representation in human superior temporal gyrus. *Nature neuroscience*, 13(11):1428–1432.

Charles-Luce, J. (1997). Cognitive factors involved in preserving a phonemic contrast. *Lang. Speech*, 40 (Pt 3):229–248.

Charles-Luce, J. and Dressler, K. M. (1999). The effects of semantic predictability in non-pathological older adults' production of a phonemic contrast. *Clinical linguistics & phonetics*, 13(3):199–217.

Charles-Luce, J., Dressler, K. M., and Ragonese, E. (1999). Effects of semantic predictability on children's preservation of a phonemic voice contrast. *Journal of Child Language*, 26(3):505–530.

Cheung, C., Hamilton, L. S., Johnson, K., and Chang, E. F. (2016). The auditory representation of speech sounds in human motor cortex. *elife*, 5:e12577.

Chomsky, C. (1970). Reading, writing, and phonology. *Harvard educational review*, 40(2):287–309.

Chomsky, N. and Halle, M. (1968). The sound pattern of English.

Cibelli, E. S., Leonard, M. K., Johnson, K., and Chang, E. F. (2015). The influence of lexical statistics on temporal lobe cortical dynamics during spoken word listening. *Brain and language*, 147:66–75.

Crone, N. E., Miglioretti, D. L., Gordon, B., and Lesser, R. P. (1998). Functional mapping of human sensorimotor cortex with electrocorticographic spectral analysis. II. event-related synchronization in the gamma band. *Brain*, 121 (Pt 12):2301–2315.

Derrick, D. and Schultz, B. (2013). Acoustic correlates of flaps in North American English. In *Proceedings of Meetings on Acoustics ICA2013*, volume 19. Acoustical Society of America.

Dresher, B. E. (2011). The phoneme. *The Blackwell companion to phonology*, pages 1–26.

Fedorov, A., Beichel, R., Kalpathy-Cramer, J., Finet, J., Fillion-Robin, J.-C., Pujol, S., Bauer, C., Jennings, D., Fennessy, F., Sonka, M., et al. (2012). 3d slicer as an image computing platform for the quantitative imaging network. *Magnetic resonance imaging*, 30(9):1323–1341.

Fischer, W. and Hirsch, I. (1976). Intervocalic flapping in English. In *Papers from the Regional Meeting of the Chicago Ling. Soc. Chicago, IL*, number 12, pages 183–198.

Flinker, A., Chang, E., Barbaro, N., Berger, M., and Knight, R. (2011). Sub-centimeter language organization in the human temporal lobe. *Brain and language*, 117(3):103–109.

Forster, K. I. (1979). Levels of processing and the structure of the language processes. *Sentence processing: Psycholinguistic studies presented to Merril Garret*, pages 27–84.

Fox, R. A. and Terbeek, D. (1977). Dental flaps, vowel duration and rule ordering in American English. *J. Phon.*, 5(1):27–34.

Gussenhoven, C. and Jacobs, H. (2017). *Understanding phonology*. Routledge.

Gwilliams, L. (2020). How the brain composes morphemes into meaning. *Philosophical Transactions of the Royal Society B*, 375(1791):20190311.

Herd, W., Jongman, A., and Sereno, J. (2010). An acoustic and perceptual analysis of /t/ and /d/ flaps in American English. *J. Phon.*, 38(4):504–516.

Hillenbrand, J., Ingrisano, D. R., Smith, B. L., and Flege, J. E. (1984). Perception of the voiced–voiceless contrast in syllable-final stops. *J. Acoust. Soc. Am.*, 76(1):18–26.

Hogan, J. T. and Rozsypal, A. J. (1980). Evaluation of vowel duration as a cue for the voicing distinction in the following word-final consonant. *J. Acoust. Soc. Am.*, 67(5):1764–1771.

Holdgraf, C. R., De Heer, W., Pasley, B., Rieger, J., Crone, N., Lin, J. J., Knight, R. T., and Theunissen, F. E. (2016). Rapid tuning shifts in human auditory cortex enhance speech intelligibility. *Nature communications*, 7(1):1–15.

Hullett, P. W., Hamilton, L. S., Mesgarani, N., Schreiner, C. E., and Chang, E. F. (2016). Human superior temporal gyrus organization of spectrotemporal modulation tuning derived from speech stimuli. *Journal of Neuroscience*, 36(6):2014–2026.

Jakobson, R., Fant, G., and Halle, M. (1963). Preliminaries to speech analysis: the distinctive features and their correlates.

Johnson, H., Harris, G., Williams, K., et al. (2007). Brainsfit: mutual information rigid registrations of whole-brain 3d images, using the insight toolkit. *Insight J*, 57(1):1–10.

Jones, D. (1967). *The Phoneme: Its Nature and Use*. Cambridge University Press.

Kajikawa, Y. and Schroeder, C. E. (2011). How local is the local field potential? *Neuron*, 72(5):847–858.

Kajikawa, Y. and Schroeder, C. E. (2015). Generation of field potentials and modulation of their dynamics through volume integration of cortical activity. *Journal of neurophysiology*, 113(1):339–351.

Kenstowicz, M. and Kisseberth, C. (2014). *Generative Phonology: Description and Theory*. Academic Press.

Leminen, A., Smolka, E., Dunabeitia, J. A., and Pliatsikas, C. (2019). Morphological processing in the brain: The good (inflection), the bad (derivation) and the ugly (compounding). *Cortex*, 116:4–44.

Leonard, M. K., Baud, M. O., Sjerps, M. J., and Chang, E. F. (2016). Perceptual restoration of masked speech in human cortex. *Nature communications*, 7(1):1–9.

Malécot, A. and Lloyd, P. (1968). The /t:d/ distinction in American alveolar flaps. *Lingua*, 19(3-4):264–272.

Marslen-Wilson, W. D. and Tyler, L. K. (2007). Morphology, language and the brain: the decompositional substrate for language comprehension. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 362(1481):823–836.

Meinhardt, E., Mai, A., Bakovic, E., and McCollum, A. (2020). Questioning to resolve transduction problems. *Proceedings of the Society for Computation in Linguistics*, 3.

Mercier, M. R., Bickel, S., Megevand, P., Groppe, D. M., Schroeder, C. E., Mehta, A. D., and Lado, F. A. (2017). Evaluation of cortical local field potential diffusion in stereotactic electro-encephalography recordings: a glimpse on white matter signal. *Neuroimage*, 147:219–232.

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

Mesgarani, N., Cheung, C., Johnson, K., and Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*, 343(6174):1006–1010.

Munding, D., Dubarry, A.-S., and Alario, F.-X. (2016). On the cortical dynamics of word production: A review of the MEG evidence. *Language, Cognition and Neuroscience*, 31(4):441–462.

Münte, T. F., Say, T., Clahsen, H., Schiltz, K., and Kutas, M. (1999). Decomposition of morphologically complex words in English: evidence from event-related brain potentials. *Brain Res. Cogn. Brain Res.*, 7(3):241–253.

Nourski, K. V., Steinschneider, M., Rhone, A. E., Kovach, C. K., Kawasaki, H., and Howard III, M. A. (2019). Differential responses to spectrally degraded speech within human auditory cortex: an intracranial electrophysiology study. *Hearing research*, 371:53–65.

Pattamadilok, C., Morais, J., Colin, C., and Kolinsky, R. (2014). Unattentive speech processing is influenced by orthographic knowledge: Evidence from mismatch negativity. *Brain and Language*, 137:103–111.

Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *J. Neurosci. Methods*, 162(1):8–13.

Peirce, J. W. (2008). Generating stimuli for neuroscience using PsychoPy. *Front. Neuroinform.*, 2:10.

Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., and Fosler-Lussier, E. (2007). Buckeye corpus of conversational speech (2nd release). *Columbus, OH: Department.*

Revoile, S., Pickett, J. M., Holden, L. D., and Talkin, D. (1982). Acoustic cues to final stop voicing for impaired-and normal hearing listeners. *J. Acoust. Soc. Am.*, 72(4):1145–1154.

Sarrett, M. E., McMurray, B., and Kapnoula, E. C. (2020). Dynamic EEG analysis during language comprehension reveals interactive cascades between perceptual processing and sentential expectations. *Brain and language*, 211:104875.

Schiller, N. O. (2020). Neurolinguistic approaches in morphology. *Oxford Research Encyclopedia, Linguistics*, pages 1–23.

Sereno, J. A. and Jongman, A. (1997). Processing of English inflectional morphology. *Mem. Cognit.*, 25(4):425–437.

Sharf, D. J. (1962). Duration of Post-Stress intervocalic stops and preceding vowels. *Lang. Speech*, 5(1):26–30.

Toscano, J. C., McMurray, B., Dennhardt, J., and Luck, S. J. (2010). Continuous Perception and Graded Categorization: Electrophysiological Evidence for a Linear Relationship Between the Acoustic Signal and Perceptual Encoding of Speech. *Psychol. Sci.*, 21(10):1532–1540.

Twaddell, W. F. (1935). On defining the phoneme. *Language*, 11(1):5–62.

Zue, V. W. and Laferriere, M. (1979). Acoustic study of medial /t, d/ in American English. *J. Acoust. Soc. Am.*, 66(4):1039–1050.

Chapter 3

Cross-frequency Variation in Speech Encoding

3.1 Introduction

3.1.1 Oscillatory dynamics of speech processing

In recent decades, oscillations have been proposed as a fundamental mechanism of cognition generally (Bastos et al., 2012; Buzsáki, 2006; Siegel et al., 2012; Womelsdorf et al., 2007) and of speech and language processing specifically (Giraud and Poeppel, 2012; Lewis et al., 2015). In particular, interest in neural oscillations has emerged in language research as part of a broader paradigm shift, as researchers come to the consensus that language results from the complex interplay of neural activity across numerous areas in both hemispheres, coordinated into several distributed networks (Chai et al. 2016; Poeppel and Embick 2017, inter alia). Because oscillations offer both a plausible mechanism for coordinating activity across such spatially distant regions (Bastos and Schoffelen, 2016; Palva and Palva, 2012) and because oscillations themselves may subserve some speech processing functions (i.e., parsing the speech stream, Kösem et al. 2016; Ding et al. 2016), oscillations have been critical to the ongoing reconceptualization of language representation in the brain.

Biophysiologically, oscillations are thought to reflect the coordinated post-

synaptic fluctuations of neural populations that may vary from thousands to millions of neurons in size (Cohen, 2017). When excitatory and inhibitory activity balance one another in the right way, these fluctuations synchronize at particular frequencies, resulting in band-limited oscillations (Buzsáki et al., 2012). The amplitude of these oscillations fluctuates over time in a manner that is thought to reflect the coordination of particular computations in the neural populations driving the oscillatory behavior, and is often time- and/or phase-locked to psychological events of interest (Klimesch, 1999).

The majority of oscillation-related language research to date focuses on some subset of the five classical frequency bands: delta (1-2 Hz), theta (3-7 Hz), alpha (8-12 Hz), beta (13-30 Hz) and gamma (30-200 Hz),[1] and as this work has developed, several bands—namely, delta, theta, and gamma bands—have emerged as having particular significance for speech processing and comprehension. Activity within and across these bands appears to be coordinated in a manner that supports different functions in the language processing cascade.

The simplest vision of the relationship between oscillatory bands and language structures ascribes the processing of a linguistic phenomenon to a band of like frequency. And indeed, lower frequency bands (delta, theta) have been shown to entrain to lower frequency language dynamics, such as syllable rate (Assaneo and Poeppel, 2018) and underlying phrase structure (Ding et al., 2016), while higher frequency bands (gamma, high gamma) entrain to higher frequency dynamics, such as speech sound processing (Mesgarani et al., 2014). However, this narrative is complicated by work showing entrainment to phrase structure in high gamma activity (Nelson et al., 2017) and increases in gamma band power during semantic processing (Wang et al., 2012; Peña and Melloni, 2012) as well as encoding of speech sound category, phoneme-like information in delta and theta bands (Di Liberto et al., 2015). Thus, higher frequencies are not solely sensitive to higher frequency language structure, and lower frequencies are not solely sensitive to lower frequency language structure; nor does the band frequency that entrains to a particular linguistic unit scale straightforwardly with that linguistic unit's de-

[1] That is, relatively little language research has been done that calculates individual- or group-specific bands.

gree of abstraction (e.g., if higher bands tracked purely acoustics and lower bands tracked syntactic dependencies).

It seems therefore that whatever function(s) band-limited oscillations subserve in language processing must be more complex than one based purely on an alignment of neural and linguistic rhythms. The unified role (if there is to be one) that each of these bands plays in the pageant of language processing remains unclear, yet as work continues to accrue, more complex theories have emerged concerning how activity at different frequencies supports language processing. In particular, several accounts focus on interactions observed between delta, theta, and gamma bands that are specific to language. In addition to its association with attention- and prediction-based modulations of auditory processing (Lakatos et al., 2008; Nozaradan et al., 2011; Stefanics et al., 2010; Golumbic et al., 2013), delta band activity has been implicated in the processing of abstract language structure, namely simple phrase structure parsing (Ding et al., 2016). Theta band activity has been implicated in a wide range of speech sound processing tasks, including phonemic restoration (Riecke et al., 2009, 2012; Strauß et al., 2014; Sunami et al., 2013), consonant identification (Ten Oever and Sack, 2015), and word boundary parsing (Kösem et al., 2016); and contains categorical phoneme-level information (Di Liberto et al., 2015). These studies, among others, illustrate that theta activity is capable of coding for information with far higher temporal precision than the central frequency of the band itself. Higher frequency activity, in gamma and high gamma bands, tracks detailed speech sound information (Mesgarani et al., 2014; Hullett et al., 2016), and is modulated by attention (Mesgarani and Chang, 2012; Pasley et al., 2012; Golumbic et al., 2013) and higher order language knowledge, such as lexical knowledge (Cibelli et al., 2015; Hannemann et al., 2007; Mai et al., 2016), language context (i.e., native vs. foreign language, see Peña and Melloni 2012), and word-level parsing (Kösem et al., 2016; Ding et al., 2016).

