Title

Cross-language speech computations in the human temporal cortex

ISBN

9798270226237

Author

Bhaya-Grossman, Ilina

Publication Date

2025-12-17

Peer reviewed|Thesis/dissertation

Cross-language speech computations in the human temporal cortex

by

Ilina Bhaya-Grossman

DISSERTATION

Submitted in partial satisfaction of the requirements for degree of

DOCTOR OF PHILOSOPHY

in

Bioengineering

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO
AND
UNIVERSITY OF CALIFORNIA, BERKELEY

Approved:

DocuSigned by:
Edward F. Chang
4B5B40824E04415...

Edward F. Chang
Chair

Signed by:
Keith Johnson
DocuSigned by:
489...

Keith Johnson

DocuSigned by:
Matthew Leonard
DocuSigned by:
46C...

Matthew Leonard

DocuSigned by:
Frederic Theunissen
E44C370166F64C2...

Frederic Theunissen

Committee Members

To my family, friends, and colleagues, without whom I would have neither the voice nor the language to complete this work.

Acknowledgements

I would like to extend my deepest gratitude to all the people who have made this work possible. In my mind there are two outcomes to a PhD: the science itself, and the scientist who emerges through the process. Everyone mentioned below has contributed immensely to both.

First and foremost, I would like to thank all the participants who have taken part in our research over the years. This work quite literally would not exist without their time, effort, and trust.

I am profoundly thankful to my mentors, each of whom I've found to be exceptional in their own right. Edward Chang (Eddie), my advisor, has shared with me his boundless curiosity and incredible appreciation for the natural world. His genuine awe for language neuroscience, and his remarkable ability to communicate its beauty to others, have fundamentally shaped my own love of science and storytelling.

Matthew Leonard (Matt) has impressed upon me the importance of slowing down and deliberating. His careful and often copious feedback, on everything from manuscript text to life decisions, has undoubtedly made me a more critical writer, scientist, and thinker. Yulia Oganian has shared with me her perseverance, warmth, and excitement for the obscure (formant) corners of acoustic science. And finally, I am beyond thankful to have worked closely with Keith Johnson, whose office I always left with a renewed sense of joy in science and a large stack of papers to read.

I also indebted to my thesis and qualifying committees, including Frederic Theunissen, Christoph Kirst, Christoph Schreiner, and Michael Yartsev, who together ensured that my PhD work remained rigorous, analytical, and grounded. I thank Jim Magnuson and the Basque Center on Cognition, Brain, and Language (BCBL) for hosting me during a research stay, an unforgettable PhD experience that broadened my scientific perspective.

I am grateful for the formative experiences I had at UC Berkeley as an undergraduate. In Bob Knight's lab, under the generous guidance of Arjen Stolk, I was first introduced to computational neuroscience and learned the grit required to see an empirical study through. It was also at Berkeley, through the Knight Lab and courses taught by Susanne Gahl in the Linguistics Department, that I developed an enduring interest in neuroscience, communication, and language.

I have learned as much from collaborators as from mentors. I am especially grateful to Yizhen Zhang, whose collaboration on the neural representations of words has been among the greatest joys of my PhD, and to Alex Silva, with whom I had the pleasure of co-teaching a medical mini-course on speech processing. I would also like to thank the many others who made meaningful contributions to my PhD work, including Laura Gwilliams, Jon Gauthier, Sarah Harper, and Marcos Lobato-Scharfhausen.

I am fortunate to have had labmates who made even the dreariest days in the lab feel bright. Deborah Levy showed me the creativity and humanity inherent in scientific discovery. Sean Metzger, Jessie Liu, Aria Lin, Quinn Greicius, Ashley Qin, Shailee Jain, and the entire past and present Chang Lab made my time in the PhD endlessly fun, through their collective intellect, humor, and kindness.

Thank you to the UCSF Neurosurgery media team, Todd Dubnicoff, Ilona Garner, and Alejandra Canales, and to the broader UCSF media teams, who have helped me share my scientific stories and reach a wider audience.

I also thank lab administrators Viv Her and Susanna Lam, who patiently handled every reimbursement and logistical hurdle; the Bioengineering program staff, especially Victoria Ross and Rocio Sanchez; and my graduate advisors, Karunesh Ganguly and Sri Nagarajan, for their steady guidance and support. The BioE community, including Amanda Meriwether, Preethi Raghavan, Kevin Joslin, Niroshan Anandasivan and Caleb Rux, has been a constant source of balance and friendship throughout my PhD.

My family's contribution to this work is beyond words, but I will make a brief attempt nonetheless. To my immediate and extended family: thank you for instilling in me optimism, humor, and work ethic. To my parents: thank you for enrolling me in a Chinese language immersion program, without which I may never have found an interest in language or speech, or neuroscience. And thank you for demonstrating to me firsthand that to be a scientist is not about choosing a career but choosing a way of being. I am especially grateful to my mother: thank you for trying, as best you can, to make sure I do not repeat your mistakes (and for reluctantly forgiving me when I do). And finally, to Norman: thank you for teaching me what it means to be fearless, for showing me how to run towards new ideas, and for sharing in both the joy and frustration of finding one's purpose.

Contributions

This dissertation contains material that has been previously published in peer-reviewed journals. Namely, Chapter 1 and Chapter 4 are directly adapted from:

> Bhaya-Grossman, I., & Chang, E. F. (2022). Speech Computations of the Human Superior Temporal Gyrus. *Annual Review of Psychology*, *73*, 79–102.

Personal contributions: I generated figures and wrote the first manuscript draft. With Edward F. Chang, I edited and refined the figures and manuscript.

Chapter 2 is directly adapted from:

> Oganian, Y.*, Bhaya-Grossman, I.*, Johnson, K., & Chang, E. F. (2023). Vowel and formant representation in the human auditory speech cortex. *Neuron*, *111*(13), 2105-2118.e4.

* Denotes equal author contributions.

Supplementary material is not included in this adaptation but is available online.

Personal contributions: I implemented the analysis for and generated figures for Experiment 1 and a subset of analysis for Experiment 2. With Yulia Oganian, I constructed stimuli for Experiment 2, collected the data, edited all figures, and wrote and edited the original manuscript.

Chapter 3 is directly adapted from:

> Bhaya-Grossman, I., Leonard, M. K., Zhang, Y., Gwilliams, L., Johnson, K., Lu, J., & Chang, E. F. (2025). Shared and language-specific phonological processing in the human temporal lobe. *Nature*, 1–12.

Personal contributions: I implemented the data analysis for and generated all main and supplementary figures. I also wrote the original manuscript. With my co-authors, I revised and edited all figures and the original manuscript.

Cross-language speech computations in the human temporal cortex

Ilina Bhaya-Grossman

Abstract

Understanding speech is a rapid and seemingly effortless feat that humans perform every day. Yet this ability relies on complex neural computations that integrate language-specific knowledge in order to transform external acoustic signals into internal linguistic representations. This dissertation investigates how neural computations in the human temporal cortex, particularly in the superior temporal gyrus (STG), support speech processing across languages.

In the first chapter, we introduce the theoretical and empirical foundations for studying speech processing in the STG and describe how high-resolution intracranial brain recordings, such as electrocorticography (ECoG), have enabled its functional description. Based on prior work, we argue that the STG performs fundamentally nonlinear and dynamic speech computations, including categorization, normalization, and the extraction of temporal structure. Open questions remain as to the extent to which speech computations in the STG are shared across speakers of different languages or language-experience dependent.

In the second chapter, we focus on vowel perception to understand how abstract language-specific speech categories emerge from continuous acoustic variation. Vowels are acoustically cued by formants, resonance frequencies determined by vocal tract configuration, and serve as a fundamental building block across spoken languages. Using ECoG recordings from Spanish speakers listening to continuous speech, we demonstrate that in the STG, the predominant encoding of vowels is acoustic. Specifically, we find that neural responses recorded from single electrode contacts are highly selective to restricted zones

within the two dimensional formant space. Only by aggregating across these local neural responses can vowel categories be decoded. A second experiment in which participants listened to synthetic formant-based sounds revealed that STG neural responses exhibit general formant sensitivity that extends beyond the natural speech range.

In the third chapter, we look beyond single sounds (e.g. vowels) to address how language experience affects speech computations in the human STG. We collected ECoG recordings from speakers of Spanish, English and Mandarin as they listened to speech in both their native and an unfamiliar foreign language. Across participants, the STG exhibited shared acoustic-phonetic encoding for basic speech sound features such as vowel formants and consonants. However, enhanced neural encoding for language-specific sequence and word-level and features was observed when listening to native speech versus foreign speech. In Spanish-English bilinguals, language-specific neural encoding was observed for both familiar languages within the same STG neural populations. Together, these findings demonstrate that the encoding of acoustic-phonetics is integrated with language-specific word-level information within the STG.

In the fourth chapter, we synthesize prior work and the empirical findings in this dissertation into an updated model of speech processing in the human STG. We propose that the STG performs multi-scale, recurrent computations to link acoustic-phonetic features with language-specific structures, like words. This model emphasizes the STG's role as a critical interface between auditory and language systems, ultimately enabling the remarkable human capacity for speech understanding.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ANOVA	analysis of variance
AUC	area under curve
ECoG	electrocorticography
fMRI	functional magnetic resonance imaging
F0	fundamental frequency
F1	first formant frequency
F2	second formant frequency
HFA	high frequency activity
HG	Heschl's gyrus
HGA	high gamma activity
LME	linear mixed effects model
MTG	middle temporal gyrus
PAC	primary auditory cortex
PCA	principal component analysis
STG	superior temporal gyrus
STS	superior temporal sulcus
TRF	temporal receptive field
VOT	voice onset time

Chapter 1: Introduction

Speech perception relies on a set of transformations that convert complex, acoustic signals into discrete and interpretable linguistic units. It is remarkable that humans are able to so easily recognize and respond to speech input, despite the fact that few physical properties of the acoustic signal can accurately identify a speaker's intended message. That is, speech sounds and the content they correspond to vary substantially depending on language, temporal and phonological context, speaker identity, speech rate, and more (Ladefoged & Johnson, 2014; Liberman et al., 1952, 1967). Determining the mechanisms by which the human brain overcomes these challenges to reliably comprehend the speech signal has been the subject of spirited debate over the past century and, following in this tradition, is the focus of the current dissertation.

Speech sounds can be described by acoustic properties of the physical sound wave, or articulatory properties that relate to the way in which the sound was formed in the vocal tract. *Phonological computation* refers to the processes that translate these features into meaningful elements of language. These computations are often *nonlinear* in that they result in perceptual correlates that are invariant to physical properties of acoustic input. They are also *dynamic*; speech representations vary as a function of time, depending heavily on adjacent word and sentence-level context. These key qualities of phonological computation give rise to a distinct form of auditory perception that facilitates speech processing.

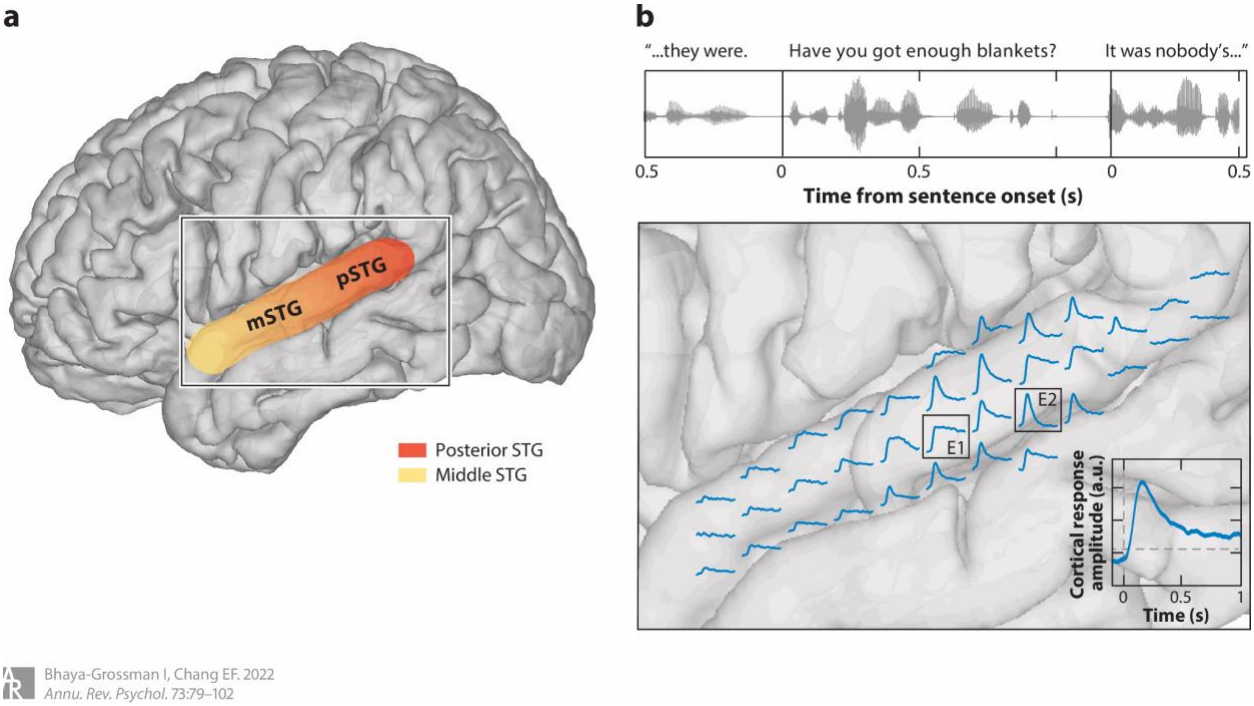
Classic anatomical lesion studies have long implicated the posterior region of the *superior temporal gyrus (STG)* (**Fig. 1.1A**) as a critical locus for speech perception and language comprehension (Geschwind, 1970; Wernicke, 1874). The STG contains part of the non-primary auditory cortex and encompasses the cytoarchitectonically defined Brodmann area 22, and lateral aspects of Brodmann areas 41 and 42. Deficits in auditory word comprehension are associated with focal injury to the dominant (usually left)

hemisphere STG (Hillis et al., 2017). Transient functional lesions of the same area, using electrocortical stimulation, demonstrate acute behavioral impairments in phonemic and word perception (Boatman, 2004; Boatman et al., 2000; Leonard et al., 2019; Roux et al., 2015), and sentence comprehension (Matsumoto et al., 2011), without clear effects on acoustic pure tone discrimination. Converging evidence from neuroimaging studies also suggest that the anterior and posterior regions of the STG occupy a special role in the neural analysis of speech. The STG is more strongly responsive to speech over non-speech sounds (Benson et al., 2001; Binder et al., 2000; Humphries et al., 2014; Zatorre et al., 1992), as well as intelligible over non-intelligible speech (Overath et al., 2015). In contrast, the primary auditory cortex does not appear to exhibit similar selectivity or specialization for speech sounds (Hamilton et al., 2020).

While numerous lesion and functional imaging studies have suggested that the STG plays a crucial role in phonological processing, they do not reveal the cortical mechanisms that underlie it. It thus remains unclear how the spatial, temporal, and spectral properties of neural activity, as a neural code, represent specific properties of speech. With the advancing technologies that allow for dynamic measurement of brain activity at high resolutions, researchers are able to address what elements of speech are represented by neural activity and how.

In this dissertation, we describe recent discoveries that contribute to our mechanistic understanding of speech processing enabled by high-resolution electrophysiology. Direct neurophysiological recordings allow for high spatial and temporal resolution at the millimeter and millisecond scale. These parameters are critical because adjacent recording electrodes even a few millimeters apart show highly distinct tuning (**Fig. 1.1B**) and speech units occur on the order of tens of milliseconds. The ability to simultaneously record neural activity from a densely sampled cortical region allows researchers to characterize patterns of activity at multiple scales of resolution. This is important for understanding speech representation at single electrodes and the emergent properties that arise from responses spread across the population of

electrodes. In particular, we focus on the intracranial electrocorticogram (ECoG) recording of electrical activity directly from the cortical surface (Chang, 2015; Parvizi & Kastner, 2018). The high frequency band of the ECoG signal, also called high gamma activity (50-200 Hz, HGA, also referred to as high frequency activity, HFA), recorded from a single electrode contact is thought to reflect highly local neuronal spiking and dendritic activity from thousands of underlying neurons (Crone et al., 2001; Leszczyński et al., 2020; Ray & Maunsell, 2011; Steinschneider et al., 2008; Towle et al., 2008). *Local neural populations* refer to the activity recorded at a single electrode contact.



Bhaya-Grossman I, Chang EF. 2022
Annu. Rev. Psychol. 73:79–102

Figure 1.1. ECoG recording enables high resolution recording of neural activity in nonprimary auditory cortex.
A. This panel illustrates the anatomical boundary of the STG. Color gradient represents the functionally differentiated posterior and middle regions of the STG (Hamilton et al., 2020; Ozker et al., 2017; Yi et al., 2019a). **B.** Example sentences from the TIMIT corpus shown in the top panel, where time from the most recent sentence onset is marked (Garofolo et al., 1993). Single electrode activity is aligned to the onset of speech and averaged across all corpus sentences. The cortical responses to the speech stimulus across the STG reveal a wide array of response profiles, even between responses recorded 4-8 millimeters apart (slow sustained cortical response for electrode labeled E1, rapid response to sentence onset for electrode labeled E2).

STG Encoding of Phonological Units

Phonological units are elemental speech sounds organized within a linguistic hierarchy. At the lowest level of this hierarchy is the phonetic feature. *Phonetic features* are defined by either specific articulatory or acoustic properties. The simultaneous presence of several phonetic features uniquely identifies minimally contrastive units of speech that make up the inventory of sounds in a given language, *phonemes (Halle & Chomsky, 1968)*. For instance, a /b/ is a sound produced with the features of 'voicing' (vibrating vocal cords), 'bilabial' (movement of the two lips), and 'plosive' (the burst of air released after an articulatory closure). Alone, any one of these phonetic features is not linguistically meaningful. In combination, however, they give rise to all consonant and vowel sounds.

Consonants and vowels are classes of speech sounds with key acoustic differences and can be described by distinct sets of articulatory features. *Consonants* are speech sounds in which the airstream through the vocal tract is partly obstructed. They are characterized by the location ('place of articulation') and manner in which airflow through the mouth is constricted by articulators ('manner of articulation'). *Vowels* are characterized by an open and unobstructed configuration of the upper vocal cavity, which is shaped by the location of the tongue in the mouth (high/low, front/back) and the extent to which the lips are rounded during production of the vowel sound.

In natural speech, consonant and vowel sounds occur in highly organized repeated motifs that define a *syllable*. The components of a syllable are the vowel nucleus, which can be optionally preceded by an initial position consonant(s) and/or followed by a final position consonant(s). The sequencing of consonants and vowels in a syllable is also organized by the general crescendo-decrescendo trajectory of sound loudness according to principles of *sonority*. Because vowels are produced by a relatively more open vocal tract configuration, they are louder and possess peak sonority. When syllables are strung together in longer utterances, this alternating structure of consonants and vowels contributes to the rhythm

of speech. The peaks and valleys of these loudness modulations make up the *amplitude envelope* of speech.

Linguists have long sought to pin down a single unit of speech perception that may provide insight as to how speech is mentally represented (e.g. phonemic or syllabic segments); many perceptual and acoustically informed candidates have been proposed. Psycholinguistic studies attempting to address this question have yielded largely inconsistent results (Healy & Cutting, 1976; Massaro, 1974; Savin & Bever, 1970). Further studies have suggested that the basic perceptual units of speech depend on age and native language experience (Cutler et al., 1986; Cutler & Otake, 1994; Nittrouer et al., 2000).

Relatedly, neuroscientists have devoted substantial work to understanding whether there exists in the brain a clear representation of a single unit of speech perception. ECoG recordings have revealed that local neural populations in the STG selectively represent acoustic-phonetic features, but not single phonemes or syllables (Mesgarani et al., 2014). Neural populations in the STG are particularly sensitive to the manner of articulation, an articulatory feature strongly differentiates vowels and consonants, and consonant categories (e.g. fricative vs plosive). Tuning to acoustic-phonetic features – rather than phonemes – is observed at both single ECoG electrode contacts (thousands of neurons), and single neurons in the STG (Chan et al., 2014; Lakertz et al., 2021).

Acoustic-phonetic features precisely characterize the human phonetic inventory, accounting for the similarities observed across speech sounds in different languages (Halle & Chomsky, 1968). Together, the distribution of local responses throughout STG to phonetic features contains the information necessary to decode abstract phonemic categories across languages. Different phonological units, such as the acoustic-phonetic feature and the phoneme, may thus be realized at two different spatial scales (the local and population ensemble, respectively).

Categorization and Normalization

Speech perception requires that complex acoustic signals with continuous spectral and temporal detail be reduced to a finite set of phonological units that make up words. In this way, it can be understood as mapping a set of acoustically distinct sounds to the same language-specific, phonological class, *categorization* (Holt & Lotto, 2010). Not only are these mappings many-to-one, but they are also one-to-many as the same speech sound can map to different phonemic classes depending on the phonological context and the physiology of the speaker (Johnson, 2008; Mann, 1980; Mann & Repp, 1980). The process by which a listener's representation of a speech sound becomes invariant to speaker identity is called *speaker normalization* (Johnson, 2008). Categorization and normalization are both nonlinear operations, in that they result in representations of speech sounds that are predictably distinct from the physical stimulus.

Categorical perception of speech refers to a behavioral phenomenon where sounds are discriminable across different sound categories, but not within a category. This too is a fundamentally nonlinear perceptual property: sensitivity to acoustic change depends on how similar a sound is to a defined sound category (Tuller et al., 2011). This can be observed via *psychometric function*, a tool to describe nonlinearities in speech perception, which allows researchers to quantify the extent to which incremental changes to an acoustic signal alter perceptual experience.

Categorization, categorical perception, and speaker normalization are critical features of speech perception that allow for stable perceptual experiences across a wide range of vocalized sounds. We speculate that categorization of speech could be implemented neuro-physiologically in several distinct ways. One possibility is that local neural populations selectively respond to a single phoneme, suppressing speaker and context-dependent acoustic differences that exist within a phonemic category. Alternatively, local neural populations may respond selectively to specific acoustic-phonetic features

while patterns of activity across multiple local populations exhibit the higher-order representation of phonemes. Current research on the STG, including work presented in this dissertation, supports the latter theory; while local neural populations on their own reliably encode acoustic-phonetic features or speaker normalized prosodic features, the activity distributed across populations contains information necessary to give rise to a categorical representation. We review the evidence for this claim, focusing our discussion on the encoding of consonant and vowel categories.

Consonants

In a classic example of categorical perception, Liberman et al showed that when participants are asked to identify synthesized speech sounds that contain smoothly and continuously changing spectral content among the plosive sound categories (/ba/, /ga/, and /da/), their perception of phoneme category changes abruptly rather than gradually (Liberman et al., 1957). By replicating this experimental paradigm with simultaneous ECoG recording, researchers were able to determine whether the brain faithfully represents physical stimulus characteristics, or the nonlinear patterns that correspond to the listener's perceptual experience (Chang et al., 2010). Although the spectral content of the acoustic stimulus was altered linearly (forming a graded acoustic continuum), participants reported categorical perception of three distinct phonemes, suggesting non-linear perceptual processes. Consistent with the behavioral results, neural activity recorded from the STG patterned in the same nonlinear and categorical manner. That is, pooled neural responses were more dissimilar when compared between categories than when compared within. The underlying dimensions driving the neural activity were the onset frequency of the second spectral peak (labeled F2 in **Fig. 1.2**) and the magnitude of the spectral transition, which corresponds to lingual articulatory movements during speech production (Lindblom & Sussman, 2012). We use this example to speculate as to how a combination of linear and nonlinear responses to acoustic-phonetic features could represent phonemic categories through a spatial code (**Fig. 1.2**).

Temporal cues can also be perceived as categorical. Voice-onset time (VOT) is defined as the length of time between a stop consonant release (the burst) and the onset of voicing, or vocal fold vibration. In English, listeners are able to exploit this temporal cue in order to discriminate between voiced and voiceless stop consonants, such as the difference between /ba/ and /pa/ (Liberman et al., 1958; Lisker & Abramson, 1964). When the stop consonants VOT is simultaneous, it is perceived as voiced (/b/, /d/, /g/), whereas a longer VOT (~50ms) is perceived as voiceless (/p/, /t/, /k/). When listening to stimuli with incremental linearly-spaced VOT changes, participants are highly sensitive to changes across a category boundary, and generally insensitive to stimulus changes within a category.

Brain areas including the STG are sensitive to differences in VOT (Blumstein et al., 2005; N. P. Fox et al., 2017, 2020; Mesgarani et al., 2014; Steinschneider et al., 1999, 2011). Functional magnetic resonance imaging (fMRI) studies of VOT perception reveal that the activation of left posterior STG is sensitive to how close a VOT stimulus is to a known category boundary (Blumstein et al., 2005; Hutchison et al., 2008), suggesting that temporal speech cues are also categorically encoded. Until recently it was unclear whether temporal information that discriminates phoneme categories (e.g., /ba/ vs /pa/) is encoded in the amplitude (similar to spectral cues), or the relative timing of the neural response. Evidence from a recent ECoG experiment demonstrates that local neural populations on middle and posterior bilateral STG encode the temporal VOT cue in the amplitude of the response, with focal neural populations categorically preferring either voiced or voiceless sounds (N. P. Fox et al., 2020). Notably, neural populations that selectively respond to one category still show sensitivity to differences among VOT values, but only within the preferred category. This study demonstrates that a predominantly temporal acoustic cue can be mapped onto spatially distinct neural populations in the STG selective for distinct categories of voicing (Liberman et al., 1958; Lisker & Abramson, 1964).

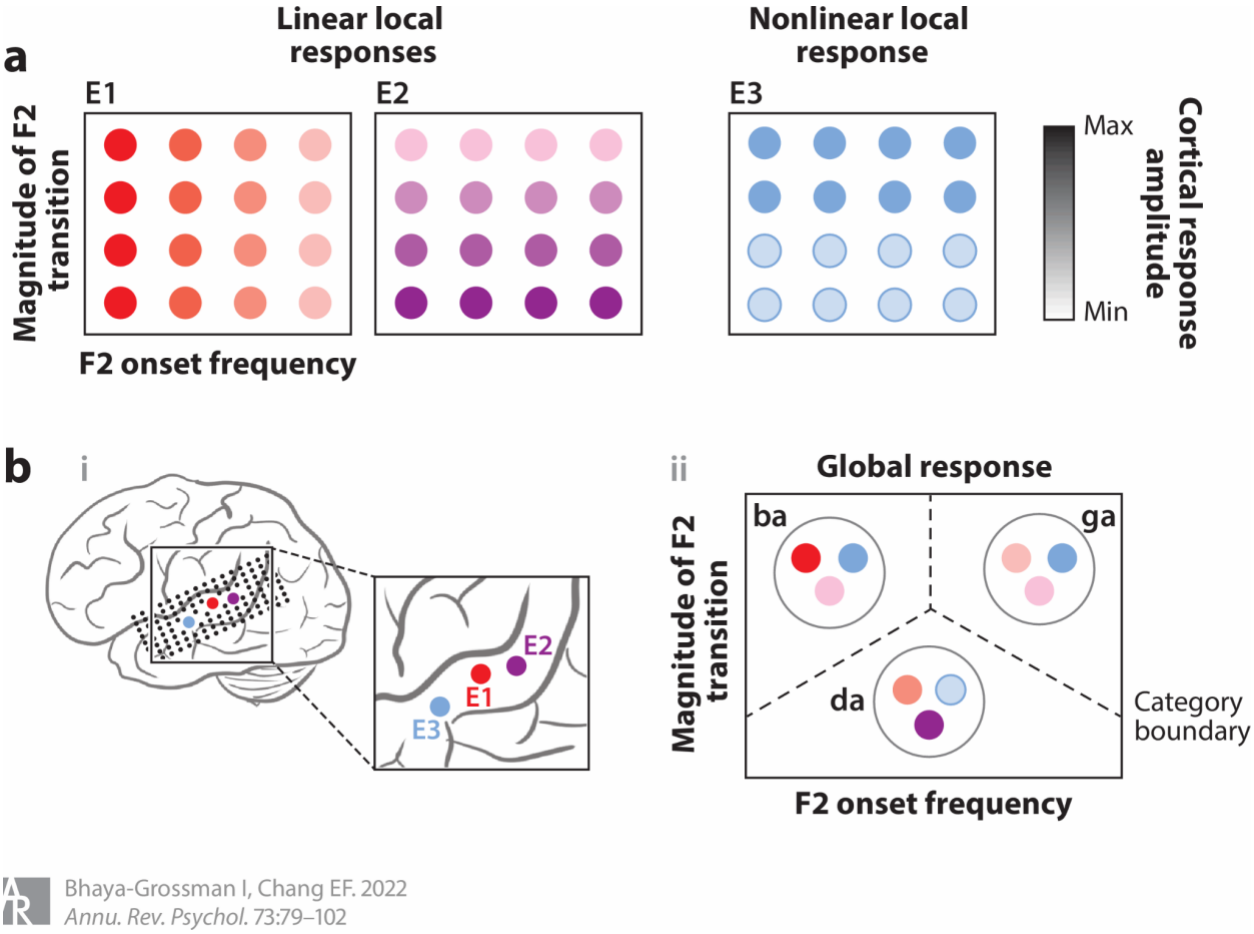


Figure 1.2. Patterns of activity across the STG allow for the categorization of phonological units. **A.** Single electrodes (selected electrodes illustrated as colored circles) respond to incremental acoustic change, showing graded linear or abrupt nonlinear monotonic tuning to certain spectral features (e.g. F2 onset frequency or magnitude of F2 transition). Single electrode responses do not prefer a phonemic category, but are tuned more generally to auditory cues such as the example acoustic-phonetic features shown on in this panel. **B.** Schematic depiction of the categorical neural encoding of speech sounds, derived from patterns of activity across the population. Information distributed across electrodes (illustrated on the brain as the set of colored circles) can be used to determine the phonemic category of presented speech sounds and perceptual experience of the listener. Further, overlapping functionality in the neural code (blue, purple responses) may be important for retaining within category sensitivities.

Vowels

Vowel sounds are generated by the different shapes of the oral cavity, giving rise to resonances at particular frequencies, known as formants. The formants are defined by the way in which air resonates in the vocal tract and are therefore dependent on vocal tract shape. The first formant (F1) corresponds to resonance at lower frequencies and the second formant (F2) at higher frequencies. Formants are roughly

correlated with the articulatory dimensions that characterize vowel sounds (the closeness of the tongue to the roof of the mouth to F1, height, the degree of tongue front-backness to F2, backness). Vowels are not perceived as categorically as consonants (Pisoni, 1973), however the perceptual vowel space is warped (nonlinearly) towards category prototypes, known as the *perceptual magnet effect (Kuhl, 1991; Kuhl et al., 1992)*. In other words, two speech sounds that are both similar to a single vowel prototype are more difficult to distinguish than those that correspond closely to two separate vowel prototypes. As a result of this effect, vowel sounds are largely perceived discretely as phonemes, and most listeners are unaware of the underlying two-dimensional formant space.

Importantly, the range of formant values that map to a particular vowel category is variable and dependent on speaker characteristics (e.g. vocal tract length) (Ladefoged & Johnson, 2014). Speakers with longer vocal tracts produce resonant frequencies that are on average lower than speakers with shorter vocal tracts, these resonant frequencies include F1 and F2. In contrast, speakers with shorter vocal tracts produce a higher range of formant frequencies. Vowel identity cannot be identified based on absolute formant values alone, rather listeners must consider *relative formant frequencies*, or how the formant frequencies for a vowel sound compare to those for other vowels produced by the same speaker (Ladefoged & Broadbent, 1957). As a result, reliable vowel discrimination and categorization processes rely on *speaker normalization*, the process by which listener's representations of relevant linguistic content discards irrelevant speaker-dependent properties (Johnson, 2008).

There is evidence that vowel category information, beyond physical acoustic features, is represented in the STG and surrounding regions of the auditory cortex. Similar to the neural encoding of consonants, STG neural populations do not respond preferentially to a single vowel category nor a narrow-band frequency range (Mesgarani et al., 2014). Rather, STG cortical responses to vowel sounds rely on complex spectral integration, representing the first two formants in combination (see Chapter 2). Alone, these responses are unable to fully discriminate between speech sounds corresponding to distinct vowel

categories. However, the population distribution of single electrode responses contains information necessary to decode vowel identity. The critical question then becomes: do neural populations in the STG represent speaker identity and absolute formant information separately? Or does the neural activity recorded from local populations additionally represent *speaker normalized* formants that are nonlinearly transformed for the purposes of vowel identification? In an ECoG study addressing the question of speaker normalization, subjects were asked to categorize synthesized vowel sounds where carrier sentences that preceded the synthesized vowels were produced by either a low or high pitch speaker. In order to test the effects of speaker normalization, experimenters incrementally increased the absolute value of F1 in the target vowel sound. Comparing psychometric and corresponding neurometric functions derived from STG activity revealed initial evidence that behavioral and neural responses to vowel sounds represented the speaker-normalized formant value (Sjerps et al., 2019). These findings are extended in Chapter 2 and together they demonstrate that the neural encoding of vowel sounds in the STG is normalized for speaker type.

Speech dynamics

Natural speech generally consists of sequences, and the structure of the speech signal as it unfolds across time is critical to communication. Here we consider how speech sounds in the STG are dynamically represented over the course of syllables, words, phrases, and sentences.

One might assume that the speech signal can be treated as "beads on a string", or linear sequences of context-invariant linguistic units. That is, the perception of a single speech element, like a phoneme, is independent of the elements surrounding it in time. If this is the case, there is no need for segmentation as discrete groupings of syllables and words are made explicit (Marslen-Wilson & Welsh, 1978). However, many speech scientists have demonstrated that this may be an unfounded assumption: critical temporal cues that occur at the scale of the syllable, phrase, and sentence can greatly improve speech comprehension and facilitate the binding of phonetic sequences (Shannon et al., 1995; Zeng et al., 2005).

Temporal landmarks for syllable onset timing

As described previously, the syllabic modulations of the speech envelope are heavily influenced by fluctuations in the aperture of the vocal tract during speaking. In particular, open configurations are associated with the greatest loudness or *sonority*, and are most pronounced during vowel articulation. Conversely, the lowest sonority sounds are plosive and fricative consonants created with oral constriction. As a result, the phases of the speech envelope correlate to syllable timing: rhythmic quasi-periodic fluctuations in the amplitude of the envelope are aligned to the alternating sequences of consonant and vowel sounds that make up syllables. Early psychophysics work has demonstrated that the duration and magnitude of change in the speech envelope is critical for the perception of phonological ordering within a syllable (Chistovich, 1980). This may explain why the speech envelope is necessary for speech intelligibility (Drullman, 1995), representing the underlying rhythms and syllabic stress patterns of speech. However, by itself it is not sufficient for speech comprehension, especially in the presence of noise (Hopkins & Moore, 2009; Lorenzi et al., 2006; Moon & Hong, 2014).

Extensive literature shows that brain activity continuously tracks the speech envelope (Ahissar et al., 2001; Drennan & Lalor, 2019; Liégeois-Chauvel et al., 2004; Nourski et al., 2009; Peelle & Davis, 2012). Previous neurophysiology studies have discovered subcortical neural responses that integrate across spectral frequencies to detect temporal onsets (Heil & Neubauer, 2001). Intracranial recording of the STG has allowed researchers to pinpoint which local cortical populations are involved and which envelope features most saliently contribute to the maintenance of the neural representation. A recent ECoG study revealed that HGA recorded from local speech-responsive neural populations in the middle STG respond selectively to the discrete temporal landmarks of *peakRate*, or the time points at which there is maximal change in envelope amplitude (Oganian & Chang, 2019).