Recent computational models have attempted to show that cross-frequency coupling between low- and high-frequency bands could provide a tractable schema for unifying single-band results into the larger picture of language processing. In particular, phase-amplitude coupling between theta and gamma bands has been used

to model segmentation of speech into syllable units by Ghitza (2011), Hyafil et al. (2015), and Shamir et al. (2009). These models, while tantalizing, are limited by their reliance on theta as the "master oscillator" coordinating all higher-frequency activity; that is, they are limited by the assumption that neural rhythms and linguistic rhythms are in more-or-less one-to-one correspondence, as described above. So while it is unlikely that these models correctly describe the totality of ways that oscillations support language processing, it may still be the case that these models accurately describe oscillatory mechanisms at certain times in certain regions. It may be the case that multiple cross-frequency coupling regimes underpin language processing —of which theta-gamma coupling would be just one— with different regimes supporting different language processes in spatially distinct areas, as explored by Fontolan et al. (2014) and Park et al. (2015). Relatedly, it may be the case that different oscillatory sources use the same frequency band to encode different linguistic units, as suggested by Ding et al. (2016). And moreover, variation in the phase of coupling between two bands may provide an additional dimension by which the brain tracks a particular perceptual feature or signals a change in oscillatory regime, etc. (Jensen et al., 2012, 2014; Kösem et al., 2014; Ng et al., 2013; Panzeri et al., 2010). Complex coupling schemes thus offer the most promising path towards understanding the contributions of band-limited activity in language processing.

To date, many effects in this field are reported for single bands, and subsequent work clusters around previous results, investigating a particular phenomenon in a particular band. While immanently reasonable, this mode of progress has a tendency to bias the field towards associating certain bands with certain phenomena and contributing to theories that, although capable of explaining data from many papers, capture a relatively small proportion of the diversity of results reported for any individual band. For example, a relatively large literature exists on theta entrainment to syllable rate (see Poeppel and Assaneo 2020 for a review), and consequently, modeling and theoretical work on theta-gamma coupling focuses on how theta rhythms may support syllable parsing (see Ghitza 2011; Poeppel and Assaneo 2020, etc.), while theta's role in other language phenomena, such track-

ing phonemic identity (Di Liberto et al., 2015) or phrase structure (Ding et al., 2016), are left more-or-less unincorporated into band-based theories of linguistic computation and cognition.

The study of cross-frequency coupling in language processing specifically is relatively recent, with less than a handful of experimental results currently reported in the literature. The few pioneering articles report unexpected results that do not map neatly onto the authors' predictions (Brennan and Martin, 2020).

A more thorough understanding of the eliciting conditions for activity in any individual band will drive stronger theories of the eliciting conditions for activity across bands. To that end, this chapter examines the relative explanatory power of acoustic and phonological category information for each of the six canonical functional bands. Moreover, in the context of this dissertation specifically, the analysis presented here bolsters the work presented in the previous chapter by providing a more general analysis of the extent to which neural activity is shaped by categorical speech sound identity above and beyond the concurrent continuously varying acoustic signal.

3.2 Methods

3.2.1 Linear mixed effects models

Linear mixed effects (LME) models extend simple linear models, allowing for both fixed and random effects to be modeled. They are particularly useful for contexts with hierarchically structured non-independence between datapoints, such as a dataset of formant values with multiple measurements sampled per vowel quality per speaker, or for longitudinal datasets with multiple measurements taken from the same unit over time, such as an EEG dataset containing multiple band power measurements sampled per electrode contact.

The correlated errors that arise (e.g., within speaker or within channel) in these types of datasets violate the assumptions of standard ANOVA and regression models. Thus, to deal with the non-independence of the measurements using a simple linear model, either datapoints must be averaged within the maximal node of each

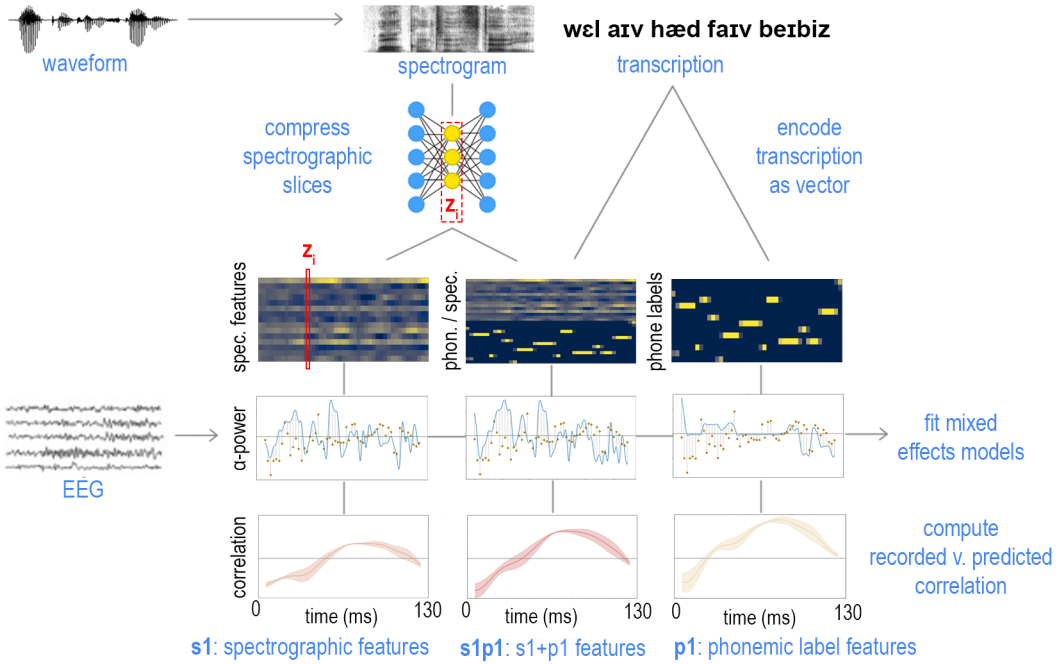

Figure 3.1: LME approach models neural activity as a combination of spectrographic and phonemic label features. For each stimulus waveform, spectrograms are computed and time-aligned phonemic-level transcriptions are assigned (top row). Spectrograms are compressed into 128-dimensional vectors using a generative adversarial autoencoder network, and transcriptions are one-hot encoded. Three classes of model are created from these features: a purely spectrographic model (left column), a purely phonemic label model (right column), and a model containing both feature sets (middle column). For each band of neural activity, mixed effects models are fit. Model weights are used to reconstruct a predicted response for each band, and the correlation between the predicted and recorded neural response is calculated.

hierarchy, or separate models must be fit for each maximal node. Averaging within categories allows the analyst to make statistical generalizations about the entire dataset, but risks losing meaningful variability within groups. Conversely, creating individual regressions for each category avoids the loss of variability through averaging, but at the expense of being able to make generalizations about the whole dataset. LME models balance these two alternatives by fitting weights both within and across groups.

A simple linear model, given by the equation $y = \beta x_i + \alpha$, assumes that a true regression line β exists, and estimates its value $\hat{\beta}$. A mixed effects model assumes

instead that β is a random normal variate with its own mean μ and standard deviation σ, modeled through the random-effects coefficients, u. The total mixed effects model is given by the equation

$$y = X\beta + Zu + \eta$$

where y is an $N \times 1$ column vector corresponding to$N$ observations of the response variable; X is a $N \times p$ matrix containing the values for each of p predictors for each response value; β is a $p \times 1$ column vector of regression coefficients for the fixed effects; Z is a design matrix for q random effects and J groups, with size $N \times qJ$; u is a $qJ \times 1$ vector of regression coefficients for the random effects; and η is the $N \times 1$ vector of residuals, accounting for the components of y that are not explained by the fixed and random terms.

When fitting the model, rather than estimating u directly, as is done for β, it is generally assumed that u is a normally distributed variate with a mean of zero and a variance of G. It is through this assumption that random effects model the variability of β: the estimated $\hat{\beta}$ is the mean of the assumed β distribution, and the mean of u is assumed to be zero to model that the random effects are simply deviations from the fixed effect. This leaves only the variance G to estimate.

When fitting a model with only one random intercept, G is simply a 1×1 matrix containing the variance of the random intercept. However, in general, for a model with n random effects, G is an $n \times n$ variance-covariance matrix. As a variance-covariance matrix, G is square, symmetric, positive semidefinite, and contains numerous redundant elements. Given these properties, rather than estimating G directly, we estimate θ, a matrix that contains only non-redundant elements (e.g., through a triangular Cholesky factorization of $G = LDL^T$), typically parameterized in such as way as to ensure that the estimates are stable (e.g., positive, such as by taking the natural log). Regardless of the precise parameterization, G is some function of θ, and θ is estimated in place of G.

The final element of a mixed effects model that must be estimated is the variance-covariance matrix of the residuals, R. Most commonly, the residual structure is given as $R = I\sigma_\eta^2$, where I is the identity matrix, and σ_η^2 is the residual

variance. This structure assumes that all observations are conditionally independent and that their residual variance is homogeneous. Altogether, a mixed effects model of form

$$y = X\beta + Zu + \eta$$

is fit through the estimation of $\hat{\beta}$, $\hat{\theta}$, and $\hat{R}$.

3.2.2 Spectrographic Features

For each timepoint of the response, a set of features representing the preceding acoustic context was calculated. First, each excerpt was resampled to 16 kHz, and a spectrogram was computed for the downsampled file using the `spectrogram` function from the `scipy.signal` Python library. For each excerpt, this function calculated consecutive Fourier transforms over segments of the excerpt waveform that were 128 time bins (=8ms) in length with no zero padding. Segments were windowed using a Hann function and had 64 time bins of overlap with one another. This function resulted in a spectrogram with 65 frequency bins. The DC component of the spectrogram was removed, and the remaining 64 frequency bins were log transformed and then pairwise-averaged twice, resulting in a spectrogram with 16 frequency bins. Time bins were then pairwise-averaged three times, such that each time bin of the final spectrogram represents a portion of time eight times longer than the initial resolution (=64ms). Finally, the full spectrogram, representing the entire excerpt, was split into overlapping 16×16 chunks, such that each time bin of the total excerpt spectrogram—from the sixteenth bin to the final bin—could be mapped to a 16×16 (=1024ms) chunk representing the spectrographic content immediately preceding that time bin.

The final sample rate of the spectrogram time bins was 15.625Hz. However, all preprocessed neural data had sample rates of 510Hz, meaning that each spectrogram time bin corresponded to between 32 and 33 timepoints of the neural data. Therefore, to place the spectrographic features and neural response in one-to-one correspondence with one another, for each spectrographic time bin, the samples of

the neural response that took place during that time bin were averaged together.

As 16×16 chunks, each spectrogram would contribute 256 features to the mixed effects model, a computationally prohibitive number of features to fit over hundreds of thousands of datapoints. Thus, to further reduce the dimensionality of the spectrographic features without resorting to further averaging and loss of spectrotemporal resolution, 128-dimensional representations of each 16×16 spectrogram were drawn from the latent space of a Generative Adversarial Interpolative Autoencoder (GAIA; Sainburg et al. 2018) that had been trained on these spectrograms.

GAIA networks are a kind of neural network whose architecture is based on a combination of Autoencoder (AE) and Generative Adversarial Network (GAN) architectures. AEs are a class of network trained to take an input x_i and generate a reproduction of the input $G(x_i)$ that minimizes some error function between the input and output. In the process of performing this mapping, x_i is typically compressed into some low-dimensional representation, z_i, called the latent representation. The portion of the network that maps x_i from the input domain X to z_i in the latent domain Z is called the encoder, and the portion that maps z_i to $G(x_i)$ is the decoder.

GANs are a class of network that combines two subnetworks, a generator and a discriminator. The generator samples from the latent layer z_i and is trained to produce a sample $G(z_i)$ of the same data type as X. The discriminator then takes both $G(z_i)$ and some x_i as input and is trained to identify which of the samples is a true member of the input set X and which is generated. The generator and discriminator networks are adversarial to one another: as the generator becomes more successful at reproducing the statistics of X, it becomes more difficult for the discriminator to correctly differentiate between x_i and $G(z_i)$, and as the discriminator becomes more successful at discriminating between x_i and $G(z_i)$, the generator is pushed to reproduce the distribution of X more and more faithfully.

A GAIA network combines these two architectures by using AEs for both the generator and discriminator of a GAN. One of the strengths of using a GAIA model to reduce the dimensionality of the spectrographic datasets derives from

properties of its latent layer, Z. Both GANs and AEs are commonly used for the properties of their latent layer, which acts as a low-dimensional representation of the input domain that preserves core dimensions of its variability. For example, the dimensions in Z often correspond to features that are meaningful within the semantics of the input domain, such as whether or not a person is wearing glasses for a dataset of face images, or whether or not one crosses the stem of the number seven in a dataset of written digits. In typical AEs, the distributional properties of Z are not constrained, meaning that if one were to sample a point z_j in Z that does not correspond exactly to the encoding of some training item but does lie between two points that *are* encodings of real datapoints x_i and x_k, it's possible that z_j may not look like a plausible sample of X when mapped through the decoder. Since the most desirable property of a network's latent space is its ability to act as a meaningful compression of the input domain, the non-convexity of the standard AE latent space and the resultant regions of uninterpretable z-space are not ideal. Since the generator networks of GANs sample the distribution of Z directly, they can be constrained to avoid non-convex latent space, reducing the likelihood of decoding a sample from z-space that does not resemble the input domain at all.

An additional strength of the GAIA architecture is its ability to generate sharp, detailed reconstructions of input data. On their own, AEs tend to produce blurry images because they use pixel-wise loss functions to reproduce the input data, and these functions tend to minimize error by smoothing over sharp contrasts, like edges, in the input data. Since GANs are trained instead to generate samples that could plausibly come from X, they generally produce much sharper images because the discriminator network could use smoothed edges to successfully determine whether a sample was drawn from X or $G(X)$.

However, on their own, GANs contain no mapping from X to Z, meaning that real input data have no direct representation in Z. The latent space of a GAN represents the statistics of X, but not X itself. For this reason, a GAN alone is not suitable for mapping specific spectrographic segments into lower dimensional space. Incorporating AEs into a GAN using a GAIA network architecture solves this problem. In a GAIA model, the discriminator performs two optimizations,

minimizing the pixel-wise loss between an input x_i and its reconstruction $D(x_i)$ by the discriminator AE, and maximizing the loss between a generated sample $G(x_i)$ and its reconstruction $D(G(x_i))$ by the discriminator network. As in a standard GAN, the generator works adversarially to minimize pixel-wise error between x_i and $D(G(x_i))$ in order to produce generated samples that are as similar to the true data as possible.

The GAIA model was built in Tensorflow 2.2 as a class of `tensorflow.keras.Model`. The network's architecture is shown schematically in Figure 3.2. The AE supporting the generator role of the GAN contained two basic subnetworks, an encoder and a decoder.

The encoder contained an input layer for batches of 16×16 tensors followed by two convolutional layers with 32 and 64 filters, respectively. Both convolutional layers had RELU activation, a kernel size of 3, and a 2×2 stride. The output of the second convolutional layer was flattened into a single dimensional tensor using a `tf.keras.layers.Flatten` layer and its output was sent to a densely-connected layer of 128 units.

The decoding subnetwork was symmetrical to the encoder, taking the output of the 128-dimensional z layer as the input to a densely-connected layer. The densely-connected layer fed into a `tf.keras.layers.Reshape` layer that returned its input into a multi-dimensional tensor. This layer was then followed by three deconvolutional layers. The first two deconvolutional layers had 64 and 32 filters, respectively, and each used a RELU activation function, had a kernel size of 3, and a 2×2 stride. The final deconvolutional layer had one filter, a kernel size of 3, a 1×1 stride, and used a sigmoid activation function.

Like the generator, the discriminator network was an AE, but with a different structure. The discriminator AE was a UNET, an end-to-end fully-convolutional AE architecture originally developed by Ronneberger et al. (2015) for biomedical image segmentation.

The encoder for this architecture contains an input layer followed by blocks of convolutional and pooling layers. The first block contained two convolutional layers with RELU activation functions, 3×3 strides, and 32 filters each, followed

by a pooling layer where each unit of the layer downsampled its input by taking the maximum value of its 2×2 window over the input. This block was followed by a block containing convolutional layers with 64 filters each but that was otherwise identical to the first block. The two blocks were followed by a final convolutional layer with 128 filters.

Like the generator, the discriminator UNET's decoding network was symmetrical to its encoding network. The output of the final convolutional layer of the encoder was the input to another 128-filter convolutional layer. The output of that layer was fed into the first of two blocks of upsampling and convolutional layers. This block began with an upsampling layer that, for each value of the input, output a 2×2 tensor of that value. The upsampling layer was followed by two convolutional layers with 64 filters each. The convolutional layers of the second block had 32 filters each, but the second block was otherwise identical to the first block. The final layer of the decoder was a convolutional layer with one filter, a kernel size of 3, a 1×1 stride, and used a sigmoid activation function.

The Adam optimizer (`tf.keras.optimizers.Adam`) was used to minimize the pixel-wise error for the generator network. This optimization algorithm is a method of stochastic gradient descent that adapts the learning rate for each parameter of the model based on the first- and second-order moments (i.e., mean and variance) of the gradients. The Adam optimizer for the generator network was initialized with a learning rate of 0.001 and an exponential decay rate of 0.5 for the first moment estimates, and all other parameters for the optimizer were set to their defaults.

The optimizer used for the discriminator network was a Root Mean Square Propagation (RMSProp) optimizer (`tf.keras.optimizers.RMSprop`), a stochastic gradient descent algorithm that, like Adam, maintains adaptive learning rates for each model parameter, but adapts learning rates based only on the mean of recent gradient magnitudes. It was initialized with a learning rate of 0.001 as well, and all other optimizer parameters were set to their defaults.

As shown in the schematic, the overall logic of the GAIA architecture is as follows: a standard autoencoder is trained to reproduce its inputs after passing

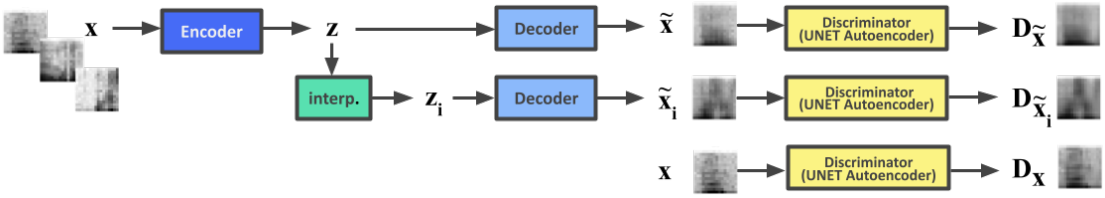

Figure 3.2: Schematic representation of the GAIA architecture used to reduce the dimensionality of the spectrographic features used in the LME models.

them through a low-dimensional z layer. The 128-dimensional space of the z layer is sampled linearly, sampling two points in the latent space as well as evenly spaced samples—interpolations—between those two points. Both the interpolations drawn from z and the low-dimensional representations of the original inputs are passed through the decoder network, yielding sets of 16×16 spectrograms—$\tilde{x}_i$ and $\tilde{x}$, respectively. These reconstructed spectrograms ($\tilde{x}_i$ and $\tilde{x}$) are then passed through the UNET autoencoder alongside original spectrograms (x), and the network is trained to discriminate between the three types of UNET reconstructions—reconstructions of original inputs ($D(x)$), reconstructions of the first AE's reconstructions ($D(\tilde{x})$), and reconstructions of the interpolations drawn from z that do not correspond to any of the original inputs ($D(\tilde{x}_i)$).

Prior to training, each spectrogram was normalized such that its minimum value was zero, and its maximum value was one. In total, across all excerpts, the dataset contained 68,271 spectrograms, which were randomly split into training and testing sets that contained 85% and 15% of the total dataset, respectively. The GAIA model was trained on batches of 4096 spectrograms, with 14 batches per training epoch, and early stopping was implemented such that training halted when the smoothed generator loss of an epoch[2] exceeded that of the tenth-most recent epoch.

The model trained for a total of 83 epochs. After training, all spectrograms were run through the encoder portion of the generator network, and each spectrogram's

[2]Lists of each epoch's loss values were saved during training to aid in determining when to stop training. To prevent local perturbations in the generator loss from causing training to stop inappropriately early, these lists were smoothed using a Hann function with a window length of 30.

128-dimensional latent space representation was saved for use in the mixed effects model. In all models, this feature set is referred to as the **s1** feature set because it contains spectrographic features.

3.2.3 Phonemic Label Features

During preprocessing, neural data were temporally registered to stimuli metadata. For each phone the participant heard, the start time, stop time, and duration of the phone were registered, alongside the phone's identity, the identity of the word it was part of, the word's part of speech, and the excerpt it was part of. From this information, a timeseries of one-hot coded features was generated for the set of unique phone labels. In other words, each timepoint of the neural response was assigned a vector of features, where each feature represented a phone label; the value of the feature corresponding to the current phone label was set to 1, and all other feature values were set to 0. Following this feature assignment, the matrix of phone label features was temporally averaged in the same way that neural data were temporally averaged to attain a one-to-one mapping with the spectrographic features. Because each vector of phone label features was sparse and binary, the resultant values in each vector summed to 1. In this way, the values in each feature vector corresponded to the proportion of time within that window where a particular phone was played for the participant. In all models, this feature set is referred to as the **p1** feature set because it contains phone label features.

3.2.4 Models

Families of LME model were fit for seven different neural response types, each derived from the preprocessed raw recorded data. Six of these response types were the z-scored power for different EEG frequency bands, and the seventh was broadband LFP. The frequency bands used in these models were delta (1–4Hz), theta (4–7Hz), alpha (8–12Hz), beta (13–30Hz), gamma (31–50Hz), and high gamma (70–150Hz) bands; broadband LFP contained frequencies 0.1–170Hz.

For each of these response types, seven LME models were fit—three of interest and four shuffle conditions. The three models of interest contained either only spectrographic features (**s1**), only phonemic label features (**p1**), or both spectrographic and phonemic features (**s1p1**). For both the **s1** and **p1** feature sets, two forms of shuffled feature sets were also created. Feature sets **s2** and **p2** were created by shuffling **s1** and **p1** features within excerpts, and feature sets **s3** and **p3** were created by shuffling **s1** and **p1** features across the entire recording session.

All **s**-series models contained 128 features, each corresponding to a dimension of the latent space constructed from the GAIA network trained on excerpt spectrograms as described in Section 3.2.2. However, based on the number of excerpts that the participant listened to, the number of phonemic label features in each model differed slightly. For participants who completed at least four blocks of the passive listening task, there were 72 unique labels included in the **p**-series models. Models for participants who completed fewer blocks had at least 60 phonemic features. Labels that did not occur in the first two task blocks were uncommon phones and extraneous labels such as oɪ, nasalized vowels like ãũ and ẽɪ, and laughter.

3.2.5 Model evaluation

Relative goodness of model fit was evaluated using the Akaike Information Criterion (AIC). The AIC is an estimate of prediction error, calculated for each model as

$$AIC = 2k - 2ln(\hat{L})$$

Where k is the number of estimated parameters in the model and $\hat{L}$ is the maximum value of the model's likelihood function.

The way in which the AIC functions is clear from this definition. If two models have the same number of parameters, and one has greater likelihood than the other, its AIC will be smaller. For this reason, higher quality models have smaller AIC values. However, the term that takes the model likelihood into account is offset by a term that includes the number of the parameters estimated in the model. The

number of model parameters drives up the value of the AIC, such that for two models with the same likelihood, the model with fewer parameters will have the lesser AIC value. In this way, the AIC balances model fit and model simplicity, deterring the selection of a model with higher likelihood that achieves its goodness of fit through overfitting.

For a set of models with AIC values $AIC_{min}, AIC_1, AIC_2, \ldots AIC_{max}$, to assess whether one model has meaningfully better fit quality than another on the basis of their AIC values, we calculate the relative likelihood of each model, also known as its Akaike weight (AICw). The AICw of a model i is calculated as $exp((AIC_{min} - AIC_i)/2)$, where AIC_{min} is the lowest AIC value of the set of models and AIC_i is the AIC value for model i. Thus, if the AIC value for a model is only 1 unit greater than the minimum AIC value of the set of models, that model is $exp(-1/2) = 0.61$ times as likely to minimize information loss when predicting data. If the AIC value for a model is 10 units greater than the minimum AIC value of the set of models, that model is $exp(-10/2) = 0.007$ times as likely to minimize information loss when predicting data.

In practice, if a model has an AIC value less than two units greater than than the minimum AIC value, there is considered to be substantial support for that model. In such a case, the currently available data are insufficient to select between that model and the top ranked model. Either more data would need to be gathered to attempt to distinguish the two models, or a multimodel created as the weighted average of the two models may be used for subsequent statistical inference. If the model's AIC value is $3 \sim 7$ units greater than the minimum, it is considered to have less support, and if the model has an AIC value more than 10 units greater than the minimum, it is considered unlikely to be the model that minimizes the prediction loss. In either of these two cases, it is typically reasonable to conclude that the model with the lowest AIC value is in fact the model that minimizes prediction error (Burnham & Anderson 2003).

The AICw is similar to the perhaps more familiar likelihood-ratio test (LRT), but more appropriate for the sets of models constructed in this chapter. Like the AICw, the LRT can be expressed as a difference between log likelihoods:

$$\lambda_{LR} = -2(ln(\hat{L}_0) - ln(\hat{L}_i))$$

where $\hat{L}$ is once again the maximum value of the model's likelihood function. In this case, however, the two models must be in a nested relationship with one another; the i-th model must be transformable into the 0-th model through restrictions on the values of its parameters. This requirement ensures that λ_{LR} is χ^2-distributed when the null hypothesis cannot be rejected and the i-th model cannot be said to differ substantially from the 0-th model. The AICw does not place this restriction on the structure of the models it evaluates, and thus is more appropriate for the models presented here.

3.3 Results

3.3.1 Model Comparison

Results from the coronal tap, plural, and past tense comparisons provide evidence for distinct contributions of categorical phonemic knowledge and acoustic processing to the neural response to speech. Here we propose and assess a basic model of the relationship between these two feature types using linear mixed effects models. For each participant, for each of the seven neural response types, seven LME models were fit. Each model fit neural response with electrode channel and excerpt speaker as random effects and either only spectrographic features, only phonemic label features, or both spectrographic and phonemic label features as fixed effects. Figure 3.1 illustrates how these three base models where constructed. Four additional models were also created, in which the phonemic or spectrographic features had either been shuffled within each excerpt or shuffled across the full recording session. Models were then compared within participant and neural response type using the Akaike Information Criterion (AIC; Akaike 1998), and all best-fit models carried 100% of the cumulative model weight and had an AIC score >200 lower than other models.

Power in δ-, θ-, α-, and β-bands was best fit by linear mixed effects models

that included both spectrographic features and phonemic labels (**s1p1**). For these bands, the **s1p1** model was the best fit for all participants. These results suggest that power in these bands is driven in part by phonemic category information that is not reducible to speech acoustics. For power in γ- and high γ-bands, however, the best fit model varied across individuals. For γ-power, eight participants' data were best fit by the model that included only spectrographic features (**s1**), and two participants' data were best fit by the **s1p1** model that included both spectrographic and phonemic label features (SD013, SD018). Similarly, for high γ-power, eight participants' data were best fit by the **s1** model that included only spectrographic features, and the remaining participants data were best fit by the **s1p1** model that included both spectrographic and phonemic label feature sets (SD011, SD013). These results suggest that power in frequencies above 30Hz are primarily driven by speech acoustics rather than phonemic category information.

For broadband LFP, all participants' data were best fit by the **s1p1** model that included both spectrographic features and phonemic labels. Given the $1/f$ structure of LFP data, this neural measure is dominated by lower frequencies; thus, the fact that the best fit model for lower frequency bands is also the best fit model for the broadband LFP is somewhat expected. What is notable, however, about the match between the best fit models for lower frequency band power and LFP is that LFP and band power are different kinds of neural measures. LFP is a measure of extracellular voltage fluctuation over time, and band power measures are derived from LFP through a series of non-linear transformations: LFP activity is filtered into logarithmically increasing frequency bands and the analytic amplitude of those bands is computed and then averaged across bands. On their own, the LFP results demonstrate that broadband LFP contains phonemic category information that is not reducible to speech acoustics. Combined with the results from lower frequency band power, the LFP results further suggest that this information is robust to the set of transformations involved in deriving band power from LFP.