Although peakRate is primarily an acoustically derived temporal landmark, it has critical implications for the extraction of syllabic information. In English, for example, peakRate events closely align with vowel onsets, marking the transition from the syllabic onset to nucleus. Further, the magnitude of peakRate events, or steepness of the envelope change, cues syllabic stress. As a result, peakRate events operate as key landmarks in the speech signal, around which syllables are organized. Neural responses in the STG reflect both the timing and magnitude of the peakRate event, effectively encoding information about syllable structure, stress and speech rate (Oganian & Chang, 2019; Ohala, 1975). These findings suggest that at the local neural population scale, there exists a discrete event-based neural encoding of syllable timing and stress as opposed to one that continuously tracks the envelope amplitude. Syllable information, including both the timing, stress, and acoustic-phonetic content, can be thought of as distributed across several functionally-distinct local neural populations.

This work demonstrates that there exists an efficient neural code for extracting key temporal landmarks in the speech signal such as peakRate. We speculate that these response types provide the temporal context information critical for speech processing.

Word-level contextual dynamics

A highly illustrative example of the brain's ability to use linguistic knowledge, specifically word-level context, to understand noisy acoustic input is the case of perceptual restoration, where speech segments are entirely unavailable in the input (Remez et al., 1981; Warren, 1970; Warren & Sherman, 1974). When a burst of noise (e.g. a cough or white noise) completely obscures a phoneme segment, listeners perceive the replaced sound and are generally unable to determine at which point in the word or phrase they heard the noise (Warren, 1970). This is known as *phoneme restoration*, since the masked noise does not impair intelligibility and the removed phoneme is perceptually restored (Samuel, 1987). Here, we review evidence that the STG exhibits properties of phonological restoration and argue that local neural populations employ dynamic and contextual encoding mechanisms to overcome perceptual challenges

early in speech processing. Importantly, the mechanisms that facilitate robust speech perception are not specific to adverse listening conditions; rather, they reflect fundamental aspects of speech perception more generally.

A key question is whether this phenomenon reflects real-time modulation of perceptual representations (i.e., lexical and contextual knowledge changes how the physical sound is interpreted by the perceptual system; (McClelland & Elman, 1986), or a post-hoc interpretation of the speech segment based on similar contextual cues (Norris et al., 2000). In a recent study, listeners were presented with stimuli where a critical phonemic segment was completely removed and replaced with noise. The possible restorations in English were compatible with only two distinct words, e.g. "fa*c*tor" and "fa*s*ter", which participants reported hearing across repeated presentations of the sounds. ECoG responses on bilateral STG showed that neural populations tuned to specific acoustic-phonetic features (e.g., /k/ vs /s/) also showed responses to the completely ambiguous noise that were consistent with what they reported hearing on each trial. The authors used population neural responses across electrodes to reconstruct the acoustic spectrogram of the noise, and showed that it contained more high-frequency power when the noise was perceived as a fricative /s/ than when it was perceived as a plosive /k/. Crucially, these representations were evident at the same latency as when listeners were presented with unambiguous fricatives and plosives, demonstrating real-time restoration (Leonard et al., 2016).

Furthermore, the authors found that neural activity in the STG and the left inferior frontal cortex could be used to predict what listeners would report hearing. Remarkably, this activity was most robust ~300ms before the onset of the noise. The fact that listeners only reported hearing sounds that were consistent with real English words (e.g., "faster" and "factor", not "fanter") suggests that these predictive modulatory signals may reflect lexical biases embedded within the STG dynamics and possibly influences from other language brain areas. Together, this represents another instance in which the patterns of neural activity in

the STG reflect the perceptual experiences of the listener in addition to features of the physical acoustic signal.

In addition to noisy environments, speech stimuli in which fine grained spectral or temporal details of acoustic signals are scrambled or removed, is used to study robustness in speech perception. Human subjects are able to almost perfectly identify linguistic content in the presence of noisy or degraded spectral information and do so by relying strongly on temporal context (Shannon et al., 1995). One commonly used degraded stimulus is sine wave speech (SWS), where the formants of natural speech are replaced with pure tones and the rest of the spectral detail in the spectrogram is removed(Remez et al., 1981). In most cases, listeners cannot understand SWS, and often do not even recognize it as speech. However, after hearing the original unfiltered version, the SWS sounds are suddenly intelligible, a phenomenon known as a perceptual pop-out effect. Several fMRI studies have attempted to determine the properties of SWS representation in the STG and have found that the superior temporal sulcus (STS), directly adjacent to the STG, as well as surrounding posterior STG regions shows higher neural activation when SWS is intelligible compared to when it is not (Benson et al., 2006; Dehaene-Lambertz et al., 2005; Möttönen et al., 2006). While greater neural activation indicates that these localized regions are selectively performing computations on intelligible speech, it is difficult to determine the content of these computations from temporally coarse activation alone.

In a recent study, Holdgraf et al. (2016) performed a similar experiment with filtered speech stimuli (in which spectral or temporal modulations are removed from the speech signal) using ECoG recordings. Subjects presented with the filtered speech signal before and after having listened to the corresponding unfiltered speech reported an increase in intelligibility after having heard the unfiltered speech signal. Rather than a simple change in the magnitude of neural activity in the STG, the authors found that the neural response to filtered speech once it is intelligible more closely resembles the activity elicited in the unfiltered condition (Holdgraf et al., 2016). Similar to the perceptual restoration effect described above,

patterns of neural activity in the STG in response to intelligible speech are warped, such that the phonologically relevant spectrotemporal features of sound are amplified. These results are consistent with the idea of a "speech mode" in which perceptual and neural systems actively engage speech knowledge that enables robust perception (Dehaene-Lambertz et al., 2005). Of note, other ECoG studies in which subjects were presented SWS or vocoded speech, did not find enhanced STG responses dependent on intelligibility, although other areas including frontal cortex did seem to be sensitive to this difference (Khoshkhoo et al., 2018; Nourski et al., 2019). These inconsistent results may be due to a number of differences in experimental design and analysis, and raise a natural question, to what extent and in which contexts is neural activity in the STG preferentially modulated by intelligibility and language-specific comprehension?

Finally, Chapter 3 of this dissertation provides evidence that language-specific knowledge is integrated into neural responses to speech in STG particularly at the word level, findings corroborated by (Zhang et al., 2025)). We therefore conclude that word-level knowledge is dynamically integrated into the phonological representation of speech in this higher-order auditory area. Natural speech corpora as well as controlled experimental paradigms expose the mechanisms by which word level information is combined with acoustic, phonetic, and syllabic speech representations. In Chapter 4, we discuss the implications of these findings for models of speech processing, and determine that the neural responses within the STG are strongly modulated by linguistic knowledge in relevant speech contexts.

Chapter 2: Vowel and formant representation in the human temporal cortex

Abstract

Vowels, a fundamental component of human speech across all languages, are cued acoustically by formants, resonance frequencies of the vocal tract shape during speaking. An outstanding question in neurolinguistics is how formants are processed neurally during speech perception. To address this, we collected high-density intracranial recordings from the human speech cortex on the superior temporal gyrus (STG) while participants listened to continuous speech. We found that two-dimensional receptive fields based on the first two formants provided the best characterization of vowel sound representation. Neural activity at single sites was highly selective for zones in this formant space. Furthermore, formant tuning is adjusted dynamically for speaker-specific spectral context. However, the entire population of formant-encoding sites was required to accurately decode single vowels. Overall, our results reveal that complex acoustic tuning in the two-dimensional formant space underlies local vowel representations in STG. As a population code, this gives rise to phonological vowel perception.

Introduction

Vowels are a significant component of all the world's languages, and they play a critical role in our ability to comprehend speech (Fogerty & Humes, 2012; Fogerty & Kewley-Port, 2009). Vowel sounds are produced when the vocal fold vibration is unobstructed, allowing a clear passage of air through the mouth, shaped by different positions of the jaw, tongue, and lips. For instance, the vowel sound /i/ (as in "heed") is produced with the tongue close to the front of the mouth, whereas /a/ (as in "had") is produced with the tongue further back. These articulatory positions create distinct vowel sounds, distinguished acoustically by the value of the lowest two vocal tract resonance frequencies, which are known as the first and second formants (F1 and F2). Formants are considered the primary acoustic cues to vowel identity (Ladefoged & Johnson, 2014).

Sounds with different vowel identities can be very similar in absolute formant values when produced by different speakers. For example, /o/ (as in "hoed") produced by a speaker with a long vocal tract (and thus low voice) could have the same absolute formant values as /u/ (as in "who'd") produced by a speaker with a short vocal tract (and thus high voice). As a result, the accurate mapping of formant values to vowel identity requires a normalization operation that computes formant frequencies relative to the speaker's voice (Johnson & Sjerps, 2021). In sum, rapid and correct mapping of formants to vowel identity relies on both the precise identification of the formant frequencies as well as their interpretation, given the speaker context.

In linguistics, the mental representation of formants, specifically the independence of F1 and F2 and how this representation supports vowel normalization, has been long debated. One fundamental theory suggests that they are independently extracted and represented, mapping individual vowel sound segments onto coordinates in the two-dimensional F1-F2 space (Nearey, 1989; Strange, 1989b). To account for the speaker context, proponents of this theory have advocated for an explicit normalization of both F1 and F2 values using features such as speaker pitch or higher formants. Alternatively, it has been

suggested that vowel identification relies on the *relationship* between formants, such as the one-dimensional distance between F1 and F2 (Syrdal & Gopal, 1986). This theory of vowel perception is bolstered by the fact that the relationship between F1 and F2 is more consistent across speakers than the independent absolute values, thus implicitly accounting for the speaker context (J. D. Miller, 1989; Peterson, 1961). Examining neural responses to vowel sounds in the human speech cortex provides us with a unique opportunity to dissociate these theoretical models and elucidate the mechanisms that underlie speaker-normalized vowel identification.

Several studies have shown that neural activity in the human auditory cortex is sensitive to vowel sounds and that it discriminates vowel identity in highly controlled acoustic contexts (Bonte et al., 2014; Formisano et al., 2008; Levy & Wilson, 2020). In the primary auditory cortex (PAC), a tonotopically organized area (Moerel et al., 2012) that is activated regardless of whether the presented stimulus is human speech (e.g., pure tones), this sensitivity is driven by narrow tuning to individual formant frequency bands (Hamilton et al., 2021; Khalighinejad et al., 2021). By contrast, in the human speech cortex on the lateral superior temporal gyrus (STG), most of the neural representations are spectrally complex and broadband (Hamilton et al., 2021). Furthermore, the STG shows stronger activity in response to speech and other complex natural sounds (e.g., music (Leaver & Rauschecker, 2010; S. Norman-Haignere et al., 2015) than in response to other sound stimuli, and this activity is more closely reflective of perceptual processing (Bhaya-Grossman & Chang, 2022; Yi et al., 2019a). It thus remains an open question what neural representation in the STG underlies vowel perception in continuous, natural speech (Kohonen & Hari, 1999; Obleser & Eisner, 2009; Rauschecker & Scott, 2009). Specifically, it is not yet clear whether formants are represented separately or in combination or, furthermore, whether this representation is tuned to narrow-band frequencies, possibly centered on single vowel categories (Shestakova et al., 2004), or broad formant ranges.

To address this, we utilized high-density direct intracranial recordings of neural activity from the surface of the human STG. The highly resolved spatial scale afforded by this recording technique was critical, as

neighboring cortical sites that are just a few millimeters apart can differ significantly in their spectral tuning (Chang et al., 2010; N. P. Fox et al., 2020). The high temporal resolution of intracranial recordings allowed us to examine neural responses at the temporal scale of a single vowel sound. We used natural speech stimuli produced by a wide variety of speakers, which allowed us to record neural responses to a large set of vowel sounds, spanning the entire formant space.

With this approach, we addressed four primary research questions. First, we asked how neural responses recorded at single electrodes in the STG were tuned to F1 and F2 in natural, continuous speech. We analyzed two-dimensional formant-receptive fields of neural responses to determine whether neural tuning to F1 and F2 frequency ranges in the human speech cortex is independent and to characterize the mathematical properties of these tuning functions (Bertalmío et al., 2020; Fischer et al., 2009). Second, we asked how vowel information can be extracted from the neural representation of formants using a population decoding approach. Third, we asked how and to what extent formant-receptive fields in the STG are normalized for speaker properties. Finally, we used a controlled set of artificial vowel-like sounds extending beyond the natural vowel formant space with experimentally decorrelated F1 and F2 to definitively test the independence of formant encodings.

We found that neural responses on most single-electrode sites in the STG were jointly tuned to both F1 and F2, resulting in heightened sensitivity to a specific zone within the vowel formant space. This sensitivity was nonlinear and sigmoidal along each of the separate formant dimensions (F1 and F2). However, the location of heightened sensitivity did not coincide with boundaries between single vowels, and we did not find single-electrode sites with selectivity for a single vowel. Rather, single vowels could only be decoded at the population level, when information from differently tuned electrode sites was pooled together. Comparisons between neural responses produced by speakers with different vocal tract lengths showed that electrodes in the STG contain normalized, not absolute, formant representations, distinguishing the STG from the narrow-band frequency tuning in PAC. Although formant tuning on many active electrodes showed inverse tuning to the two formants (e.g., tuned to high F1 and low F2)

when presented with natural speech, decorrelating the formants using a set of artificial vowels revealed a set of electrodes with the same direction of tuning to both formants (e.g., high F1 and high F2). This confirms that there exists a range of formant-encoding types in the STG and that F1 and F2 are neurally represented as coordinates in a two-dimensional formant space rather than as a ratio or distance.

Results

STG cortical activity sensitive to vowel formant differences

In Experiment 1, Spanish monolingual patients (n = 8) listened to naturally produced Spanish sentences (**Fig. 2.1A**) while we recorded neural activity from the lateral surface of the STG using high-density ECoG electrode grids. The Spanish vowel system is well suited to study the vowel representation for multiple reasons. Spanish has only 5 vowel sounds (as opposed to, e.g., up to 20 in some English dialects (Hagiwara, 2004) that span a large range of formant values. Thus, Spanish vowel categories are clearly and easily separated in the acoustic formant space: the median formant values for single vowel instances correspond strongly with their vowel label (**Fig. 2.1B-C**; vowel clustering: median silhouette score = 0.099, permutation test with 500 repetitions; $p < 0.002$). Moreover, although F1 and F2 tend to be inversely correlated across languages, this is less the case for Spanish than for English (Spanish $r = -0.01$, $p = 0.31$ in our stimulus set; English $r = -0.21$, $p < 0.0001$ calculated from speech stimuli used in prior studies, e.g., Mesgarani et al.).

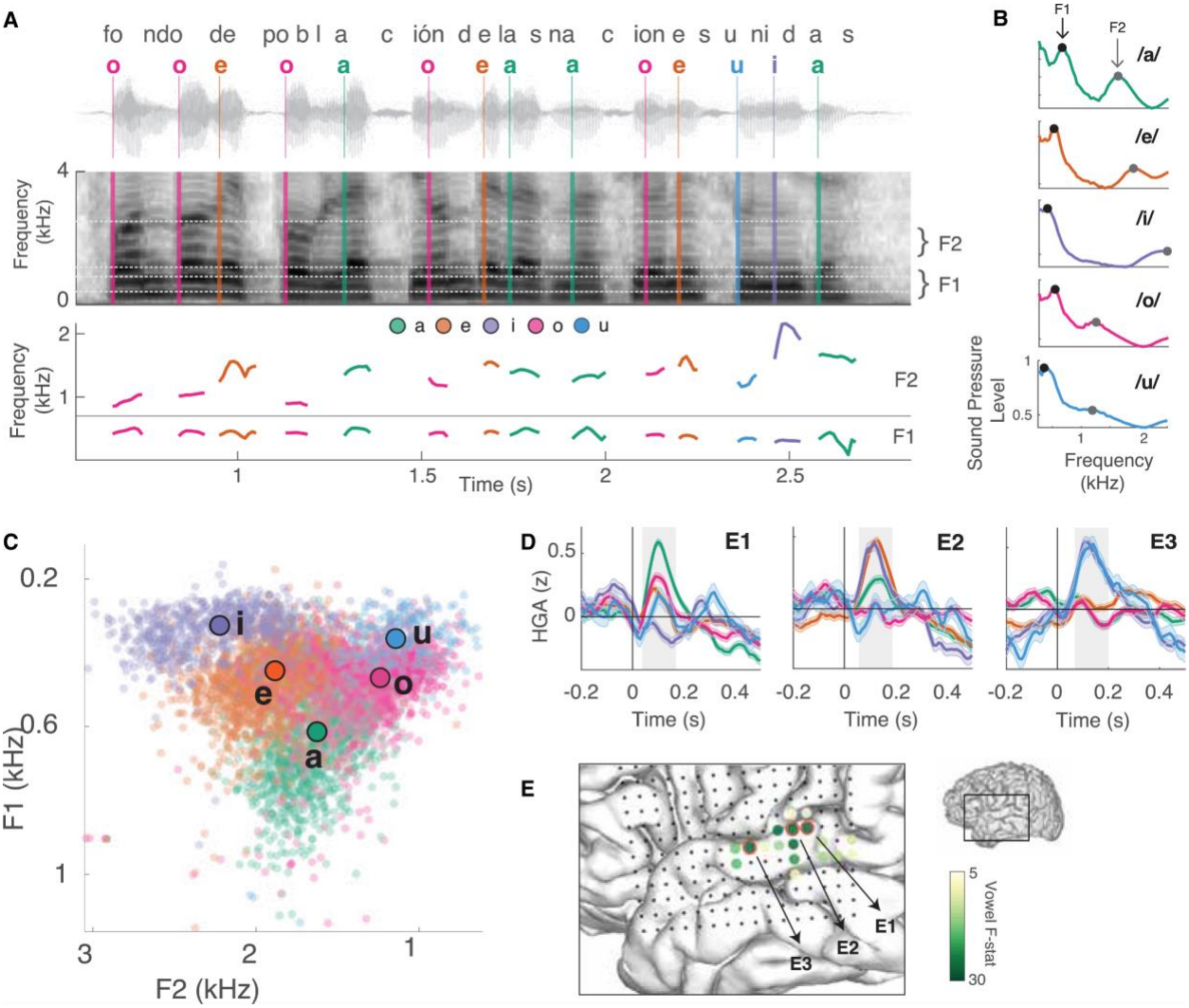


Figure 2.1. Neural activity on single electrodes in bilateral human STG is sensitive to vowels. **A.** Example Spanish stimulus sentence waveform (top), spectrogram (middle), and extracted formant trajectories (bottom). Vertical lines mark vowel onsets, colored by vowel identity. **B.** Median frequency spectrum for a single speaker per vowel. Formant spectral peaks are marked by filled circles. **C.** Median F1 and F2 values across all vowel instances in our stimulus set. **D.** Differential average HGA responses to single vowel categories on three example STG electrodes. Error bars indicate standard error of the mean, gray shaded area marks time window of averaging for electrode selection. **E.** Vowel discriminative electrodes for a single participant. Gray circles demarcate all grid electrodes, example electrodes from D are marked in red.

In our analyses, we focused on evoked response amplitudes in the high gamma range (HGA). We found that evoked HGA responses on a subset of STG electrodes discriminated between vowel categories (n = 116 of 359 speech responsive, range: 2–26 per participant, one-way peak F-statistic across vowel categories > 5, **Fig. 2.1E**). Responses peaked at about 100–150 ms post vowel onset, with different

response magnitudes corresponding to different vowels. For example, each of the three prototypical example electrodes E1–E3 (**Fig. 2.1D**) exhibited its strongest response for a different vowel, with graded response magnitudes for all other vowels.

Notably, none of these electrode responses shows a preference for a single vowel. That is, we did not observe instances where there was selectivity to a single vowel and no response to all other vowel sounds. All subsequent analyses included electrodes that discriminated between vowels.

Non-linear monotonic tuning to vowel formant frequencies

We first asked how vowel formants are represented by vowel-discriminating populations. Specifically, we evaluated the following three alternative hypotheses: narrow-band nonlinear frequency tuning centered on vowel categories (**Fig. 2.2A**, left panel), linear monotonic encoding of an entire vowel formant range (**Fig. 2.2A**, middle panel), or nonlinear monotonic encoding of formants within a limited dynamic range (**Fig. 2.2A**, right panel). In the case of narrow-band frequency tuning, we expect neural populations to preferentially respond to a narrow range of formant frequencies, as is typical for frequency tuning in primary auditory cortices (Bitterman et al., 2008; Hamilton et al., 2021). This model implies that the maximal neural response could be located in the center of the vowel's formant range. By contrast, in the case of monotonic formant encoding, we expect neural responses to increase across the range of possible formant values, with the maximal response located at the edges of the formant range. Although linear encoding implies equal sensitivity to formant differences across the entire range, nonlinear encoding would result in a heightened sensitivity to a narrower range of formant frequencies and little to no sensitivity to frequencies outside of this range.

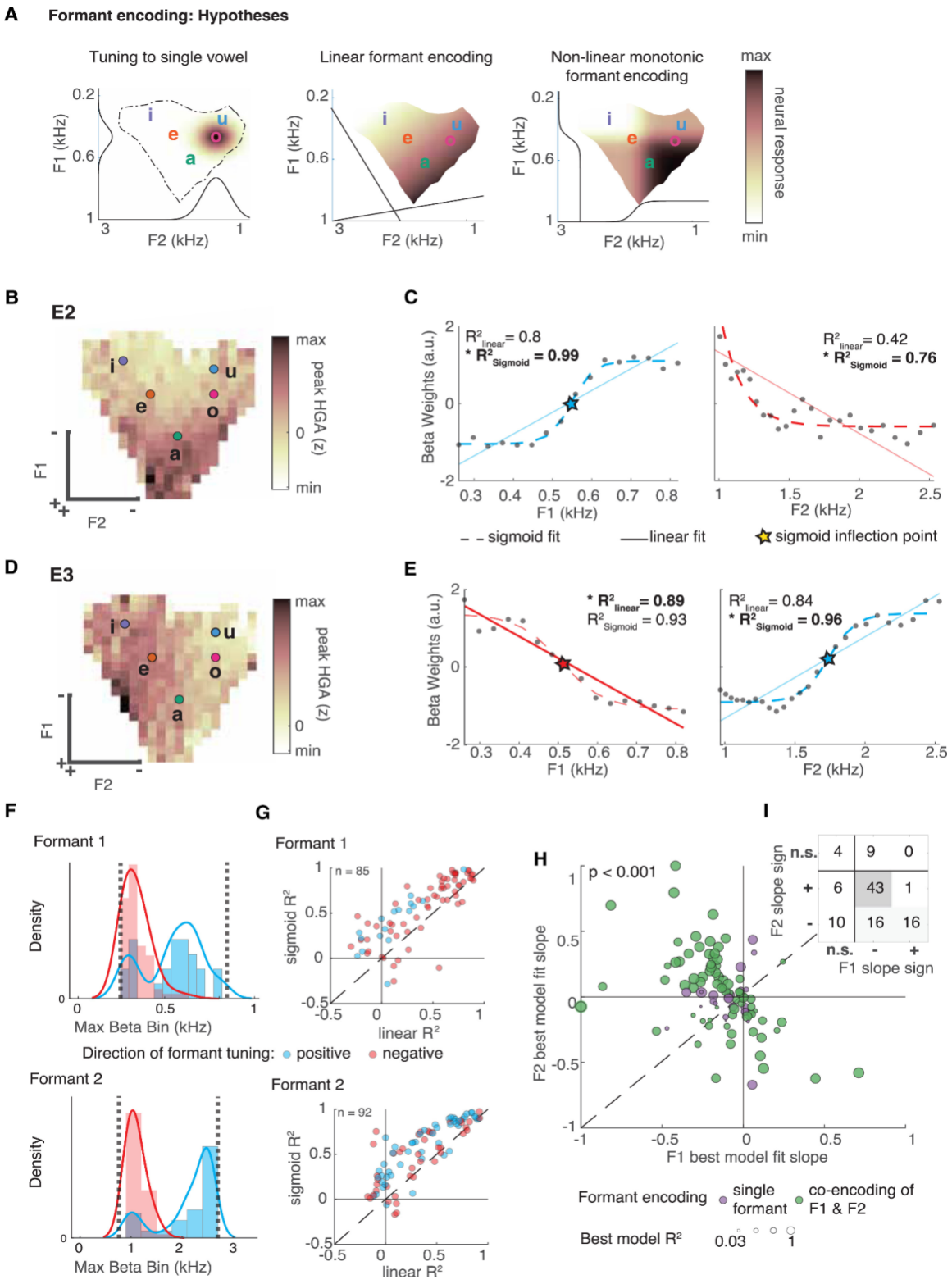


Figure 2.2. Nonlinear monotonic encoding of vowel formant frequencies in human STG. **A.** Three alternative hypotheses for encoding of vowel formants on single electrodes. **B-D.** Example electrodes' formant receptive fields. *(caption continues on next page)*

(caption continued from previous page) **C-D.** Formant tuning curves of the electrodes in B-D. Red: decrease in beta weight as formant values increase; blue: increase in beta weight as formant values increase **F.** Frequency of maximal beta weights in F1 and F2 ranges. **G.** Comparison between linear and sigmoid model fits for F1 (top) and F2 (bottom) **H.** Distribution of slopes for F1 and F2 across vowel-responsive electrodes. P value indicates significance of F1∼F2 slope **I.** Inset showing the total number of electrodes that encode F1 and F2 and the direction of their slopes, corresponding to the number of electrodes in each quadrant in H.

We tested which of these models captured neural activity best separately for every single electrode. First, to distinguish between narrow non-monotonic formant tuning and monotonic encoding, we examined whether neural responses peaked at the edges or in the center of the formant frequency range (**Fig. 2.2F**). Then, to discriminate between linear and nonlinear encoding, we compared linear and sigmoidal models of neural responses using cross-validated R2. We estimated the neural response to vowel formants using feature temporal receptive field (F-TRF) modeling (Gill et al., 2006; Theunissen et al., 2001). Model features of interest were the spectro-temporal content of the speech signal in the formant frequency ranges. The model produced a regression weight time series (beta weights) for each frequency bin. For the main analysis, we extracted mean beta weights in a 50 ms window around the peak in the beta-weight time course (125–175 ms) and fit formant-encoding models to these values.

For the representative electrodes E2 (**Fig. 2.2B-C**) and E3 (**Fig. 2.2D-E**), we found that responses were the strongest toward one end of the vowel formant space, suggesting monotonic encoding of formant frequencies. On E2, response magnitudes and model beta weights increased with increases in F1 but decreased with increases in F2. By contrast, in electrode E3, beta weights were the strongest for high F1 and low F2 values.

Across all vowel-discriminating electrodes, we found that neural responses peaked near the boundaries of the vowel formant space, reflected in the bimodal distribution of maximal beta values for both formants (**Fig. 2.2F**; mixed-effects F1: beta = −0.21, SE = 0.025, $t(96) = -8.64$, $p < 0.001$; F2: beta = −0.99, SE = 0.082, $t(96) = -12.1$, $p < 0.001$). We thus focused our attention on the comparison between the linear and sigmoidal encoding of vowel formants. Across electrodes, we found that cross-validated R2 values were

higher for the sigmoidal model than the linear model on 82.4% of electrodes for F1 and 81.5% of electrodes for F2. We found a mixture of preferences for high and low formant values in both F1 and F2. This shows that STG neural populations have limited dynamic ranges: each local population represents a subspace of the vowel formant space, such that a representation of the entire range emerges across the entire population.

Finally, we wanted to characterize the extent to which both formants are jointly encoded at a single electrode. Across electrodes, we found an inverse correlation between tuning to F1 and F2 (**Fig. 2.2H**; mixed-effects beta = −0.61, SE = 0.10, $t(103) = -5.90$, $p < 0.001$). That is, electrodes with a preference for high F1 values also preferred low F2 values, and vice versa, as previously found for an English-language dataset (Mesgarani et al., 2014). Here, we found that this trend was driven by two main patterns. First, most electrodes jointly encoded both formants (76 of the 105 electrodes), with a preference for negative tuning in F1 and positive tuning in F2 (43 of the 105 electrodes). By contrast, although a small subset of electrodes (n = 16) had negative tuning to both formants, only a single electrode in our dataset had positive tuning to both formants. This raises the following question: does the lack of electrode sites with positive tuning to both formants reflect a specialization of STG for speech sound harmonics? Alternatively, it might be a confound of the limited formant frequency ranges in natural speech. We will address this question with Experiment 2 below.

Overall, these results show that neural activity at single-electrode sites is only discriminative in a subspace of the overall vowel formant space. However, across electrodes, the range of formant tuning at the population level should be sufficient to represent the entire vowel space with high fidelity. We directly test this in the next analysis step using population decoding.

Discrimination between vowel categories at the population level

We evaluated whether single vowel categories are represented at the population level by comparing the accuracy of vowel decoding on different electrode subsets using a linear support vector machine (SVM)

approach. First, we compared classifier accuracies derived from single electrodes versus from the entire electrode population (**Fig. 2.3A**). We found that decoding from the entire electrode population was significantly more accurate when considering all pairwise comparisons (average improvement: 2.2%–9.1% correct averaged across pairwise comparisons, t test showed a significant difference between all single electrodes versus the entire population accuracies with $p < 0.0001$; **Fig. 2.3A**). Notably, the best single electrodes did not necessarily show selective responses to a single vowel (**Fig. 2.3A**, right panel).

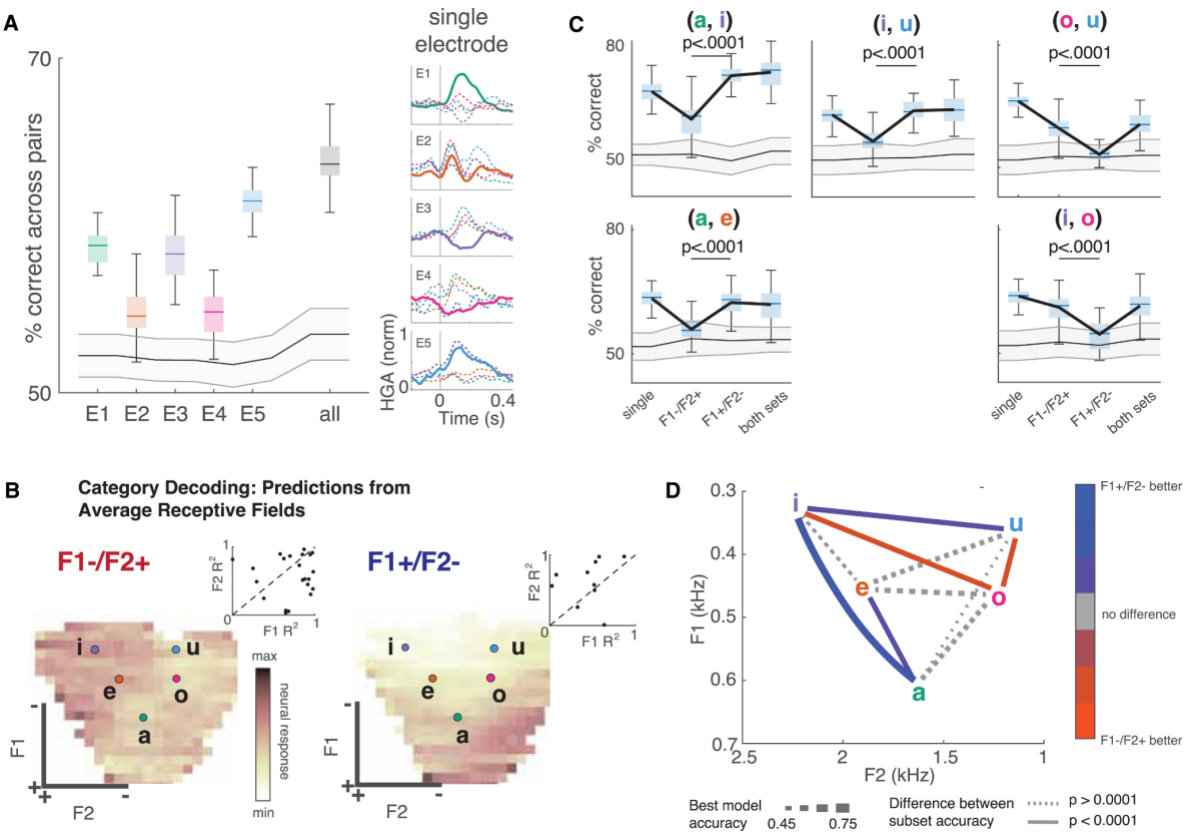


Figure 2.3 Emergent vowel representation at the population level.
A. Vowel decoding accuracy from the best single electrode per vowel (five unique electrodes across four subjects) as compared with the decoding accuracy using all electrodes. All accuracies averaged across ten pairwise comparisons. Right: average high gamma amplitude (HGA) response to each vowel in five electrodes used in single electrode decoding in the leftmost panel. Solid line corresponds to the vowel for which the selected electrode shows best average accuracy, dashed lines correspond to all other categories. Gray shaded bar indicates empirical chance performance over repetitions. Error bars indicate standard error of the mean across all pairwise decoding. **B.** Average formant receptive fields for electrodes with different tuning types and corresponding F1/F2 sigmoidal model R2. **C.** Decoding accuracy from five exemplary vowel pairs separated by electrode sub-group (best single electrode, each tuning type, both tuning types; p value reported for significant comparisons between sub-group accuracies). **D.** Summary of decoding accuracy across vowel pairs.

The improvement in classification accuracy from a single electrode to the population could be due to an increase in the signal-to-noise ratio. However, this improvement may also reflect complementary vowel encoding. That is, single electrodes represent only select regions of the vowel formant range in high detail, and as a result, pooling information across electrodes with sensitivity to different select regions may be key to decoding vowel categories across all pairwise comparisons. To determine whether this was the case, we split electrodes into the two dominant tuning subsets, namely F1−/F2+ and F1+/F2− type electrodes. Formant-receptive fields averaged across electrodes with the same directionality of F1/F2 tuning (F1−/F2+: n = 36 electrodes; F1+/F2−: n = 15 electrodes; **Fig. 2.3B**) suggested that each set would be critical for a subset of comparisons. For instance, average receptive fields suggested that F1+/F2− populations would be able to discriminate between /a/, /e/, and /i/, whereas this would not be the case for the F1−/F2+ population.

Decoding performed separately on the two subsets of electrodes confirmed this prediction, as can be seen in **Fig. 2.3C** for five exemplary comparisons ($p < 0.0001$ for significant comparisons between decoding accuracies on different subsets). Moreover, this analysis clearly demonstrates that increased decoding accuracies at the population level reflect the addition of informational content rather than increases in the signal-to-noise ratio. This is because the accuracy of decoders based on the best electrode subset and both electrode subsets together did not significantly differ. Finally, **Fig. 2.3D** summarizes the accuracies of all pairwise comparisons, showing that the two subsets of electrodes discriminate different sets of vowel pairs. The F1−/F2+ electrodes show significantly higher accuracy for the /i/–/o/ and /o/–/u/ pairwise classification, whereas the F1+/F2− electrodes show a higher accuracy for the /i/–/a/, /i/–/u/, and /e/–/a/ classification. However, in conjunction, these two electrode sets contain the complementary information necessary to significantly discriminate between all vowel pairs.

Shifting dynamic ranges underlie normalized vowel representations

It is well established that the mapping between vowel formants and vowel categories depends on a speaker's vocal tract length, which is correlated with voice pitch (Johnson, 2020; Johnson & Sjerps, 2021). We thus wanted to determine whether the formant tuning of neural populations in STG shifts with speaker voice quality during listening to continuous speech.