As a model selection tool, the AIC straightforwardly aids the selection of the most probable model from a set of models. More generally, however, the AIC facilitates model ranking, and given the partially-nested structure of the models

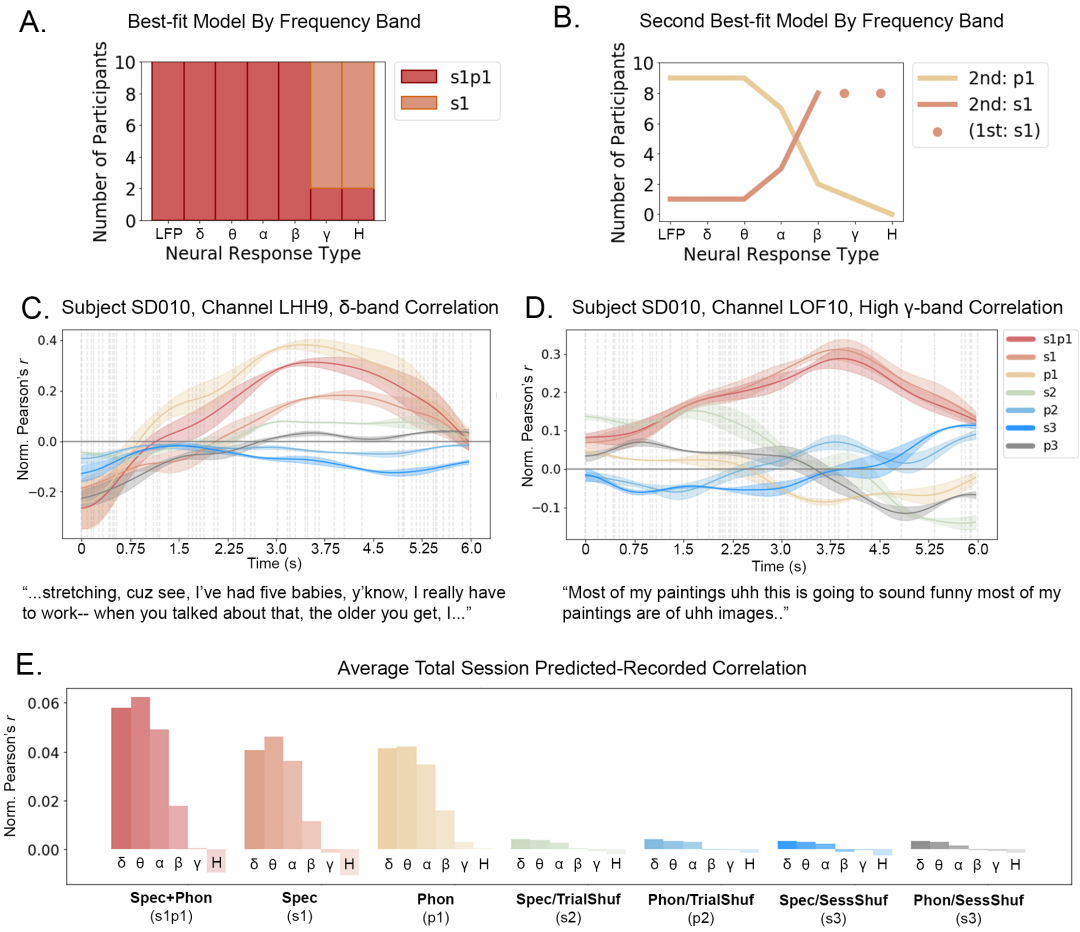

Figure 3.3: Phonemic information dominates lower frequency band responses to speech. (A) For LFP and δ–β-bands, models that include phonemic label information in addition to spectrographic information (**s1p1**, red) best fit the neural response to speech for all participants. At higher frequencies, most participants' data are best fit by models containing solely spectrographic features (**s1**, orange). (B) Models containing phonemic label information alone (**p1**, yellow) are more likely to provide a better fit for neural response at lower frequencies than at higher frequencies, where models containing spectrographic information alone (**s1**, orange) are more likely to provide a better fit. (C) Time-varying correlation between predicted and recorded neural response for the δ-power response of one electrode, illustrating a period of sustained advantage of the **p1** model (yellow) over other models, including the **s1p1** (red) model. (D) Time-varying correlation between predicted and recorded neural response for the high γ-power response of one electrode, illustrating a period of sustained advantage of the **s1** model (orange) and **s1p1** (red) models over models that do not contain unshuffled spectrographic information. For (C) and (D), grey vertical lines indicate phoneme boundaries for the transcript provided below each plot. (E) Predicted-Recorded correlations averaged across all speech selective electrodes for all participants. In the aggregate, the **s1p1** model (red) outperforms **s1** (orange) and **p1** (yellow) models, and all three contentful models (**s1p1**, **s1**, **p1**) outperform all four shuffled models.

under comparison in this study, the relative ranking of the **s1** and **p1** models is worthy of assessment. Especially where **s1p1** was the best-fit model, assessing the relative fit of **s1** and **p1** models provides insight concerning which of the two component feature sets of the **s1p1** model contributed the most to its goodness-of-fit. For each subject and response type, the model with the second lowest AIC score was >100 times as likely as the third ranked model. As shown in Figure 3.3B, models containing only phonemic label features were most likely to be the second ranked model for lower frequency bands. However, with increasing band frequency, the **s1** model becomes more likely to provide a better fit for the data until, for γ- and high γ-bands, it is the best fit model overall. These results appear to provide support for a larger-scale generalization that phonemic labels better explain power at lower frequencies, while acoustic features excel at explaining power in higher frequencies.

3.3.2 Significant parameters

Significant parameters for each of the models were considered to be those for which a Bonferroni-corrected confidence interval for the parameter did not include zero. Interval sizes were adjusted from the standard threshold for significance ($\alpha = 0.05$), using a Bonferroni correction to account for the number of fixed effects in each of the models. That is, for **p**-series models, a $100 \cdot (1 - 0.05/72) = 99.93\%$ confidence interval was calculated; for **s**-series models a $100 \cdot (1 - 0.05/128) = 99.96\%$ confidence interval was calculated; and for the **s1p1** model a $100 \cdot (1 - 0.05/200) = 99.98\%$ confidence interval was calculated. For subjects and response types for which the **s1p1** model best fit the data, the parameters of the model that were found to be significant overwhelmingly overlapped with the parameters that were found to be significant in the corresponding **s1** and **p1** models. These observations are corroborated by the distribution of significant parameters in each of the models.

3.4 Discussion

3.4.1 Validation of Di Liberto et al. (2015)

In fitting mixed effects models containing acoustic features and phonemic label features across the six major functional bands, this study validates results found in scalp EEG work by Di Liberto et al. (2015). Di Liberto et al. (2015) fit multivariate temporal response functions (mTRFs) to bandpassed scalp EEG data using spectrographic features, phonemic label features, and both spectrographic and phonemic label features. As in this study, they found that models containing both spectrographic and phonemic label features provided the best fit model for δ and θ bands. Furthermore, in these lower frequency bands, their phonemic label model outperformed their spectrographic model. Conversely, in higher bands (α, β, γ) the model containing only spectrographic features outperformed the model containing only phonemic label features.

This chapter improves on the statistical methods used in Di Liberto et al. (2015) to make stronger claims about the relative fit of different models. Di Liberto et al. (2015) assessed model performance using Pearson correlation between their recorded and predicted responses only. For this reason, they were incapable of directly assessing whether the enhanced performance of their model that contained both spectrographic and phonemic label features was driven by the fact that it had more free parameters than all other models (Di Liberto et al., 2015, pg. 2459). Through use of the AIC in this study, differing numbers of model parameters were taken into account during model selection, thereby facilitating stronger claims concerning model quality. In this way, the findings reported in this chapter both validate and bolster Di Liberto et al.'s (2015) claims.

In addition to validating Di Liberto et al.'s (2015) main results, this study also replicated one of the finer details of their work. In particular, Di Liberto et al. (2015) found that the overall correlation between predicted and recorded responses was strongest in the lowest frequency bands and dropped to a fraction of that magnitude as center band frequency increased. This effect can be seen clearly in Figure 3.3E and to a lesser extent for the MNE models in Chapter 4, Figure

4.2C. Di Liberto et al. (2015) attributed this steep dropoff in the predictive ability of their models to the lower signal-to-noise ratio of these bands when recorded from the scalp. Observing this effect in intracranial recordings is not necessarily inconsistent with Di Liberto et al.'s (2015) interpretation, but the sources of noise may be different. For example, it may be the case that higher frequency signals have lower fidelity in white matter sites due to the distance of these sites from signal generators.

Another, unrelated possibility concerns the temporal resolution of the phonemic and spectrographic features relative to the intrinsic temporal resolution of the neural response features. As mentioned in §3.1.1, oscillatory models of language processing often propose roles for particular neural bands by appealing to shared temporal granularity between linguistic and neural rhythms. By such accounts, activity in higher frequency bands is expected to encode the rapidly fluctuating spectrotemporal dynamics of speech acoustics, while lower frequency bands are expected to encode more slowly varying linguistic units, such as syllables. While this study did find that spectrotemporal features better accounted for the activity in higher frequency bands than phonemic label features, all models performed worse at predicting high frequency band power than low frequency band power.

The steep decline in predictive power as band frequency increases is unexpected for the models containing spectrographic features. Beyond the possibly lower signal-to-noise ratio of high frequency activity when recorded from white matter, the unexpectedly poor predictive power of the high frequency models may stem from any of several temporal mismatches between the stimulus and neural response. For example, the models in this this study were fit using only one temporal alignment of stimulus and response features. It is possible that different neural frequency bands respond to stimulus features at different latencies. It is also possible that the temporal resolution of the spectrographic features (15.625Hz) and resultant level of spectral detail was too low to adequately capture variance in the neural response at higher frequencies. Relatedly but distinctly, rapid fluctuations in band power may play a larger role in stimulus encoding at higher frequencies, and as a result, downsampling the neural response to match the sample rate

of stimulus features may have abolished more stimulus-dependent information at those frequencies.

3.4.2 Validation of ERP Case Studies

The results of this study broadly corroborate the findings reported in Chapter 2 for the coronal tap, regular past tense, and regular plural comparisons. For each of those comparisons, for all neural response bands, a significant number of sites exhibited sensitivity to categorical phonemic identity over acoustic (tap comparison) or morphological (plural and past tense comparisons) identity. In this study, models containing categorical phonemic label features outperformed other models for all participants in δ- through β-bands, and for two participants in γ- and high γ-bands. These results generally indicate that phonemic category information does not merely recapitulate spectrographic information but captures some portion of neural variance not otherwise available in the speech acoustics. This category-related activity is consistent with the observation of sites sensitive to phonemic identity in Chapter 2; however, the variability in best-fit model across participants for γ- and high γ-bands raises questions for this interpretation.

3.4.3 Cross-participant variability in LME model fits

The variability in the best-fit LME model across participants for γ-power and high γ-power data is not easily explainable. It may result from differential electrode coverage across participants. The fact that both SD013's γ-power and high γ-power were best fit by the **s1p1** model would be consistent with such an explanation, since electrode coverage is a factor that is fixed for each participant, and we might therefore expect coverage-based differences to be consistent within participants. However, the fact that only SD018's γ-power and only SD011's high γ-power are best fit by the **s1p1** model reveals the true complexity of a coverage-based explanation for the variability in these results. For the best fit model for SD018's γ-power to diverge from that of all other participants for only that power band would require coverage that differs from other participants' in such a way

that it differentially impacts γ-power only. Similarly, the fact that the best fit models for SD013 differ from the majority in only two bands requires one to posit a state of affairs in which SD013's coverage differs in a way that substantially impacts only two bands.

Individual variation presents another possible explanation. Whereas a coverage-based explanation assumes that results across participants ought to be the same and only differ because non-homologous sets of neurons were recorded across participants, an explanation based on individual variation would allow for the possibility that even with functionally identical coverage, individuals may still vary in the way that they process speech sounds. From this perspective, the consistency across participants in lower bands is merely artifactual of the small number of participants whose data are reported here, and perhaps if data from a greater numbers of individuals were available, some variability in the best-fit model would be observed in each band. However, while individual variability most certainly exists in speech sound processing, it seems unlikely that information about speech sound categories and their acoustics, being both quite close to the sensory periphery and fundamental to speech perception, would vary substantially across individuals. This is an empirical question, though, and while its answer ought to be pursued, it is not a question that this dataset is well-suited to address. Given the limitations of the dataset, arriving at a satisfactory account for the cross-participant variability of these results will require follow-up work.

3.4.4 Implications of feature compression

It is worth reiterating that the spectrographic features used in these models were drawn from the latent space of a variational autoencoder. In this sense, like the phonemic label features, the spectrographic features used in this study were also abstracted from surface acoustics. However, whereas the phonemic label features represent a linguistically-informed compression of the acoustic signal, the compressed spectrographic features are purely informed by acoustic information. It is remarkable then, with all their compression, that weightings of these feature sets are still capable of reconstructing neural response with such high fidelity,

with correlation of over 0.4 at times. In this way, the ability of these features to account for a significant portion of neural variance validates the use of such nonlinear compression methods for neural encoding analyses.

Material from Chapter 3 has been submitted for publication as "The neural response to speech is language specific and irreducible to speech acoustics" with authors Anna Mai, Stephanie Riès, Sharona Ben-Haim, Jerry Shih, & Timothy Gentner. The dissertation author was the primary investigator and author of this paper.

Bibliography

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected papers of Hirotugu Akaike*, pages 199–213. Springer.

Assaneo, M. F. and Poeppel, D. (2018). The coupling between auditory and motor cortices is rate-restricted: Evidence for an intrinsic speech-motor rhythm. *Science advances*, 4(2):eaao3842.

Bastos, A. M. and Schoffelen, J.-M. (2016). A tutorial review of functional connectivity analysis methods and their interpretational pitfalls. *Frontiers in systems neuroscience*, 9:175.

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., and Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron*, 76(4):695–711.

Brennan, J. R. and Martin, A. E. (2020). Phase synchronization varies systematically with linguistic structure composition. *Philosophical Transactions of the Royal Society B*, 375(1791):20190305.

Buzsáki, G. (2006). *Rhythms of the Brain*. Oxford University press.

Buzsáki, G., Anastassiou, C. A., and Koch, C. (2012). The origin of extracellular fields and currents—EEG, ECoG, LFP and spikes. *Nature reviews neuroscience*, 13(6):407–420.

Chai, L. R., Mattar, M. G., Blank, I. A., Fedorenko, E., and Bassett, D. S. (2016). Functional network dynamics of the language system. *Cerebral Cortex*, 26(11):4148–4159.

Cibelli, E. S., Leonard, M. K., Johnson, K., and Chang, E. F. (2015). The influence of lexical statistics on temporal lobe cortical dynamics during spoken word listening. *Brain and language*, 147:66–75.

Cohen, M. X. (2017). Where does EEG come from and what does it mean? *Trends in neurosciences*, 40(4):208–218.

Di Liberto, G. M., O' Sullivan, J. A., and Lalor, E. C. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current Biology*, 25(19):2457–2465.

Ding, N., Melloni, L., Zhang, H., Tian, X., and Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature neuroscience*, 19(1):158–164.

Fontolan, L., Morillon, B., Liegeois-Chauvel, C., and Giraud, A.-L. (2014). The contribution of frequency-specific activity to hierarchical information processing in the human auditory cortex. *Nature communications*, 5(1):1–10.

Ghitza, O. (2011). Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input rhythm. *Frontiers in psychology*, 2:130.

Giraud, A.-L. and Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. *Nature neuroscience*, 15(4):511–517.

Golumbic, E. M. Z., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., Goodman, R. R., Emerson, R., Mehta, A. D., Simon, J. Z., et al. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party" . *Neuron*, 77(5):980–991.

Hannemann, R., Obleser, J., and Eulitz, C. (2007). Top-down knowledge supports the retrieval of lexical information from degraded speech. *Brain research*, 1153:134–143.

Hullett, P. W., Hamilton, L. S., Mesgarani, N., Schreiner, C. E., and Chang, E. F. (2016). Human superior temporal gyrus organization of spectrotemporal modulation tuning derived from speech stimuli. *Journal of Neuroscience*, 36(6):2014–2026.

Hyafil, A., Giraud, A.-L., Fontolan, L., and Gutkin, B. (2015). Neural cross-frequency coupling: connecting architectures, mechanisms, and functions. *Trends in neurosciences*, 38(11):725–740.

Jensen, O., Bonnefond, M., and VanRullen, R. (2012). An oscillatory mechanism for prioritizing salient unattended stimuli. *Trends in cognitive sciences*, 16(4):200–206.

Jensen, O., Gips, B., Bergmann, T. O., and Bonnefond, M. (2014). Temporal coding organized by coupled alpha and gamma oscillations prioritize visual processing. *Trends in neurosciences*, 37(7):357–369.

Klimesch, W. (1999). EEG alpha and theta oscillations reflect cognitive and memory performance: A review and analysis. *Brain research reviews*, 29(2-3):169–195.

Kösem, A., Basirat, A., Azizi, L., and van Wassenhove, V. (2016). High-frequency neural activity predicts word parsing in ambiguous speech streams. *Journal of neurophysiology*, 116(6):2497–2512.

Kösem, A., Gramfort, A., and Van Wassenhove, V. (2014). Encoding of event timing in the phase of neural oscillations. *Neuroimage*, 92:274–284.

Lakatos, P., Karmos, G., Mehta, A. D., Ulbert, I., and Schroeder, C. E. (2008). Entrainment of neuronal oscillations as a mechanism of attentional selection. *science*, 320(5872):110–113.

Lewis, A. G., Wang, L., and Bastiaansen, M. (2015). Fast oscillatory dynamics during language comprehension: Unification versus maintenance and prediction? *Brain and language*, 148:51–63.

Mai, G., Minett, J. W., and Wang, W. S.-Y. (2016). Delta, theta, beta, and gamma brain oscillations index levels of auditory sentence processing. *Neuroimage*, 133:516–528.

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

Mesgarani, N., Cheung, C., Johnson, K., and Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*, 343(6174):1006–1010.

Nelson, M. J., El Karoui, I., Giber, K., Yang, X., Cohen, L., Koopman, H., Cash, S. S., Naccache, L., Hale, J. T., Pallier, C., et al. (2017). Neurophysiological dynamics of phrase-structure building during sentence processing. *Proceedings of the National Academy of Sciences*, 114(18):E3669–E3678.

Ng, B. S. W., Logothetis, N. K., and Kayser, C. (2013). EEG phase patterns reflect the selectivity of neural firing. *Cerebral Cortex*, 23(2):389–398.

Nozaradan, S., Peretz, I., Missal, M., and Mouraux, A. (2011). Tagging the neuronal entrainment to beat and meter. *Journal of Neuroscience*, 31(28):10234–10240.

Palva, S. and Palva, J. M. (2012). Discovering oscillatory interaction networks with M/EEG: Challenges and breakthroughs. *Trends in cognitive sciences*, 16(4):219–230.

Panzeri, S., Brunel, N., Logothetis, N. K., and Kayser, C. (2010). Sensory neural codes using multiplexed temporal scales. *Trends in neurosciences*, 33(3):111–120.

Park, H., Ince, R. A., Schyns, P. G., Thut, G., and Gross, J. (2015). Frontal top-down signals increase coupling of auditory low-frequency oscillations to continuous speech in human listeners. *Current Biology*, 25(12):1649–1653.

Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., Shamma, S. A., Crone, N. E., Knight, R. T., and Chang, E. F. (2012). Reconstructing speech from human auditory cortex. *PLoS biology*, 10(1):e1001251.

Peña, M. and Melloni, L. (2012). Brain oscillations during spoken sentence processing. *Journal of cognitive neuroscience*, 24(5):1149–1164.

Poeppel, D. and Assaneo, M. F. (2020). Speech rhythms and their neural foundations. *Nature reviews neuroscience*, 21(6):322–334.

Poeppel, D. and Embick, D. (2017). Defining the relation between linguistics and neuroscience. In *Twenty-first century psycholinguistics: Four cornerstones*, pages 103–118. Routledge.

Riecke, L., Esposito, F., Bonte, M., and Formisano, E. (2009). Hearing illusory sounds in noise: the timing of sensory-perceptual transformations in auditory cortex. *Neuron*, 64(4):550–561.

Riecke, L., Vanbussel, M., Hausfeld, L., Başkent, D., Formisano, E., and Esposito, F. (2012). Hearing an illusory vowel in noise: suppression of auditory cortical activity. *Journal of Neuroscience*, 32(23):8024–8034.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Shamir, M., Ghitza, O., Epstein, S., and Kopell, N. (2009). Representation of time-varying stimuli by a network exhibiting oscillations on a faster time scale. *PLoS computational biology*, 5(5):e1000370.

Siegel, M., Donner, T. H., and Engel, A. K. (2012). Spectral fingerprints of large-scale neuronal interactions. *Nature Reviews Neuroscience*, 13(2):121–134.

Stefanics, G., Hangya, B., Hernádi, I., Winkler, I., Lakatos, P., and Ulbert, I. (2010). Phase entrainment of human delta oscillations can mediate the effects of expectation on reaction speed. *Journal of Neuroscience*, 30(41):13578–13585.

Strauß, A., Kotz, S. A., Scharinger, M., and Obleser, J. (2014). Alpha and theta brain oscillations index dissociable processes in spoken word recognition. *Neuroimage*, 97:387–395.

Sunami, K., Ishii, A., Takano, S., Yamamoto, H., Sakashita, T., Tanaka, M., Watanabe, Y., and Yamane, H. (2013). Neural mechanisms of phonemic restoration for speech comprehension revealed by magnetoencephalography. *brain research*, 1537:164–173.

Ten Oever, S. and Sack, A. T. (2015). Oscillatory phase shapes syllable perception. *Proceedings of the National Academy of Sciences*, 112(52):15833–15837.

Wang, L., Zhu, Z., and Bastiaansen, M. (2012). Integration or predictability? a further specification of the functional role of gamma oscillations in language comprehension. *Frontiers in psychology*, 3:187.

Womelsdorf, T., Schoffelen, J.-M., Oostenveld, R., Singer, W., Desimone, R., Engel, A. K., and Fries, P. (2007). Modulation of neuronal interactions through neuronal synchronization. *science*, 316(5831):1609–1612.

Chapter 4

A Relational Model of Invariance

4.1 Introduction

Speech is a highly-structured auditory signal that carries necessary and sufficient information for language comprehension among neurotypical, hearing populations. Speech perception, accordingly, involves the integration of both bottom-up auditory processing and top-down language knowledge. Top-down language knowledge itself comes in two broad forms: knowledge that may hold of languages generally (=*general language knowledge*) and knowledge that is specific to particular languages (=*specific language knowledge*). General language knowledge may include awareness about general language structure (e.g., words may modify other words) or about how language is produced (e.g., different speakers have different voices, taller signers have larger signing spaces, etc.), while specific language knowledge requires extensive experience with a particular language (e.g., adjectives follow the nouns they modify in Hebrew, word-final consonants are devoiced in German, etc.). This study focuses on the role that specific language knowledge plays in speech perception, asking at what point linguistic units of speech are processed in a language-specific manner and what general characteristics of speech acoustics support the perception of those units.

4.1.1 The problem of invariance

Understanding how an acoustic signal supports what is ultimately perceived as language is central to speech perception research, and the idea that some form of acoustic invariance plays a role in the perceptual experience of invariance in speech has long challenged researchers (Liberman et al. 1952; Fant 1960; Cole and Scott 1974b,a; Klatt 1979; Blumstein and Stevens 1979, inter alia). On the one hand, the acoustic characteristics of speech are highly constrained, both by articulatory constraints on what sounds can be produced and by auditory constraints on what sounds can be perceived. And arguably more interestingly, within the confines of those physical constraints, there is great structure: certain configurations of articulators produce stable acoustic characteristics, such as the stability of the correlation between rising formant transitions and the release of a labial constriction or the attenuation of energy between the second and third formants during the production of a lateral sound. These kinds of articulo-acoustic dependencies are undoubtedly foundational to our ability to form linguistic sound categories, relating what we hear to what we learn to produce. On the other hand, the realization of a sound in a particular category may vary substantially from speaker to speaker, such as in the case of one speaker's high tone having the same fundamental frequency as another speaker's low tone (see Mandarin, Lee 2009), or even in a single speaker, certain articulo-acoustic cues may overlap greatly with the cues to other categories, such as a speaker who produces a distribution of high back rounded vowels that overlaps almost entirely with their production of mid back rounded vowels (see Kinande, Gick et al. 2006).

For these reasons, all current theories that set out to link speech acoustics to perception and comprehension include some role for surrounding acoustic context, including speaker, speech rate, and phonological context (see Holt and Lotto 2010 or Casserly and Pisoni 2010 for a review). And each of these features may also differ in how they are deployed in different languages. Distinct languages with superficially similar sound categories may make use of different acoustic cues to aid in the recognition of those sounds. For example, listeners appear to rely more on voice quality than f0 to identify falling tones in White Hmong (Garellek et al.,

2013), while in Cham, f0 is the primary cue used for tone identification despite the availability of voice quality cues (Brunelle, 2012). Moreover, within a single language, the primary acoustic cues to certain contrasts may change over time, as in the case of Korean tonogenesis (Silva, 2006), or even disappear, as may have been the case Tutrugbu's vowel inventory (Essegbey and McCollum, 2018). Thus, the problem remains of how an organ with the same general capacities across all these populations manages to extract and replicate the patterns of segments that are consistent with the grammar of a particular language. That is, what general characteristics of extant surface acoustic patterns support the extraction of linguistically meaningful categorical units?

This study offers a broad starting point towards characterizing how the brain responds to acoustic stimuli in a way that supports sound categorization. Through examination of how first- and second-order statistical moments of speech acoustics contribute to observed changes in band power across different frequencies, this study begins to address the relationship between the acoustic processing of spectrotemporally complex features and the linguistically relevant categories those features belong to. In this way, instead of focusing on literal invariances of the acoustic signal, this approach focuses on which aspects of the acoustic signal drive invariances in the neural response, placing the ultimate onus for linguistic structure on the physiological characteristics of the brain and peripheral sensory organs rather than the linguistic signal itself.

Receptive field estimation approaches use mathematically explicit hypotheses about the nature of the relationship between stimulus and response to determine what features of the stimulus drive a response. Thus, they are immanently well-suited to answer the questions posed above, since the effects of hypotheses used by different approaches can be compared straightforwardly through their explicit definitions. In the following section, we review different approaches to receptive field estimation and provide background for the most promising approaches for LFP speech data. Then, we describe how these models were fit, characterize the resultant receptive fields, and assess the contribution of model type, acoustic structure, and the availability of category-level phonemic information to the goodness-of-fit.

4.1.2 Receptive field estimation

The process of transforming a sensory signal into a neurobiological signal that can be used to inform an organism's behavior involves numerous, non-linear transformations. In the case of sound, when a mechanical wave reaches the cochlea, it is transduced into electrical energy through its interaction with stereocilia, the polarized epithelial cells that line the spiral of the cochlea. This electrical signal then passes via the auditory nerve through several brainstem nuclei before it reaches the neocortex. At each step, a cell's response to the signal that it receives is particular to the features of that signal, and different cells respond in different ways to the inputs that they receive. For example, one of the most salient organizing principles of the early auditory system is its tonotopy, a property that in part emerges from the fact that cells in this pathway respond selectively to certain frequencies. Information at these stages may inform reflexes, such as the acoustic startle response, but quintessentially human behaviors such as those involving music and language require further transformation of sensory signals in neocortex.