In line with previous work, we found that vowel formant frequencies increase with speaker pitch (**Fig. 2.4A**) and that normalizing for speaker pitch reduces the formant variance within vowel categories (**Fig. 2.4B**). We hypothesized that STG encoding of formants would reflect speaker-normalized rather than absolute formant frequencies. To test how speaker pitch affects the representation of vowel formants in human STG, we refit F-TRF models separately on subsets of data with low and high (>170 Hz) pitch levels and estimate the sigmoidal fits to F1 and F2 model weights separately for each of the models. If the STG representation of vowel formants reflects absolute formant frequencies, we expect to see no difference in formant tuning between the models (**Fig. 2.4C**, left). By contrast, if STG representations are normalized for speaker properties, we expect to see a shift in STG dynamic ranges, matching the shift in the vowel formant space between single models (**Fig. 2.4C**, right).

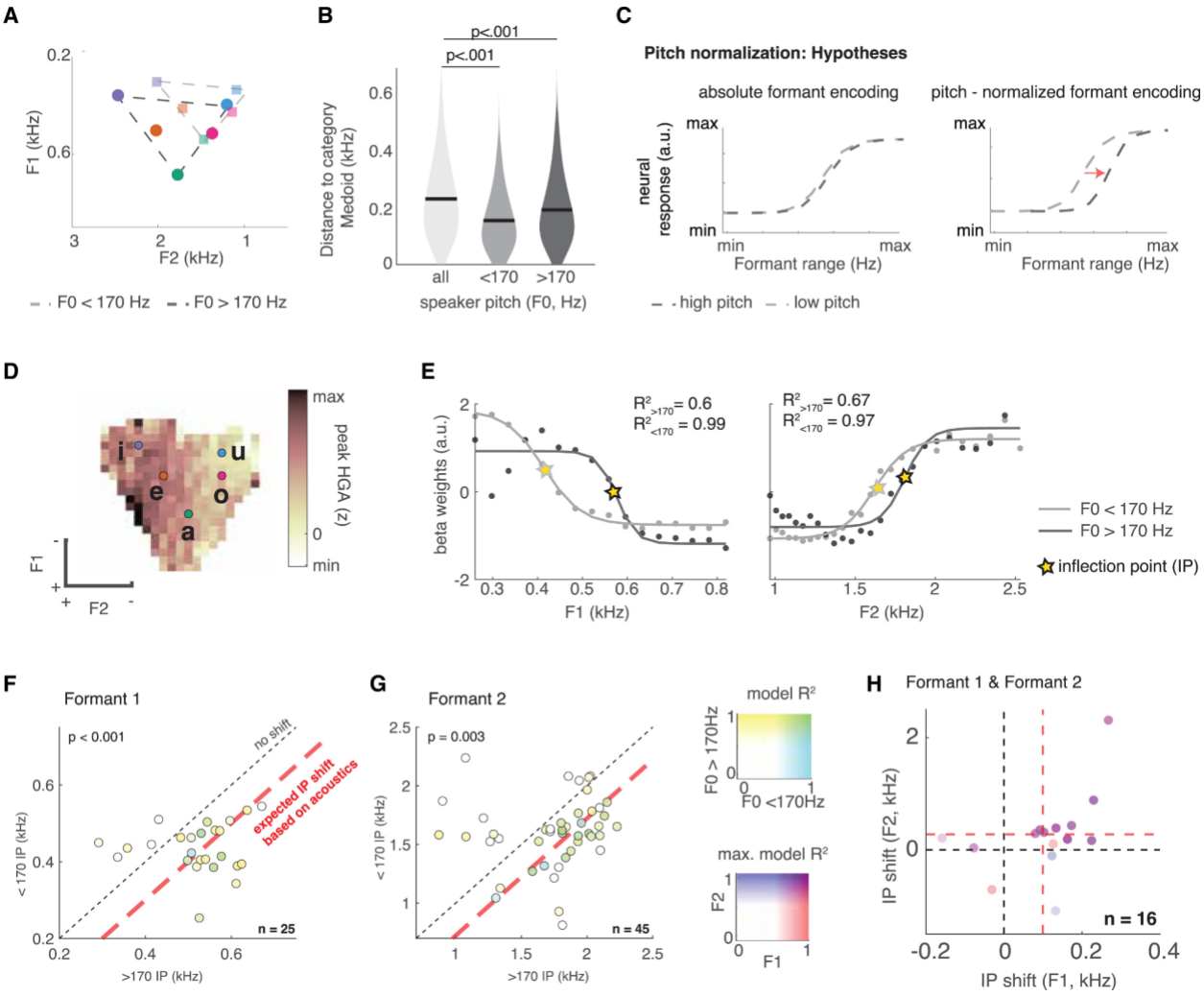


Figure 2.4 Shifting dynamic ranges underlie normalization of vowel representation for speaker pitch.
A. Vowel formants shift between speakers with short and long vocal tracts (and thus high and low pitch). **B.** Formant normalization in two groups by high/low pitch reduces formant variability within vowel categories. **C.** Hypotheses for absolute (left) and pitch-normalized (right) encoding of vowel formants. **D.** Formant receptive field for an exemplary electrode. **E.** Formant frequency tuning for F1 (left) and F2 (right) on the same exemplary electrode, split by speaker pitch. **F-G.** Tuning curve inflection points (IPs), calculated separately for sentences with high and low speaker pitch, across all F1 (F.) and F2 (G.) encoding electrodes. Note some electrodes included in (H.) are located outside axis limits in (F.) and (G.). P values indicate significance of inflection point shift. **H.** IP shift for F1 and F2 on electrodes encoding both formants. P values indicate significance of inflection point shift.

Fig. 2.4D-E shows the tuning curves for low- and high-pitch speakers on a single example electrode. For this exemplary electrode, neural responses shifted their dynamic ranges to higher frequencies in response to vowels produced by speakers with a high pitch, in line with speaker normalization for both F1 and F2.

To quantify the magnitude and extent of this shift on STG across all electrodes with significant formant frequency tuning (n = 105), we focused on electrodes with robust frequency tuning curves for vowels produced by both high- and low-pitch speakers (F1 n = 25, F2 n = 45 [unique electrodes n = 58], 55.24% of the electrodes). We extracted the sigmoidal fit inflection points (IPs) for each subset of speakers separately and assessed the difference in IP between models across electrode sites using linear mixed-effects modeling.

We found a systematic and robust shift of tuning curve IPs on these electrodes toward higher formant ranges with increases in the speaker pitch, with a most robust shift present on electrodes with overall good model fits for single formants (mixed-effects F1: pitch beta = 35.38, SE = 9.46, $t(111) = 3.7$, $p < 0.001$; F2: pitch beta = 95.03, SE = 31.7, $t(138) = 2.99$, $p = 0.003$; **Fig. 2.4F-G**). Remarkably, the average magnitude of the IP shift across electrodes mirrored the difference in the average formant frequency between high- and low-pitch speakers (red lines in **Fig. 2.4F-G**). Finally, we also found that electrodes with a robust encoding of F1 and F2 (n = 16) also showed speaker normalization for both formants in **Fig. 2.4H**. Taken together, these analyses show that formant normalization for speaker voice characteristics is a general feature of formant-encoding neural populations in the human STG.

Vowel formant encoding as general complex frequency tuning

Our analysis of the representation of vowel formants in STG raised two questions. First, are the opposite directions of preference for F1 and F2 due to the limited range and covariance between F1 and F2 in natural speech? That is, when presented with complex harmonic sounds with a broader range of F1 and F2 values, will neural populations show sensitivity to formant values with the same direction of tuning? Second, does each formant value affect the neural responses independently (**Fig. 2.5A**, top and middle row, joint independent encoding), or are co-encoding neural populations additionally integrating across F1 and F2, i.e., do responses to each formant depend on the value of the other formant (**Fig. 2.5A**, bottom, joint interactive encoding)? Notably, the qualitative patterns differ between independent and interactive

co-encoding: the latter shows a u-shaped tuning along one of the diagonals. Finally, we wanted to know whether STG encoding of the formant structure in sounds would continue outside the boundaries of the human vowel formant space. Alternatively, STG may contain separate neural populations encoding the sound harmonic structure in speech and non-speech frequency ranges.

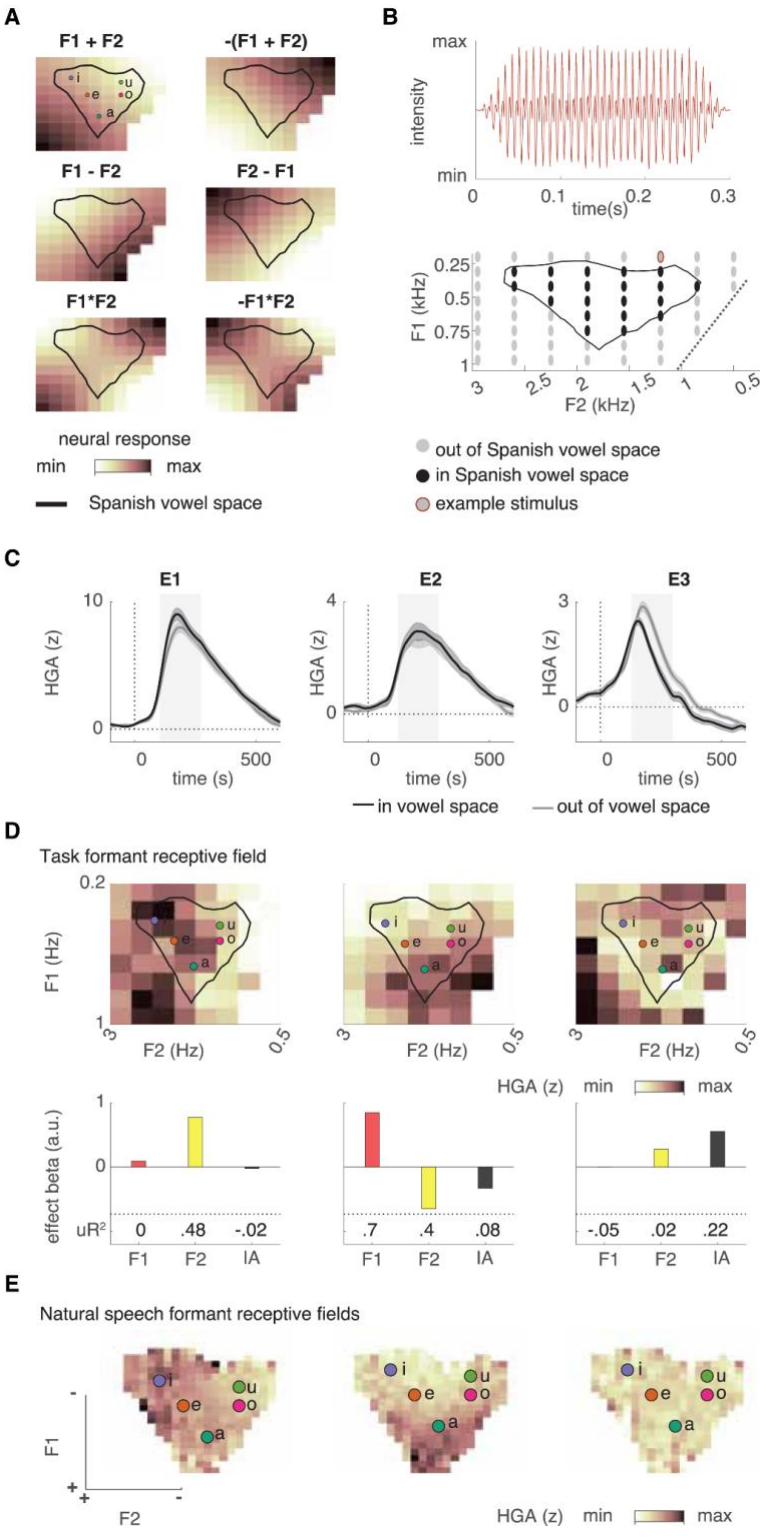


Figure 2.5 Vowel formant tuning emerges from complex frequency tuning on STG.
A. Schematic of independent (top two rows) and interactive encoding of F1 and F2 (bottom row). **B.** Stimulus design for synthetic vowel task; Top: example token waveform. Bottom: F1 and F2 values task **C.** Mean responses (± SEM) to stimulus tokens that fall within and outside the Spanish vowel space on three example STG electrodes. *(caption continues on the next page)*

(caption continued from previous page) Dark background marks peak area for averaging and further analyses. **D.** Formant receptive fields (top) and linear model effect R2 (bottom) for the same example electrodes. Black outline in formant receptive fields shows the vowel formant range in continuous Spanish sentences in the natural speech task. **E.** Formant receptive fields derived from the DIMEx corpus of Mexican Spanish (Pineda et al., 2010) for the same example electrodes.

To test these questions, we presented a group of ECoG patients (n = 8) with a set of isolated artificial vowel-like sounds with F1 and F2 values inside and outside the natural vowel space (black outline in **Fig. 2.5B**). We report all results across English and Spanish native speakers but note that no systematic differences between native speakers of the two languages were found. Notably, this task was language-independent, as vowels across languages fall within the same space due to the physical constraints on vowel production. We chose to use isolated harmonic sounds rather than consonant-vowel combinations to reduce any effects of coarticulation and ensure that tokens outside the vowel formant range could be perceived as non-linguistic. We found robust evoked HGA responses to single stimulus tokens on a subset of STG electrodes, which stereotypically peaked 150–200 ms after the stimulus onset, in line with responses to natural speech. We first tested whether electrodes responded differently to tokens within and outside the natural formant range (black vs. gray in **Fig. 2.5B**). Then, to model the encoding of F1 and F2, we extracted peak HGA for each stimulus token and evaluated the effects of F1, F2, and their linear interaction (IA, F1 × F2) on the peak HGA magnitude. Across all 8 patients, analyses were focused on 160 electrodes (10–33 per patient), for which the best linear regression model explained at least 10% of the total response variance.

Fig. 2.5C-E shows patterns of neural responses for three exemplary electrodes. Electrode E1 responded equally to tokens inside and outside the natural formant range (**Fig. 2.5C**, left), with the strongest responses to high F2 values (**Fig. 2.5D**, left) and no effect of F1 or the interaction term (**Fig. 2.5D**, bottom). By contrast, responses on electrode E2 were highest for the combination of low F1 and high F2 values, as supported by the significant interaction term on this electrode (**Fig. 2.5D**, middle). Finally, electrode E3 responded more strongly to tokens outside the vowel formant range (**Fig. 2.5C**, right), which was due to a significant positive interaction effect (**Fig. 2.5D**, right). On all three electrodes, the pattern of

responses is generally smooth around the vowel space boundaries, suggesting that any difference between the responses inside and outside this space is due to spectral and not speech tuning. Crucially, we found a high overlap between formant receptive fields derived from the synthetic vowel tokens and from natural speech stimuli, suggesting that both stimuli drive neural responses on STG to the same degree and in a comparable manner (**Fig. 2.5E**).

Formant encoding in synthetic vowel sounds and natural speech

Across all electrodes, we found that the main effects of F1 and F2 explained the most unique variance on single electrodes (F1: R2 median = 0.07, max = 0.71; F2: R2 median = 0.07, max = 0.72), with a minor but significant contribution of interaction terms (F1 × F2: R2 median = 0.03, max = 0.32; **Fig. 2.6A**). Notably, effect magnitudes for F1 and F2 were not correlated ($r = -0.1$, $p = 0.2$, mixed-effects model $t(158) = -1.48$, not significant [n.s.]; **Fig. 2.6B**), suggesting that each contributes independently to neural responses. By contrast, main and interaction effect magnitudes were negatively correlated ($r = -0.52$, $p < 0.001$, mixed-effects model $t(158) = -6.4$, $p < 0.001$; **Fig. 2.6C**). That is, electrodes with large interaction effects had little independent contribution of F1 and F2 main effects and low R2 overall. This suggests that encoding of F1 and F2 and integration across formants are implemented by distinct STG populations with the independent joint encoding of F1 and F2 dominating neural response patterns.

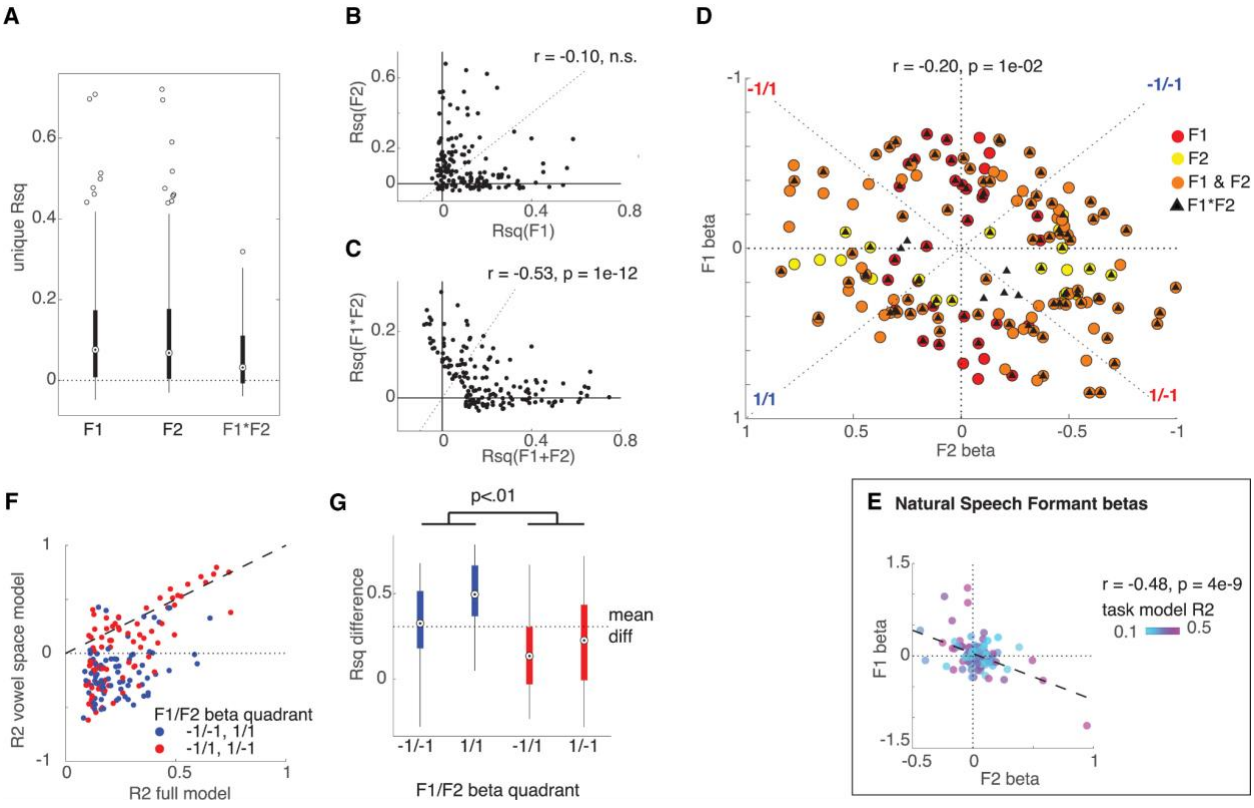


Figure 2.6 Comparison between formant encoding in synthetic vowel sounds and in natural speech. **A.** Effect R^2 distribution across electrodes. Error bars indicate standard error of the mean. **B-C.** Correlation between effect R^2 for the main effects of F1 and F2 (B.) and between main effects and the linear interaction effect (C.). **D-E.** Across all electrodes, the direction of effects for F1 and F2 is less correlated in the synthetic vowel task (D.) than in the natural speech corpus (E.). **F.** Model R^2 for model fit on the entire vowel task stimulus set vs. only on stimuli with formants located within the natural formant space. **G.** R^2 values for models fit on the full and reduced stimulus sets. Error bars indicate standard error of the mean. P value indicates significance of interaction of model and tuning direction.

In a second step, we asked whether the STG representation of vowel formants is tailored to formant ranges found in natural speech. Unlike in natural speech, in our synthetic vowel stimuli, we found only a weak negative correlation between F1 and F2 model weights ($r = -0.2$, $p = 0.01$, mixed-effects model: $t(158) = -3.31$, $p = 0.001$; **Fig. 2.6D**). Importantly, the range of tuning to different combinations of F1 and F2 values (n: +/+ 36, −/− 44, +/− 28, −/+ 52 tuning direction) in our data speaks in favor of the independent co-encoding of F1 and F2, rather than differential joint encoding.

We hypothesized that this discrepancy was due to the naturally limited range of F1 and F2 values in natural speech. To test this, we reran our analyses on a subset of stimuli with formants falling within the

natural formant frequency range (as marked in **Fig. 2.5B**). **Fig. 2.6F** shows the model R2 for full and subset models by electrode formant tuning. We found that in addition to the expected overall lower model R2 with a subset of stimuli (main effect of model: $b = -0.21$, $SE = 0.04$, $t(316) = -5.94$, $p < 0.001$) and marginally higher R2 values in electrodes with opposite tuning to F1 and F2 (main effect of tuning directions: $b = 0.16$, $SE = 0.08$, $t(316) = 2.02$, $p = 0.04$), this reduction was more pronounced for electrodes with the same direction of tuning for both formants (interaction of model and tuning direction: $b = -0.19$, $SE = 0.05$, $t(316) = -3.67$, $p < 0.001$; **Fig. 2.6G**). Overall, this shows that populations with the same directionality of tuning for both formants are not as strongly activated by vowels but rather are tuned to other harmonic sounds.

Discussion

This study provides a comprehensive account of the representation of vowels in the human speech cortex on the lateral STG. We found that vowel responses at local electrode sites in the STG were best characterized by complex two-dimensional receptive fields, defined by the first two formant frequencies, and normalized for speaker voice characteristics. Along each formant range (or receptive field axis), electrode sites showed nonlinear, monotonic frequency tuning, with a high sensitivity to a specific zone in the natural formant space. Because the spectral location of this zone often did not correspond to any single vowel, discrimination between vowel categories at single-electrode sites was unreliable. However, when neural population response patterns were aggregated across electrode sites, vowel categories could be decoded with high accuracy. Finally, neural responses to artificial vowel-like sounds with experimentally de-correlated F1 and F2 values showed that the complex tuning to formants in the STG independently represents F1 and F2 and extends beyond the natural vowel formant range.

The primary objective of this study was to use neural data to adjudicate linguistically informed theoretical models of vowel representation. Two fundamental models of vowel representation have been proposed, one in which the mental representation of vowels is described by a one-dimensional relationship between

F1 and F2 (F1–F2) (Syrdal & Gopal, 1986) and another in which it is described by a two-dimensional coordinate in space (F1, F2) (Nearey, 1989; Strange, 1989b). Although the neural responses to vowel sounds in this and earlier studies (Diesch & Luce, 1997; Mesgarani et al., 2014; Obleser et al., 2003) seem to support the former theory, the results from our synthesized vowel experiment suggest that this interpretation is misleading because of the limited range of formants in natural speech. By presenting subjects with formant combinations that extended beyond the range found in natural speech, we show that although most of the electrode sites inversely encode F1 and F2 (high F1 values and low F2 values, or vice versa), there exist electrode sites with other joint encoding patterns (e.g., high F1 values and high F2 values) or encoding to only a single formant. The fact that the encoding of formants on individual electrode sites spans all possible tuning combinations suggests that neural populations on the human speech cortex represent F1 and F2 as two distinct dimensions of the vowel space, rather than a relationship between them (Monahan & Idsardi, 2010; Syrdal & Gopal, 1986).

Our second objective was to determine how the neural encoding of formants gives rise to the representation of vowel information and speculate as to why the encoding of formants should organize in this way (Kohonen & Hari, 1999; Obleser & Eisner, 2009). Although previous studies found categorical vowel representations in the auditory cortex (Levy & Wilson, 2020; Scharinger et al., 2011), our decoding analysis revealed no evidence for the categorical representation of vowels at local electrode sites on the STG. That is, electrode sites responded strongly to subareas of the two-dimensional formant space, but these zones were not centered on single vowel categories. However, population-level decoding, where neural responses at single electrode sites were aggregated, allowed for the robust decoding of vowel categories. This is in line with an accumulating number of ECoG studies on the neural representation of consonants that also reported that specific discrete phonemes can only be reliably decoded at the population level (Chang et al., 2010; N. P. Fox et al., 2020; Mesgarani et al., 2014).

Our decoding results add to the existing evidence that a distributed representation of vowel sounds on the lateral STG is organized as a heterogeneous spatial code (Bhaya-Grossman & Chang, 2022; Bonte et al.,

2014; Formisano et al., 2008; Obleser et al., 2010; Yi et al., 2019a). Electrode sites that jointly encoded F1 and F2 belonged to two spatially interspersed encoding types: those with the strongest neural responses to high F1 and low F2 and vice versa. Together, the two encoding types represented the complementary formant information necessary to decode vowel categories. Since most of the electrode sites were nonlinearly tuned to F1 and F2 (**Fig. 2.2**), resulting in maximal neural responses at the edges of the formant range, formant encoding on the STG is also likely the basis for nonlinear, categorical vowel perception and perceptual magnet effects (Feldman & Griffiths, 2007; Iverson & Kuhl, 1995; Kuhl, 1991).

The third objective of this study was to determine the degree to which speaker normalization takes place on the lateral STG. Building on previous work that used active, isolated word-level tasks (Johnson, 2020; Sjerps et al., 2019), we found that local electrode sites on the STG represented vowel formant frequencies normalized for the speaker even when subjects passively listened to natural speech. These results strongly support the notion that neural representations of speech on the lateral STG are context-sensitive and contradict recent findings suggesting that these representations reflect the absolute frequency content of harmonic sounds (Daube et al., 2019). This discrepancy may be in part due to the different spatial resolutions of ECoG and EEG scalp recordings (Mukamel & Fried, 2012).

To perform speaker normalization in natural speech, the auditory system can draw on several distinct and often co-occurring acoustic cues (Fujisaki & Kawashima, 1968), such as the distributions of F1 and F2 for the speaker (Sjerps et al., 2019), the ratio between F1 and F3 (Monahan & Idsardi, 2010), and the average F0 of the speaker (Johnson, 1988). We did not experimentally isolate these cues and thus are unable to make a definitive argument regarding the cue used by neural populations on the STG to initiate speaker normalization. However, our results did show that the degree of normalization on single neural populations matched the distance between the average formants of speakers with low and high pitches, differing from previous ECoG studies that report only partial normalization (Sjerps et al., 2019). It is possible that the larger magnitude normalization effects observed in natural speech are due to the

presence of co-occurring acoustic cues versus only speaker F1 values that were manipulated in the previous study (Sjerps et al., 2019).

We can briefly speculate about the mechanisms that may account for the observed normalization effect (Johnson & Sjerps, 2021). First, these effects may be explained by the general auditory mechanisms that give rise to adaptation and contrast-enhancing sensory representations in both speech and non-speech contexts (Huang & Holt, 2012; Laing et al., 2012); examples of such mechanisms include stimulus-specific adaptation and gain control (Johnson & Sjerps, 2021; Pérez-González & Malmierca, 2014; Rabinowitz et al., 2011) or a critical band behavior sensitive to the density of harmonics (Fishman et al., 2000). It is also possible that speaker normalization reflects an integration of spatially interspersed but functionally distinct cortical areas encoding talker identity and might involve other parts of the temporal cortex, such as the superior temporal sulcus (Bonte et al., 2014; Formisano et al., 2008; Okada et al., 2010; von Kriegstein et al., 2003). We note that the possibilities listed here as mechanisms for normalization processes are not mutually exclusive and may both play a role in causing the effects we observe.

Finally, we were interested in understanding to what extent STG representation of complex harmonic sounds is specialized for human vowels. Although it is well established that vowel representation in PAC relies on narrow-band frequency tuning of tonotopically organized neural populations that are not specifically tuned to human speech (Fisher et al., 2018; Obleser et al., 2006; Ohl & Scheich, 1997), the vowel representation in non-primary auditory areas has not been fully explored. Here, we show that vowel-discriminating neural populations in the STG encode F1 and F2, with nonlinear, sigmoidal tuning along each separate formant dimension. Notably, our artificial vowel data showed that such representations also support the encoding of other harmonic sounds in the environment, namely, formant combinations outside those found in naturally produced vowels. The specialization for different harmonic ranges may support discrimination of non-speech complex harmonic sounds and possibly underlies recent

findings of distinct neural populations for speech and music in this area (Boebinger et al., 2021; S. Norman-Haignere et al., 2015).

Importantly, this study relied on two passive listening paradigms, one in which subjects listened to natural speech and one in which subjects listened to synthetic sound stimuli. We thus cannot exclude the possibility that task-dependent effects modulated the recorded neural responses. It is known that when subjects attend to and comprehend natural, meaningful speech, speech-relevant receptive fields are enhanced (Mesgarani & Chang, 2012), and furthermore, that task demands and complexity critically alter connectivity patterns across speech and language cortical networks (Liljeström et al., 2015; Saarinen et al., 2015). Task-specific enhancement effects may reflect top-down inputs from prefrontal areas (Fritz et al., 2013) via task-dependent oscillatory phase alignment (Bonte et al., 2009) or selective enhancement of receptive fields for task relevance (Atiani et al., 2014; Fritz et al., 2010). However, previous work comparing data from active and passive listening paradigms found qualitatively similar fMRI responses across task conditions in PAC and STG (Vannest et al., 2009) and comparable receptive fields in ECoG (N. P. Fox et al., 2020), suggesting that the effect of an active versus a passive task paradigm is more likely to fine-tune existing representations rather than alter them completely.

The current study on the neural representation of vowel formants in the human speech cortex leaves several important questions open. First, we focused our analysis on discrete vowel categories and static formant values, and as a result, we did not explore the effects of vowel duration or formant temporal dynamics on neural activity in the STG (Hillenbrand et al., 1995; Jenkins et al., 1983; Kuwabara, 1985; Strange, 1989a; Strange et al., 1976). We believe that this study lays the groundwork for the future exploration of the neural encoding of such complex formant dynamics. Second, we did not explicitly address the extent to which language experience affects the neural vowel representation in the current work. Language experience influences vowel recognition (Dufour et al., 2013; R. A. Fox et al., 1995; Näätänen et al., 1997), and a targeted paradigm is needed to address this. Finally, unlike past findings of functional asymmetries in speech processing across hemispheres (Boemio et al., 2005; Hickok & Poeppel,

2007a; Schönwiesner et al., 2005; Zatorre, 2001), but in line with previous ECoG studies (Mesgarani et al., 2014; Towle et al., 2008), we did not observe hemispheric differences with respect to any of our findings. This indicates at least some level of bilateral involvement in vowel processing. However, alternative neuroimaging modalities with bilateral coverage in single subjects are better suited to make claims about hemispheric asymmetries.

In conclusion, the results of this study demonstrate the broad and complex tuning of local electrode sites on the lateral STG to the formants in natural vowel sounds. This tuning is best described by nonlinear, two-dimensional formant receptive fields that adapt to the speaker's voice. Although most local electrode sites jointly encode the first two formants, without a strong preference for a single vowel, vowel categories can be extracted from the neural responses if aggregated across the population. Overall, we provide a comprehensive account of the representation of vowels in non-tonotopic areas of the auditory parabelt instrumental in the sensory processing of speech.

Methods

Participants

The study was approved by the University of California, San Francisco Committee on Human Research and all participants gave informed written consent before experimental testing. Fifteen (7 female) patients were implanted with 256-channel, 4-mm electrode distance, subdural ECoG grids as part of their treatment for intractable epilepsy. Electrode grids were placed over the peri-Sylvian region of one of the patients' hemispheres, as determined by clinical assessment. Eight Spanish-native speakers (5 male, 4 LH) with little to no knowledge of English participated in the DIMEx corpus speech experiment. Eight participants (2 Spanish-native; 6 English-native, 7 LH) listened to the synthesized vowel tokens of Experiment 2. Participants in the synthesized vowel experiment also listened to the DIMEx corpus. Two Spanish-native participants took part in both experiments. All participants had normal hearing and left-dominant language functions.

Neural data acquisition and preprocessing

Data acquisition methods closely follow those reported in our previous work (Hamilton et al., 2021; Oganian & Chang, 2019). ECoG signals were recorded with a multichannel PZ2 amplifier, connected to an RZ2 digital signal acquisition system [TuckerDavis Technologies (TDT), Alachua, FL, USA], with a sampling rate of 3052 Hz. The audio stimulus was split from the output of the presentation computer and recorded in the TDT circuit, time-aligned with the ECoG signal. In addition, the audio stimulus was recorded with a microphone and also input to the RZ2. Data were online referenced in the amplifier without any further re-referencing.

All data analyses were based on the analytic amplitude of neural responses in the high gamma range (HGA; 70 to 150 Hz), which is closely related to local neuronal firing and tracks neural activity at the

temporal scales of natural speech (Mukamel & Fried, 2012; Ray & Maunsell, 2011). Offline preprocessing of the data included (in this order) downsampling to 400 Hz, notch-filtering of line noise at 60, 120, and 180 Hz, extraction of the analytic amplitude in the high-gamma frequency range (70 to 150 Hz, HGA), exclusion of bad channels, and exclusion of bad time intervals.

HGA was extracted using eight band-pass filters [Gaussian filters, logarithmically increasing center frequencies (70 to 150 Hz) with semi-logarithmically increasing bandwidths] with the Hilbert transform. The high-gamma amplitude was calculated as the first principal component of the signal in each electrode across all eight high-gamma bands, using principal components analysis. Bad channels were defined by visual inspection as channels with excessive noise. Bad time points were defined as time points with noise activity in HG band, which typically stemmed from movement artifacts, interictal spiking, or non-physiological noise.

Last, the HGA was downsampled to 100 Hz, and z-scored relative to the mean and SD of the data within each experimental block. All further analyses were based on the resulting time series.

Electrode localization

For anatomical localization, electrode positions were extracted from postimplantation computer tomography scans, coregistered to the patients' structural magnetic resonance imaging and superimposed on three-dimensional reconstructions of the patients' cortical surfaces using a custom-written imaging pipeline (Hamilton et al., 2017). Freesurfer was used to create a 3d model of the individual subjects' pial surfaces, run automatic parcellation to get individual anatomical labels, and warp the individual subject surfaces into the cvs_avg35_inMNI152 average template.

Experiment 1: Continuous speech (DIMEx)

Stimuli and procedure

Participants passively listened to a selection of 500 Spanish sentences from the DIMEx corpus (Pineda et al., 2004, 2010), spoken by a variety of native Mexican-Spanish speakers. Data in this task were recorded in five blocks of approximately 7-min duration each. Four blocks contained distinct sentences, and one block contained 10 repetitions of 10 sentences. Sentences were 2.5 to 8.03 s long and presented with an intertrial interval of 800 ms. The repeated block was used for validation of temporal receptive field models (TRF; see details below).

All stimuli were presented at a comfortable ambient loudness (~70 dB) through free-field speakers (Logitech) placed approximately 80 cm in front of the patients' head using custom-written MATLABR2016b (MathWorks, www.mathworks.com) scripts. Speech stimuli were sampled at 16000 Hz for presentation in the experiment. Participants were asked to listen to the stimuli attentively and were free to keep their eyes open or closed during the stimulus presentation.

Stimulus spectrograms were calculated using the NSL toolbox (http://nsl.isr.umd.edu/downloads.html) for Matlab. Continuous formant values were extracted using the praat software (Boersma & Weenink, n.d.), https://www.fon.hum.uva.nl/praat/). We found that median formant values discriminate between vowel categories with a high accuracy. Thus all depictions of vowel tokens in two dimensional formant space reflect the token's median formant values.

Electrode selection

Analyses included electrodes located on the STG, for which the per electrode peak HGA response (over a contiguous window of 50 ms) after vowel onset significantly discriminated between vowel categories (using a one-way F-test and a non-corrected threshold of $p < 0.001$). Final analyses included 122 electrodes, 2 to 26 within single patients (median = 12). Selected electrodes were equally distributed

across hemispheres, with no hemispheric differences in vowel discriminability or electrode location along the anterior-posterior axis of the STG. For each of the below analysis, a subset of electrodes from this set were selected based on relevant criteria.