As in earlier pre-cortical processing of sound, one of the principle organizing features of early cortical processing is tonotopy; however, once in cortex, temporal features of the sensory source also begin to play a larger organizational role. With each successive synapse ascending through the auditory system, the features of the sensory source that are represented at that point grow in complexity, such that in higher auditory areas in humans, there is evidence for distinct populations that respond preferentially to instrumental music or speech (Norman-Haignere et al., 2015), and that despite variability in the acoustic source, behaviorally relevant categories of sounds elicit similar responses (Chang et al., 2010; Toscano et al., 2010). In this way, higher auditory responses are directly tied to perception and subsequent behavior, and thus characterizing how the nervous system transforms sensory inputs into these neurobiological signals is essential to understanding the neural basis of the human behaviors that rely on these sensory signals, such as language.

One way of characterizing the nervous system's transformation of sensory signals involves determining how higher-level sensory neurons respond relative to the

sensory source. From this approach, the sets of acoustic features that elicit a response from a particular cell constitute that cell's *receptive field*, and that receptive field in turn provides some indication of the transformations of the sensory signal that the nervous system must perform. However, determining the receptive fields of higher-level sensory neurons remains difficult for two primary reasons. First, the response functions of higher-level sensory neurons rely on multidimensional transformations of inputs which themselves are not well-characterized. Second and inter-relatedly, higher-level sensory neurons are often highly selective in their responses. They do not, in general, respond to broadband noise stimuli, tone sweeps, or other spectrotemporally simple inputs that would make the project of determining the components of their response functions simpler. Rather, structured, often spectrotemporally complex, ecologically valid stimuli are required to elicit a response. However, even when presented with such stimuli, it is difficult to predict when a particular neuron will respond, because the specific combination of features that drives its response function, in general, do not occur frequently.

The difficulties inherent to the project of calculating complex receptive fields spawn related computational challenges. Given the high dimensional natures both of a cell's response function and of the sensory signal, standard statistical models often fail to recover plausible receptive fields. Relative to the number of model parameters, available data often fail to sample sufficiently over the stimulus space and response distribution, resulting either in woefully over- or under-fit models. To confront these challenges successfully, thus requires computational methods that are capable of estimating multidimensional transformations of input data; robust to the biases of highly correlated natural stimuli; and resistant to overfitting.

A number of methods designed to meet these demands are currently used to estimate complex receptive fields, and although these methods differ in a number of respects, they are unified in their approach to receptive field estimation as a kind of dimensionality reduction problem. Since higher-order sensory neurons respond sparsely to features of the sensory input, these methods assume that the set of components driving a cell's response is smaller than the set of all components that would be necessary to describe the entire stimulus space. From this perspective,

the act of estimating a cell's receptive field amounts to finding a lower-dimensional space that captures its behavior, and correspondingly, the computational methods used to estimate receptive fields in this manner may be familiar from other computational contexts in which some form of dimensionality reduction is desired.

In the context of receptive field estimation, these methods can be roughly divided into those that correlate the neural response directly to components of the stimulus, s_i, versus those that additionally correlate the neural response to pairwise products $s_i s_j$ of stimulus components as well. These two broad families of methods are known as linear and quadratic methods, respectively.

Perhaps the simplest linear method, the spike-triggered average (STA), calculates an average for each dimension of the stimulus, weighted by the corresponding observed response distribution. The formula for the STA, given in (4.1.2), results in a receptive field that represents the average stimulus that co-occurs with an increase in a cell's neural activity.

(3) The Spike-triggered Average

$$\omega^{STA} = \frac{1}{N_{spk}} \sum_{t=1}^{N_{samp}} y_t s_t$$

In this formulation, for time index t, s_t is a vector of features that represents the stimulus at t, y_t represents the neural response at t, N_{samp} represents the total number of samples, and N_{spk} represents the total number of spikes.

The corresponding quadratic method is the spike-triggered covariance (STC). The primary advantage of the STC over the STA is that the STC can estimate more than one component of the receptive field. This is particularly useful for neurons in higher-order sensory areas, many of which have been shown to contain cells that are selective for more than one distinct feature of the sensory stimulus. The reconstruction of these so-called *multidimensional* or *multicomponent receptive fields* benefits from models that include quadratic features, such as STC models because they are capable of estimating multiple receptive field components. The STC extends the STA by calculating not only the the first-order statistical moment—the average value of the stimulus vector—that correlates with an increase in neural

response (spiking activity), but also the the second-order moment, the average value of the stimulus covariance. The STC is calculated as the difference between the spike-triggered stimulus covariance matrix,

(4)
$$C^{spk} = \frac{1}{N_{spk}} \sum_{t=1}^{N_{samp}} y_t s_t s_t^T$$

and the stimulus covariance matrix,

(5)
$$C^{stim} = \frac{1}{N_{samp}} \sum_{t=1}^{N_{samp}} y_t s_t s_t^T .$$

This results in the difference matrix $C^{diff} = C^{spk} - C^{stim}$ from which a set of estimated features are extracted through eigendecomposition. Decomposing C^{diff} into $\Omega\Lambda\Omega^T$ through eigendecomposition produces eigenvectors, Ω and their corresponding weights, the eigenvalues Λ, and in this way, the STC captures multiple components of the stimulus that drive neural response.

However, spike-triggered methods like the STA and STC are limited by their strict assumptions about the stimulus distribution. The STA requires the stimulus distribution to be zero-centered and spherically symmetric, and the STC requires stimuli be drawn from an uncorrelated, zero-centered Gaussian distribution. These restrictions render spike-triggered methods generally unsuitable for modeling the receptive fields of higher-order sensory neurons because, as previously mentioned, such cells tend not to respond to noise stimuli of the kind that would be necessary to satisfy the distributional assumptions of these methods.

For this reason, instead of pursuing spike-triggered methods for receptive field reconstruction in this work, an information-theoretic method is pursued instead. The earliest such approach of this kind, is a linear method known as the Maximally Informative Dimensions (MID) approach, which estimates components of a unit's receptive field by searching for the set of vectors that maximizes the mutual information for each spike, based on empirical estimates of the posterior distribution (= the probability of a spike given a certain stimulus vector) and the prior distribution (= the probability of a given stimulus vector itself). In relying on distributions

derived from the available data, information-theoretic approaches like MID circumnavigate the distributional assumptions of spike-triggered methods that are unrealistic for higher-order sensory receptive fields. However, as a linear method that does not assume a stimulus distribution, MID falls prey to the curse of dimensionality, since the proportion of the lower dimensional (receptive field) space that would be filled by stimuli projected into that space decreases exponentially as the number of components being fit by the model grows. As a result, although linear MID can reconstruct complex receptive fields from non-Gaussian stimuli, it is typically only successful at reconstructing up to three or four components (Rowekamp and Sharpee, 2011). Since cells far earlier in sensory pathways, such as the retinal ganglion cells (RGCs) of salamanders, have been shown to respond to as many as eight independent stimulus features (Fairhall et al., 2006; Kaardal, 2017), and higher-order sensory neurons are generally assumed to have receptive field properties of equal if not greater complexity than those closer to the sensory periphery, dimensionality reduction techniques that are capable of stably reconstructing but scant handfuls of receptive field components are likely inadequate for the reconstruction of higher-order receptive fields.

As a quadratic information-theoretic method, maximum noise entropy (MNE) is an approach to receptive field reconstruction that both averts the curse of dimensionality as well as the assumption that the stimulus space is Gaussian distributed. Like other information-theoretic methods, the stimulus distribution is estimated empirically and therefore need not be assumed *a priori*. However, unlike the linear MID approach, numerous (> 4) components can be stably recovered from MNE models because MNE models constrain the properties of the low dimensional (receptive field) space using a data-derived minimal model of the space. Furthermore, unlike MID models, MNE models minimize, rather than maximize, the mutual information between the response and the stimulus by maximizing the noise entropy of the stimulus. This is accomplished by rewriting the mutual information between the stimulus and the response ($I(y;\mathbf{s})$ as the difference between the response entropy,

$$H_{resp} = -dyP(y)log(P(y))$$

And the stimulus noise entropy, given as

$$H_{noise} = -dydsP(y|s)P(s)log(P(y|s))$$

.

Since the probability of the response $P(y)$ is calculable from the recorded response itself, the value of H_{resp} is fixed, and consequently, $I(y;\mathbf{s}) = H_{resp} - H_{noise}$ will be minimized by maximizing the value of H_{noise}.

Optimization of H_{noise} is not possible without the imposition of a model for the conditional response probability $P_{mod}(y|\mathbf{s})$; else, the set of models that could generate the observed set of correlations between the stimulus and response is theoretically infinite. Thus, to ensure that H_{noise} is able to be maximized, a model for $P_{mod}(y|\mathbf{s})$ is constructed that, while being consistent with the observed correlations, is otherwise as unconstrained as possible, such that it is minimally biased with respect to the observed data. This minimal model is constructed by constraining model statistics to match the response-weighted moments of the stimulus distribution, $\langle y \rangle_y, \langle y\mathbf{s} \rangle_{y,\mathbf{s}}, \langle y\mathbf{s}\mathbf{s}^\mathbf{T} \rangle_{y,\mathbf{s}}$, etc. With these minimal, data-driven constraints, the conditional probability $P_{mod}(y|\mathbf{s})$ is given by the logistic function

$$P(y|\mathbf{s}) = \frac{1}{1 + e^{-z(\mathbf{s})}},$$

where $z(\mathbf{s}) = a + \mathbf{h}^T\mathbf{s} + \mathbf{s}^T\mathbf{J}\mathbf{s}$, and the weights $a, \mathbf{h}$, and $\mathbf{J}$ are determined by minimizing the negative log likelihood

$$L(a, \mathbf{h}, \mathbf{J}) = -\frac{1}{N_{samp}} \sum_t \left[y_t log(P(y|\mathbf{s}_t)) + (1 - y_t) log(1 - P(y|\mathbf{s}_t)) \right],$$

defined here as an average over samples. This log likelihood function is convex, and thus is guaranteed not only to converge but to converge on its global minimum

in polynomial time.[1]

The quadratic polynomial $z(\mathbf{s}) = a + \mathbf{h} \cdot \mathbf{s} + \mathbf{s}^T J \mathbf{s}$ defines a set of quadratic surfaces that correspond to the minimal model's contours of constant probability. When $z(\mathbf{s})$ (and consequently J) is diagonalized, then, the resultant eigenvectors $\{\mathbf{z}_i\}$ correspond to the principal axes of the constant probability surfaces, and the magnitude of their respective eigenvalue $\{\beta_i\}$ corresponds to the curvature of the surface in that dimension, which in turn corresponds to the degree of selectivity for the feature defined by that dimension.

Altogether, MNE models enable the stable reconstruction of multicomponent receptive fields for higher order sensory neurons. To date MNE models have been used to characterize the receptive fields of putative single neurons based on spiking activity in visual and auditory areas of nonhuman animals (Kozlov and Gentner, 2016; Clemens et al., 2012; Rowekamp and Sharpee, 2017; Atencio and Sharpee, 2017). This chapter builds on those successes, showing not only that MNE models can successfully reconstruct receptive fields from human intracranial LFP activity, but also that phonemic category information aids in the reconstruction of higher-order sensory receptive fields. Section 4.2 describes the sets of models that were constructed and how they were compared to one another; Section 4.3 describes the results of those comparisons; Section 4.4 discusses the significance of these results; and Section 4.5 concludes.

4.2 Methods

4.2.1 Input features

Input features were created for each of the two conditions of interest. For the unlabeled spectrogram condition, 256-dimensional (16x16) spectrograms were created for each timebin of the neural response to represent the 1024ms of acoustic context preceding that timebin. These spectrograms were created using the proce-

[1]This is another advantage of MNE models over MID models, whose optimization functions are not convex and therefore are vulnerable to converging on local rather than global optima (Fitzgerald et al., 2011).

dure described in §2.2. For the labeled spectrogram condition, 8-bit binary codes (i.e., [01010110]) were created for each possible ARPABET label present in the Buckeye Corpus. Then, two labels (totaling 16 bits) were assigned to each spectrogram. For 1024ms windows containing only one phone-level label, the 16-bit label given to that spectrogram was the simple concatenation of the code for that ARPABET label with itself. For spectrograms containing more than one phone-level label, the 16-bit label given to that spectrogram was the simple concatenation of the code for the label that occupied the plurality of that 1024ms window and the runner-up label. Then the mean and standard deviation of each spectrogram was calculated, and the top row of each spectrogram was replaced with a 1x16 vector representing its 16-bit label. For this row vector, "0" entries of the 16-bit label were replaced with the value one standard deviation below the spectrogram's mean, and "1" entries were replaced with the value one standard deviation above the spectrogram's mean. In this way, spectrographic labels were designed to have minimal impact on the prexisting variance and covariance structure of the spectrograms.

4.2.2 Neural response measures

Models were fit for seven response types: delta, theta, alpha, beta, gamma, and high gamma power, and broadband LFP. The preprocessing for each of these data types is identical to those reported in §2.2.

4.2.3 MNE models

MNE models were fit for each participant for each response type for each condition. As described in Section 4.1.2 and reprinted below in (6), these models fit a logistic function that models the response type (LFP or band power) as a linear combination of first- and second-order features of the stimulus (labeled or unlabeled spectrograms).

(6) Maximum Noise Entropy Model

$$P(y|\mathbf{s}) = \frac{1}{1+e^{-z(\mathbf{s})}}, z(\mathbf{s}) = a + \mathbf{h}^T\mathbf{s} + \mathbf{s}^T\mathbf{J}\mathbf{s}$$

To prevent overfitting, model parameters were estimated over four jackknives: Input and response data were split into four batches, and for each jackknife, three of the batches were used as training data, and the remaining batch was used for testing. The log loss between the response and the weighted input was minimized using `fmin_cg`, a nonlinear conjugate gradient algorithm from the scientific computing library `scipy.optimize`, and early stopping was implemented such that if the log loss on the testing data did not decrease for ten consecutive epochs, training was halted. The optimized weights for each of the four jackknives were then averaged, and this set of mean weights were used in all subsequent work.

For each channel, the fit models were used to generate predicted neural responses. First-order predictions were generated using the fit linear features only, with $z(\mathbf{s}) = a + \mathbf{h}^T\mathbf{s}$, and second-order predictions were fit using the complete model, with $z(\mathbf{s}) = a + \mathbf{h}^T\mathbf{s} + \mathbf{s}^T\mathbf{J}\mathbf{s}$. Then for each predicted response, Pearson's r was calculated to assess the degree of correlation between the recorded and predicted responses. These values were then normalized using the Fisher Z-Transformation such that prediction quality could be compared across model types and conditions.

4.2.4 Control

To ensure that each model fit better than would be expected by chance, model predictions were also generated from the shuffled features of the fit models. For the linear model, the entries of the 256-dimensional vector feature representing the power-responsive average were shuffled, and predictions were generated using the shuffled linear features only, with $z(\mathbf{s}) = a + \mathbf{h}^T\mathbf{s}$. For the quadratic model, the linear feature was shuffled as described for the linear model. Additionally, the quadratic model's J-matrix was eigendecomposed, its eigenvalues were shuffled, and the J-matrix was then recomposed using the shuffled eigenvalues. Predictions for the shuffled quadratic models were then generated as expected, with $z(\mathbf{s}) = a + \mathbf{h}^T\mathbf{s} + \mathbf{s}^T\mathbf{J}\mathbf{s}$. Shuffling in this way leaves the quadratic features (the eigenvectors of the J-matrix) intact, altering only the weight afforded to each feature in the predicted response. This manner of shuffling therefore provides a

relatively liberal estimate of the quality of response that would be expected by chance, since it preserves the structure of the second-order features in their entirety, merely reweighting their contribution to the predicted response.

4.3 Results

4.3.1 Model evaluation

In Chapter 3, all best-fit mixed effects models included spectrographic features, highlighting the pervasive importance of acoustic information in speech processing. However, in lower frequency bands in particular, categorical phonemic information also played a substantial role, explaining variability in the neural data that was not otherwise explained by spectrographic information. This observation raises two questions: First, which aspects of the stimulus are most important in explaining the neural response to speech? And second, what is the nature of the relationship between spectrographic and phonemic label features? Maximum Noise Entropy (MNE) models provide the analytical tools to address these questions because they allow one to test specific claims about the nature of the link between the stimulus and the neural response. Here, by comparing the fit of linear models to quadratic models, we determine the extent to which relationships between pairs of features impact neural fit, and by comparing the fit of models given phonemic label information to models fit with solely spectrographic information, we determine the role that phonemic information plays relative to speech acoustics.

To assess whether model order and/or the availability of phonemic label information contributed to the quality of MNE model fits, five mixed effects models were fit. These models predicted the Fisher Z-transformed Pearson correlation between predicted and observed neural responses on the basis of the neural response type, model order, the availability of phonemic label information, the interaction of model order with the availability of label information, and whether the model feature weights were shuffled. Subject and channel were included as random effects. The formulas for these models are given in Table 4.1. In these models, `model` is a binary feature referring to whether the predicted fit was generated using only

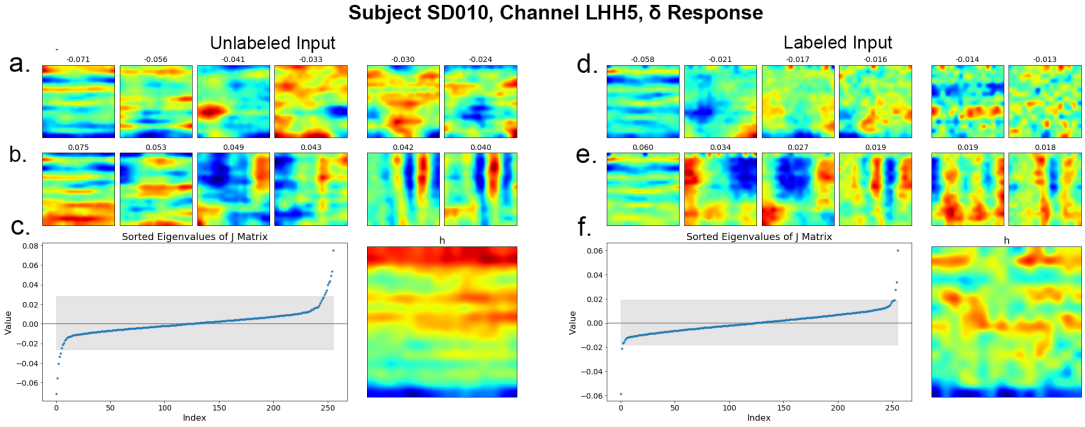

Figure 4.1: Reconstructed receptive fields reflect available phonemic and spectrographic information. (A) Receptive fields corresponding to the six most negative eigenvalues for a single electrode whose δ-power response was fit with purely spectrographic information. (B) Receptive fields of the six most positive eigenvalues for the same electrode as (A). (C) Left: Eigenvalues of the fit model, ordered by their value. Grey shaded box encompasses the bottom 95% of eigenvalue magnitudes. Right: The power-dependent average receptive field. (D) Receptive fields for the six most negative eigenvalues reconstructed from the δ-power response of the same electrode shown in (A) using a model that was fit with both spectrographic and phonemic label information. (E) Receptive fields of the six most positive eigenvalues for the same electrode as (D). (F) Left: Eigenvalues of the fit model, ordered by their value. Grey shaded box encompasses the bottom 95% of eigenvalue magnitudes. Right: The power-dependent average receptive field.

the linear MNE feature or predicted using both the linear and quadratic features; the feature `label` indicates whether or not the MNE model was fit using labeled or unlabeled spectrographic features; and the feature `shuffle` indicates whether or not the predicted fit was generated using MNE features whose weights had been shuffled. The model containing only the `shuffle` parameter as a fixed effect was used as a baseline.

Model selection was determined using the AIC. Of the five models, the model containing `model`, `label`, and `model:label` features best fit the data, having an AIC score 54.75 units lower than the next best model and carrying 100% of the cumulative model weight. The next-best fit model included `label` and `shuffle` features and had an AIC score 1.60 units less than the model that additionally included the `model` feature. Similarly, the baseline model containing only the

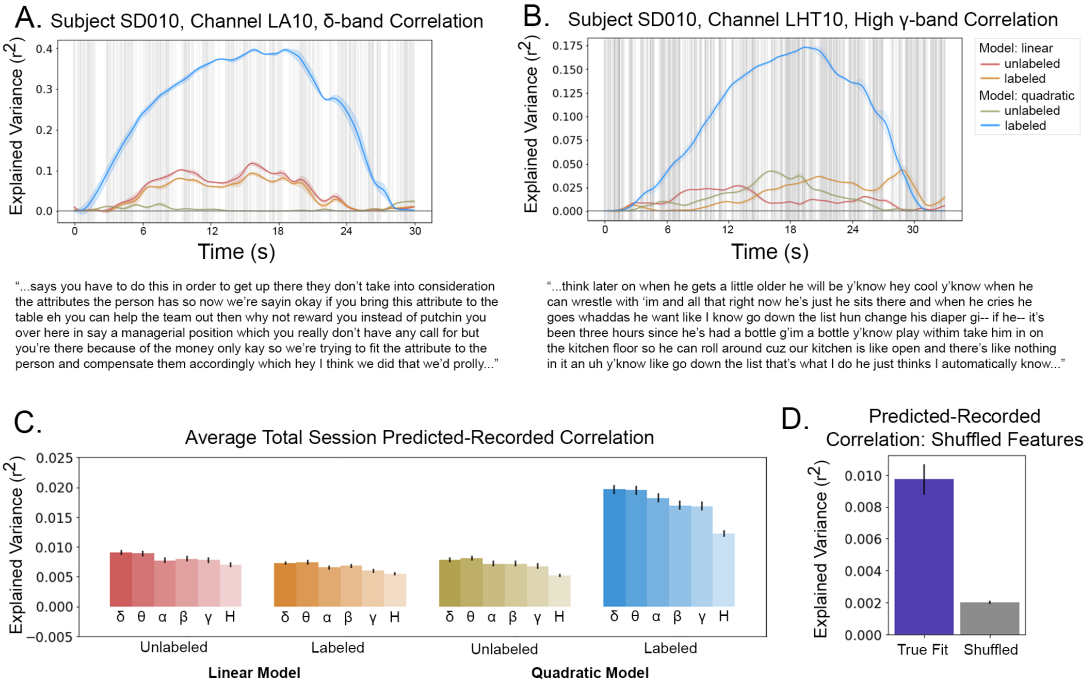

Figure 4.2: Combined with phonemic label information, stimulus covariance structure best predicts the neural response to speech across all response bands. Time-varying correlation between predicted and recorded neural response for the (A) δ-power and (B) high γ-power responses of single electrodes, illustrating periods of sustained advantage of the labeled quadratic model (blue) over other models. Grey vertical lines indicate phoneme boundaries for the transcripts p rovided below each plot. (C) Average correlation between predicted and recorded response for all subjects, broken out by model type and neural response band. (D) Average correlation between predicted and recorded response for all true fit models vs. all shuffled models.

`shuffle` feature also outperformed a model that was minimally augmented by the addition of the `model` feature, garnering an AIC score of 1.59 units lower than the model containing both `model` and `shuffle` features. Together, these results suggest that while MNE model complexity alone does not positively impact prediction quality, its interaction with the label status of the spectrograms used to fit the MNE models is substantial. In other words, when phonemic label information is available, information in the stimulus covariance can be used to more accurately predict neural response. Without phonemic label information, the stimulus covariance does not contribute substantially to model goodness-of-fit.

Table 4.1: Fixed effects and AIC scores for MNE models. Models arranged in descending order of fit. Subject and channel were included as random effects for all models.

AIC	Delta	AICw	Model Fixed Effects
-274494.0	0.00	1.00	`model + label + model:label + shuffle + band`
-274439.2	54.75	0.00	`label + shuffle + band`
-274437.8	56.35	0.00	`model + label + shuffle + band`
-274348.0	145.92	0.00	`shuffle + band`
-274346.4	147.52	0.00	`model + shuffle + band`

The failure of the `model` feature to contribute independently to goodness-of-fit was not predicted. *Ceteris paribus*, the higher dimensionality of the quadratic model is expected to recover a higher proportion of the variability in the neural response than the linear model, as has been demonstrated in Kozlov and Gentner (2016). However, whereas Kozlov and Gentner (2016) recorded from songbird auditory areas only, the data in this study were recorded from speech-selective electrodes, the majority of which were not localized to auditory areas. Given that the receptive fields in this study were estimated on the basis of primarily spectrographic input features, it may be the case that the higher dimensionality of the quadratic model does not confer a benefit in non-auditory areas where spectrographic information contributes less to neural response. Nevertheless, the fact that the interaction feature (`model:label`) does substantially improve model goodness-of-fit suggests that information about spectrographic covariance is useful for predicting neural response when it is reinforced by categorical phonemic label information. In this sense, the auditory stimulus covariance structure and phonemic identity synergistically impact neural response, jointly providing information available in neither feature independently.

4.3.2 Model performance for Catalan speech

Both the LME models fit in Chapter 3 and the MNE models fit in this chapter indicated that categorical phonemic information contributes to the prediction of the neural response to natural speech. Next we asked whether the utility of

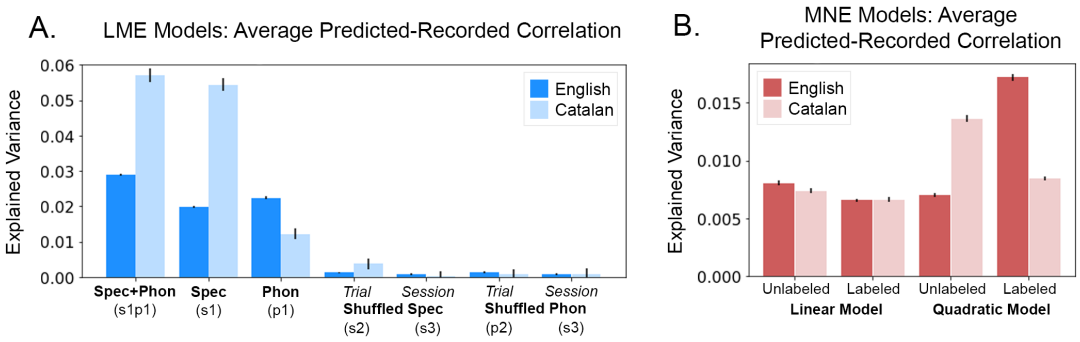

Figure 4.3: Phonemic labels improve model fits for English but not Catalan data. (A) Average correlation between the response predicted by each of the LME models and the actual recorded response. (B) Average correlation between the response predicted by each of the MNE models and the actual recorded response.

phonemic label information in these models requires specific language knowledge. That is, it could be the case that phonemic label information plays a significant role in these models simply because it reinforces acoustic information. If this were the case, we would expect the addition of phonemic label information to improve model fit both when participants were listening to a language they understand and are familiar with (i.e., English) and when listening to a language they do not understand and are unfamiliar with (i.e., Catalan). Alternatively, phonemic label information could play a significant role in these models because it provides language-specific category membership information that is not otherwise available in the speech acoustics. If this were the case, phonemic label information should only contribute significantly towards explaining neural activity while participants are listening to speech in a language that they are familiar with (English) and not while listening to speech in a language they are unfamiliar with and do not understand (Catalan).

To assess whether this was the case, families of mixed effects models were fit on the predicted-recorded correlation values for the predictions of each of the seven LME models and each of the four MNE models.

For the LME models, each model for assessing the role of specific language knowledge predicted the Fisher Z-transformed Pearson correlation between pre-

dicted and observed neural responses to particular excerpts on the basis of the LME model type (`model`: **s1p1**, **s1**, **p1**, etc.) and the band of neural activity. A second model had an additional feature (`lang`) for the language being spoken in the excerpt (English or Catalan), and a third model additionally included an interaction term for excerpt language and model type (`lang:model`). For each of these three models, subject and channel were included as random effects. The formulas for the three models in this family are given in Table 4.2.