Feature temporal receptive field analysis

We fit neural data with a linear temporal receptive field (F-TRF) model with different sets of speech features as predictors. In this model, the neural response at each time point [HGA(t)] is modeled as a weighted linear combination of features (f) of the acoustic stimulus (X) in a window of 600 ms before that time point, resulting in a set of model coefficients, b1…, d for each feature f, with d = 60 for a sampling frequency of 100 Hz and inclusion of features from a 600 ms window (See previous work, (Mesgarani et al., 2014).

$$HGA(t) = \sum_{f=1}^{F} \sum_{k=1}^{L} b(k,f)X(f,t-k)$$

The models were estimated separately for each electrode, using five-fold cross-validation (80% train, 20% test). The regularization parameter was estimated using a 10-way bootstrap procedure on the training dataset for each electrode separately. Then, a final value was chosen as the average of optimal values across all electrodes for each patient. For all models, predictors and dependent variables were z-scored and scaled to between −1 and 1 before entering the model. This approach ensured that all estimated beta values were scale free and could be directly compared across predictors, with beta magnitude interpreted as an index for the contribution of a predictor to model performance.

Predictors in the model included spectral ranges that spanned the 5-95th percentile of frequency values for the first two formants for all vowels (F1: 250 - 800 Hz, F2: 1000 - 2500 Hz), as well as sentence onsets, vowel onsets, and predictors for timing and magnitude of peakRate, a marker of rising envelope edge dynamics (for peakRate feature structure see Oganian and Chang81). Vowel and sentence onset predictors were timed to onsets of the respective phonemes in the speech.

Linear and sigmoidal model comparisons on F-TRF betas

Linear and sigmoidal curves were fit to the mean beta weights around the beta peak in a 600 ms window (125-175ms), as estimated by the F-TRF model described above. Curve fitting for all model types was leave-one-out cross-validated (16 and 26 times for F1 and F2 bins respectively). This allowed us to directly compare models with different numbers of free parameters. Corresponding model R2 values were calculated based on the average error derived from the cross-validated predictions. R2 values were calculated as 1 - RSS/SST, where RSS is defined as the residual sum of squares and SST is defined as the total sum of squares. Note, this coefficient of determination measure (R2) could be negative if the regression predictions are further from the true value than a model that predicts the sample average. Bayesian Information Criterion (BIC) was used to validate the R2 values. According to this metric, the sigmoidal model outperformed the linear model on 31.4 % of electrodes for F1 and 25.7% of electrodes for F2. To test for bimodality, effects of tuning direction on the frequency value of maximal beta weight were assessed using linear mixed effects modeling with fixed effects of tuning direction, and random intercepts and slopes for subject and electrode (Maximal Beta Position ~Tuning Direction + (1 | Subject) + (1 | Electrode:Subject)).

Vowel decoding

Binary SVM linear classifiers (using pre-built function from Matlab Statistics and Machine Learning Toolboxes) were trained on neural data to distinguish pairs of vowel (all possible pairs of the five Spanish vowel categories, /a/, /e/, /i/, /o/, /u/) from a fixed window of 50 ms around the time point of peak discriminability between vowels (centered at approximately 150 ms post-vowel onset). Four types of classifiers were constructed: population level classifiers, two sets of population-subset classifiers, and single electrode classifiers. Population level classifiers were trained and tested on the output of principal component analysis (PCA) applied to neural activity from a population of electrodes (n = 54) spanning 4 subjects with sufficiently overlapping stimulus sets. To ensure that the pairwise decoding accuracy was comparable across pairs, data were subset to contain equal numbers of samples per vowel, equal to the

number of samples for the least frequent vowel (/u/) resulting in approximately 110 samples per vowel. Population-subset classifiers were also trained and tested on the output of a PCA and included electrodes with either F1+/F2- (n = 40) or F1-/F2+ (n = 14) individual tuning profiles derived from previous analysis. Single electrode classifiers were trained and tested on the neural activity recorded at single electrodes (n=105). Each classifier was 5-fold cross-validated and reported accuracy measures were averaged across each of the cross-validation sets. Significance testing was based on classification accuracies for data with permuted vowel labels, based on 50 permutations.

Speaker normalization

To determine the extent to which cortical responses to formants depend on speaker physiology, separate F-TRF models were were fit to neural data using subsets of speakers with an average fundamental frequency across the sentence of either less than or greater than 170 Hz, using the same predictors as the model described above. A total of 233 sentences (4840 vowel instances) were used in the low pitch speaker model and 267 sentences (5750 vowel instances) were used in the model corresponding to high pitch speakers. Curve fitting was performed on the speaker subset F-TRF model beta weights in the same way as described above. Inflection point shifts were calculated using the parameters derived from each of the F-TRF model types.

Effects of speaker subset on inflection point shift was determined using linear mixed effects modeling (using the Matlab Statistics and Machine Learning Toolbox) with fixed effects of model R2, speaker subset, and their interaction, and random intercepts and slopes for speaker within electrode (Inflection Point Position ~ Model R2 * SpeakerSubset + (1 | Subject) + (1 | Electrode:Subject)).

Experiment 2: Synthetic vowels

Stimuli and procedure

Participants passively listened to a set of synthesized vowel tokens. Stimuli were synthesized using an online version of the Klatt vowel synthesizer (www.source-code.biz/klattSyn/), with fixed pitch (250 Hz), F3 (3.1 kHz) and F4 (3.3 kHz). F1 and F2 were varied orthogonally to cover the entire vowel formant range as well as values outside those occurring in natural speech. For F1, we selected 10 values between 200 and 100 Hz (200, 255, 310, 420, 530, 640, 750, 860, 915, 970); for F2 we selected 10 values between 500 and 3000 Hz (500, 650,850,1200,1550, 1900, 2250, 2600, 2800, 2950). Stimuli covered all F1/F2 combinations with F2 larger than F1. Data in this task were recorded in five blocks of approximately 4-min duration each, resulting in 8 to 10 repetitions per token in each patient. Tokens were 300 ms long and were presented with an average intertrial interval of 800 ms, randomly sampled from a uniform distribution between 700 and 900 ms.

All stimuli were presented at a comfortable ambient loudness (~70 dB) through free-field speakers (Logitech) placed approximately 80 cm in front of the patients' head using custom-written MATLABR2016b scripts and psychtoolbox (Brainard, 1997; Pelli, 1997). Stimuli were sampled at 16 kHz for presentation in the experiment. Participants were asked to listen to the stimuli attentively and were free to keep their eyes open or closed during the stimulus presentation.

Electrode selection

Analyses included electrodes located on the STG, which showed robust evoked responses to vowel stimuli, defined as electrodes for which the best linear model, either with only main effects or with main effects and interaction terms, explained at least 10 % of the variance. Analyses contained 160 electrodes, 10 to 33 within single patients.

Analysis of variance

As neural responses had a stereotypical evoked response peaking between 100 and 400 ms after stimulus onset, we focused our analyses on the mean HGA in that time window. Single time point analyses of the entire HGA time course produced qualitatively the same results.

We modeled the main effects of F1, F2 and their linear interaction (F1× F2) onto evoked HGA responses to synthetic vowel onset using linear models. To assess the unique variance for main effects, we compared a main effect model (HGA ~ F1 + F2) to models containing only one formant (e.g., HGA ~ F1). To assess the unique variance explained by the interaction term, a full model (HGA ~ F1 + F2 + F1× F2) was compared to the model containing main effects only (HGA ~ F1 + F2).

For comparability across electrodes, predictors and HGA were z-scored prior to model fitting. For comparison to natural speech data, S-TRF models were fitted to natural speech response data on the same electrodes, using the same procedures as described above. To assess the effect of the formant range onto electrode response properties, models were fitted twice: First using all stimulus tokens, and second using only the subset of stimuli with formant values within the DIMEx vowel formant range.

Analyses across electrodes

For the correlations between beta weights across electrodes the model:

$$beta(F2) \sim beta(F1) + (1|subj) + (1|el:subj)$$

For the comparison between models on full data and reduced stimulus sets the model was

$$R^2 \sim model * tuningDirection + (1|subj) + (1|el:mod)$$

Quantification and statistical analysis

We used R2 as a metric of model fit for model comparisons between linear and sigmoidal models.

Analysis across electrodes

Analysis across electrodes were conducted using Pearson's correlations across all electrodes as well as with mixed-effects modeling with random intercepts and slopes for subjects and electrodes. Mixed effects models fit using the matlab function fitlme. Specific models are listed in the corresponding sections above.

Chapter 3: Shared and language-specific speech encoding in the human temporal cortex

Abstract

All spoken languages are produced by the human vocal tract, which defines the limited set of possible speech sounds. Despite this constraint, however, there exists incredible diversity in the world's 7000 spoken languages, each of which are learned through extensive experience hearing speech in language-specific contexts (Evans & Levinson, 2009). It remains unknown which elements of speech processing in the brain depend on daily language experience and which do not. In this study, we recorded high-density cortical activity from adult participants with diverse language backgrounds as they listened to speech in their native language and an unfamiliar foreign language. We found that, regardless of language experience, both native and foreign languages elicited similar cortical responses in the superior temporal gyrus (STG), associated with shared acoustic-phonetic processing of foundational speech sound features (Mesgarani et al., 2014; Yi et al., 2019b), such as vowels and consonants. However, only during native language listening did we observe enhanced neural encoding in the STG for word boundaries, word frequency, and language-specific sound sequence statistics. In a separate cohort of bilingual participants, this encoding of word and sequence level information appeared for both familiar languages within the same individual and within the same STG neural populations. These results suggest that experience-dependent language processing involves dynamic integration of both shared acoustic-phonetic and language-specific sequence and word level information in the STG.

Introduction

When listening to speech in a foreign language, we hear a fast, continuous, and uninterpretable stream of speech sounds (Bosker & Reinisch, 2017; Cutler, 2015b). However, when we listen to speech in a language we know, we hear these sounds as words. This specialized perception of speech in one's native languages develops over the course of language acquisition (Cutler, 2015b; Cutler et al., 1983; Kuhl et al., 1992, 1997; Saffran, Newport, et al., 1996; Stager & Werker, 1997; Werker & Tees, 1999). A central question in speech neuroscience is how and to what extent the adult brain's processing of speech is specific to one's native languages.

Prior work has identified the human STG, a non-primary auditory area, as crucial for speech perception (Hamilton et al., 2021; Hickok & Poeppel, 2007b; Rauschecker & Scott, 2009; Yi et al., 2019b). Both native and foreign or unfamiliar speech elicit a broadly similar pattern of activation in the bilateral middle STG (Binder et al., 1994, 2000; Démonet et al., 1992; Lipkin et al., 2022; Malik-Moraleda et al., 2022; Mazoyer et al., 1993; Narain et al., 2003; S. V. Norman-Haignere & McDermott, 2018; Schlosser et al., 1998). As a result, models of speech processing have hypothesized that the middle STG performs complex acoustic processing which predicts identical neural processing of native and foreign speech (Fedorenko et al., 2024; Hickok & Poeppel, 2007b). However, the STG contains neural representations of speech features subject to language-specific differences, such as phonetic categories (Chang et al., 2010; N. P. Fox et al., 2020; Mesgarani et al., 2014) and lexical tone (Y. Li et al., 2021, 2023). This provides indirect evidence that the STG may play a differential role in native as compared to foreign speech perception for certain speech features. In this study, we attempt to reconcile these conflicting hypotheses by specifically asking which speech features are affected by language experience in the human temporal cortex.

To address this question, we report on a rare dataset collected over the span of ten years in which we leveraged high-density ECoG recordings from a cohort of Spanish, English, and Mandarin monolingual speakers passively listening to naturally spoken sentences in their native language and a foreign language (**Fig. 3.1A**). The high spatial and temporal resolution of direct cortical surface ECoG enabled us to ask not only about similarities and differences in neural populations activated by native and unfamiliar foreign languages, but much more specifically how transient speech features are encoded depending on language experience.

We found that the overall magnitude of neural activity in bilateral STG is similar for native and foreign languages, driven by the largely conserved encoding of acoustic-phonetic speech features. Given this result, we hypothesized that neural signatures present in ECoG recordings of known language processing might be realized at a different representational level, namely, in the enhanced encoding of sequences of acoustic-phonetic features that form larger, language-specific compositional structures, such as syllables and words. Indeed, we found that neural populations more robustly encoded the sequential structure of listeners native language, including identifying where individual words begin and end in natural speech. In a separate cohort of Spanish-English bilingual participants, we found that neural populations encoded phoneme sequence and word level information for both familiar languages, demonstrating experience-dependent mechanisms for speech processing across multiple languages. Finally, in participants with varying degrees of proficiency with English, we found that proficiency modulated the extent to which word level information in speech could be neurally decoded. We propose that these results support a neurobiological model of human speech processing wherein STG neural populations perform dynamic integration of shared acoustic-phonetic and language-specific phoneme sequence and word level information during native language listening.

Results

Shared phonetic encoding in native and foreign speech

To understand how language knowledge affects neural responses to natural speech, we recorded high-density ECoG from participants with diverse language backgrounds undergoing intracranial monitoring for epilepsy while they listened to read spoken sentences. Each participant heard sentences in both their known native language and an unknown foreign language (**Fig. 3.1A**; **Fig. 3.6**), allowing us to directly compare activity in the same neural populations.

Since our research question focuses on cross-linguistic *phonology* rather than morphology, syntax or semantics, we used as stimuli read-sentences that were designed to comprehensively span each language's phonemic inventory (Garofolo et al., 1993; A. Li et al., 2000; Pineda et al., 2004). We collected ECoG recording from 20 monolingual participants who spoke either Spanish, English, or Mandarin (14 LH; left hemisphere, 11 female; 9 Spanish speakers, 8 English speakers, and 3 Mandarin speakers), which are three of the most common languages in the world, each spoken by over half a billion people and originate from two different language families (Indo-European, Sino-Tibetan). While some participants had received exposure to other languages, they all reported little to no proficiency in the foreign language (**Table 3.1**) and that they were unable to comprehend when presented with foreign speech. All participants listened to English speech (mean across participants=20.44 minutes), Spanish speakers additionally listened to Spanish (mean=20.80 minutes), Mandarin speakers additionally listened to Mandarin (mean=23.68 minutes), and English speakers either additionally listened to Spanish or Mandarin (**Fig. 3.1A**).

First, we asked whether the same neural populations respond to both native and foreign speech. We identified 883 electrodes with significantly greater activity during speech than silence in any language

condition (high frequency activity [HFA], 70-150 Hz, see "Methods" for details; see **Fig. 3.7** for coverage). An example from a Spanish listener illustrates that the same electrodes were responsive to both native and foreign speech, particularly within the STG (**Fig. 3.1B**). Furthermore, the average response dynamics of these electrodes were highly correlated between language conditions (Pearson $r(250)=0.98$, $p<0.001$; **Fig. 3.1C**). These examples demonstrate that within an individual participant, the same neural populations are active in response to speech regardless of language knowledge.

Across all participants, a majority of speech-responsive electrodes were located in the STG, with additional responses in the middle temporal, postcentral, precentral and supramarginal gyrus (**Fig. 3.1E**). Electrodes with the strongest speech activity were active in response to both native and foreign speech (**Fig. 3.8**) and 79.84% of electrodes (82.20% median, 10.70% standard deviation across participants) showed significant responses to both languages. Single electrode HFA peaks were not consistently higher nor was the number of active electrodes significantly more when participants listened to native speech (**Fig. 3.8**). These results demonstrate a highly overlapping area of cortex activated in response to native and foreign speech (Binder et al., 2000).

We next asked whether these neural populations, which were largely active in response to both languages, also encoded the same information, independent of whether the presented language was comprehensible to the listener. All spoken languages consist of the same broad classes of acoustic-phonetic features, which describe short segments like vowels or consonants (Hamilton et al., 2018; Mesgarani et al., 2014; Yi et al., 2019b) and therefore neural tuning to these broad classes may be conserved across native and foreign speech. However, Spanish, English, and Mandarin phonology differ systematically (e.g. in how phonetic features are used to distinguish segments (Flege & Eefting, 1988; R. A. Fox et al., 1995; Hualde, 2005), which may give rise to language specific or selective neural tuning to certain phonetic features.

We evaluated whether language experience affects encoding of acoustic-phonetic features in natural speech – specifically, we asked, in cases where an acoustic-phonetic feature is present across languages, is it neurally encoded in the same way across native and foreign languages? To address this question, we fit a temporal receptive field (TRF with acoustic-phonetic features) (Theunissen et al., 2001) to each language separately, and asked to what extent acoustic-phonetic feature tuning was the same across native and foreign speech. We compared electrode feature tuning across languages by correlating the weights from TRF models that had been fit on each language separately.

We identified electrodes with significant model fits and tuning to specific acoustic-phonetic features, and observed that their response patterns were highly correlated across languages. For three example electrodes, tuning to vowel formants, changes in the speech amplitude envelope (peakRate; (Oganian & Chang, 2019), and fricative manner of articulation showed similar feature weights in known and foreign speech (**Fig. 3.1F**). Across electrodes with significant TRF model fits in both languages (R^2>0.1), feature weights across languages were highly correlated (Pearson r(1992)=0.86, p<0.001; **Fig. 3.1G**).

Overall, we found that 155 out of 166 electrodes (93.37%) had TRF feature weight profiles correlated above chance (p<0.05 compared to uncorrected non-parametric permuted distribution generated from shuffling TRF weight labels, **Fig. 3.1H**). We further tested this by training a TRF model in the native speech condition and testing in the foreign speech condition or vice versa. We found that this cross-training still produced significantly correlated model predictions (**Fig.3.9**). These results demonstrate that in addition to the same brain regions being active in response to native and foreign languages, acoustic-phonetic speech feature tuning in local neural populations is largely conserved, suggesting that these populations encode the shared auditory content relevant for general speech processing.

Finally, we considered the fact that languages systematically differ from one another in their specific inventory of speech sounds and that language-specific inventories influence speech sound perception, e.g.

categorical perception (Abramson & Lisker, 1972; Liberman et al., 1957, 1967). It is therefore possible that, despite the high resolution of the neural data and the sensitivity of TRF encoding models, neural populations could encode more subtle, categorical differences in phonology between languages (Näätänen et al., 1997). While natural speech is not ideal for testing these differences because of its inherent multivariate, complex nature, we examined two specific examples of speech sound differences across languages to test whether they are prevalent in STG activity: voice onset time (VOT) and vowel formants.

Spanish and English differ systematically in how the acoustic cue of VOT is used to determine phoneme category membership for plosive sounds (Abramson & Lisker, 1972; Flege & Eefting, 1988). We found a small but significant effect of native language on the distribution of electrodes tuned to VOT in the predicted direction (more electrodes tuned to voiced plosives in Spanish speakers), but no clear difference in how single electrodes encode plosives (**Fig. 3.10**). In addition to VOT, Spanish and English differ in the number and acoustic realization of language-specific vowel categories (Bradlow, 1995). In line with our VOT results, we found no consistent, significant differences between the neural encoding of vowels in English and Spanish speakers (**Fig. 3.11**). These results suggests that while categorical differences between phonemes can be affected by language experience (perhaps best studied with stimuli that parametrically dissociate acoustic cues, e.g. (R. A. Fox et al., 1995), the dominant pattern in STG when presented with natural speech is one of shared acoustic-phonetic representation across languages where categorical effects are overridden by the availability of other cues.

Together, these results indicate that the same STG neural populations respond to speech regardless of whether the listener is familiar with the language presented. Particularly within high-level auditory areas like STG, neural populations encode largely the same acoustic-phonetic speech content, reflecting common vocal tract articulator movement and the resulting shared phonological structure of different languages.

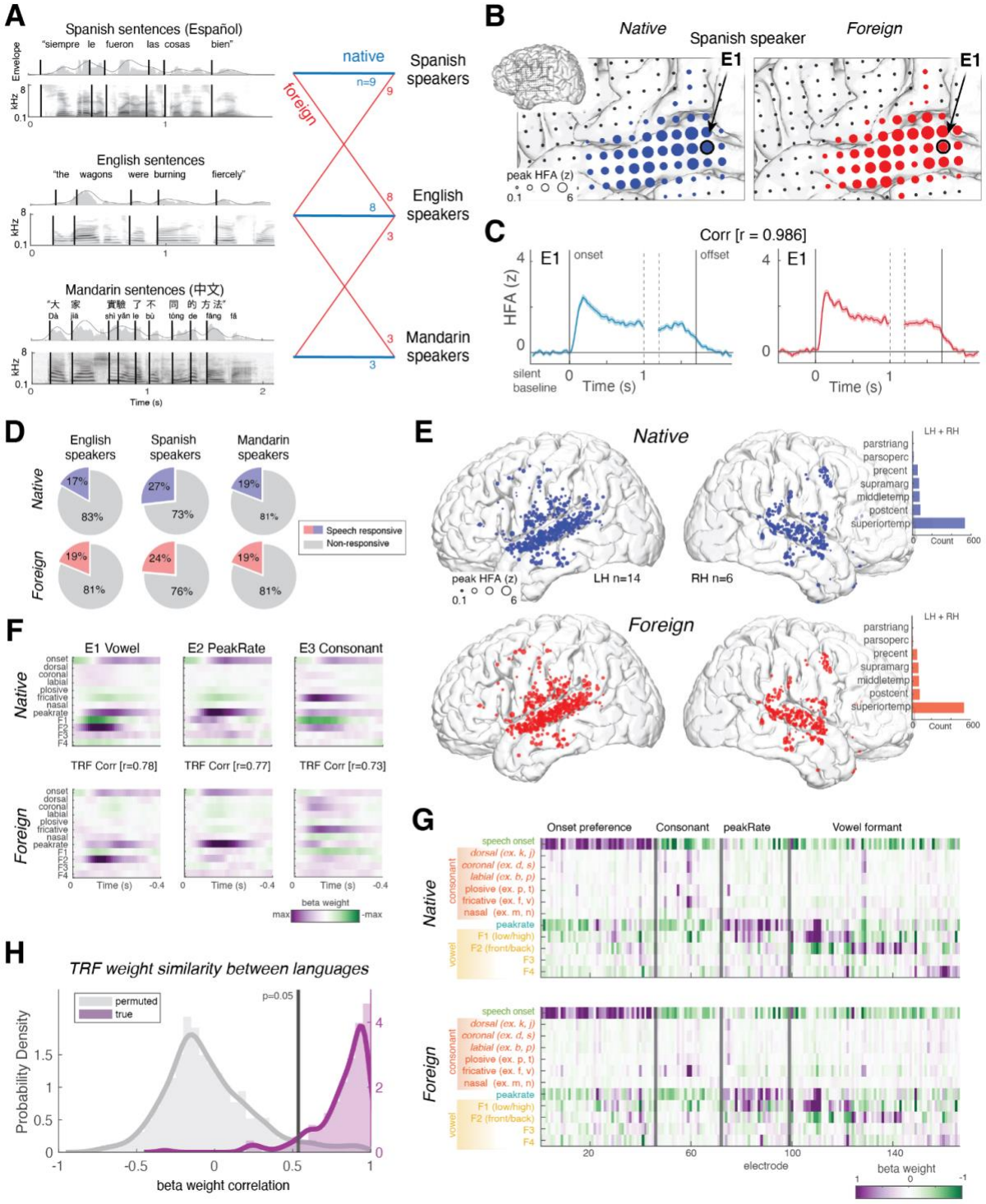


Fig 3.1. Shared acoustic-phonetic processing in STG across native and foreign speech.
A. Spanish, English, and Mandarin speaking ECoG participants passively listened to native and foreign speech. Left, sound envelope and spectrogram of example stimuli. Right, colored lines indicate speech condition, labeled with the number of participants in each group. **B.** In an example Spanish listener, the same neural populations are active in response to native and foreign speech. *(caption continues on the next page)*

(caption continued from previous page) Electrode size indicates magnitude of the peak HFA averaged sentence response, color indicates speech condition. **C.** Average HFA sentence responses for an example electrode is highly correlated across native and foreign speech (Pearson $r(250)=0.986$, $p<0.001$). Neural responses time-aligned to sentence onset and offset, color indicates speech condition. Shaded patches indicate standard error of the mean across sentences. **D.** Proportion of speech-responsive electrodes for Spanish, English, and Mandarin speakers similar across native and foreign speech, ranging from 17% to 27% across groups. **E.** Speech-responsive electrodes across 20 participants primarily localized to the STG in native (left) and foreign (right) speech conditions across hemispheres (LH: left hemisphere, RH: right hemisphere). Electrode size indicates magnitude of the peak HFA averaged sentence response. In-laid histograms show electrode counts across native and foreign speech within each anatomical region. **F.** Acoustic-phonetic temporal receptive field [TRF] weights across three example electrodes show strong correlations between native (top) and foreign (bottom) speech conditions. Different electrodes show tuning to distinct acoustic-phonetic features common to both languages. **G.** TRF weights across speech-responsive electrodes for native and foreign speech are highly correlated (Pearson $r(1992)=0.86$, $p<0.001$). Top panel columns (electrodes) ordered by the acoustic-phonetic feature showing the maximal weight. Bottom panel columns ordered identical to the panel above for visual comparison. **H.** Distribution of TRF native-foreign weight correlations across electrodes (purple) is significantly higher compared to the non-parametric permuted distribution (gray). Black vertical line indicates the 95th percentile in the permuted distribution.

Language experience enhances word encoding

Given that the brain represents individual speech sounds similarly regardless of language experience (**Fig. 3.1**), we hypothesized that experience is reflected more strongly in listeners' ability to compose these sounds into phonological sequences and words. This hypothesis can be intuited by the fact that when we hear someone speaking an unfamiliar language, we may be able to tell broadly which consonants and vowels we hear, but we find it substantially more challenging to identify where one word ends and another begins. This is due to the fact that natural speech does not have reliable cues to word boundaries (Cole et al., 1980; Cutler et al., 1997; Saffran, Aslin, et al., 1996), and that different languages like Spanish, English, and Mandarin are highly distinct at the level of speech sound sequences and words, specifically with respect to phonotactic and timing-related differences (Cuetos, Hallé, et al., 2011; Hualde, 2005). As a result, we were able to specifically test the hypothesis that when listening to one's native language, compared to an unfamiliar foreign language, language experience supports the neural encoding of phonological structure (Jacquemot et al., 2003), that is in turn used to segment the continuous speech stream into a series of auditory objects like words.

To test this hypothesis, we fit TRF models with the same acoustic-phonetic features as in **Fig. 3.1** and additional features that we predicted would be relevant for identifying the phonological structure of words in continuous speech: word boundary, word length (Shatzman & McQueen, 2006; Turk & Shattuck-Hufnagel, 2000), word frequency (Cuetos et al., n.d.; van Heuven et al., 2014), and phoneme surprisal (negative log probability (Gwilliams & Davis, 2020); for a full predictor list see "Methods"; **Fig. 3.2A**). We limited the following analyses to the Spanish and English speaking participants in order to perform a fully crossed comparison in which all participants listened to the same two languages (English and Spanish).

To evaluate the specific encoding of word-level features and phoneme surprisal independent from acoustic-phonetic tuning, we used unique R^2 as a measure of explained variance (Hamilton et al., 2021; Leonard et al., 2024). Across speech responsive electrodes, word boundary, length, and frequency explained significantly greater unique variance for native compared to foreign speech (linear mixed effects [LME] word boundary $t(243)= 4.00$, SE=0.0008, $p<0.001$; length $t(123)=3.65$, SE=0.0006, $p<0.001$); frequency $t(253)=5.18$, SE=0.0006, $p<0.001$); **Fig. 3.2B**). We also found that phoneme surprisal explained significantly greater unique variance in native compared to foreign speech ($t(29)=3.58$, SE=0.0017, $p=0.0012$; **Fig. 3.2B**), suggesting that response to sub-lexical sequence information is also dependent on language experience. These language experience-dependent effects were in contrast to no differences for acoustic-phonetic feature encoding (**Fig. 3.2C**; all $p>0.05$), consistent with the analysis of acoustic-phonetic feature weights (**Fig. 3.1E-G**). As a further control, we tested differences in unique variance explained by pitch (absolute pitch and relative pitch[54] and the continuous speech envelope, both of which showed no significant differences (pitch $t(175)=0.55$, SE=0.0027, $p=0.58$; envelope $t(473)=1.191$, SE=0.0009, $p=0.23$). This result establishes that a key neural difference between hearing one's native versus a foreign language is in how the phonological structure of speech is encoded toward the goal of perceiving words.

Next, we asked where on the lateral brain surface language experience-dependent information was predominantly encoded (we combined the word-level and phoneme surprisal features because encoding of each is strongly correlated; **Fig. 3.12**). We found that electrodes showing significant unique variance for these features were primarily located in bilateral mid-STG (**Fig. 3.2D-E**). These electrodes made up approximately 34% of all the speech responsive electrodes identified in **Fig. 3.1** in the native speech condition and approximately 12% of all the speech responsive electrodes in the foreign speech condition (**Fig. 3.2F**). Across single participants, the percentage of electrodes significantly encoding word-level and phoneme surprisal features was significantly higher in the native language (**Fig. 3.13**; signed-rank test $n=17$, $p=0.0012$).

If encoding of phonological structure is largely dependent on language experience, it is perhaps surprising that we find any significant encoding of this information in foreign speech. To understand to what extent the encoding of phonological structure in foreign speech reflects the same processes as in the native language, we compared the unique variance for word level and phoneme surprisal features across languages. A minority of electrodes encoded phonological structure only in the foreign language (lower right quadrant), potentially implicating a learning signal or tuning to novel sound structure[55]. Overall, encoding of phonological structure was significantly stronger in the native language (word-level and phoneme surprisal $t(179)=4.77$, $SE=0.0018$, $p<0.001$), including for electrodes that had significant effects in both native and foreign languages (**Fig. 3.2G**). Furthermore, the pattern of unique variances was not significantly correlated between languages (**Fig. 3.2G**; Spearman $r(98)=0.16$, $p=0.12$). Together, this suggests that while the brain may leverage native-like processing to attempt to parse foreign speech, the ability to encode phonological sequence information and words is significantly stronger in one's native language.

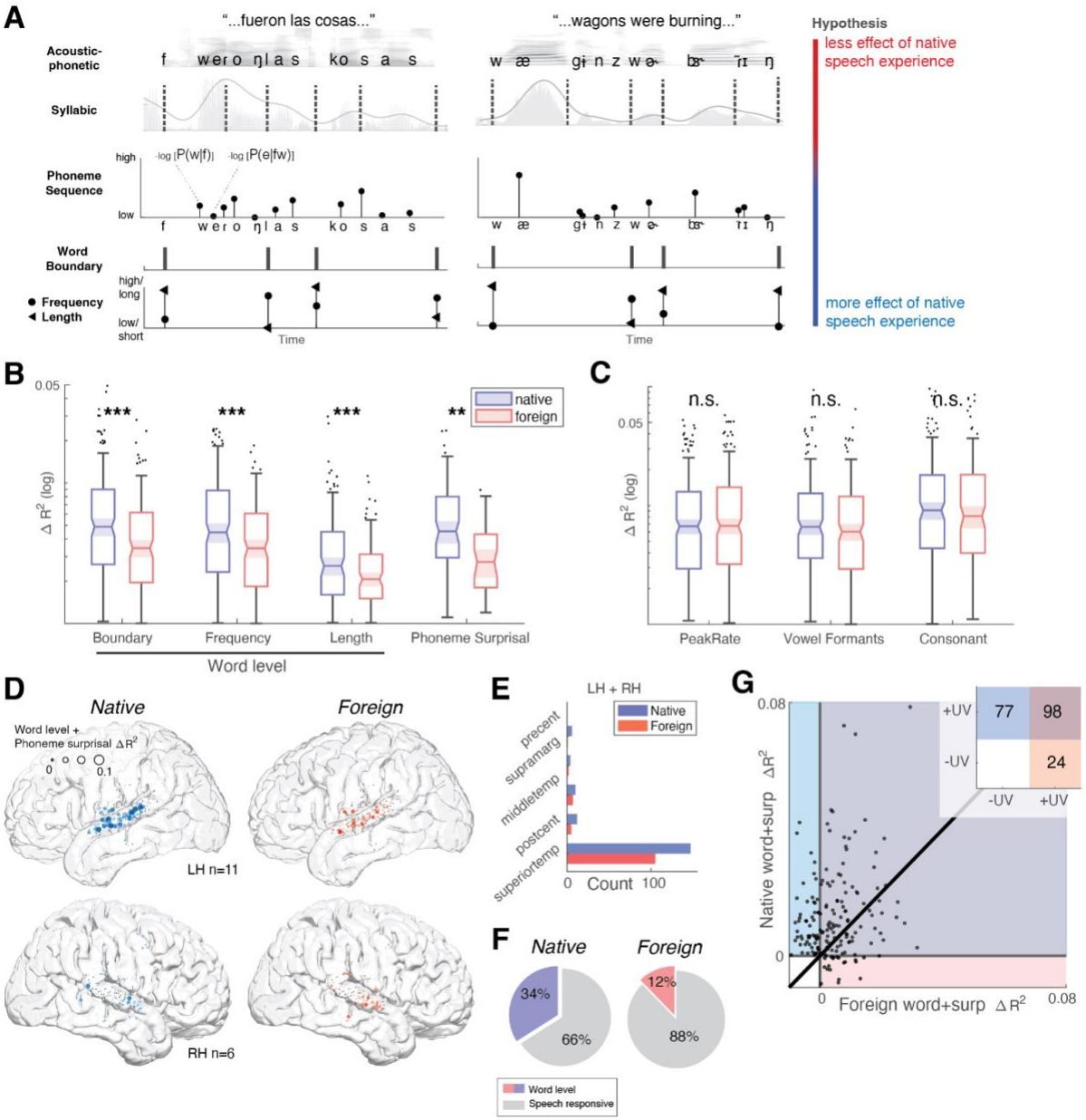


Fig 3.2. Enhanced encoding of word-level and phoneme surprisal features in native speech. **A.** Spanish and English sentences annotated with phonemes, syllable boundaries, phoneme surprisal, word boundaries, frequencies and lengths. We hypothesized neural encoding of speech representations requires increasing language experience (color bar on right). **B.** Unique variance for word level features is significantly higher in native vs. foreign speech (linear mixed effects [LME] model boundary ***p<0.001; frequency ***p<0.001; length ***p<0.001, see "Methods" for formula) and phoneme surprisal (LME **p=0.0012). Box plots show nonoutlier maximum and minimum values (whiskers), median (center line) and lower and upper quartiles (box limits), corresponding to electrodes surpassing a significance threshold for model R^2>0.05 and ΔR^2>0.001 (boundary n=246; frequency n=256; length n=126). **C.** Unique variance for acoustic-phonetic features do not significantly differ in native vs. foreign speech (LME peakRate p=0.41; formant p=0.60; consonant p=0.31, see "Methods" for formula). *(caption continues on the next page)*

(caption continued from previous page) Box plots show nonoutlier maximum and minimum values (whiskers), median (center line) and lower and upper quartiles (box limits), corresponding to electrodes surpassing a significance threshold for model R^2>0.05 and ΔR^2>0.001 (peakRate n=502; formant n=316; consonant n=270). **D**. Electrodes with positive unique variance for word and phoneme surprisal features primarily located in mid-STG. Electrode size indicates unique variance magnitude. Black scatters correspond to speech-responsive electrodes. **E.** Histograms show electrode counts with significant word and phoneme surprisal feature unique variance (permutation test vs. shuffled distribution) across native and foreign speech in each anatomically defined brain region. **F.** Pie-charts show proportion of speech responsive electrodes with significant word and phoneme surprisal feature unique variance (permutation test vs. shuffled distribution) for speech conditions across hemispheres. **G.** 98 electrodes show positive unique variance for word and phoneme surprisal in both native and foreign languages, it is not significantly correlated across languages (Spearman r(98)=0.16, p=0.12). Inlaid plot shows electrode count in each quadrant of the scatter plot.

STG segments words in native speech

When we hear speech in a foreign language, a salient challenge for listeners is identifying when individual words begin and end (*word segmentation*) (Cole et al., 1980; Lehiste, 1972). This is in part due to the fact that there are no reliable acoustic cues to word boundaries. Even prominent acoustic cues to word boundaries in English, like sharp increases in the speech envelope at word onset(Cutler, 2015a; Cutler & Carter, 1987; Cutler & Norris, 1988), are unreliable because they can also occur at syllable boundaries (**Fig. 3.3A-B**). Indeed, a decoder trained to classify word boundaries using 0.4s of the spectrogram around the boundary performs with a mean accuracy of 0.72 (median AUC=0.76 logistic regression, 20-fold cross-validated), meaning that over one in four words are incorrectly segmented based on acoustic cues alone (**Fig. 3.3C**).

Given that listeners show enhanced representation of linguistic features relevant for words in their native language (word boundary, length, and phoneme surprisal features; **Fig. 3.2B**), we hypothesized that neural activity would also be better able to discriminate between word and syllable boundaries in one's native compared to a foreign language. We first compared neural responses aligned to boundary events in individual electrodes. We found electrodes in STG that showed differential responses to word and syllable boundaries only for native speech for both Spanish and English speakers (**Fig. 3.3D-E**). Specifically, in native speech, both example electrodes show a negative deflection in amplitude

immediately after a word boundary as compared to the positive peak in amplitude at the syllable boundary (Zhang et al., 2025). Of the electrodes that discriminated between boundary events (50ms of contiguous significant difference, $p<0.01$ with Bonferroni correction), a majority did so only in native speech (**Fig. 3.3F**; 57%). Further, for electrodes that discriminated between boundary events in both languages (**Fig. 3.3F**; 24%), discrimination occurred for significantly longer in native speech (**Fig. 3.3F**; $t(455)=10.556$, $SE=0.65943$, $p<0.001$). These results suggest that language experience enhances the brain's ability to discriminate between word and syllable boundaries in continuous speech, leveraging knowledge about the phonological structure of words to overcome ambiguous acoustic cues.

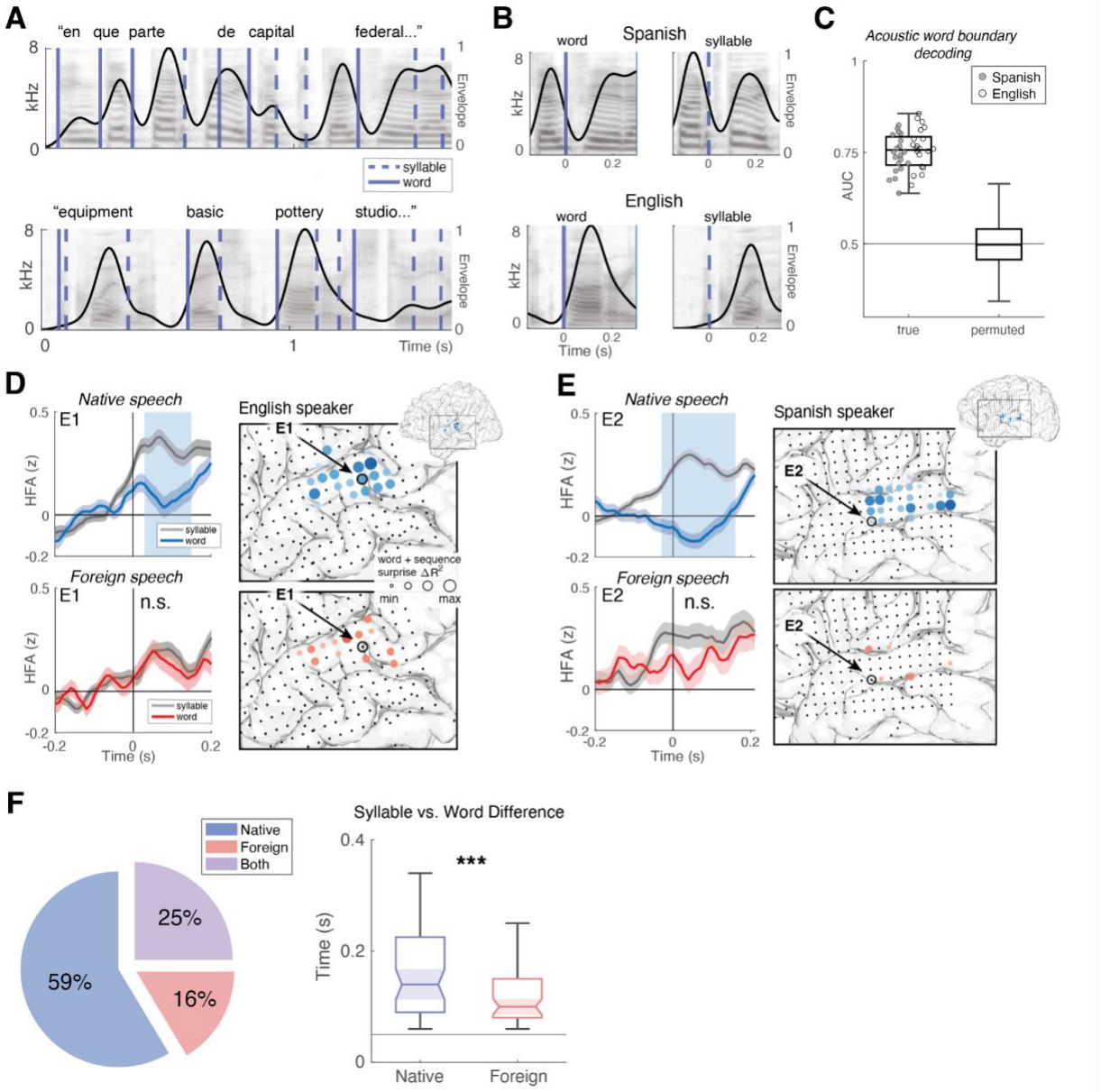


Fig 3.3. Enhanced discrimination between word and syllable boundaries in native speech. **A.** Spectrogram and speech envelope for example Spanish and English sentence stimuli. Word and syllable boundaries marked in solid and dashed lines, respectively. **B.** Spectrogram and speech envelope for example word and syllable boundaries are closely matched to one another in Spanish (top) and English (bottom), but differ across languages. Example word and syllable boundaries marked as in panel 3A. **C.** Acoustic classification of word boundaries using the spectrogram across English and Spanish speech. Each scatter point represents 1/20 folds. Permuted distribution median AUC of 0.50 (25 permutations with 20 folds each for each language). Box plots show maximum and minimum values (whiskers), median (center line) and the 25th to 75th percentiles (box limits). **D.** Example electrode (E1) from an English listener differentiates word and syllable boundaries in native (top) but not in foreign speech (bottom). Rectangular patches indicate time points with a significant difference (per-time point two-sided ANOVA $p<0.01$ with Bonferroni correction). Continuous shaded patches indicate standard error of the mean HFA response across boundaries. *(caption continues on next page)*

(caption continued from previous page) Right panels show the anatomical location of E1 and the unique variance for word and phoneme surprisal features in the example participant. **E.** Example electrode (E2) from a Spanish listener differentiates word and syllable boundaries in native (top) but not in foreign speech (bottom). Rectangular patches indicate time points with a significant difference (per-time-point two-sided ANOVA $p<0.01$ with Bonferroni correction). Continuous shaded patches indicate standard error of the mean HFA response across boundaries. Right panels show anatomical location of E2 and unique variance for word and phoneme surprisal features in the example participant. **F.** Word vs. syllable boundary discrimination is significant in more electrodes (left) and for a longer duration of the neural response (right) in native versus foreign speech (LME ***$p<0.001$, see "Methods" for formula).

To quantify the extent to which neural word segmentation was dependent on language experience, we trained a decoder to use ECoG recordings to classify word and syllable boundaries within each participant group and compared the performance (AUC) across cross-validated folds (0.2s before to 0.4s after each boundary; 20-fold cross-validated logistic regression). We found that word boundary decoding performance was significantly higher in native compared to foreign speech (native AUC=[0.77±0.052]; foreign AUC=[0.66±0.053]; **Fig. 3.4A**). We additionally found that this effect was consistent in each speech corpus separately and within single participants (**Fig. 3.14**).

While word segmentation is challenging without language experience, it is still possible for naive listeners to identify some boundaries based on salient acoustic cues(Cutler, 1990; Endress & Hauser, 2010) (**Fig. 3.4B-C**, **Fig. 3.15**). Therefore, we hypothesized that language experience particularly enhances word boundary decoding for words with unreliable acoustic cues. We characterized each word boundary according to an acoustic ambiguity index (AAI) which measures the extent to which acoustic cues could be used to classify each instance as a word or within-word syllable boundary (see "Methods" for details; **Fig. 3.4B**). We computed AAI for Spanish and English separately since acoustic and prosodic cues to word boundaries are language-specific (Cutler, 1990; Cutler et al., 1986; Tyler & Cutler, 2009) (**Fig. 3.4C**, **Fig. 3.15**) Critically, in both languages the envelope of high AAI word boundary examples (**Fig. 3.4B**) strongly resembles the average envelope aligned to within-word syllable boundaries (**Fig. 3.4C**). This suggests that if unfamiliar listeners were to use average envelope patterns alone to identify word boundaries, they would incorrectly classify the high AAI boundary trials.

Next, we binned the continuous AAI metric into three discrete bins and compared native versus foreign neural word boundary decoder performance (decoders from **Fig. 3.4A**). To do this, we randomly sampled trials to generate a distribution of AUC values in each bin and computed the pairwise difference between AUC values across speech conditions (native - foreign). We found that as acoustic cues became more ambiguous (increasing AAI), the difference in AUC between the native and foreign neural decoders increased as well, which we found to be significant using a 1-way ANOVA (**Fig. 3.4E**; Spanish F(2, 672)=12.88, p<0.001; English f(2, 672)=24.91, p<0.001). This result demonstrates that enhanced neural decoding of word boundaries in native speech is likely driven not only by sensitivity to the acoustic-phonetic cues to word boundaries, but also cues that draw on language experience and are not explicit in the acoustic signal.

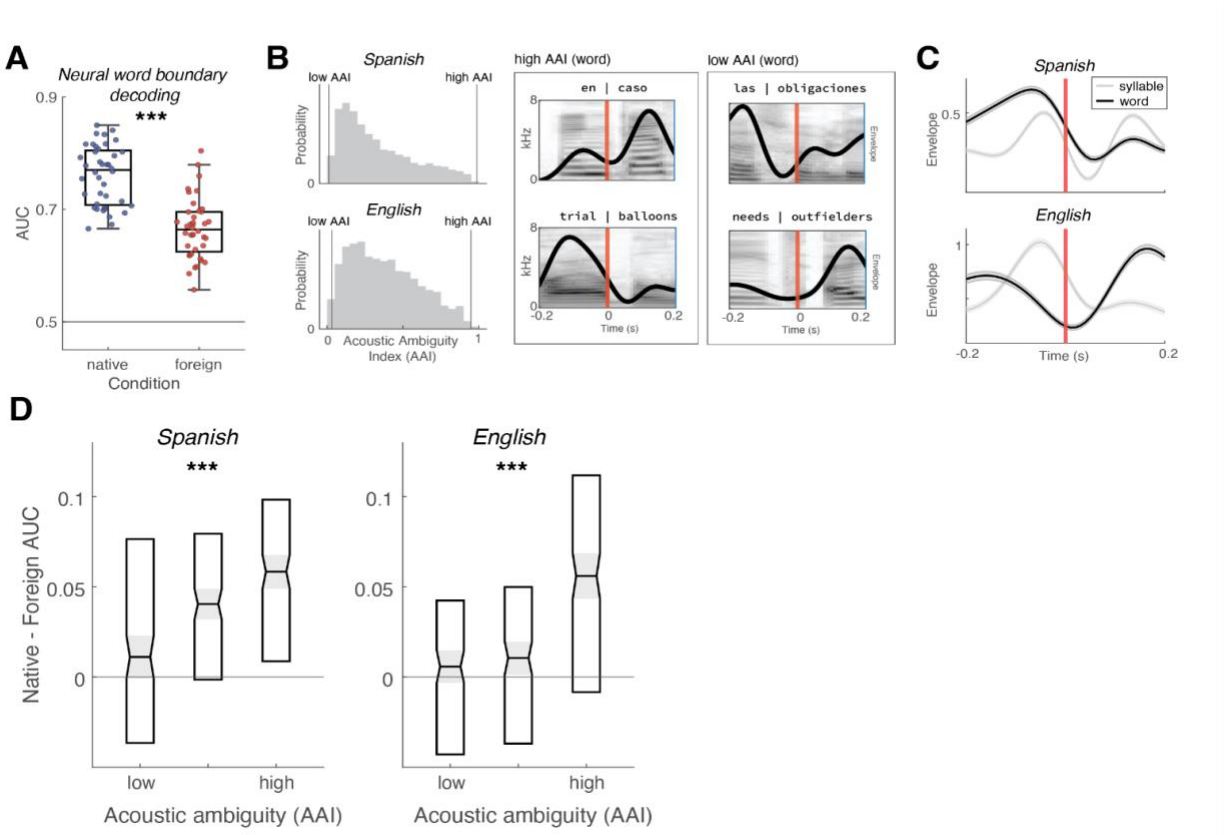


Fig 3.4. Enhanced neural word boundary decoding in native speech.
A. Neural classification of word boundaries is significantly better in native vs. foreign speech (each scatter point corresponds to the AUC of 1/20 folds combined across English and Spanish; 1-tailed 2-sample t-test ***p<0.001). Box plots show the maximum and minimum values (whiskers), median (center line) and the 25th to 75th percentiles (box limits). *(caption continues on next page)*

*(caption continued from previous page)***B.** Acoustic ambiguity indices (AAI) span a similar range in English and Spanish. AAI values were derived from an acoustic word boundary decoder trained separately on Spanish (top) and English (bottom) speech stimuli. Spectrogram and envelope of low and high AAI examples are shown on the right, e.g. "trial | balloons" and "needs | outfielders", which reflect the acoustics of highly likely and unlikely transitions at a word boundary in American English, respectively. **C.** Low AAI examples reflect the language-specific acoustic properties of word and syllable boundaries in Spanish and English as demonstrated by the average speech envelope. This resemblance can be seen by comparing low AAI examples in panel B to the average envelope aligned to the word boundary (black trace). Shaded patches around each trace are the standard error of the envelope across boundary events in the given language. **D.** Native language sensitivity to word boundaries is enhanced for ambiguous word boundaries. The difference between native and foreign neural word boundary decoding (AUC) increases as AAI increases. Box plots show median (center line) and the 25th to 75th percentiles (box limits) with each distribution corresponding to 225 pairwise AUC difference values generated by random sampling. AUC differences show a significant effect of bin (one-way ANOVA ***$p<0.001$ in both languages).

STG identifies words across languages

The vast majority of the world's population is proficient in more than one language. While it is known that in bilinguals, both familiar languages evoke neural activity in the same core set of brain regions (Cargnelutti et al., 2019; Hesling et al., 2012), it remains unclear what this activity reflects. Here, we took advantage of rare opportunities to record from bilingual participants while they listened to continuous speech in both of their familiar languages. We specifically asked to what extent encoding of both acoustic-phonetic and word-level information is the same for both languages, and how these representations differ as a function of second language proficiency.

First, we focused on recordings from 8 Spanish-English bilinguals (5 LH, 3 female) who were highly proficient in both languages (see **Table 3.3** for language profiles, **Fig. 3.7** for coverage). Consistent with the results from participants only familiar with one language (**Fig. 3.1**), we found that similar anatomical regions were active regardless of whether bilingual participants listened to English or Spanish, with the highest proportion of electrodes located in the STG (**Fig. 3.5A**). Activity in this region could be explained by encoding of acoustic-phonetic speech features (**Fig. 3.5B**), which was significantly correlated across languages (64 out of 67 electrodes, 95.52%), and had TRF feature weight profiles that were correlated above chance ($p<0.05$ compared to uncorrected non-parametric permuted distribution generated from

shuffling TRF weight labels, **Fig. 3.5C**). Thus, in the same brain of a bilingual participant, encoding of spectro-temporal speech content in STG is substantially similar in multiple languages (**Fig. 3.5B-C**).

Above, we showed that language knowledge most prominently affects representations of word-level and phoneme sequence speech information. Since Spanish-English bilinguals are familiar with the statistics of Spanish and English, we hypothesized that neural activity would reflect encoding of this information in both languages, unlike in the monolingual participants (**Fig. 3.2**). Using a linear mixed effects model, we found that word boundary, frequency, length, and phoneme surprisal unique variance was not significantly different between the two languages within bilingual participants (**Fig. 3.5D**) (boundary $t(163)=-0.05$, $SE=0.0009$, $p=0.96$; frequency $t(187)=-1.93$, $SE=0.0008$, $p=0.055$; length $t(177)=-1.20$, $SE=0.0008$, $p=0.23$; surprisal $t(73)=0.76$, $SE=0.0011$, $p=0.45$). Furthermore, encoding of word level and phoneme surprisal features was not significantly different in bilinguals as compared to the known language condition for monolinguals (**Fig. 3.4E**) (boundary $t(321)=-1.28$, $SE=0.0010$, $p=0.20$; frequency $t(346)=0.23$, $SE=0.0010$, $p=0.82$; length $t(318)=1.37$, $SE=0.0072$, $p=0.17$; surprisal $t(144)=-1.09$, $SE=0.0010$, $p=0.28$). Together, these results demonstrate that language experience – even for multiple languages in a single brain – modulates the extent to which word level and phoneme surprisal features affect neural representations in the temporal lobe.

Bilingual participants provide a unique opportunity to determine whether word level and phoneme surprisal encoding electrodes are language specific – that is, do electrodes encode these features for only one of the two familiar languages? In contrast to monolingual participants (**Fig. 3.3**), in bilingual participants, we found STG electrodes that strongly differentiated word and syllable boundaries in both Spanish and English (**Fig. 3.5F**). To test for language specificity across electrodes, we compared the corresponding unique variances for word level and phoneme surprisal features across Spanish and English. We found a large proportion of electrodes encoded word level and surprisal features for both English and Spanish, similar to results reported for monolingual participants in

Fig. 3.2. However, unlike the monolingual participants, the unique variances across languages were significantly correlated between languages (**Fig. 3.5G**; Spearman r(45)=0.31, p=0.03). This result demonstrates that word and phonological sequence information can be represented in the same populations across languages and implicates a shared mechanism for encoding phonological structure across languages in the bilingual brain.

Finally, while the key results thus far have been demonstrated with three languages that represent some of the linguistic diversity that exists across countries and cultures (English, Spanish, and Mandarin), there are many other languages and language families acquired as native languages throughout the world. Over more than a decade, we sampled some of these languages by collecting data from participants who were speakers of Russian, Arabic, and Korean (representing Indo-European, Semitic, and Sino-Tibetan language families; **Fig. 3.5H**). Furthermore, these participants had varying degrees of self-reported English proficiency (see **Table 3.4** for language profiles), allowing us to test the effects of experience across a wider array of known languages that differ from English to varying degrees. Given that the ability to identify word boundaries is one of the key phonological markers of language knowledge(Saffran et al., 1042), we asked to what extent neural word boundary decoding accuracy is graded as a function of English proficiency across these participants. Using a linear mixed effects model, we found that AUC was significantly correlated with English proficiency (**Fig. 3.5I**) ($t(253)=4.58$, $SE=0.002$, $p<0.001$), suggesting that the brain's ability to extract word-level structure from continuous speech depends on listeners' overall knowledge of the language. While these results arise from limited data, the fact that proficiency rather than typology of the native language (e.g. Russian, Korean, etc.) predicts neural word boundary decoding in English suggests that neural encoding of word boundaries may be a general speech processing mechanism that arises independently for each language that an individual learns.

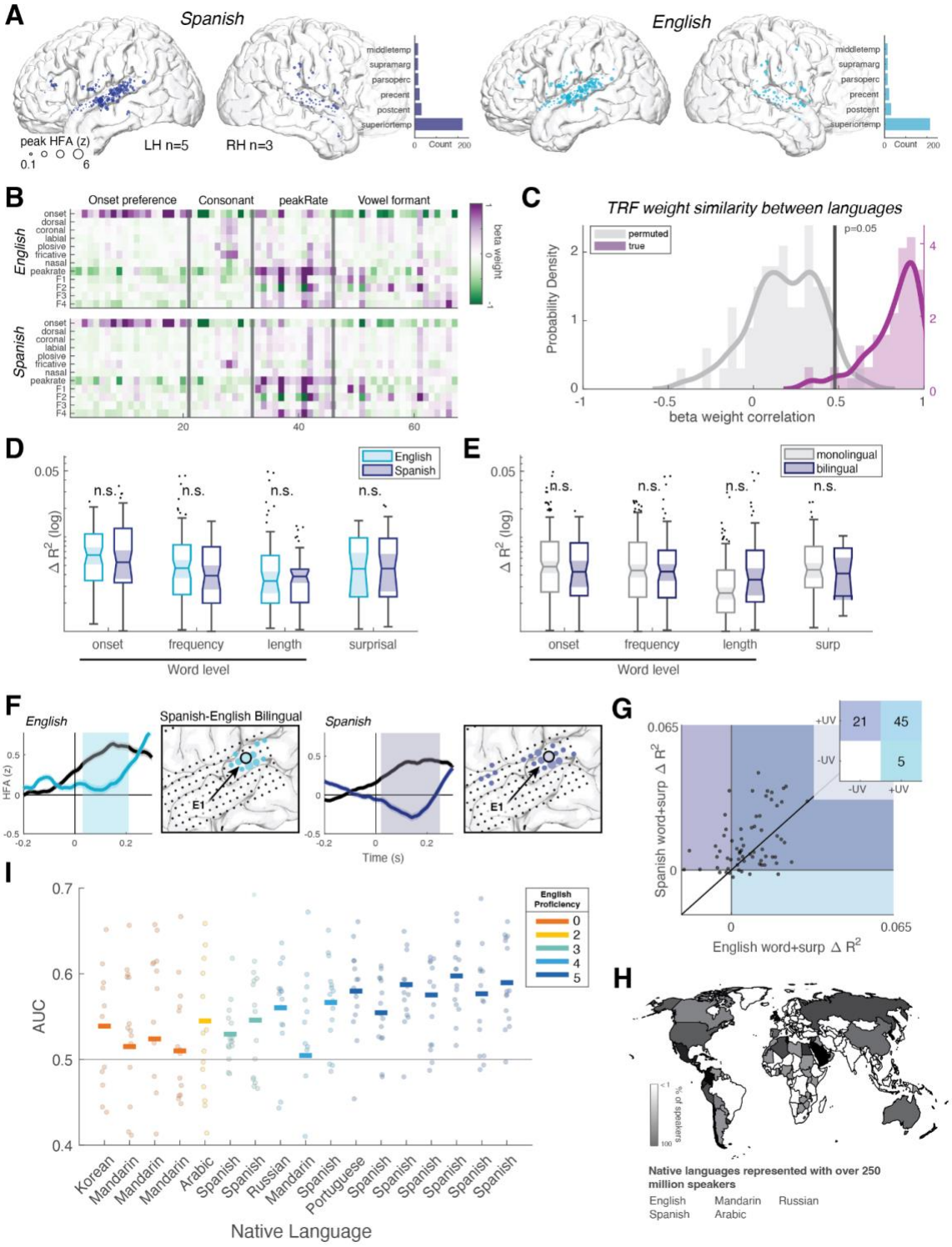


Fig 3.5. Bilingual listeners show neural encoding of word level features and phoneme surprisal for both familiar languages.

A. Speech-responsive electrodes across 8 Spanish-English bilingual participants show similar STG localization for Spanish (left) and English (right) speech across hemispheres (LH: left hemisphere, RH: right hemisphere). *(caption continues on next page)*

(caption continued from previous page) Electrode size indicates magnitude of the peak HFA averaged sentence response. In-laid histograms show electrode counts across English and Spanish speech within each anatomical region. **B.** TRF weights across speech-responsive electrodes for Spanish and English are significantly correlated (Pearson $r(804)=0.88$, $p<0.001$). Top panel columns (electrodes) ordered by the acoustic-phonetic feature showing the maximal weight. Bottom panel columns are ordered identical to the panel above for visual comparison. **C.** Distribution of TRF Spanish-English weight correlations across electrodes (purple) is significantly higher compared to the non-parametric permuted distribution (gray). Black vertical line indicates the 95th percentile in the permuted distribution. **D.** Encoding of word level and phoneme surprisal features are not significantly different across Spanish and English (LME speech condition $p>0.05$, see "Methods" for formula). **E.** Encoding of word level and phoneme surprisal features are not significantly different between bilingual and monolingual participants (LME speech condition $p>0.05$, see "Methods" for formula). **F.** Example electrode from a Spanish-English bilingual participant differentiates word and syllable boundaries in both the English speech condition (top) and the Spanish speech (bottom). **G.** 45 electrodes show positive unique variance for word level and phoneme surprisal in English and Spanish. Unique variance for word level features across languages significantly correlated (Spearman $r(45)=0.31$, $p=0.03$). Inlaid plot shows the count of electrodes in each quadrant of the scatter plot. **H.** World map indicating language backgrounds represented in the participant cohort and extent to which these languages are spoken across the world. Maps generated using MathWorks mapping toolbox and information from the World Factbook (Central Intelligence Agency, 2024). **I.** Neural word boundary classification AUC in English scales with English language proficiency for speakers of Spanish-English bilinguals and speakers of diverse languages (LME proficiency $p<0.001$, see "Methods" for formula). Each line corresponds to the mean AUC per participant, each scatter point is AUC of 1/15 folds.

Discussion

One of the remarkable attributes of the human brain is the ability to master the complexities of spoken language in just the first few years of life. Yet this skill does not transfer to all languages – we only understand languages to which we have been sufficiently exposed. This fact is underscored when we listen to a foreign language; while we may be able to identify individual speech sounds that are similar to the sounds of languages we know, it is substantially more challenging to parse the continuous input into larger chunks like words. To date, it has been unclear what drives this gap in our perception of familiar and foreign languages.

In this study, we leveraged high-resolution direct brain recordings to provide a detailed picture of the neural processes that reflect shared auditory phonological processes and those that are influenced by language experience. Due to shared human anatomy, the same fundamental articulatory features make up speech in all languages (Ladefoged & Johnson, 2014); features which the human brain is readily able to

perceive. However, with exposure that begins before birth, the brain becomes exquisitely tuned to the fine-grained, specific characteristics of the native language, including the specific inventory of speech sounds used by that language (Di Liberto et al., 2023; Kuhl et al., 1992, 1997, 2006; McCarthy et al., 2019; Werker & Lalonde, 1988). Throughout development, listeners become sensitive to how these sounds are combined to form the words that make up the language (Jusczyk & Hohne, 1997; Stager & Werker, 1997; Werker & Tees, 1999), which are also unique to each of the world's 7000 languages. Here, we demonstrate that language experience more strongly impacts this latter ability, with enhanced neural encoding of word-level features associated only with native languages.

Our results showing shared neural representations of speech in the human temporal cortex confirm its role as a processor of fundamental speech features (Mesgarani et al., 2014), independent of language experience. One surprising aspect of our results is the finding that experience-dependent neural activity also exists in these brain regions that process both native and foreign language input. This challenges the prevailing view that language processing only happens in a network of cortical regions that are selective for comprehensible speech (Fedorenko et al., 2024). Theoretical models of speech processing posit that auditory regions like the STG perform surface-level spectrotemporal extraction before sound input is linguistically transformed in higher-order cortical regions (Davis, 2016; Hickok & Poeppel, 2007a; Wilson et al., 2018). Under this view, the STG is a language-agnostic speech processor and therefore discriminates between languages as a consequence of distinct acoustic properties but not language experience of the listener. In contrast, we find evidence that representations of acoustic-phonetic, sub-lexical, and even word-level information is intermixed throughout the STG, with neural populations and representations showing varying levels of sensitivity to language experience. Ultimately, the current results are consistent with a view of the STG as a high-order auditory region, acting as an interface between speech input and linguistic knowledge (Bhaya-Grossman & Chang, 2022; Leonard & Chang, 2014; Yi et al., 2019b).

The present results demonstrate how language knowledge affects neural populations in the temporal lobe that are classically thought to process lower-level acoustic information (Fedorenko et al., 2024; Hickok & Poeppel, 2007a). These findings do not fully disambiguate the extent to which language experience-dependent knowledge is encoded elsewhere in the brain and exerts a top-down modulatory effect on STG (Davis & Johnsrude, 2007; Myers & Blumstein, 2008). However, the fact that the strongest and most prevalent encoding of language-specific word-level information is within STG suggests that these neural populations and underlying circuitry may be intrinsically sensitive to this type of learned knowledge (Yi et al., 2019b).

The present study is limited in that we did not sample from all brain regions potentially relevant to language experience. In particular, these results are based on ECoG responses recorded from the lateral surface that did not allow for the examination of the superior temporal plane such as the primary auditory cortex on Heschl's gyrus, the superior temporal sulcus (STS), or subcortical structures which can be structurally and functionally affected by language experience (Beguš et al., 2023; Hamilton et al., 2021; Kepinska et al., 2025; Wilson et al., 2018; Zhao & Kuhl, 2018). Prior recordings from this area with native English speakers listening to English suggest that these neural populations may be less sensitive to the acoustic-phonetic and word-level features examined here (Hamilton et al., 2021). It also remains to be seen whether our findings extend to learned non-speech sound stimuli, such as music (Boebinger et al., 2021; S. V. Norman-Haignere et al., 2022; Sankaran et al., 2024).

These findings have implications for understanding how language comprehension abilities arise during development, and for explaining the challenges associated with learning new languages later in life. While the adult brain retains an ability to learn new linguistic information (Golestani & Zatorre, 2004, 2009; Myers, 2014; Wang et al., 1999; Yi et al., 2021), achieving native-like proficiency with the phonology of a language after development can be difficult and depends on several different usage factors (Archibald, 2021; Eckman, 2008; Major, 2008; Strange & Shafer, 2008). Our results suggest this

challenge is in part related to learning how to parse sequences of individual speech sounds into language-specific units like words.

In bilinguals, neural encoding of language-specific knowledge (English, Spanish) is localized to the same brain regions, STG, and in some cases, even occurs within the same neural populations (**Fig. 3.5G**). These findings are consistent with theoretical models that posit a single and highly overlapping brain network within which multiple languages are processed (Abutalebi et al., 2001; Abutalebi & Green, 2007; Golestani, 2016; Green, 2003; Kim et al., 2020; Połczyńska & Bookheimer, 2021). While language-specific neural activation and impairments post-brain-injury have been found, greater neural differences between languages often correspond to greater disparities in skill-level (Kuzmina et al., 2019; Sulpizio et al., 2020), consistent with our finding that in multilinguals, English proficiency modulates the extent to which language-specific knowledge is neurally encoded (**Fig. 3.5I**) (Bice & Kroll, 2019; de Bruin, 2019; Kaushanskaya et al., 2020).

Studying speech and language in the human brain has traditionally been limited to a small number of languages, and typically with native speakers of those languages. While more work is needed to survey the vast linguistic diversity that exists across the world's languages, this work reveals that broadening the scope of inquiry to a diversity of cultural and experiential differences can give critical insights into fundamental mechanisms of neural coding for communication.

Methods

Data acquisition followed similar procedures to those conducted in prior work (Hamilton et al., 2021; Oganian et al., 2023; Oganian & Chang, 2019).

Participants

Thirty four patients (17 female) were implanted with the same type of subdural ECoG grids (Integra or PMT) with 4 mm center-to-center electrode spacing, 1.17 mm diameter contacts (recording tens to hundreds of thousands of neurons (K. J. Miller et al., 2009; Parvizi & Kastner, 2018) as part of their neurosurgical treatment either at UCSF or Huashan Hospital. All participants gave informed written consent to participate in the study prior to experimental testing. Most electrode grids were placed over the lateral surface of a single hemisphere and were centered around the temporal lobe extending to adjacent cortical areas. The precise location of electrode placement was determined by clinical assessment.

For all participants, age, gender, language background information and which hemisphere was recorded from is included in **Tables 3.1**, where each table corresponds to the respective participant group: English monolinguals, Spanish and Mandarin monolinguals, Spanish-English bilinguals, and others with diverse language backgrounds. Which participants were included in each set of analyses is listed in **Table 3.5**. ECoG recordings with the four Mandarin speakers were conducted at Huashan Hospital while patients underwent awake language mapping as part of their surgical brain-tumor treatment. ECoG recordings with all other participants were conducted at UCSF while patients underwent clinical monitoring for seizure activity as part of their surgical treatment for intractable epilepsy. The vast majority of participants in this dataset had epileptic seizure foci that were located in deep medial structures (e.g. the insula) or far anterior temporal lobe outside of the main regions of interest in this study, like the STG. All participants included in this study reported normal hearing and spoken language abilities.

All protocols in the current study were approved by the University of California, San Francisco Committee on Human Research and by the Huashan Hospital Institutional Review Board of Fudan University. Participants gave informed written consent to take part in the experiments and for their data to be analyzed. Informed consent of non-English speaking participants at UCSF was acquired via a medically certified interpreter platform (Language-Line Solutions) and communication with research staff was facilitated by either in-person or video-call based interpreters who were fluent in the participant's native language.

All participants were asked to self-report speech comprehension proficiency, age of acquisition, and frequency of use for all languages that they were familiar with. Participants who self-identified as Spanish-English bilinguals were asked to complete a comprehensive language questionnaire via an online Qualtrics survey (Kaushanskaya et al., 2020).

Neural data acquisition and preprocessing

ECoG signals were recorded with a multichannel PZ2 amplifier, connected to an RZ5 digital signal acquisition system (TuckerDavis Technologies (TDT), Alachua, FL, USA), with a sampling rate of 3 kHz. During the stimulus presentation, the audio signal was recorded in the TDT circuit, and therefore time-aligned with the ECoG signal. Audio stimulus was also recorded with a microphone and this signal was also recorded in the TDT circuit to ensure accurate time-alignment.

Offline preprocessing of the data included downsampling to 400 Hz, notch-filtering of line noise (at 60 Hz, 120 Hz, and 180 Hz for recording at UCSF and 50 Hz, 100 Hz, and 150 Hz for recording at Huashan Hospital), extracting the analytic amplitude in the high-gamma frequency range (70 to 150 Hz, HFA), and excluding extended interictal spiking or otherwise noisy channel activity through manual inspection. The Hilbert Transform was used to extract HFA using eight band-pass Gaussian filters with logarithmically increasing center frequencies (70 to 150 Hz) and semi-logarithmically increasing

bandwidths. High-gamma amplitude was calculated as in prior work (Hamilton et al., 2021; Oganian et al., 2023) as the first principal component of the signal in each electrode across all eight high-gamma bands, using principal components analysis (PCA). Last, the HFA was downsampled to 100 Hz, and z-scored relative to the mean and SD of the neural data in the experimental block. Each sentence HFA was normalized relative to the 0.5s of silent pre-stimulus baseline. All subsequent analyses were based on the resulting neural time series.

Electrode localization

For anatomical localization, electrode locations were extracted from postimplantation computer tomography scans, coregistered to the patients' structural magnetic resonance imaging and superimposed on three-dimensional reconstructions of the patients' cortical surfaces using an in-house imaging pipeline validated in prior work (Hamilton et al., 2017). Freesurfer (https://surfer.nmr.mgh.harvard.edu/) was used to create a 3D model of the individual participant's pial surfaces, run automatic parcellation to get individual anatomical labels, and warp the individual participant surfaces into the cvs_avg35_inMNI152 average template.