Model selection was then determined using the AIC. Of the three models, the model containing both `lang` and `lang:model` features in addition to `model` and `band` features best fit the data, having an AIC score 474.39 units lower than the next best model and carrying 100% of the cumulative model weight. The next-best fit model included only `lang`, `model`, and `band` features and had an AIC score 9748.75 units less than the model that included only the `model` and `band` features. These results indicate not only that the strength of correlation between the predicted and recorded responses varies based on the language that the participant is listening to, but also that difference in predicted-recorded correlation between English and Catalan excerpts also varies with model type. This is shown most clearly in Figure 4.3A: the average predicted-recorded correlation for models containing spectrographic features (**s1p1** and **s1**) is higher for Catalan excerpts than for English excerpts. However, for the model containing only phonemic label features (**p1**), the average predicted-recorded correlation is higher for English excerpts than for Catalan excerpts. Specific language knowledge is required for phonemic label information to substantially predict neural response to speech.

Similarly, to assess assessing the role of specific language knowledge on the fit of the MNE models, a family of models was fit. These models predicted the Fisher Z-transformed Pearson correlation between predicted and observed neural responses on the basis of model order (linear or quadratic), the availability of phonemic label information (yes or no), excerpt language (English or Catalan), and combinations of their interactions. Subject and channel were included as random effects. The model that included `band`, `model`, `label`, and `model:label` features as fixed effects was treated as the baseline model. The formulas for the six models in this family

Table 4.2: Fixed effects and AIC scores for mixed models that were fit on r^2 values for the correlation between the neural responses predicted by LME models and those recorded in actuality. Models arranged in descending order of fit. Subject and channel were included as random effects for all models.

AIC	**Delta**	**AICw**	**Model Fixed Effects**
-214554.6	0.00	1.00	`lang + model + lang:model + band`
-214080.2	474.39	0.00	`lang + model + band`
-204331.5	10223.14	0.00	`model + band`

are given in Table 4.3.

As before, model selection was determined using the AIC. Of the six models, the model containing the simplex features `model`, `label`, `lang` and all their interactions best fit the data, having an AIC score 775.87 units lower than the next best model and carrying 100% of the cumulative model weight. Among the models with lower fit, those including a feature for the interaction between excerpt language and label status (`lang:label`) better fit the data than models without this feature, and all models including the excerpt language feature (`lang`) better fit the data than the baseline model. These results indicate that participants' knowledge of the language they are listening to significantly impacts model fit and does so in a way that interacts primarily with the availability of categorical phonemic information. As shown in Figure 4.3B, the average predicted-recorded correlation for quadratic models is higher for Catalan excerpts than for English excerpts when the MNE model's input features are purely spectrographic, whereas for MNE models that include phonemic label information in addition to spectrographic information, the average predicted-recorded correlation is higher for English excerpts. In this way, the availability of phonemic information does not benefit MNE model fit for neural activity recorded while participants are listening to an unfamiliar language that they cannot understand. Categorical speech sound information requires specific language knowledge in order to account for neural activity.

Table 4.3: Fixed effects and AIC scores for mixed models that were fit on r^2 values for the correlation between the neural responses predicted by MNE models and those recorded in actuality. Models arranged in descending order of fit. Subject and channel were included as random effects for all models.

AIC	Delta	AICw	Model Fixed Effects							
			band	model	label	model:label	lang	lang:model	lang:label	lang:model:label
-313838.1	0.00	1.00	✓	✓	✓	✓	✓	✓	✓	✓
-313062.2	775.87	0.00	✓	✓	✓	✓	✓	✓	✓	
-313057.1	780.96	0.00	✓	✓	✓	✓	✓		✓	
-312431.7	1406.36	0.00	✓	✓	✓	✓	✓	✓		
-312426.7	1411.37	0.00	✓	✓	✓	✓	✓			
-312406.0	1432.05	0.00	✓	✓	✓	✓				

4.4 Discussion

4.4.1 Reframing the invariance problem

MNE models model the relationship between the stimulus and neural response as a linear combination of the contributions of stimulus feature sets with varying cardinalities. That is, a first-order MNE model models neural response as the contributions of individual stimulus features; a second-order model sums single feature contributions with the contributions made by pairs of stimulus features; a third-order model further includes the contributions of triplets of stimulus features; and so on.

This study specifically compares the fits of first- and second-order models. In doing so, it assesses the extent to which relationships between stimulus features contribute to model goodness-of-fit beyond the contributions of individual features themselves. In finding that second-order, quadratic models account for significantly more of the variance in neural response only when phonemic label information is available, this study shows that categorical phonemic information bootstraps acoustic information in explaining the neural response to speech. More precisely, without the availability of pairwise information about the stimulus, phonemic label

information does not contribute significantly to predicting neural response.

In this way, relationships between spectrotemporal features of the speech signal are themselves the structure from which stable phonemic neural responses arise. That is, rather than tracking absolute invariances in the speech signal, phoneme-like neural responses are crucially relational, drawing on contextualized rather than absolute properties of speech acoustics. This result underscores the importance of context in phonological theory. Although analysts posit static underlying forms for speech, all recognize that the determination of those underlying forms requires precise kinds of contextual knowledge. For example, determining the underlying form of the German word [ʁat] 'wheel' requires both local phonological knowledge that the sound [t] is word final (perhaps cued by acoustic context such as vowel lengthening) as well as global lexical knowledge that the word [ʁat] means 'wheel' and not 'council'.

This study assessed the utility of restricted forms of phonological and spectrotemporal context for predicting the neural response to speech: Only 1024ms of low-resolution spectrographic and phonemic label information were available to predict each time point of the neural response. Beyond the inherent restrictiveness of the features' low resolution, relevant domain information such as morpheme and word boundaries or the presence of stress were also not provided to the models. Very likely, the inclusion of these and other forms of higher order context, such as part of speech or word identity, would improve the fit of these models. However, the decision not to use extended features sets in this study was a principled one.

One of the primary goals of this work is to assess the extent to which sensory information is capable of accounting for the neural response to speech. In privileging the acoustic signal and providing only phonemic label features as an alternative explanatory feature set, this study is pitched precisely to assess the interaction between continuous units of speech and the categorical units of language most redundant with those continuous units. Thus, finding that both spectrographic covariance and phonemic label features together account for more of the neural response to speech than either factor independently strongly suggests that the covariance structure of speech contains information that can only be extracted

with knowledge of a language's phonemic categories. In this sense, the relationship between underlying phonemic structure and surface acoustics is a two-way street: determining the underlying form of the German word [ʁat] 'wheel' draws on local acoustic cues, but accurately interpreting local acoustic cues also draws on categorical phonological knowledge.

Thus, there are two systems of invariance at play in the neural response to speech. One is signal internal: the covariance structure of speech acoustics explains more of the neural response than the mean acoustic structure. The other constitutes phonology itself: the covariance structure of speech acoustics explains more of the neural response than the mean acoustic structure *only when categorical phonemic information is available.* The systems of invariance at play include both the mutual dependence between surface and abstract forms *and* the stability of spectrotemporal relationships within the speech stream itself.

This study provides structure to longheld understandings of the important role that relative spectrotemporal features play in speech sound identification. For example, at single time points, measures of F1:F2 ratio or spectral tilt provide key information for discriminating among vowels or among fricatives, respectively, and comparing measures across time provides further discriminatory ability. Features of spectrographic stimulus covariance have also been implicated in speaker normalization (Zhou and Hansen, 2005). In these ways, the sensory signal itself provides an exceptional degree of structure that mirrors in fine scale what linguistic categories demonstrate at a coarser scale. That is, speech sounds are built from relationships between acoustic features; phonemes from relationships between speech sounds; morphemes from relationships between phonemes; meaning from relationships between morphemes, and so on.

4.5 Conclusion

This study provides evidence that relative power between pairs of time-frequency bins supports the neural response to speech when phonemic category information is also available. Linking features of the sensory signal to linguistic categories in this

way provides structure for an early link between sensory and linguistic cognition.

Material from Chapter 4 has been submitted for publication as "The neural response to speech is language specific and irreducible to speech acoustics" with authors Anna Mai, Stephanie Riès, Sharona Ben-Haim, Jerry Shih, & Timothy Gentner. The dissertation author was the primary investigator and author of this paper.

Bibliography

Atencio, C. A. and Sharpee, T. O. (2017). Multidimensional receptive field processing by cat primary auditory cortical neurons. *Neuroscience*, 359:130–141.

Blumstein, S. E. and Stevens, K. N. (1979). Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *The Journal of the Acoustical Society of America*, 66(4):1001–1017.

Brunelle, M. (2012). Dialect experience and perceptual integrality in phonological registers: Fundamental frequency, voice quality and the first formant in cham. *The Journal of the Acoustical Society of America*, 131(4):3088–3102.

Casserly, E. D. and Pisoni, D. B. (2010). Speech perception and production. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(5):629–647.

Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., and Knight, R. T. (2010). Categorical speech representation in human superior temporal gyrus. *Nature neuroscience*, 13(11):1428–1432.

Clemens, J., Wohlgemuth, S., and Ronacher, B. (2012). Nonlinear computations underlying temporal and population sparseness in the auditory system of the grasshopper. *Journal of Neuroscience*, 32(29):10053–10062.

Cole, R. A. and Scott, B. (1974a). The phantom in the phoneme: Invariant cues for stop consonants. *Perception & Psychophysics*, 15(1):101–107.

Cole, R. A. and Scott, B. (1974b). Toward a theory of speech perception. *Psychological review*, 81(4):348.

Essegbey, J. and McCollum, A. G. (2018). Initial prominence and progressive vowel harmony in Tutrugbu.

Fairhall, A. L., Burlingame, C. A., Narasimhan, R., Harris, R. A., Puchalla, J. L., and Berry, M. J. (2006). Selectivity for multiple stimulus features in retinal ganglion cells. *Journal of neurophysiology*, 96(5):2724–2738.

Fant, G. (1960). Acoustic theory of speech production, s' -gravenhage. *Mouton and Co.*

Fitzgerald, J. D., Sincich, L. C., and Sharpee, T. O. (2011). Minimal models of multidimensional computations. *PLoS computational biology*, 7(3):e1001111.

Garellek, M., Keating, P., Esposito, C. M., and Kreiman, J. (2013). Voice quality and tone identification in White Hmong. *The Journal of the Acoustical Society of America*, 133(2):1078–1089.

Gick, B., Pulleyblank, D., Campbell, F., and Mutaka, N. (2006). Low vowels and transparency in Kinande vowel harmony. *Phonology*, 23(1):1–20.

Holt, L. L. and Lotto, A. J. (2010). Speech perception as categorization. *Attention, Perception, & Psychophysics*, 72(5):1218–1227.

Kaardal, J. T. (2017). *Decoding the computations of sensory neurons*. University of California, San Diego.

Klatt, D. H. (1979). Speech perception: A model of acoustic–phonetic analysis and lexical access. *Journal of phonetics*, 7(3):279–312.

Kozlov, A. S. and Gentner, T. Q. (2016). Central auditory neurons have composite receptive fields. *Proceedings of the National Academy of Sciences*, 113(5):1441–1446.

Lee, C.-Y. (2009). Identifying isolated, multispeaker Mandarin tones from brief acoustic input: A perceptual and acoustic study. *The Journal of the Acoustical Society of America*, 125(2):1125–1137.

Liberman, A. M., Delattre, P., and Cooper, F. S. (1952). The role of selected stimulus-variables in the perception of the unvoiced stop consonants. *The American journal of psychology*, pages 497–516.

Norman-Haignere, S., Kanwisher, N. G., and McDermott, J. H. (2015). Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. *Neuron*, 88(6):1281–1296.

Rowekamp, R. J. and Sharpee, T. O. (2011). Analyzing multicomponent receptive fields from neural responses to natural stimuli. *Network: Computation in Neural Systems*, 22(1-4):45–73.

Rowekamp, R. J. and Sharpee, T. O. (2017). Cross-orientation suppression in visual area V2. *Nature communications*, 8(1):1–9.

Silva, D. J. (2006). Acoustic evidence for the emergence of tonal contrast in contemporary Korean. *Phonology*, 23(2):287–308.

Toscano, J. C., McMurray, B., Dennhardt, J., and Luck, S. J. (2010). Continuous Perception and Graded Categorization: Electrophysiological Evidence for a Linear Relationship Between the Acoustic Signal and Perceptual Encoding of Speech. *Psychol. Sci.*, 21(10):1532–1540.

Zhou, B. and Hansen, J. H. (2005). Rapid discriminative acoustic model based on eigenspace mapping for fast speaker adaptation. *IEEE Transactions on Speech and Audio Processing*, 13(4):554–564.

Chapter 5

Concluding Remarks

Speech is a highly-structured auditory signal that carries necessary and sufficient information for language comprehension among neurotypical, hearing populations. Speech perception, accordingly, involves the integration of both bottom-up auditory processing and top-down language knowledge. As explored in Chapter 1, however, while much is known about the brain's sensitivity to various acoustic and phonological manipulations, mechanistic transformative links between auditory sensation and linguistic perception remain poorly understood.

This dissertation contributed three studies to the ongoing pursuit of links between auditory and linguistic processing of speech in the brain. Intracranial stereo electroencephalography (sEEG) was recorded while participants listened to conversational English speech, and for each study, different sets of analyses were performed to target different aspects of these links.

Chapter 2 introduced a novel comparative technique for the isolation of phonological and morphological identity from the auditory speech signal. Through structured comparisons of the neural response to sounds in allophonic relationships with one another, sites that dissociate phonemic and morphological identity from acoustic similarity were identified. These results provided crucial evidence that the neural response to speech is sensitive to abstract sub-lexical structure.

Chapter 3 provided general models to account for the unique contributions of categorical phonemic information and spectrographic information to the neural response. In lower frequency bands, phonemic category information explained a

greater proportion of the neural response variance than spectrographic information. These results indicated that the explanatory value of phonemic category features in neural encoding models is not reducible to the surface spectrotemporal features that correlate with those categories.

Chapter 4 proposed a model for the relationship between the speech stimulus and neural response that integrated both phonemic and spectrographic information. Inclusion of the stimulus covariance structure increased model accuracy only when categorical phonemic information was available, indicating that phoneme-related neural activity is likely mediated through the acoustic covariance structure of speech. Moreover, phonemic label information conferred no benefit to model fit when participants listened to speech in an unfamiliar language (Catalan).

Together these studies show that neural responses associated with categorical phonemic information are language specific and irreducible to speech acoustics, and they scout a path towards mechanistic, transformative links between auditory sensation and language cognition.