Stimuli and procedure

All participants passively listened to approx. 30 minutes of speech in a language they knew (either Spanish, English, or Mandarin Chinese) and one additional language (either Spanish, English). Speech stimuli consisted of a selection of 239 unique Spanish sentences from the DIMEx corpus (Pineda et al., 2004, 2010), spoken by a variety of native Mexican-Spanish speakers, 499 unique English sentences from the TIMIT corpus, spoken by a variety of American English speakers (Garofolo et al., 1993), and 58 unique Mandarin Chinese paragraphs from the ASCCD corpus from the Chinese LDC (http://www.chineseldc.org/), spoken by a variety of Mandarin Chinese speakers. Although the total

duration of stimuli presentation was approximately equal across languages, statistical tests that were conducted at the sentence level across corpora accounted for the differing number of sentences.

We selected these specific speech corpora, originally designed to test speech recognition systems, to meet the following requirements: 1) comprehensively span the phonemic inventories of the languages in question, 2) provide validated phonemic, phonetic, and word-level annotations, and 3) include a diverse and wide range of speakers with varying accents. These corpora were not intended to span specific semantic or syntactic structures within the languages or to capture long-range dependencies that extended across multiple sentences. Therefore we did not ask research questions about these specific speech representations.

The contents of each speech corpus were split across five blocks where each block was approximately 5-7-minutes in duration. In English and Spanish speech stimuli, four blocks contained unique sentences, and a single block contained 10 repetitions of 10 sentences. In the Mandarin Chinese speech stimuli, 4 paragraphs were repeated 6 times and repetitions were intermixed with unique paragraphs across blocks. Repeated sentences were used for validation of temporal receptive field models (TRF; see details below). Sentences were presented with an intertrial interval (ITI) of 0.4-0.5s in English and Mandarin and 0.8s in Spanish. Spanish sentences were on average 4.77s long (range: 2.5 - 8.03s), English sentences were on average 3.05s long (range: 1.99 - 3.60s), Mandarin Chinese sentences were on average 3.16s long (range: 1.17 - 11.76s). While speech blocks of different languages could be intermixed in the same recording session, each 5-7 minute block of speech consisted of only one language.

Speech stimuli were presented at a comfortable ambient loudness (~70 dB) through free-field speakers placed approximately 80 cm in front of the patients' head. All speech stimuli were presented in the experiment using custom-written MATLABR2016b scripts (MathWorks, www.mathworks.com). Participants were asked to listen to the stimuli attentively and were free to keep their eyes open or closed

during the stimulus presentation. All subsequent data analysis was performed using custom-written MATLABR2024b scripts.

Electrode selection

Subsequent analyses included all speech responsive electrodes: electrodes for which at least a contiguous 0.1s of the neural response (10 samples, 100 Hz sampling rate) time aligned to the first or last 0.5s of the spoken sentences was significantly different from that of the pre-speech silent baseline. To test for significance we used a one-way analysis of variance (ANOVA) F-test at each time point, Bonferroni corrected for multiple comparisons using a threshold of $p < 0.01$. We included the neural response aligned to the last 0.5s of each sentence to ensure that we did not exclude electrodes for which there was a significant response time aligned to the end of sentences but not the beginning.

Feature temporal receptive field analysis (F-TRF)

HFA responses to speech corpora were predicted using standard linear temporal receptive field models with various sets of speech features (Aertsen et al., 1981; Holdgraf et al., 2017; Theunissen et al., 2001). Many of the speech features used were extracted from pre-existing corrected acoustic, phonetic, phonemic and word-level annotations included with these speech corpora. In the feature temporal receptive field (F-TRF) models described below, HFA neural responses recorded from a single electrode were predicted as a linear combination of speech features that occurred at most 0.6s prior to the current time point:

$$HFA(t) = x_0 + \sum_{f=1}^{F} \sum_{l=1}^{L} \beta(l,f) X(f, t-l)$$

Where x_0 corresponds to the model intercept, F corresponds to the set of speech features included in the model, L corresponds to the number of time lags considered in the model (60 samples reflecting a total lag of 0.6s). $\beta(l,f)$ weights are therefore interpreted as the weight or importance of the specific

combination of speech feature, f, and time-lag, l, in predicting the final neural response given the set of speech features and training sentences. Models were trained and tested separately for each electrode using the same cross-validation, ridge regularization, and feature normalization protocols as reported in prior work (Hamilton et al., 2021; Oganian et al., 2023; Oganian & Chang, 2019). To derive the F-TRF model performance, we fit the trained model to a held-out set of approx. 10 sentences averaged across 10 repeats (see Stimuli and Procedure above) and then calculated the correlation-based R^2 fit. In subsequent analysis, we analyzed both the magnitude β weights for each speech feature and the overall model R^2 derived from the held-out test sentences. In addition to training and testing F-TRF models on sentences in the same language, we also trained F-TRF models on a language and tested the model on sentences from another language in order to test the generality of our model fits. This analysis followed the same training and testing procedure as the models which were trained and tested on the same language.

Selection of features for F-TRF

The speech features included in the acoustic-phonetic F-TRF model were selected from those that have been identified in previous studies to strongly drive HFA in the human temporal cortex. We included the sentence onset feature as well as the acoustic edges of envelope (peakRate; (Oganian & Chang, 2019), both features that drive neural responses in posterior STG and middle STG respectively. The consonant manner of articulation and place of articulation features were selected based on Mesgarani et. al., 2014 which included binary values at phoneme onset indicating whether the phoneme was: dorsal, coronal, labial, nasal, plosive, fricative. The vowel features were selected based on Oganian, Bhaya-Grossman et. al, 2023, which included the median formant values for F1, F2, F3, and F4 (Hz) for each vowel sound since continuous formant values were shown to drive neural responses more than categorical vowel features (e.g. high, front, low, back).

To test the effect of word-level features on the F-TRF neural prediction, We also included word onset, word frequency, absolute word length (as number of phonemes), and within-word phoneme surprisal. All

features had been pre-annotated for each speech corpus except for word frequency, phoneme surprisal, and syllable onset (in Spanish only). Word boundary and phoneme surprisal are metrics reflective of how the brain tracks sequential speech structure, with the former being a discrete, sparse marker of segmentation and the latter being a continuous measure of how predictable each phoneme is given all preceding phonemes in the current word. Word frequency and phoneme surprisal features were calculated based on the distribution of the language presented regardless of the participant language background (Spanish surprisal used in the model in Spanish speech blocks, a precise description of how these features were generated is below). In the TRF model, we excluded phoneme surprisal at word-initial positions so that the word-level features could be dissociated from the phoneme surprisal feature in time.

Model weight correlation across conditions

We determined a particular electrode's feature tuning by analyzing the model weights, where each weight was proportional to the change in electrode response given a change in that feature's value (Holdgraf et al., 2017). We estimated the F-TRF β weights corresponding to each unique speech feature for each electrode separately, as described in the model fitting procedure above (Mesgarani et al., 2014). To compare electrode encoding of the base F-TRF model across speech conditions (**Fig. 3.1**), we first selected all electrodes with $R^2 > 0.1$ for both speech conditions; this ensured that the β weight analysis would be based on TRF models that were able to explain the neural responses sufficiently well. We then extracted the mean of the 0.05s (5 samples) window around the maximum β weight averaged across features. This resulted in a single vector for each electrode per speech condition with one β weight per speech feature; to calculate weight correlation we performed a Pearson correlation across vectors of the same electrode from different speech conditions. This allowed us to determine the cross-language correlation between the relative β weights. To determine a distribution at chance level across electrodes, we performed permutation testing by shuffling the β weights for a single speech condition and repeating the Pearson correlation 10 times per electrode.

Word frequency, word length, and within-word phoneme surprisal features

Word frequency values were adopted from the SUBTLEX American English (*SUBTLEXus*, 2015) and SUBTLEX Spanish (Cuetos, Glez-Nosti, et al., 2011) frequency estimates which included a total of 74,287 and 94,341 unique words respectively. We included frequency estimates for all words in our corpora that existed in the English and Spanish SUBTLEX word sets. All word frequency measures were log-scaled (base 10). Word length was determined as the number of phonemes contained within the word onset and offset from the corpus annotations. We used the same word frequency and word lists documented in the SUBTLEX American English and Spanish corpus to construct within-word phoneme surprisal where each phoneme was associated with a surprisal value. We constructed within-word phoneme surprisal as:

$$-log[\ P(x_n|x_1 \ldots x_{n-1}) = \frac{P(x_1 \ldots x_n)}{P(x_1 \ldots x_{n-1})} = \frac{freq\ (x_1 \ldots x_n)}{freq\ (x_1 \ldots x_{n-1})}\ ]$$

where $x_1 \ldots x_{n-1}$ is all preceding phonemes within the current word and x_n is the current phoneme. We did not include surprisal values for the first phoneme of a word, only phonemes at the second position in the word onwards. $freq(x_1 \ldots x_n)$ is the count of all unique words with $x_1 \ldots x_n$ as the prefix multiplied by the frequency of each of those words as estimated by the SUBTLEX English and Spanish corpus. We additionally calculated within-word phoneme entropy (expected value of surprisal at each phoneme) as well as bi-phone and tri-phone surprisal and entropy but ultimately did not include these features in the final F-TRF model as for a majority of electrodes, these features did not explain added unique variance.

Annotation of DIMEx syllable onset

We generated the syllable onset feature for the DIMEx speech corpus through a combination of automatic syllable onset detection and manual validation. We did not annotate instances of resyllabified word onsets, e.g. instances where the consonants at the end (coda) of a syllable were pronounced as part of the onset of the following syllable. For example, the phrase *los otros* [los.ˈotɾos] would be annotated to reflect the canonical form (e.g. [los.ˈotɾos]), though it may be pronounced as [lo.ˈso.tɾos].

Unique Variance calculation

Since several of the speech features we used in the F-TRF models were highly correlated, we determined the unique contribution of each feature to driving neural responses through the calculation of unique variance. Unique variance (ΔR^2) is defined as difference between the portion of variance explained by the model with all relevant features and the model with the feature of interest (f) removed. Broadly, this is:

$$\Delta R^2\,(f) \;=\; [full\ model]\ R^2 \;-\; [model\ with\ f\ removed]\ R^2$$

The acoustic-phonetic (AP) F-TRF model included: sentence-onset, peakRate, consonant-features (dorsal, coronal, labial, nasal, plosive, fricative), vowel formants (F1, F2, F3, F4). The full F-TRF model included: sentence-onset, peakRate, consonant-features (dorsal, coronal, labial, nasal, plosive, fricative), vowel-formants (F1, F2, F3, F4), word-onset, word-frequency, word-length, phoneme-surprisal.

Unique variance for acoustic-phonetic features (**Fig. 3.2**) were calculated as follows:

$$\Delta R^2\,(sentence\ onset) \;=\; [full\ model]\ R^2 \;-\; [full\ model \;-\; sentence\ onset]\ R^2$$

$$\Delta R^2\,(peakRate) \;=\; [full\ model]\ R^2 \;-\; [full\ model \;-\; peakRate]\ R^2$$

$$\Delta R^2\,(consonant\ features) \;=\; [full\ model]\ R^2 \;-\; [full\ model \;-consonant\ features]\ R^2$$

$$\Delta R^2\,(vowel\ formants) \;=[full\ model]\ R^2 \;-\; [full\ model \;-vowel\ formants]\ R^2$$

Unique variance for word-level and phoneme surprisal features (**Fig. 3.2**) were calculated as follows:

$$\Delta R^2\,(word\ onset) \;=\; [AP\ model \;+\; word\ onset]\ R^2 \;-\; [AP\ model]\ R^2$$

$$\Delta R^2\,(word\ frequency) \;=\; [AP\ model \;+\; word\ onset \;+word\ frequency]\ R^2 \;-\; [AP\ model \;+\; word\ onset]\ R^2$$

$$\Delta R^2\,(word\ length) \;=\; [AP\ model \;+\; word\ onset \;+word\ length]\ R^2 \;-\; [AP \;+\; word\ onset]\ R^2$$

$$\Delta R^2\,(phoneme\ surprisal) \;=\; [AP\ model \;+\; phoneme\ surprisal]\ R^2 \;-\; [AP\ model]\ R^2$$

Combined unique variance for all word-level and phoneme surprisal features was calculated as:

$$\Delta R^2\,(word\ features) \;=\; [full\ model]\ R^2 \;-\; [acoustic\ phonetic\ model]\ R^2$$

To determine the significance of unique variance values for a single feature or feature family, we computed a distribution of 300 permuted unique variance values by applying random circular temporal shifts to the feature or feature family of interest. To acquire the p-value associated with the unique variance value, we summed the number of permuted unique values greater than the true unique value and divided by the total number of permutations we computed.

Statistical testing

Linear mixed effect (LME) models as computed by the fitlme function in MATLABR2024b included random effects for participant, hemisphere, as well as the interaction between participant and electrode to ensure that the effects that we found were significant across participants and hemispheres. LME models were primarily used to determine whether the unique variance of a particular feature f was significantly different between native and foreign speech conditions; LME formulas took the form:

$$\Delta R^2\,(f) \sim speech\ condition\ +\ speaker\ language$$

$$+(1 \mid hemisphere) + (1 \mid participant) + (1 \mid electrode{:}participant)$$

The same LME formula (above) was used to test significance in all instances in which TRF feature unique variance could be compared across speech conditions (e.g. **Fig. 3.2B-C** and **Fig. 3.5D**). For Pearson and Spearman correlation, two-tailed tests were performed unless otherwise noted.

Acoustic word boundary decoding

Acoustic word boundary decoding consisted of a logistic regression trained to classify whether the mel-spectrogram from 0.2s around a boundary event was a within-word syllable boundary or a word boundary. All single syllable words and words at sentence onset were removed from this analysis. In order to match the instances of word and syllable boundaries in the Spanish and English corpora, we randomly sampled 1900 instances of word and within-word syllable boundaries from each corpus. In

Spanish, we used 1273 within-word syllable onsets and 627 word onsets. In English, we used 1153 within-word syllable onsets and 747 word onsets.

For each boundary event in both corpora the 0.2s 80 band mel-spectrogram (sampled at 100Hz) was vectorized. PCA was then performed on the vectorized mel-spectrogram to reduce the total dimensionality, the components with the highest variance that cumulatively explained over 90% of the total variance were ultimately used as input to the regression, which totaled approximately 60-80 PCs depending on the language. Logistic regression was run using standard MATLABR2024b regression functions (fitclinear) with Lasso (L1) regularization and 20-fold cross-validation. The AUC, or the area under the receiving operating curve, was extracted for each held-out test fold, where chance AUC would equal 0.5.

In addition to training and testing acoustic word boundary decoding in the same language, we also trained the word boundary decoders in one language and tested on boundary events from the other language to test the extent to which acoustic cues to word boundaries overlapped between the languages. This analysis followed the same training and testing procedure as the decoders which were trained and tested on the same language. To assess the significance of the test-set AUC values, permutation testing was performed, in which word and within-word syllable boundary labels were randomly shuffled and the same logistic regression procedure was applied.

Neural encoding of word boundary

To determine whether speech responsive electrodes showed a significant difference between boundary events (within-word syllable boundaries and word boundaries), we assessed the duration of contiguous time points for which the neural response (100Hz) time aligned 0.2s before and 0.4s after word boundary events significantly differed from responses aligned to within-word syllable boundaries. To test for significance we used a one-way analysis of variance (ANOVA) F-test at each time point, Bonferroni

corrected for multiple comparisons using a threshold of $p<0.01$. We computed this for each electrode and each language separately such that every electrode was associated with the duration of significant differences value for each language.

Neural word boundary decoding consisted of a logistic regression trained to classify whether the contiguous window of neural activity 0.2s before and 0.4s after a boundary corresponded to a within-word syllable boundary or a word boundary. To compare against the acoustic word boundary decoding, all single syllable words and words at sentence onset were removed from this analysis. The size of the window used in neural decoding was larger than that of the acoustic decoding (0.2s) because the neural response dynamics are slower in latency than the acoustic dynamics. In Spanish, we used approximately 1300 within-word syllable onsets and 600 word onsets. In English, we used approximately 1150 within-word syllable onsets and 750 word onsets.

To perform neural word boundary decoding, neural responses (sampled at 100Hz) from all speech responsive electrodes from a subset of participants were concatenated into a matrix of the form (electrode x time x event), which was then vectorized across the electrode and time dimensions. The subset of participants that had the maximum overlap in trials were used for neural word boundary decoding. PCA was performed on the vectorized spatio-temporal neural data to reduce the total dimensionality, the components with the highest variance that cumulatively explained over 90% of the total variance were ultimately used as input to the regression; this was approximately 800-1000 PCs depending on the participant group. **Table 3.6** summarizes the number of {unique participants, electrodes, PCs} used to train each decoder.

Logistic regression was run using standard MATLABR2024b regression functions (fitclinear) with Lasso (L1) regularization and 20-fold cross-validation. The AUC, or the area under the receiving operating curve, was extracted for each held-out test fold. Beta linear coefficient estimates were extracted from the

trained ClassificationLinear MATLABR2024b model and the weights of single electrodes were calculated by thresholding the percentile weight values and converting from PC space back to the original (electrode x time) space.

To characterize the temporal dynamics of decoding performance (**Fig. 3.14**), we performed sliding-window neural word boundary decoding. To do this, we used the same set of speech responsive electrodes as for the fixed 0.6s window decoding, but used a sliding window of 0.02s within this 0.6s of neural response to perform the decoding. The same trials and electrodes were used for decoding across all 0.02s windows within the same participant group and language.

We also performed single participant neural word boundary decoding (**Fig. 3.14**) using all speech responsive electrodes per participant. Single participant word boundary decoding was calculated similarly to the participant group decoding (with the same time window, with PCA) but included electrode responses from a single participant only. The number of electrodes included in the decoding varied by participants (electrodes: median = 47, range = {21, 79}). For statistical analysis comparing decoding performance across speech conditions in single participants (**Fig. 3.14C**), a paired Wilcoxon signed-rank test was used.

For the single participants that spoke English and at least one other language (**Fig. 3.5**), we followed the same procedure as above for calculated English word boundary decoding AUC. To assess the significance of the effect of English proficiency on neural word boundary classification performance (in English), a linear model was used,

$$AUC \sim proficiency \ + (1 \mid number\ of\ electrodes)$$

Acoustic Ambiguity Index (AAI)

The acoustic ambiguity index (AAI) is based on the acoustic word boundary decoder that takes in the spectrogram values using 0.2s around each boundary event and outputs whether the trial corresponds to a word boundary or a within-word syllable boundary. As above, single syllable words and words at sentence onset were removed from this analysis. For each single trial, the decoder outputs both a predicted label (0, 1) and a posterior probability score (0-1). The scores of the held-out test trials are extracted as one of the outputs of the MATLABR2024b ClassificationLinear model predict function. The per-trial value: abs(correct label - posterior probability score) is then calculated to produce a continuous measure of how accurate the classifier is on single trials, where larger values indicate that the classifier is likely predicting incorrectly. The corresponding neural word-boundary decoding is produced after, 1) calculating the difference between the probability scores for single trials from neural decoding and correct label and 2) comparing this distribution between known and foreign groups for each bin.

Extended Figures and Tables

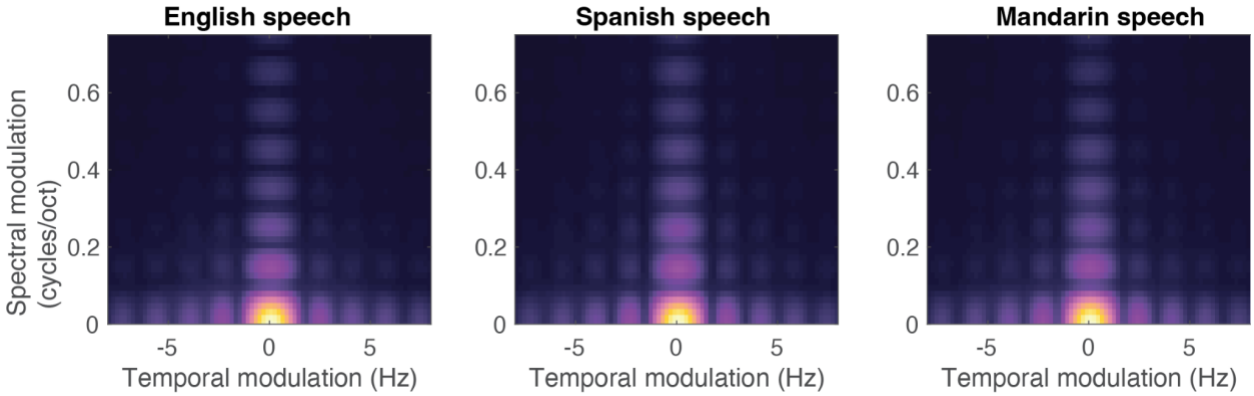


Fig 3.6. Similar modulation spectrum across speech corpora.
Highly similar modulation spectra (qualitative) for English speech stimuli (TIMIT), Spanish speech stimuli (DIMEx), and Mandarin speech stimuli (ASCCD).

Table 3.1. Spanish proficiency, frequency of use, and age of acquisition for self-reported English monolingual participants.
Self-rated proficiency profiles of Spanish for English monolingual participants. (L1: First learned language). Proficiency scale as: 1- single words, 2-basic conversation, 3-fluent, 4-very good, 5- native. Frequency of use scale as using this language: 1-daily, 2-weekly, 3-monthly, 4-irregular, 5-never.

ID	Hemisphere	Age	Handedness	Gender	L1	Spanish Proficiency	Spanish Frequency of Use	Learned Spanish at Age
E01	L	49	R	F	English	0		
E02	L	37	R	F	English	3	irregular	14
E03	L	23	R	M	English	1	irregular	14
E04	L	47	A	F	English	1	irregular	18
E05	R	29	R	M	English	0		
E06	L	31	R	F	English	0		
E07	L	47	R	F	English	1	monthly	14
E08	R	36	R	F	English	3	daily	13

Table 3.2. English proficiency, frequency of use, and age of acquisition for self-reported Spanish and Mandarin monolingual participants.
Self-rated proficiency profiles of English for Spanish and Mandarin monolingual participants. (L1: First learned language, NC: Not collected). Proficiency scale as: 1- single words, 2-basic conversation, 3-fluent, 4-very good, 5- native. Frequency of use scale as using this language: 1-daily, 2-weekly, 3-monthly, 4-irregular, 5-never.

ID	Hemisphere	Age	Handedness	Gender	L1	English Proficiency
S01	L	28	R	F	Spanish	0
S02	R	45	L	F	Spanish	0
S03	L	35	NC	M	Spanish	0
S04	R	59	R	M	Spanish	0
S05	R	23	R	M	Spanish	0
S06	L	39	L	F	Spanish	0
S07	L	NC	R	M	Spanish	0
S08	R	36	R	M	Spanish	0
S09	L	41	R	M	Spanish	0
M01	L	NC	R	M	Mandarin	0
M02	L	51	R	M	Mandarin	0
M03	L	46	R	F	Mandarin	0

Table 3.3. Spanish and English proficiency, frequency of use, and age of acquisition for self-reported Spanish-English bilinguals.
Self-rated Spanish and English proficiency profiles of Spanish-English multilingual participants (Hem: hemisphere, Hand: Handedness, Gend: Gender, Dom: Dominant Language, FoU: Frequency of Use, AoA: Age of Acquisition, Span: Spanish, Eng: English, Cat: Catalan, NC: Not collected). Proficiency scale as: 1- single words, 2-basic conversation, 3-fluent, 4-very good, 5- native. Frequency of use scale as using this language: 1-daily, 2-weekly, 3-monthly, 4-irregular, 5-never.

ID	Hem	Age	Hand	Gend	Dom	L1	L2	Spanish Prof	Spanish FoU	Spanish AoA	English Prof	English FoU	English AoA
B01	L	28	R	M	Span	Span	Eng	5	daily	1	5	daily	1
B02	R	33	R	F	Eng	Span	Eng	5	daily	1	5	daily	5
B03	R	27	A	F	Eng	Span	Eng	3	daily	1	5	daily	1
B04	R	23	R	M	Span	Span	Eng	5	daily	1	3	irregular	5
B05	L	23	R	M	Span	Span	Eng	5	daily	1	4	daily	3
B06	L	22	R	M	Eng	Span	Eng	3	daily	3	5	daily	3
B07	L	26	R	F	Span	Span	Eng	5	daily	1	5	daily	5
B08	L	32	R	M	Span	Span	Eng	5	daily	1	5	daily	4

Table 3.4. English proficiency, frequency of use, and age of acquisition for participants with diverse language backgrounds.
Self-rated English proficiency profiles of diverse multilingual participants (Hem: hemisphere, Hand: Handedness, Gend: Gender, Dom: Dominant Language, FoU: Frequency of Use, AoA: Age of Acquisition, Span: Spanish, Eng: English, Cat: Catalan, Port: Portuguese, Mand: Mandarin, Kor: Korean, Ara: Arabic, Russ: Russian, NC: Not collected). Proficiency scale as: 1- single words, 2-basic conversation, 3-fluent, 4-very good, 5- native. Frequency of use scale as using this language: 1-daily,2-weekly, 3-monthly, 4-irregular, 5-never.

ID	Hem	Age	Hand	Gend	Dom	L1	L2	L3	English Prof	English FoU	English AoA
D01	L	NC	R	F	Ara	Ara	Fren	Eng	2	daily	15
D02	L	33	R	F	Russ	Russ	Eng		4	daily	8
D03	L	NC	NC	F	Kor	Kor			0		
D04	L	53	R	M	Mand	Mand	Eng		4	daily	15
D05	Bil	22	R	M	Port	Port	Span	Eng	5	daily	5
D06	R	42	R	F	Span	Span	Catal	Eng	3	irregular	8

Table 3.5. Summary of participant groups in each analysis.
Participants cohorts included in each analysis within each of the main figures.

Figure	Analysis	Participant Groups
Fig. 1	Comparison of speech-responsive sites (B-E), Acoustic-phonetic TRF analysis (F-H)	Spanish (S01-S09) English (E01-E08) Mandarin (M01-M03)
Fig. 2	Comparison of word-level and phoneme surprisal TRF analysis (B-G)	Spanish (S01-S09) English (E01-E08)
Fig. 3	Word vs. syllable boundary differences across conditions (D-F)	Spanish (S01-S09) English (E01-E08)
Fig. 4	Exploration of word vs. syllable boundary decoding (A-D)	Spanish (S01-S09) English (E01-E08)
Fig. 5	Comparing acoustic-phonetic and word-level TRF analysis in two known languages (multilingual; A-F)	Spanish-English Bilingual (B01-B08) Spanish (S01-S09) English (E01-E08)
Fig. 5	Word vs. syllable boundary decoding as a function of English proficiency (H-I)	Spanish-English Bilingual (B01-B08) Mandarin (M01-M03) Diverse languages (D01-D06)

Table 3.6. Summary of participants, electrodes, PCs used in each decoder.
The number of participants, PCs, and trials included in each neural word boundary classification run across speaker types (Spanish, English monolinguals) and speech condition (English, Spanish speech).

	Spanish monolingual	English monolingual
Spanish speech	{4, 189, 1005}	{4, 174, 952}
English speech	{3, 141, 879}	{5, 179, 936}

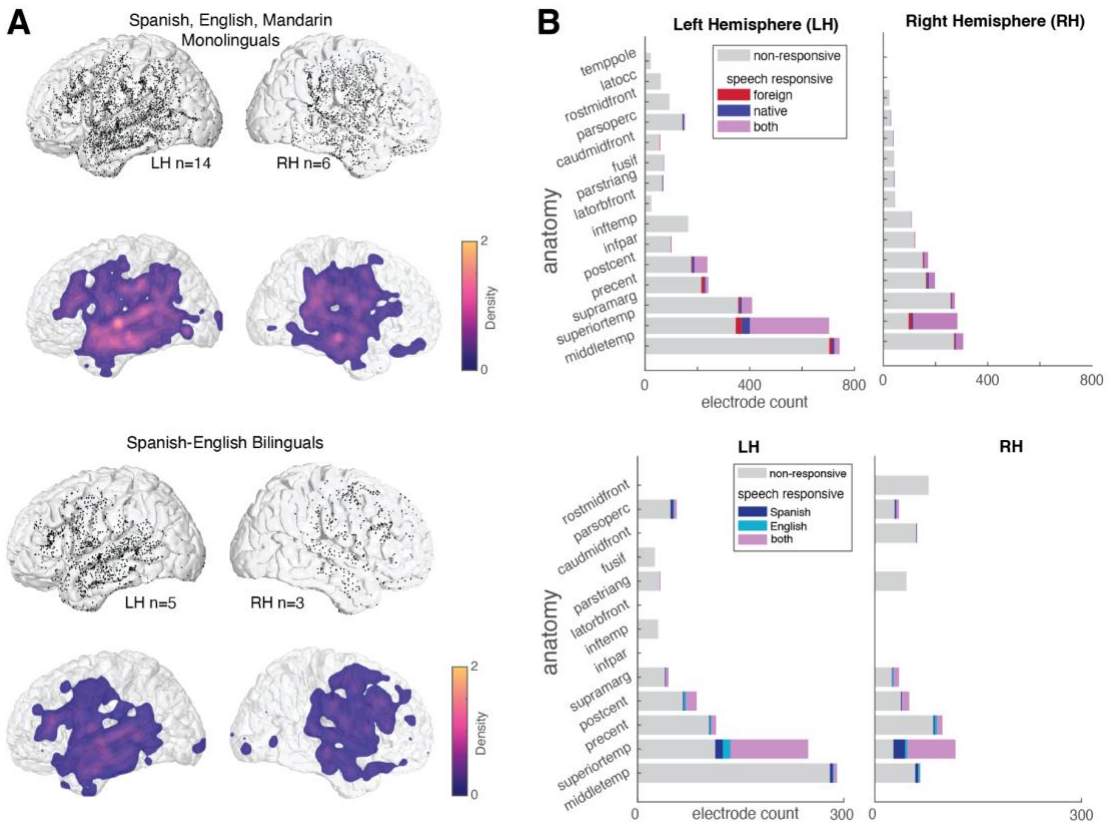


Fig 3.7. High density lateral surface coverage of the human temporal lobe in English, Spanish, Mandarin speakers and Spanish-English bilinguals.
A. Wide and dense coverage of the superior temporal lobe across left and right hemispheres across participants. Anatomical location (MNI) of all electrodes plotted prior to subselecting for speech responsive electrodes. **B.** Highest counts of electrodes in the left and right hemisphere superior temporal and middle temporal cortices bilaterally. Highest counts of speech-responsive electrodes in superior-temporal cortex. In Spanish, English and Mandarin monolinguals, a majority of speech-responsive electrodes respond to both foreign and native speech, while a minority of electrodes show selective responses to either foreign or native speech. Similarly, in Spanish-English bilinguals, a majority of speech-responsive electrodes respond to both English and Spanish speech.

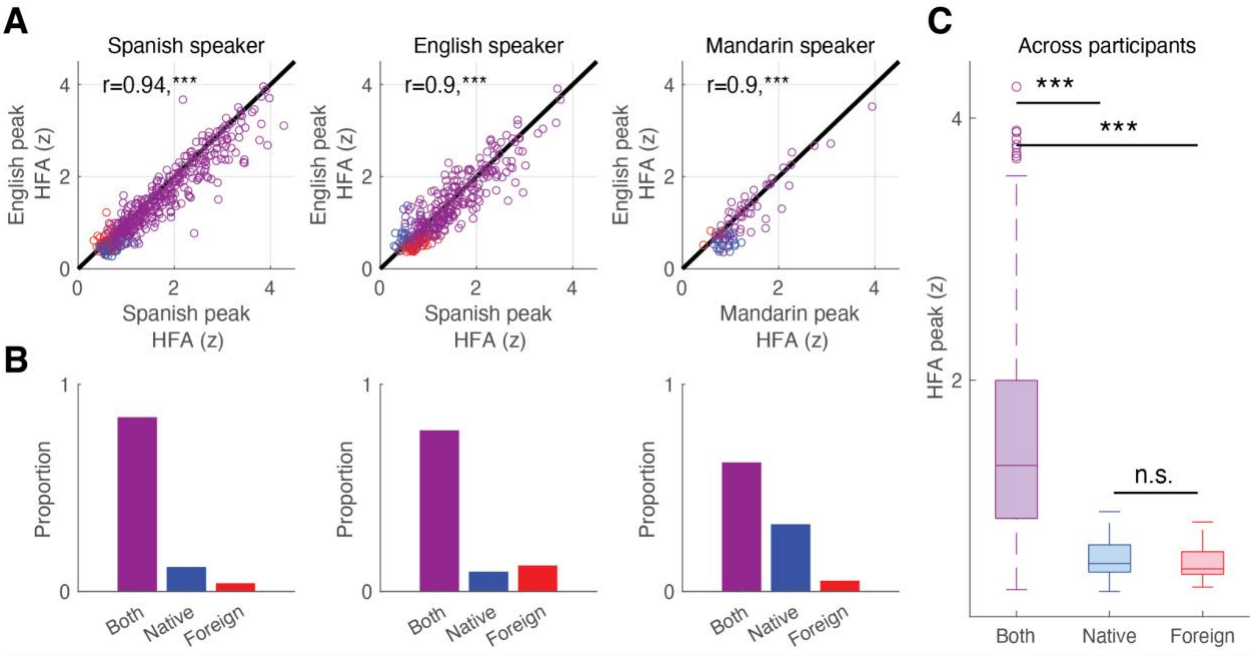


Fig 3.8. Similar peak HFA response across native and foreign speech conditions.
A. Across Spanish, English, and Mandarin speakers, a majority of electrodes are responsive to both native and foreign speech conditions (purple) as opposed to only native (blue) or only foreign (red) speech. Each scatter point corresponds to a single electrode. HFA peak response is calculated as the peak magnitude of the average response across all sentences in the same language. Average HFA peak magnitudes are correlated across speech conditions (Pearson $r(472)=0.94$; $r(334)=0.9$; $r(77)=0.9$, ***$p<0.001$ for all participant groups). **B.** Spanish, English, and Mandarin speakers show no consistent bias in the proportion of electrodes that are only responsive for only either native or foreign speech conditions (blue, red bars). **C.** Speech responsive electrodes to one language but not the other showed significantly smaller HFA amplitude responses. Electrodes responsive to only foreign or native speech showed no consistent bias in HFA amplitude and no consistent anatomical location (data not shown). Two-sided Wilcoxon rank sum test ($n=[705; 113; 65]$) determined significance between the distribution of HFA peak magnitudes across groups (***$p<0.001$). Box plots show the maximum and minimum values (whiskers), median (center line) and the 25th to 75th percentiles (box limits).

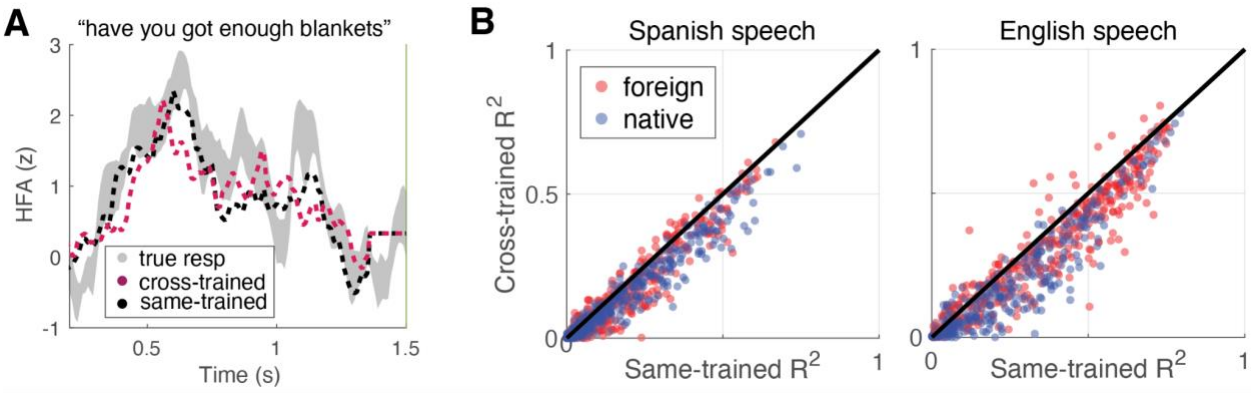


Fig 3.9. Same and Cross-language Model Predictions.
A. Highly similar predicted neural response to an example English sentence from a model trained on another language (Spanish) as opposed to the same language (English). Dashed lines indicate a TRF predicted response. **B.** Cross-language TRF prediction performance (e.g. training on Spanish speech, testing on English speech) is correlated with same-language prediction TRF performance across both Spanish (Pearson r(883)=0.96, ***p<0.001) and English speech (Pearson r(883)=0.94, ***p<0.001).

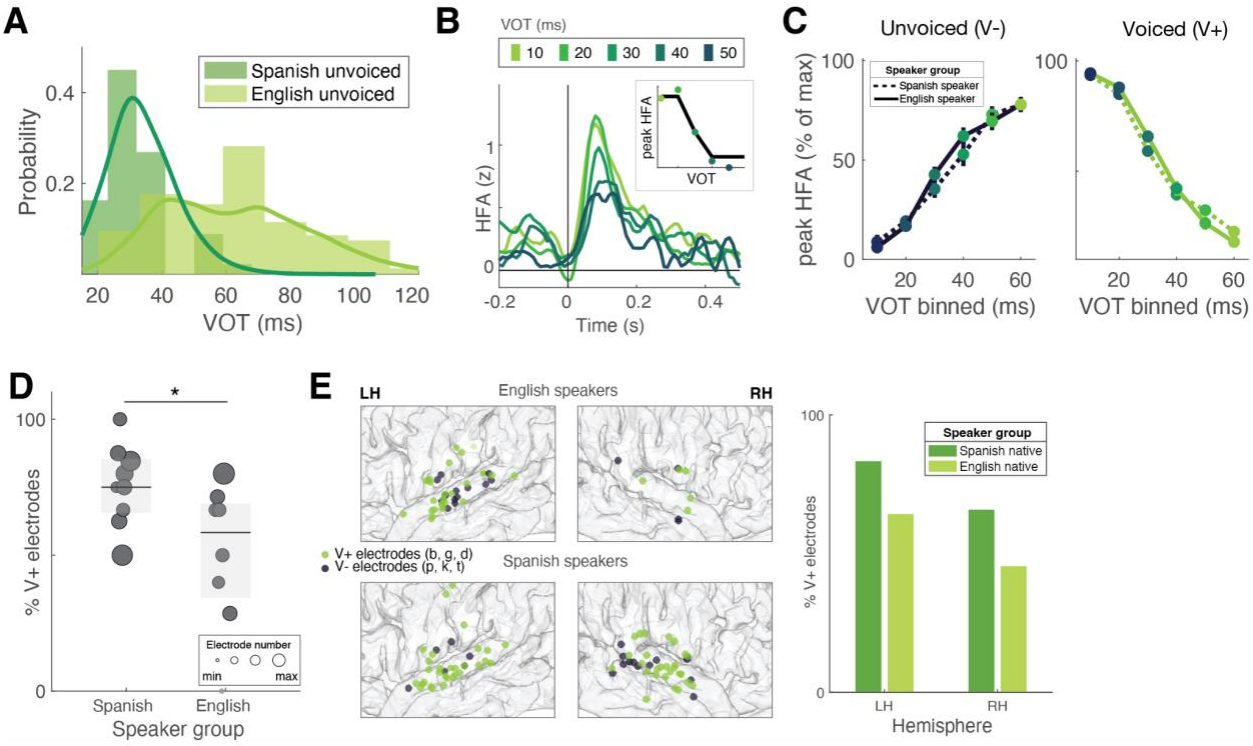


Fig 3.10. Voice Onset Time Encoding Across Languages.
A. VOT values of unvoiced stop-consonants in the Spanish (dark green) and English (light green) show systematic differences in duration. This systematic difference along with prior research on the neural encoding of voice onset time (VOT) suggest that language experience may influence how sounds are processed in the STG. **B.** Electrodes are sensitive to different VOT values (0:10:60ms). The top-right inset shows the VOT tuning function for this electrode, where each scatter point corresponds to the peak HFA response to the VOT bin. **C.** Consistent with prior research, average single electrode tuning curves either showed positive (V-, preferring unvoiced) or negative (V+, preferring voiced) tuning direction. We found no significant difference between tuning curves across VOT sensitive electrodes in Spanish and English speakers (two-tailed Wilcoxon rank sum test, V+ n=[57, 32], V- n=[20, 21]). Error bars indicate standard error of the mean across electrodes within each VOT bin for each speaker group. **D.** Proportion of V+ electrodes was significantly higher ($p<0.05$) in Spanish speakers as opposed to English speakers. Size of each scatter corresponds to the number of electrodes within each participant that show significant tuning for VOT. P-value corresponds to language background in a linear mixed effect model ($t(128)=2.11$, SE= 0.0176, $p=0.0367$). Box plots show the median (center line) and the 25th to 75th percentiles (box limits). **E.** Intermixed V+ and V- electrodes in English and Spanish speakers. Both left and right hemispheres showed a higher proportion of V+ electrodes in Spanish speakers, suggesting that language experience effects may arise at the population level(Chang et al., 2010). While multiple overlapping cues present in natural continuous speech may obscure language-specific contrastive features, these effects can be isolated in controlled stimuli(Hitczenko & Feldman, 2022).

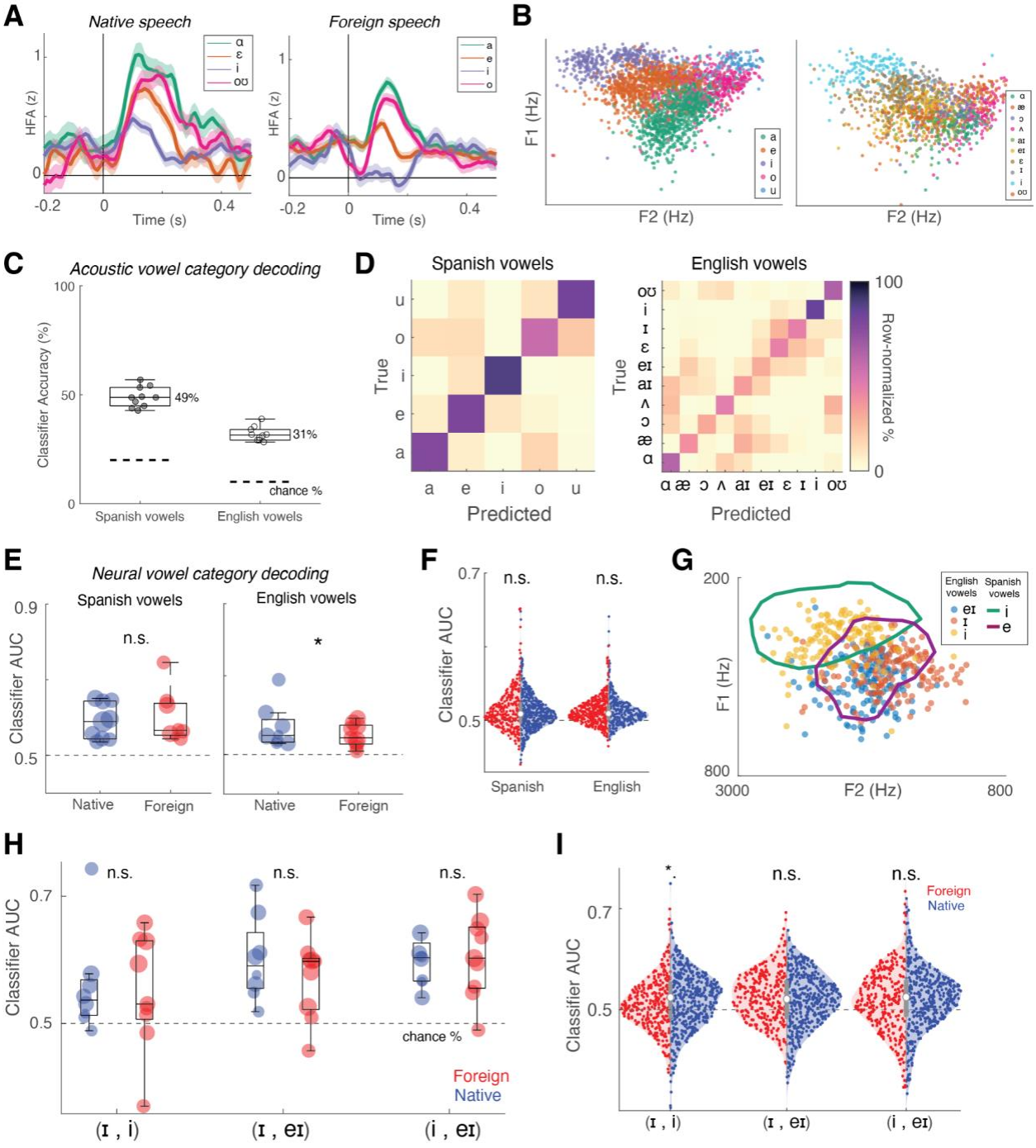


Fig 3.11. Vowel Category Encoding Across Languages.
A. Example electrode response to approximately equivalent vowel categories across languages (English, Spanish) shows qualitatively similar response patterns in native and foreign speech conditions. Shaded patches indicate standard error of the mean HFA response across vowels. **B.** Formant frequencies (Hz) from each vowel instance are shown in the scatter and colored by the vowel category they belong to. Spanish (left) and English (right) vowel spaces include a different number of vowels, and contain different formant space distributions for each category. **C.** Acoustic vowel category decoding (10-fold linear discriminant analysis) accuracy using the formant values of vowel sounds across English and Spanish speech corpora. *(caption continues on the next page)*

(caption continued from previous page) Each scatter point represents 1/10 validation folds. Chance accuracy was calculated based on the number of vowel categories included from the particular language (Spanish: 1/5, English 1/10). Vowel categories with more than 100 instances in the speech corpus were included. Box plots show the maximum and minimum values (whiskers), median (center line) and the 25th to 75th percentiles (box limits). **D.** Confusion matrix resulting from acoustic vowel category decoding exhibits a clear diagonal, with off-diagonal confusions reflecting overlap in F1/F2 space (e.g. /a/ and /e/). E. Single participant neural vowel category decoding (10-fold linear discriminant analysis; each scatter corresponds to a single participant) separated by speech condition shows no significant difference in Spanish (LME t(168)=0.49, SE=0.0198, p=0.63, see "Methods" for formula) and a weak effect in English (LME t(168)=1.98, SE=0.0095, *p=0.049). Box plots show the maximum and minimum values (whiskers), median (center line) and the 25th to 75th percentiles (box limits). F. Single electrode neural vowel category decoding (10-fold linear discriminant analysis; each scatter corresponds to a single electrode) separated by language shows not significant effect of speech condition (two-sided Wilcoxon rank sum test Spanish n=[453, 302], p=0.23); English n=[292, 416], p=0.86). **G.** Vowel formants (F1, F2) for instances of "i", "I", and "eɪ" in English are subsumed by only two vowel categories in Spanish, "i" and "e". 95th percentile contours for Spanish vowel formants shown as colored lines. **H.** Pairwise neural decoding AUC (10-fold logistic regression) for three English vowel categories, "i", "I", and "eɪ" shows no significant difference between participant groups (LME t(168)=-0.029, p= 0.98; t(134)=-0.87, p=0.38; t(134)=0.84, p=0.40, see "Methods" for formula). Each scatter point corresponds to the mean AUC for a single participant (colored by speech condition). Box plots show the maximum and minimum values (whiskers), median (center line) and the 25th to 75th percentiles (box limits). **I.** Pairwise neural decoding AUC (10-fold logistic regression) for the same English vowel categories, shows no significant difference between single electrodes across participant groups. Each scatter point corresponds to the mean AUC for a single electrode (two-sided Wilcoxon rank sum test n=[416,292], p=0.02; p=0.08; p=0.24).

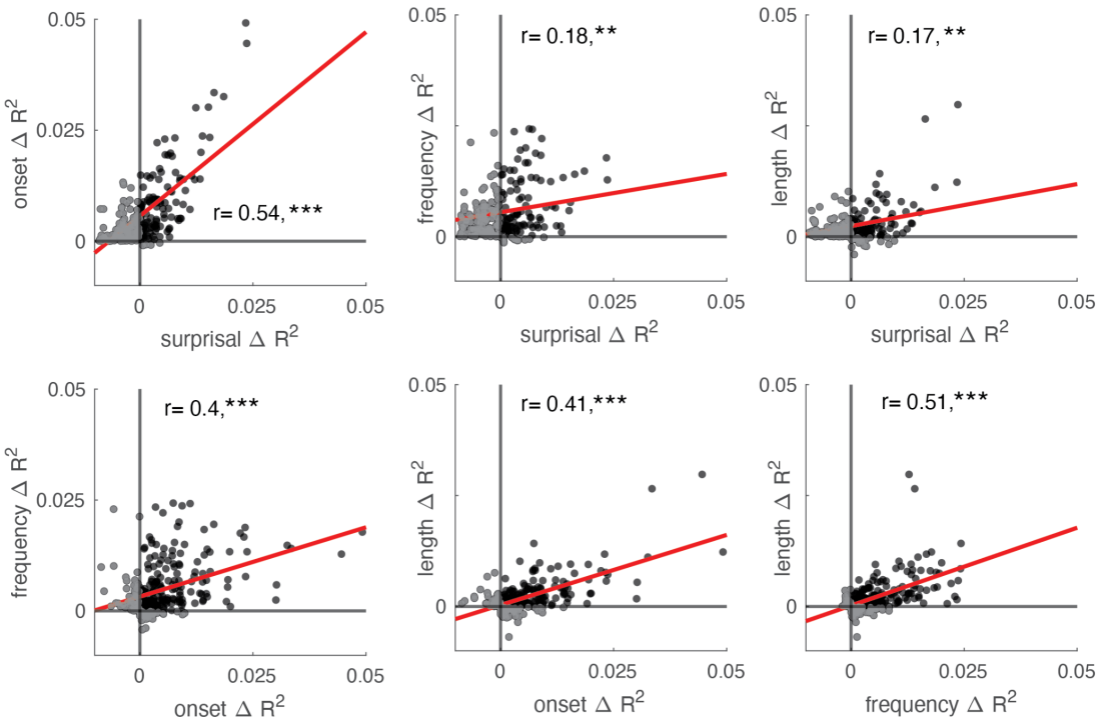


Fig 3.12. Comparing Word and Phoneme Surprisal Unique Variances.
Significant correlations between positive unique variance values for word onset vs. surprisal (Pearson r(257)=0.54, ***p<0.001), frequency vs. surprisal (Pearson r(268)=0.18, **p=0.0033), length vs. surprisal (Pearson r(250)=0.17, **p=0.009), word onset vs. frequency (r(303)=0.4, ***p<0.001), word onset vs. length (Pearson r(301)=0.41, ***p<0.001), and length vs. frequency (r(293)=0.51, ***p<0.001) in the native speech condition.

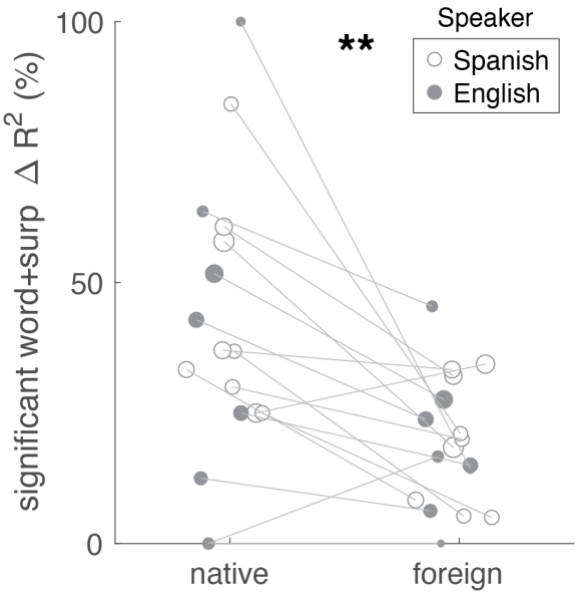


Fig. 3.13. Proportion of Word-level Electrodes Across Conditions.
Proportion of electrodes with significant word and phoneme surprisal unique variance for single participants across native and foreign speech conditions. Significance determined by a 1-tailed paired Wilcoxon signed-rank test (n=17, **p=0.0012).

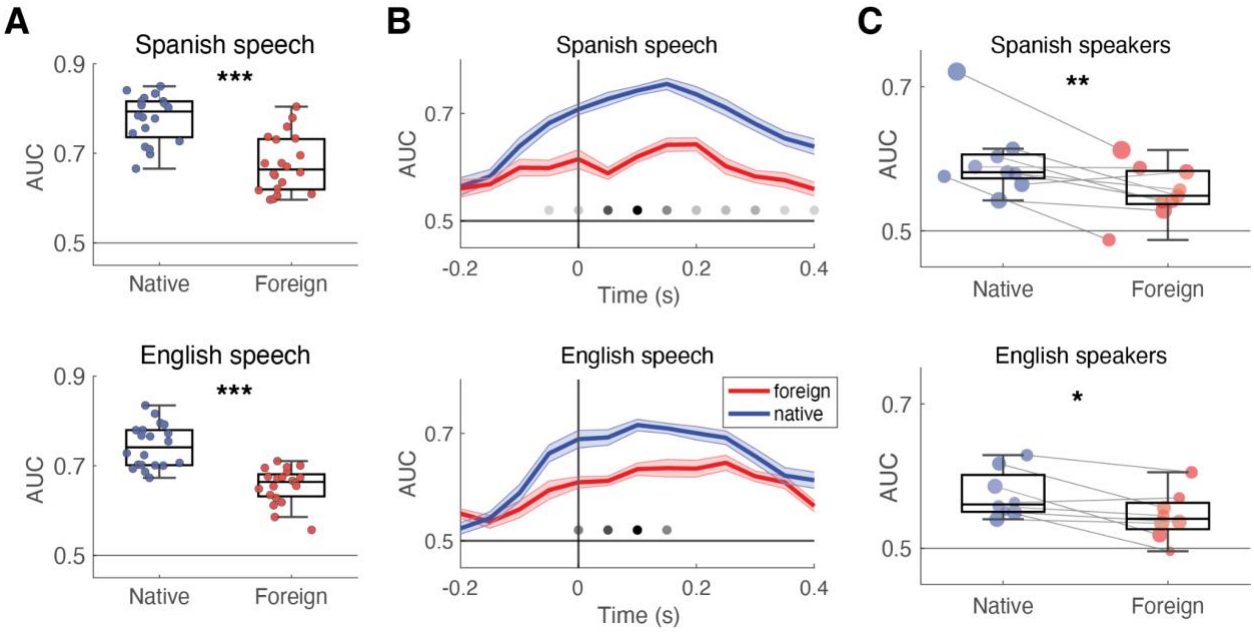


Fig. 3.14. Characterization of Neural Word Boundary Decoding.
A. Neural classification of word boundaries is significantly better in native vs. foreign speech (each scatter point corresponds to the AUC of 1/20 folds shown separately for English and Spanish; 1-tailed 2-sample t-test ***p<0.001). Box plots show the maximum and minimum values (whiskers), median (center line) and the 25th to 75th percentiles (box limits). **B.** Temporally-resolved neural classification of word boundaries (increments of 0.05s and 0.2s before to 0.4s after boundary, mean across 20 folds) is significantly better in native vs. foreign speech approximately 0.1s after the boundary event (per-time point two-sided ANOVA p<0.01 with Bonferroni correction show significant time-points as black scatter points). Continuous shaded patches indicate standard error of the mean across decoding folds for each time point. **C.** Neural classification of word boundaries is significantly better in native vs. foreign speech when considering single participant decoding performance (each scatter corresponds to median AUC across 20 folds per participant separated by group, size corresponds to the number of electrodes used for classification). P-values derived from a 1-tailed paired Wilcoxon signed-rank test (Spanish speakers n=9, **p=0.0098; English speakers n=8, *p=0.012). Box plots show the maximum and minimum values (whiskers), median (center line) and the 25th to 75th percentiles (box limits).

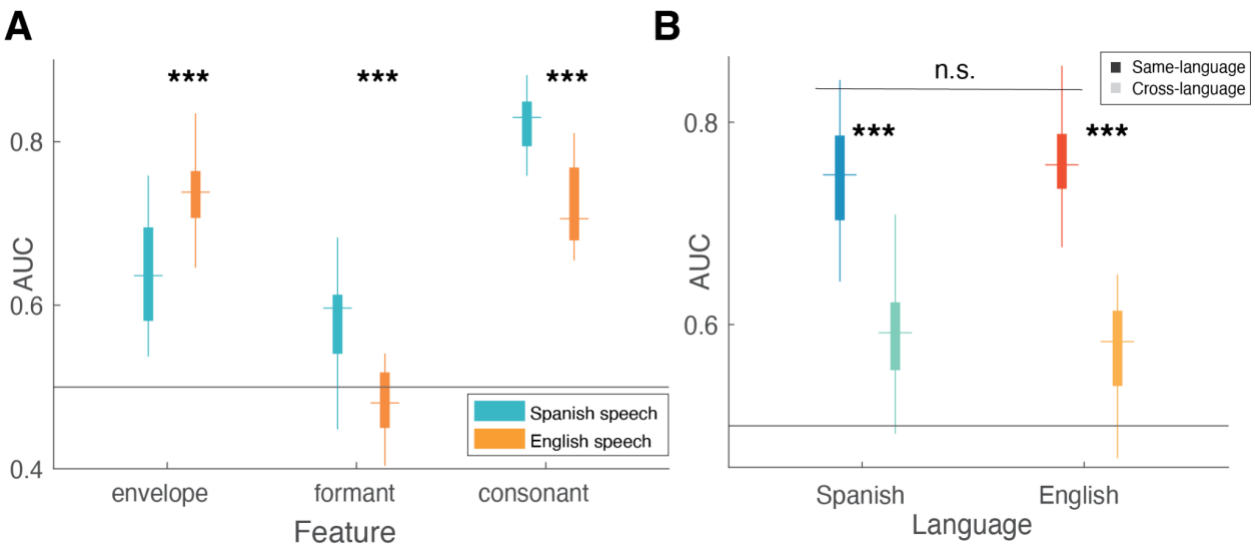


Fig. 3.15. Characterization of the Acoustic Cues to Word Boundary.
A. In Spanish and English speech, different acoustic-phonetic features cue word and syllable boundaries. We used three distinct acoustic-phonetic features: continuous envelope, continuous formants, and consonant features around each boundary event to classify word vs. syllable boundary. All features we tested show significantly different AUC across English and Spanish (2-tailed Wilcoxon signed-rank test, for all features n=20, ***p<0.001). Box plots show the maximum and minimum values (whiskers), median (center line) and the 25th to 75th percentiles (box limits). **B.** Acoustic classification of word boundary trained and tested on the spectrogram from the same language shows significantly higher AUC than a classifier trained on one language and tested on the other (2-tailed Wilcoxon signed-rank test, same-language vs. cross-language n=20, ***p<0.001; English vs. Spanish n=20, p=0.37), suggesting acoustic cues to word boundaries are distinct in Spanish and English. Box plots show the maximum and minimum values (whiskers), median (center line) and the 25th to 75th percentiles (box limits).

Chapter 4: An integrative, neural model of phonological processing

Speech perception, like visual recognition, is a complex process that involves multiple distinct and overlapping levels of representation that unfold across time and space in the brain. Some of the dominant models of speech processing have been inspired by hierarchical and feedforward neurobiological models of the visual system (**Fig. 4.1A**, for example see (Blumstein, 2009). For instance, two dual stream models (Hickok & Poeppel, 2004, 2007a; Rauschecker & Scott, 2009) propose parallel systems in which the ventral pathway operating along the temporal lobe controls the mapping from acoustic signal to word or meaning while the dorsal stream enables the sensorimotor transformations necessary for articulation. In the Hickock & Poeppel (H/P) model, a ventral processing stream begins with spectrotemporal analysis in the bilateral STG, then phonological processing in the bilateral superior temporal sulcus (STS), and lexical processing in the middle and inferior temporal gyri.

In contrast, in the Rauschecker & Scott (R/S) model, spectrotemporal analysis is thought to occur in the posteromedial primary auditory cortex and auditory word-form recognition is achieved as activity flows anterior and lateral towards the anterior STG. The R/S ventral processing stream runs parallel to the visual ventral stream in the infratemporal cortex (Jasmin et al., 2019; Rauschecker & Scott, 2009). Despite being highly influential, these models propose very different anatomical trajectories of ventral stream cortical processing. This question remains unresolved as few studies have documented a clear, causal transformation from sound to lexical representation along either trajectory. To our knowledge, there is no direct evidence that hierarchically organized linguistic representations (e.g. phonemes, syllables, morphemes, words) are mapped onto adjacent cortical regions.

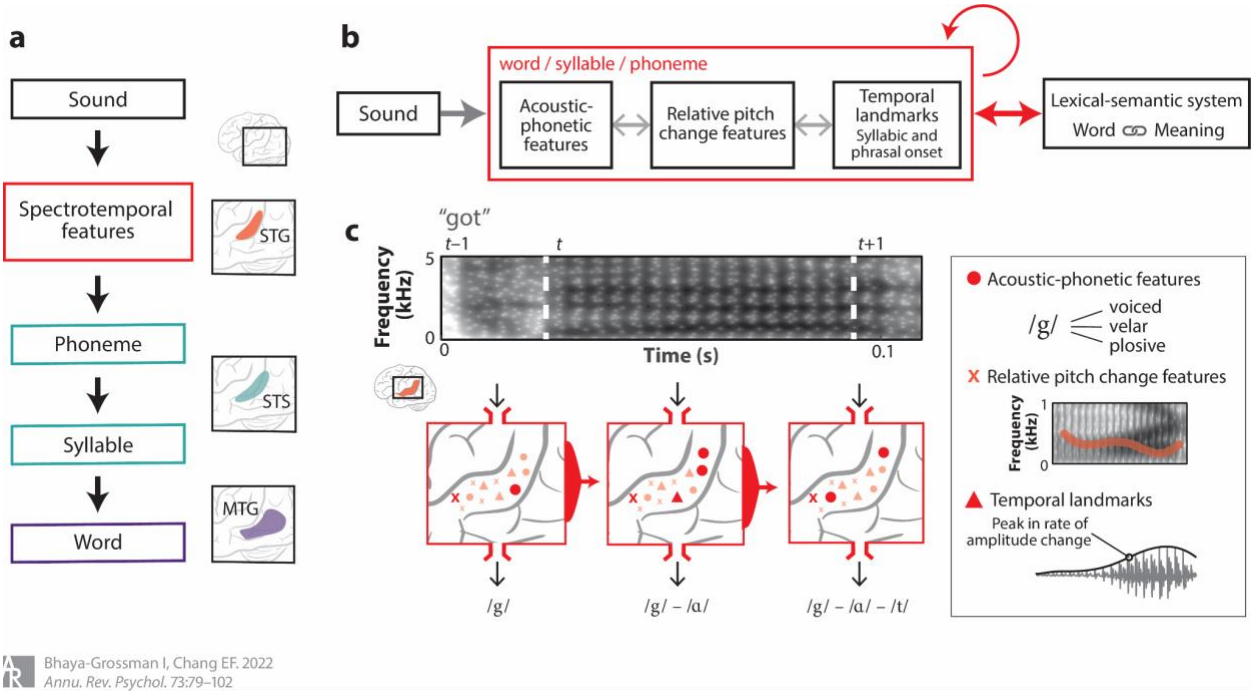


Figure 4.1. A new model of phonological analysis.
A. Classical model of auditory word recognition in which primarily serial, feed-forward, hierarchical processing takes place. The first processing step is spectrotemporal analysis through which relevant features are extracted. Spectro-temporal features are grouped into phonemic segments which are then sequentially assembled into syllables. Finally, the distributed lexical-semantic system links the auditory word form and meaning level content. In the classic models of auditory word recognition, each processing step is assigned to an approximate anatomical location (the schematic here shows one example of these assignments). The neural representation of speech becomes increasingly higher-order (e.g. spectral features to phonological sequence to lexical item) as it is moved through successive brain areas. **B.** An alternative recurrent, multi-scale, and interactive model of auditory word recognition that more closely aligns with the presented neurophysiological evidence. Acoustic signal inputs are analyzed concurrently by local processors with selectivity for acoustic phonetic features, salient temporal landmarks (e.g., peakRate), and prosodic features that occur over phonemic segments. The light gray bidirectional arrows indicate that local processors interact with one another. Recurrent connectivity indicates an integration of temporal context and sensitivity to phonological sequences by binding inputs over time during word processing. Anticipatory top-down, word-level information arises from the lexical-semantic system and the internal dynamics of ongoing phonological analysis. **C.** Three local neuronal populations (*circles*, *triangles*, and *crosses*) on the STG encode relative (speaker-normalized) formant values, relative pitch changes, and the magnitude of peakRate events. In addition to being functionally diverse, these populations likely show distinct electrophysiological signatures (i.e., sustained versus rapid responses). The encoding of normalized spectral content (formants and pitch) suggests the presence of a context-sensitive mechanism that enables rapid retuning to speaker-specific spectral bands. Together, this set of neural responses and the responses at the previous time step define a neural state from which the appropriate word form can be decoded. Every sound segment is processed by the STG in a highly specific context that is sensitive to both temporal and phonological information.

Neural activity in spatially distinct regions of the STG has been shown to be sensitive to specific spectrotemporal fluctuations (Hullett et al., 2016; Santoro et al., 2017; Schönwiesner & Zatorre, 2009).

Nonetheless, we have also established through the results presented in this dissertation that the STG is crucially involved in complex phonological processes that cannot be accounted for by a purely spectrotemporal filter, namely phonetic-acoustic categorization, speaker normalization, and sequence and word-level segmentation. In the H/P model, phonological processes are assumed to take place largely in the bilateral STS. In the R/S model, phonological-relevant encoding is observed through middle and anterior STG. While in these models, information is assumed to travel from primary auditory to non-primary auditory cortex, recent studies have found that transient functional lesioning of the primary auditory cortex via electrical stimulation or focal ablation of the primary auditory cortex does not impair speech comprehension (Hamilton et al., 2020). The same stimulation procedure targeting STG impairs speech comprehension without impairing tone discrimination (Boatman, 2004). The double dissociation between primary and nonprimary auditory cortices as well as the highly distributed nature of speech feature encoding throughout STG pose significant challenges to prevailing anatomically defined hierarchical stream models of cortical processing.

Current neurobiological models of speech processing rely on assigning a psychological or linguistic level of representation to a given brain region. However, as is clear from evidence presented in this dissertation, the functional organization of brain computations need not align with the levels of representation assumed in linguistics and psychology. While there is little evidence for a dedicated cortical processing stream from phonemes to syllables to words, there do exist important functional subdivisions of the STG, for example that correspond to temporal landmarks at different timescales described previously (Yi et al., 2019a).

As an alternative to the models described above, we seek to describe one that is explanatory but does not have strong commitments to predefined linguistic units (**Fig. 4.1B**). To account for the neurophysiological evidence presented in this dissertation, we propose that the acoustic signal is analyzed concurrently by a set of local processing units with selectivity for acoustic phonetic features, salient

temporal landmarks (e.g. peakRate, onsets), and prosodic features (**Fig. 4.1C**). Language-specific sequence and word boundary responses, which require a dynamic and time-dependent representation of speech sound sequences, can occur if the "top-down" word level information can in real-time modulate phonological analysis. This predictive capacity may be computationally realized through recurrent connections that embed temporal context into the processing state (Elman, 1990; Jordan, 1997).

In contrast to a feedforward stream through cortical areas, the proposed model emphasizes an overlapping, distributed processing of speech that is sensitive to language-specific knowledge and skill (**Fig. 4.2A**). While local cortical sites are categorically tuned to acoustic-phonetic features, larger and more abstract speech elements such as phonemes and syllables are captured in the spatial patterns of activity that occur both within and across these sites. As far as we know, the predictive capabilities of the STG are also embedded within a largely distributed network. These capabilities may be neuro-physiologically implemented through the recurrent connections that exist to link layers within and across cortical columns (Douglas & Martin, 2007; Yi et al., 2019a).

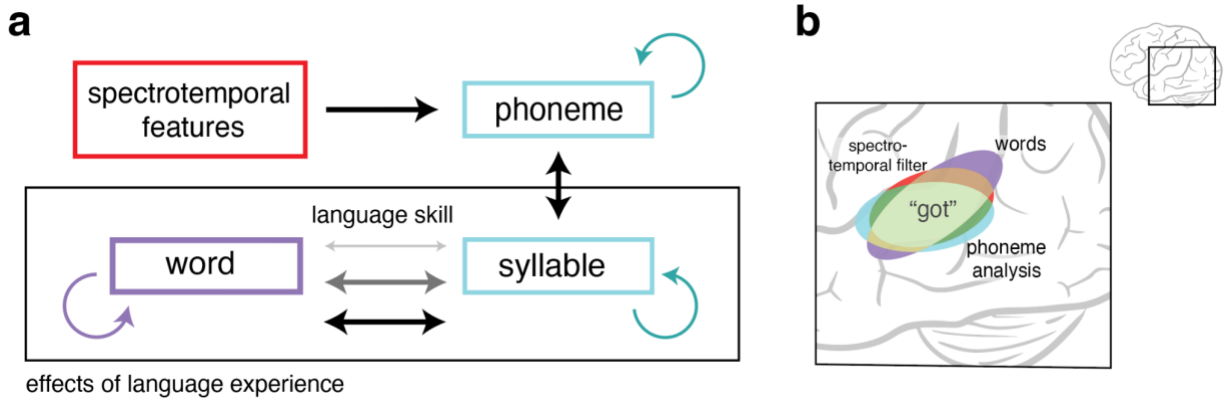


Figure 4.2. Language-specific, integrated phonological analysis.
A. Expanding on the alternative interactive model of auditory word recognition presented in **Fig. 4.1**, we propose that language experience has the greatest effect at the syllable and word-level where the strength of local interaction and recurrent connectivity is dependent on language-specific knowledge and skill. **B.** Depiction of the idea that spectrotemporal filtering, phoneme analysis, and syllable / word-level processing can co-occur all within a single brain region, the STG.

Emerging Principles and Open Questions

Here, we review the key principles that guide phonological computations in the STG, as evidenced by prior work as well as the findings presented in this dissertation. We contend that neural populations in the STG exhibit non-linear and dynamic representations of speech sounds. STG neurons are not passive sound filters or feature detectors, but rather act with "interpretive" function by integrating prior linguistic knowledge and recent temporal context in real time (**Fig. 4.1B, Fig. 4.2A**).

Local acoustic-phonetic tuning exhibit key nonlinearities

STG responses to speech are nonlinear, resulting from a set of transformations that enable the perceptual system to extract linguistic content. In categorization, neural representations show stronger sensitivity to acoustic changes that occur across phonetically-relevant category boundaries, rather than within the same category. In speaker normalization, neural responses become invariant to speaker characteristics that do not contribute meaningfully to phonological content. Chapter 2, characterizing vowel formant representation in the STG, reveals how both categorization and normalization of acoustic-phonetic features may be neurally implemented.

Dynamic integration of language-specific knowledge

STG cortical activity not only represents the instantaneous acoustic content of the speech signal, but is dependent on both language context and language-specific experiences of the listener. Chapter 3, describing aspects of shared and language-specific speech processing across speakers of different languages, reveals how neural responses in the STG are modulated by word-level information only in contexts where speech is comprehended.

Acoustic-phonetic and higher-order linguistic representations co-exist in the STG

While the neural responses recorded from single electrodes represent acoustic-phonetic features, single electrodes and patterns of electrode responses across the STG are also relevant for representing phoneme, syllable, and word-level information, as evidenced in Chapter 3 (**Fig. 2B**). In this way, the neural implementation of phonological analysis is made up of a distribution of functionally heterogeneous units that, in combination, perform the computations required for speech comprehension. However, it remains unclear how perceptual units emerge from this distributed neural code: How are the local neural populations that encode distinct acoustic and linguistic properties (vowels, consonants, peakRate, word boundary) ultimately integrated into a cohesive and experience-based reflection of speech input?

Conclusions

Decades of neuroscientific research have shown that the human STG plays an essential and requisite role in speech perception. It is thus not surprising that prior work and the empirical findings in this dissertation provide evidence for the STG interfacing between auditory and language systems. Several important principles govern patterns of cortical activity in the STG, and suggest that this brain region may be performing more complex speech computations than originally assumed. Technological and experimental advances have enabled researchers to document the character of functionally heterogeneous neural populations in non-primary auditory cortex. Despite this significant progress, there remain many important challenges that lay ahead in order to fully elucidate the neural mechanisms that support speech perception in the STG; these future discoveries will likely have a far-reaching impact in a wide variety of disciplines, including psychology, neuroscience, and linguistics.

References

Abramson, A. S., & Lisker, L. (1972). Perception of voice timing in Spanish stop consonants. *The Journal of the Acoustical Society of America*, *52*(1A_Supplement), 175–175.

Abutalebi, J., Cappa, S. F., & Perani, D. (2001). The bilingual brain as revealed by functional neuroimaging. *Bilingualism (Cambridge, England)*, *4*(2), 179–190.

Abutalebi, J., & Green, D. (2007). Bilingual language production: The neurocognition of language representation and control. *Journal of Neurolinguistics*, *20*(3), 242–275.

Aertsen, A. M. H. J., Olders, J. H. J., & Johannesma, P. I. M. (1981). Spectro-temporal receptive fields of auditory neurons in the grassfrog. *Biological Cybernetics*, *39*(3), 195–209.

Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., & Merzenich, M. M. (2001). Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(23), 13367–13372.

Archibald, J. (2021). Ease and difficulty in L2 phonology: A mini-review. *Frontiers in Communication*, *6*. https://doi.org/10.3389/fcomm.2021.626529

Atiani, S., David, S. V., Elgueda, D., Locastro, M., Radtke-Schuller, S., Shamma, S. A., & Fritz, J. B. (2014). Emergent selectivity for task-relevant stimuli in higher-order auditory cortex. *Neuron*, *82*(2), 486–499.

Beguš, G., Zhou, A., & Zhao, T. C. (2023). Encoding of speech in convolutional layers and the brain stem based on language experience. *Scientific Reports*, *13*(1), 6480.

Benson, R. R., Richardson, M., Whalen, D. H., & Lai, S. (2006). Phonetic processing areas revealed by sinewave speech and acoustically similar non-speech. *NeuroImage*, *31*(1), 342–353.

Benson, R. R., Whalen, D. H., Richardson, M., Swainson, B., Clark, V. P., Lai, S., & Liberman, A. M. (2001). Parametrically dissociating speech and nonspeech perception in the brain using fMRI. *Brain*

and Language, *78*(3), 364–396.

Bertalmío, M., Gomez-Villa, A., Martín, A., Vazquez-Corral, J., Kane, D., & Malo, J. (2020). Evidence for the intrinsically nonlinear nature of receptive fields in vision. *Scientific Reports*, *10*(1), 16277.

Bhaya-Grossman, I., & Chang, E. F. (2022). Speech Computations of the Human Superior Temporal Gyrus. *Annual Review of Psychology*, *73*, 79–102.

Bice, K., & Kroll, J. F. (2019). English only? Monolinguals in linguistically diverse contexts have an edge in language learning. *Brain and Language*, *196*(104644), 104644.

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

Binder, J. R., Rao, S. M., Hammeke, T. A., Yetkin, F. Z., Jesmanowicz, A., Bandettini, P. A., Wong, E. C., Estkowski, L. D., Goldstein, M. D., & Haughton, V. M. (1994). Functional magnetic resonance imaging of human auditory cortex. *Annals of Neurology*, *35*(6), 662–672.

Bitterman, Y., Mukamel, R., Malach, R., Fried, I., & Nelken, I. (2008). Ultra-fine frequency tuning revealed in single neurons of human auditory cortex. *Nature*, *451*(7175), 197–201.

Blumstein, S. E. (2009). Auditory word recognition: evidence from aphasia and functional neuroimaging. *Language and Linguistics Compass*, *3*(4), 824–838.

Blumstein, S. E., Myers, E. B., & Rissman, J. (2005). The perception of voice onset time: an fMRI investigation of phonetic category structure. *Journal of Cognitive Neuroscience*, *17*(9), 1353–1366.

Boatman, D. (2004). Cortical bases of speech perception: evidence from functional lesion studies. *Cognition*, *92*(1-2), 47–65.

Boatman, D., Gordon, B., Hart, J., Selnes, O., Miglioretti, D., & Lenz, F. (2000). Transcortical sensory aphasia: revisited and revised. *Brain: A Journal of Neurology*, *123 (Pt 8)*, 1634–1642.

Boebinger, D., Norman-Haignere, S. V., McDermott, J. H., & Kanwisher, N. (2021). Music-selective neural populations arise without musical training. *Journal of Neurophysiology*, *125*(6), 2237–2263.

Boemio, A., Fromm, S., Braun, A., & Poeppel, D. (2005). Hierarchical and asymmetric temporal

sensitivity in human auditory cortices. *Nature Neuroscience*, *8*(3), 389–395.

Boersma, P., & Weenink, D. (n.d.). *Praat: doing phonetics by computer*. [Computer Program]. Retrieved August 23, 2022, from https://www.fon.hum.uva.nl/praat/

Bonte, M., Hausfeld, L., Scharke, W., Valente, G., & Formisano, E. (2014). Task-dependent decoding of speaker and vowel identity from auditory cortical response patterns. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *34*(13), 4548–4557.

Bonte, M., Valente, G., & Formisano, E. (2009). Dynamic and task-dependent encoding of speech and voice by phase reorganization of cortical oscillations. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *29*(6), 1699–1706.

Bosker, H. R., & Reinisch, E. (2017). Foreign languages sound fast: Evidence from implicit rate normalization. *Frontiers in Psychology*, *8*, 1063.

Bradlow, A. R. (1995). A comparative acoustic study of English and Spanish vowels. *The Journal of the Acoustical Society of America*, *97*(3), 1916–1924.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436.

Cargnelutti, E., Tomasino, B., & Fabbro, F. (2019). Language Brain Representation in Bilinguals With Different Age of Appropriation and Proficiency of the Second Language: A Meta-Analysis of Functional Imaging Studies. *Frontiers in Human Neuroscience*, *13*, 154.

Central Intelligence Agency. (2024). *The CIA world factbook 2024-2025*. Skyhorse Publishing.

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., & Cash, S. S. (2014). Speech-specific tuning of neurons in human superior temporal gyrus. *Cerebral Cortex* , *24*(10), 2679–2693.

Chang, E. F. (2015). Towards large-scale, human-based, mesoscopic neurotechnologies. *Neuron*, *86*(1), 68–78.

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

Chistovich, L. A. (1980). Auditory processing of speech. *Language and Speech*, *23*(1), 67–73.

Cole, R. A., Jakimik, J., & Cooper, W. E. (1980). Segmenting speech into words. *The Journal of the Acoustical Society of America*, *67*(4), 1323–1332.

Crone, N. E., Boatman, D., Gordon, B., & Hao, L. (2001). Induced electrocorticographic gamma activity during auditory perception. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology*, *112*(4), 565–582.

Cuetos, F., Glez-Nosti, M., Barbon, A., & Brysbaert, M. (2011). SUBTLEX-ESP: frecuencias de las palabras espanolas basadas en los subtitulos de las peliculas. *Psicologica: Revista de Metodologia Y Psicologia Experimental*. https://go.gale.com/ps/i.do?id=GALE%7CA406053433&sid=googleScholar&v=2.1&it=r&linkaccess=abs&issn=02112159&p=AONE&sw=w

Cuetos, F., Hallé, P. A., Domínguez, A., & Segui, J. (2011). Perception of Prothetic/e/in# sC Utterances: Gating Data. *ICPhS*, 540–543.

Cuetos, Glez-Nosti, Barbón, & Brysbaert. (n.d.). SUBTLEX-ESP: Spanish word frequencies based on film subtitles. *Psicologica: Revista de Metodologia Y Psicologia Experimental*. https://www.redalyc.org/pdf/169/16923102001.pdf

Cutler, A. (1990). Exploiting prosodic probabilities in speech segmentation. In *Cognitive models of speech processing: Psycholinguistic and computational perspectives* (pp. 105–121). MIT Press.

Cutler, A. (2015a). Lexical Stress in English Pronunciation. In *The Handbook of English Pronunciation* (pp. 106–124). John Wiley & Sons, Inc.

Cutler, A. (2015b). *Native listening*. MIT Press.

Cutler, A., & Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech & Language*, *2*(3), 133–142.

Cutler, A., Dahan, D., & van Donselaar, W. (1997). Prosody in the comprehension of spoken language: a literature review. *Language and Speech*, *40 (Pt 2)*, 141–201.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1983). A language-specific comprehension strategy.

Nature, *304*(5922), 159–160.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1986). The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language*, *25*(4), 385–400.

Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(1), 157–177.

Cutler, A., & Otake, T. (1994). Mora or phonemes? Further evidence for language-specific listening. *Journal of Memory and Language*, *33*(6), 824–844.

Daube, C., Ince, R. A. A., & Gross, J. (2019). Simple Acoustic Features Can Explain Phoneme-Based Predictions of Cortical Responses to Speech. *Current Biology: CB*, *29*(12), 1924–1937.e9.

Davis, M. H. (2016). The neurobiology of lexical access. In *Neurobiology of Language* (pp. 541–555). Elsevier.

Davis, M. H., & Johnsrude, I. S. (2007). Hearing speech sounds: top-down influences on the interface between audition and speech perception. *Hearing Research*, *229*(1-2), 132–147.

de Bruin, A. (2019). Not All Bilinguals Are the Same: A Call for More Detailed Assessments and Descriptions of Bilingual Experiences. *Behavioral Sciences*, *9*(3), 33.

Dehaene-Lambertz, G., Pallier, C., Serniclaes, W., Sprenger-Charolles, L., Jobert, A., & Dehaene, S. (2005). Neural correlates of switching from auditory to speech perception. *NeuroImage*, *24*(1), 21–33.

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

Diesch, E., & Luce, T. (1997). Magnetic fields elicited by tones and vowel formants reveal tonotopy and nonlinear summation of cortical activation. *Psychophysiology*, *34*(5), 501–510.

Di Liberto, G. M., Attaheri, A., Cantisani, G., Reilly, R. B., Ní Choisdealbha, Á., Rocha, S., Brusini, P., & Goswami, U. (2023). Emergence of the cortical encoding of phonetic features in the first year of life. *Nature Communications*, *14*(1), 7789.

Douglas, R. J., & Martin, K. A. C. (2007). Recurrent neuronal circuits in the neocortex. *Current Biology: CB*, *17*(13), R496–R500.

Drennan, D. P., & Lalor, E. C. (2019). Cortical Tracking of Complex Sound Envelopes: Modeling the Changes in Response with Intensity. *eNeuro*, *6*(3). https://doi.org/10.1523/ENEURO.0082-19.2019

Drullman, R. (1995). Temporal envelope and fine structure cues for speech intelligibility. *The Journal of the Acoustical Society of America*, *97*(1), 585–592.

Dufour, S., Brunellière, A., & Nguyen, N. (2013). To what extent do we hear phonemic contrasts in a non-native regional variety? Tracking the dynamics of perceptual processing with EEG. *Journal of Psycholinguistic Research*, *42*(2), 161–173.

Eckman, F. R. (2008). 4. Typological markedness and second language phonology. In *Studies in Bilingualism* (pp. 95–115). John Benjamins Publishing Company.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Endress, A. D., & Hauser, M. D. (2010). Word segmentation with universal prosodic cues. *Cognitive Psychology*, *61*(2), 177–199.

Evans, N., & Levinson, S. C. (2009). The myth of language universals: language diversity and its importance for cognitive science. *The Behavioral and Brain Sciences*, *32*(5), 429–448; discussion 448–494.

Fedorenko, E., Ivanova, A. A., & Regev, T. I. (2024). The language network as a natural kind within the broader landscape of the human brain. *Nature Reviews. Neuroscience*. https://doi.org/10.1038/s41583-024-00802-4

Feldman, N. H., & Griffiths, T. L. (2007). A rational account of the perceptual magnet effect. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *29*. https://escholarship.org/content/qt18f7t8f7/qt18f7t8f7.pdf

Fischer, B. J., Anderson, C. H., & Peña, J. L. (2009). Multiplicative auditory spatial receptive fields created by a hierarchy of population codes. *PloS One*, *4*(11), e8015.

Fisher, J. M., Dick, F. K., Levy, D. F., & Wilson, S. M. (2018). Neural representation of vowel formants

in tonotopic auditory cortex. *NeuroImage*, *178*, 574–582.

Fishman, Y. I., Reser, D. H., Arezzo, J. C., & Steinschneider, M. (2000). Complex tone processing in primary auditory cortex of the awake monkey. II. Pitch versus critical band representation. *The Journal of the Acoustical Society of America*, *108*(1), 247–262.

Flege, J. E., & Eefting, W. (1988). Imitation of a VOT continuum by native speakers of English and Spanish: evidence for phonetic category formation. *The Journal of the Acoustical Society of America*, *83*(2), 729–740.

Fogerty, D., & Humes, L. E. (2012). The role of vowel and consonant fundamental frequency, envelope, and temporal fine structure cues to the intelligibility of words and sentences. *The Journal of the Acoustical Society of America*, *131*(2), 1490–1501.

Fogerty, D., & Kewley-Port, D. (2009). Perceptual contributions of the consonant-vowel boundary to sentence intelligibility. *The Journal of the Acoustical Society of America*, *126*(2), 847–857.

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

Fox, N. P., Leonard, M., Sjerps, M. J., & Chang, E. F. (2020). Transformation of a temporal speech cue to a spatial neural code in human auditory cortex. *eLife*, *9*. https://doi.org/10.7554/eLife.53051

Fox, N. P., Sjerps, M. J., & Chang, E. F. (2017). Dynamic emergence of categorical perception of voice-onset time in human speech cortex. *The Journal of the Acoustical Society of America*, *141*(5), 3571–3571.

Fox, R. A., Flege, J. E., & Munro, M. J. (1995). The perception of English and Spanish vowels by native English and Spanish listeners: a multidimensional scaling analysis. *The Journal of the Acoustical Society of America*, *97*(4), 2540–2551.

Fritz, J. B., David, S., & Shamma, S. (2013). Attention and Dynamic, Task-Related Receptive Field Plasticity in Adult Auditory Cortex. In Y. E. Cohen, A. N. Popper, & R. R. Fay (Eds.), *Neural Correlates of Auditory Cognition* (pp. 251–291). Springer New York.

Fritz, J. B., David, S. V., Radtke-Schuller, S., Yin, P., & Shamma, S. A. (2010). Adaptive, behaviorally

gated, persistent encoding of task-relevant auditory information in ferret frontal cortex. *Nature Neuroscience*, *13*(8), 1011–1019.

Fujisaki, H., & Kawashima, T. (1968). The roles of pitch and higher formants in the perception of vowels. *IEEE Transactions on Audio and Electroacoustics*, *16*(1), 73–77.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., & Pallett, D. S. (1993). *DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1* (NASA STI/Recon Technical Report N, Vol. 93, p. 27403). adsabs.harvard.edu. https://ui.adsabs.harvard.edu/abs/1993STIN...9327403G

Geschwind, N. (1970). The organization of language and the brain. *Science*, *170*(3961), 940–944.

Gill, P., Zhang, J., Woolley, S. M. N., Fremouw, T., & Theunissen, F. E. (2006). Sound representation methods for spectro-temporal receptive field estimation. *Journal of Computational Neuroscience*, *21*(1), 5–20.

Golestani, N. (2016). Neuroimaging of phonetic perception in bilinguals. *Bilingualism (Cambridge, England)*, *19*(4), 674–682.

Golestani, N., & Zatorre, R. J. (2004). Learning new sounds of speech: reallocation of neural substrates. *NeuroImage*, *21*(2), 494–506.

Golestani, N., & Zatorre, R. J. (2009). Individual differences in the acquisition of second language phonology. *Brain and Language*, *109*(2-3), 55–67.

Green, D. W. (2003). 9. Neural basis of lexicon and grammar in L2 acquisition: The convergence hypothesis. In *The Lexicon–Syntax Interface in Second Language Acquisition* (pp. 197–218). John Benjamins Publishing Company.

Gwilliams, L., & Davis, M. (2020). Extracting language content from speech sounds: An information theoretic approach. *The Auditory Cognitive Neuroscience of Speech Perception*. https://hal.archives-ouvertes.fr/hal-03013496/

Hagiwara, R. E. (2004). An Acoustic Analysis of Vowel Variation in New World English (review). In *Language* (Vol. 80, Issue 4, pp. 903–903). https://doi.org/10.1353/lan.2004.0191

Halle, M., & Chomsky, N. (1968). *The sound pattern of English*. Harper & Row.

Hamilton, L. S., Chang, D. L., Lee, M. B., & Chang, E. F. (2017). Semi-automated Anatomical Labeling and Inter-subject Warping of High-Density Intracranial Recording Electrodes in Electrocorticography. *Frontiers in Neuroinformatics*, *11*, 62.

Hamilton, L. S., Edwards, E., & Chang, E. F. (2018). A Spatial Map of Onset and Sustained Responses to Speech in the Human Superior Temporal Gyrus. *Current Biology: CB*, *28*(12), 1860–1871.e4.

Hamilton, L. S., Oganian, Y., & Chang, E. F. (2020). Topography of speech-related acoustic and phonological feature encoding throughout the human core and parabelt auditory cortex. In *Cold Spring Harbor Laboratory* (p. 2020.06.08.121624). https://doi.org/10.1101/2020.06.08.121624

Hamilton, L. S., Oganian, Y., Hall, J., & Chang, E. F. (2021). Parallel and distributed encoding of speech across human auditory cortex. *Cell*, *184*(18), 4626–4639.e13.

Healy, A. F., & Cutting, J. E. (1976). Units of speech perception: Phoneme and syllable. *Journal of Verbal Learning and Verbal Behavior*, *15*(1), 73–83.

Heil, P., & Neubauer, H. (2001). Temporal integration of sound pressure determines thresholds of auditory-nerve fibers. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *21*(18), 7404–7415.

Hesling, I., Dilharreguy, B., Bordessoules, M., & Allard, M. (2012). The neural processing of second language comprehension modulated by the degree of proficiency: a listening connected speech FMRI study. *The Open Neuroimaging Journal*, *6*, 44–54.

Hickok, G., & Poeppel, D. (2004). Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition*, *92*(1-2), 67–99.

Hickok, G., & Poeppel, D. (2007a). The cortical organization of speech processing. *Nature Reviews. Neuroscience*, *8*(5), 393–402.

Hickok, G., & Poeppel, D. (2007b). The cortical organization of speech processing. *Nature Reviews. Neuroscience*, *8*(5), 393–402.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American

English vowels. *The Journal of the Acoustical Society of America*, *97*(5 Pt 1), 3099–3111.

Hillis, A. E., Rorden, C., & Fridriksson, J. (2017). Brain regions essential for word comprehension: Drawing inferences from patients. *Annals of Neurology*, *81*(6), 759–768.

Hitczenko, K., & Feldman, N. H. (2022). Naturalistic speech supports distributional learning across contexts. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(38), e2123230119.

Holdgraf, C. R., de Heer, W., Pasley, B., Rieger, J., Crone, N., Lin, J. J., Knight, R. T., & Theunissen, F. E. (2016). Rapid tuning shifts in human auditory cortex enhance speech intelligibility. *Nature Communications*, *7*, 13654.

Holdgraf, C. R., Rieger, J. W., Micheli, C., Martin, S., Knight, R. T., & Theunissen, F. E. (2017). Encoding and Decoding Models in Cognitive Electrophysiology. *Frontiers in Systems Neuroscience*, *11*, 61.

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

Hopkins, K., & Moore, B. C. J. (2009). The contribution of temporal fine structure to the intelligibility of speech in steady and modulated noise. *The Journal of the Acoustical Society of America*, *125*(1), 442–446.

Hualde, J. I. (2005). *The Sounds of Spanish*. Cambridge University Press.

Huang, J., & Holt, L. L. (2012). Listening for the norm: adaptive coding in speech categorization. *Frontiers in Psychology*, *3*, 10.

Hullett, P. W., Hamilton, L. S., Mesgarani, N., Schreiner, C. E., & Chang, E. F. (2016). Human Superior Temporal Gyrus Organization of Spectrotemporal Modulation Tuning Derived from Speech Stimuli. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *36*(6), 2014–2026.

Humphries, C., Sabri, M., Lewis, K., & Liebenthal, E. (2014). Hierarchical organization of speech perception in human auditory cortex. *Frontiers in Neuroscience*, *8*, 406.

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

Iverson, P., & Kuhl, P. K. (1995). Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. *The Journal of the Acoustical Society of America*, *97*(1), 553–562.

Jacquemot, C., Pallier, C., LeBihan, D., Dehaene, S., & Dupoux, E. (2003). Phonological grammar shapes the auditory cortex: a functional magnetic resonance imaging study. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *23*(29), 9541–9546.

Jasmin, K., Lima, C. F., & Scott, S. K. (2019). Understanding rostral-caudal auditory cortex contributions to auditory perception. *Nature Reviews. Neuroscience*, *20*(7), 425–434.

Jenkins, J. J., Strange, W., & Edman, T. R. (1983). Identification of vowels in "vowelless" syllables. *Perception & Psychophysics*, *34*(5), 441–450.

Johnson, K. (1988). *Processes of speaker normalization in vowel perception* (M. Beckman (ed.)) [The Ohio State University]. https://www.proquest.com/dissertations-theses/processes-speaker-normalization-vowel-perception/docview/303700096/se-2

Johnson, K. (2008). Speaker normalization in speech perception. *The Handbook of Speech Perception.*, *708*, 363–389.

Johnson, K. (2020). The ΔF method of vocal tract length normalization for vowels. *Laboratory Phonology*, *11*(1). https://doi.org/10.5334/labphon.196

Johnson, K., & Sjerps, M. J. (2021). Speaker Normalization in Speech Perception. In *The Handbook of Speech Perception* (pp. 145–176). Wiley. https://doi.org/10.1002/9781119184096.ch6

Jordan, M. I. (1997). Chapter 25 - Serial Order: A Parallel Distributed Processing Approach. In J. W. Donahoe & V. Packard Dorsel (Eds.), *Advances in Psychology* (Vol. 121, pp. 471–495). North-Holland.

Jusczyk, P. W., & Hohne, E. A. (1997). Infants' memory for spoken words. *Science (New York, N.Y.)*, *277*(5334), 1984–1986.

Kaushanskaya, M., Blumenfeld, H. K., & Marian, V. (2020). The Language Experience and Proficiency Questionnaire (LEAP-Q): Ten years later. *Bilingualism* , *23*(5), 945–950.

Kepinska, O., Dalboni da Rocha, J., Tuerk, C., Hervais-Adelman, A., Bouhali, F., Green, D., Price, C. J., & Golestani, N. (2025). Auditory cortex anatomy reflects multilingual phonological experience. In *eLife*. https://doi.org/10.7554/eLife.90269.3

Khalighinejad, B., Patel, P., Herrero, J. L., Bickel, S., Mehta, A. D., & Mesgarani, N. (2021). Functional characterization of human Heschl's gyrus in response to natural speech. *NeuroImage*, *235*(118003), 118003.

Khoshkhoo, S., Leonard, M. K., Mesgarani, N., & Chang, E. F. (2018). Neural correlates of sine-wave speech intelligibility in human frontal and temporal cortex. *Brain and Language*, *187*, 83–91.

Kim, S. Y., Liu, L., Liu, L., & Cao, F. (2020). Neural representational similarity between L1 and L2 in spoken and written language processing. *Human Brain Mapping*, *41*(17), 4935–4951.

Kohonen, T., & Hari, R. (1999). Where the abstract feature maps of the brain might come from. In *Trends in Neurosciences* (Vol. 22, Issue 3, pp. 135–139). https://doi.org/10.1016/s0166-2236(98)01342-3

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics*, *50*(2), 93–107.

Kuhl, P. K., Kiritani, S., Deguchi, T., Hayashi, A., Stevens, E. B., Dugger, C. D., & Iverson, P. (1997). Effects of language experience on speech perception: American and Japanese infants' perception of /ra/ and /la/. *The Journal of the Acoustical Society of America*, *102*(5), 3135–3136.

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science*, *9*(2), F13–F21.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, *255*(5044), 606–608.

Kuwabara, H. (1985). An approach to normalization of coarticulation effects for vowels in connected speech. *The Journal of the Acoustical Society of America*, *77*(2), 686–694.

Kuzmina, E., Goral, M., Norvik, M., & Weekes, B. S. (2019). What influences language impairment in bilingual aphasia? A meta-analytic review. *Frontiers in Psychology*, *10*, 445.

Ladefoged, P., & Broadbent, D. E. (1957). Information Conveyed by Vowels. *The Journal of the Acoustical Society of America*, *29*(1), 98–104.

Ladefoged, P., & Johnson, K. (2014). *A course in phonetics*. Nelson Education.

Laing, E. J. C., Liu, R., Lotto, A. J., & Holt, L. L. (2012). Tuned with a Tune: Talker Normalization via General Auditory Processes. *Frontiers in Psychology*, *3*, 203.

Lakertz, Y., Ossmy, O., Friedmann, N., Mukamel, R., & Fried, I. (2021). Single-cell activity in human STG during perception of phonemes is organized according to manner of articulation. *NeuroImage*, *226*, 117499.

Leaver, A. M., & Rauschecker, J. P. (2010). Cortical representation of natural complex sounds: effects of acoustic features and auditory object category. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *30*(22), 7604–7612.

Lehiste, I. (1972). The timing of utterances and linguistic boundaries. *The Journal of the Acoustical Society of America*, *51*(6B), 2018–2024.

Leonard, M. K., Baud, M. O., Sjerps, M. J., & Chang, E. F. (2016). Perceptual restoration of masked speech in human cortex. *Nature Communications*, *7*, 13619.

Leonard, M. K., Cai, R., Babiak, M. C., Ren, A., & Chang, E. F. (2019). The peri-Sylvian cortical network underlying single word repetition revealed by electrocortical stimulation and direct neural recordings. *Brain and Language*, *193*, 58–72.

Leonard, M. K., & Chang, E. F. (2014). Dynamic speech representations in the human temporal lobe. *Trends in Cognitive Sciences*, *18*(9), 472–479.

Leonard, M. K., Gwilliams, L., Sellers, K. K., Chung, J. E., Xu, D., Mischler, G., Mesgarani, N., Welkenhuysen, M., Dutta, B., & Chang, E. F. (2024). Large-scale single-neuron speech sound encoding across the depth of human cortex. *Nature*, *626*(7999), 593–602.

Leszczyński, M., Barczak, A., Kajikawa, Y., Ulbert, I., Falchier, A. Y., Tal, I., Haegens, S., Melloni, L.,

Knight, R. T., & Schroeder, C. E. (2020). Dissociation of broadband high-frequency activity and neuronal firing in the neocortex. *Science Advances*, *6*(33), eabb0977.

Levy, D. F., & Wilson, S. M. (2020). Categorical Encoding of Vowels in Primary Auditory Cortex. *Cerebral Cortex* , *30*(2), 618–627.

Li, A., Lin, M., Chen, X., Zu, Y., Sun, G., Hua, W., Yin, Z., & Yan, J. (2000). Speech corpus of Chinese discourse and the phonetic research. *Sixth International Conference on Spoken Language Processing*. https://www.isca-archive.org/icslp_2000/li00k_icslp.pdf

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431–461.

Liberman, A. M., Delattre, P. C., & Cooper, F. S. (1958). Some Cues for the Distinction Between Voiced and Voiceless Stops in Initial Position. *Language and Speech*, *1*(3), 153–167.

Liberman, A. M., Delattre, P., & Cooper, F. S. (1952). The role of selected stimulus-variables in the perception of the unvoiced stop consonants. *The American Journal of Psychology*, *65*(4), 497–516.

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, *54*(5), 358–368.

Liégeois-Chauvel, C., Lorenzi, C., Trébuchon, A., Régis, J., & Chauvel, P. (2004). Temporal envelope processing in the human left and right auditory cortices. *Cerebral Cortex* , *14*(7), 731–740.

Liljeström, M., Kujala, J., Stevenson, C., & Salmelin, R. (2015). Dynamic reconfiguration of the language network preceding onset of speech in picture naming. *Human Brain Mapping*, *36*(3), 1202–1216.

Lindblom, B., & Sussman, H. M. (2012). Dissecting coarticulation: How locus equations happen. *Journal of Phonetics*, *40*(1), 1–19.

Lipkin, B., Tuckute, G., Affourtit, J., Small, H., Mineroff, Z., Kean, H., Jouravlev, O., Rakocevic, L., Pritchett, B., Siegelman, M., Hoeflin, C., Pongos, A., Blank, I. A., Struhl, M. K., Ivanova, A., Shannon, S., Sathe, A., Hoffmann, M., Nieto-Castañón, A., & Fedorenko, E. (2022). Probabilistic

atlas for the language network based on precision fMRI data from >800 individuals. *Scientific Data*, *9*(1), 529.

Lisker, L., & Abramson, A. S. (1964). A Cross-Language Study of Voicing in Initial Stops: Acoustical Measurements. *Word & World*, *20*(3), 384–422.

Li, Y., Anumanchipalli, G. K., Mohamed, A., Chen, P., Carney, L. H., Lu, J., Wu, J., & Chang, E. F. (2023). Dissecting neural computations in the human auditory pathway using deep neural networks for speech. *Nature Neuroscience*, *26*(12), 2213–2225.

Li, Y., Tang, C., Lu, J., Wu, J., & Chang, E. F. (2021). Human cortical encoding of pitch in tonal and non-tonal languages. *Nature Communications*, *12*(1), 1–12.

Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., & Moore, B. C. J. (2006). Speech perception problems of the hearing impaired reflect inability to use temporal fine structure. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(49), 18866–18869.

Major, R. C. (2008). 3. Transfer in second language phonology: A review. In *Studies in Bilingualism* (pp. 63–94). John Benjamins Publishing Company.

Malik-Moraleda, S., Ayyash, D., Gallée, J., Affourtit, J., Hoffmann, M., Mineroff, Z., Jouravlev, O., & Fedorenko, E. (2022). An investigation across 45 languages and 12 language families reveals a universal language network. *Nature Neuroscience*, *25*(8), 1014–1019.

Mann, V. A. (1980). Influence of preceding liquid on stop-consonant perception. *Perception & Psychophysics*, *28*(5), 407–412.

Mann, V. A., & Repp, B. H. (1980). Influence of vocalic context on perception of the [zh]-[s] distinction. *Perception & Psychophysics*, *28*(3), 213–228.

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, *10*(1), 29–63.

Massaro, D. W. (1974). Perceptual units in speech recognition. *Journal of Experimental Psychology*, *102*(2), 199–208.

Matsumoto, R., Imamura, H., Inouchi, M., Nakagawa, T., Yokoyama, Y., Matsuhashi, M., Mikuni, N.,

Miyamoto, S., Fukuyama, H., Takahashi, R., & Ikeda, A. (2011). Left anterior temporal cortex actively engages in speech perception: A direct cortical stimulation study. *Neuropsychologia*, *49*(5), 1350–1354.

Mazoyer, B. M., Tzourio, N., Frak, V., Syrota, A., Murayama, N., Levrier, O., Salamon, G., Dehaene, S., Cohen, L., & Mehler, J. (1993). The cortical representation of speech. *Journal of Cognitive Neuroscience*, *5*(4), 467–479.

McCarthy, K. M., Skoruppa, K., & Iverson, P. (2019). Development of neural perceptual vowel spaces during the first year of life. *Scientific Reports*, *9*(1), 19592.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. In *Cognitive Psychology* (Vol. 18, Issue 1, pp. 1–86). https://doi.org/10.1016/0010-0285(86)90015-0

Mesgarani, N., & 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., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*, *343*(6174), 1006–1010.

Miller, J. D. (1989). Auditory-perceptual interpretation of the vowel. *The Journal of the Acoustical Society of America*, *85*(5), 2114–2134.

Miller, K. J., Sorensen, L. B., Ojemann, J. G., & den Nijs, M. (2009). Power-law scaling in the brain surface electric potential. *PLoS Computational Biology*, *5*(12), e1000609.

Moerel, M., De Martino, F., & Formisano, E. (2012). Processing of natural sounds in human auditory cortex: tonotopy, spectral tuning, and relation to voice sensitivity. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *32*(41), 14205–14216.

Monahan, P. J., & Idsardi, W. J. (2010). Auditory sensitivity to formant ratios: Toward an account of vowel normalisation. *Language and Cognitive Processes*, *25*(6), 808–839.

Moon, I. J., & Hong, S. H. (2014). What is temporal fine structure and why is it important? *Korean Journal of Audiology*, *18*(1), 1–7.

Möttönen, R., Calvert, G. A., Jääskeläinen, I. P., Matthews, P. M., Thesen, T., Tuomainen, J., & Sams,

M. (2006). Perceiving identical sounds as speech or non-speech modulates activity in the left posterior superior temporal sulcus. *NeuroImage*, *30*(2), 563–569.

Mukamel, R., & Fried, I. (2012). Human intracranial recordings and cognitive neuroscience. *Annual Review of Psychology*, *63*, 511–537.

Myers, E. B. (2014). Emergence of category-level sensitivities in non-native speech sound learning. *Frontiers in Neuroscience*, *8*, 238.

Myers, E. B., & Blumstein, S. E. (2008). The neural bases of the lexical effect: an fMRI investigation. *Cerebral Cortex (New York, N.Y.: 1991)*, *18*(2), 278–288.

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

Narain, C., Scott, S. K., Wise, R. J. S., Rosen, S., Leff, A., Iversen, S. D., & Matthews, P. M. (2003). Defining a left-lateralized response specific to intelligible speech using fMRI. *Cerebral Cortex* , *13*(12), 1362–1368.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *The Journal of the Acoustical Society of America*, *85*(5), 2088–2113.

Nittrouer, S., Miller, M. E., Crowther, C. S., & Manhart, M. J. (2000). The effect of segmental order on fricative labeling by children and adults. *Perception & Psychophysics*, *62*(2), 266–284.

Norman-Haignere, S., Kanwisher, N. G., & McDermott, J. H. (2015). Distinct Cortical Pathways for Music and Speech Revealed by Hypothesis-Free Voxel Decomposition. In *Neuron* (Vol. 88, Issue 6, pp. 1281–1296). https://doi.org/10.1016/j.neuron.2015.11.035

Norman-Haignere, S. V., Feather, J., Boebinger, D., Brunner, P., Ritaccio, A., McDermott, J. H., Schalk, G., & Kanwisher, N. (2022). A neural population selective for song in human auditory cortex. *Current Biology: CB*, *32*(7), 1470–1484.e12.

Norman-Haignere, S. V., & McDermott, J. H. (2018). Neural responses to natural and model-matched

stimuli reveal distinct computations in primary and nonprimary auditory cortex. *PLoS Biology*, *16*(12), e2005127.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: feedback is never necessary. *The Behavioral and Brain Sciences*, *23*(3), 299–325; discussion 325–370.

Nourski, K. V., Reale, R. A., Oya, H., Kawasaki, H., Kovach, C. K., Chen, H., Howard, M. A., 3rd, & Brugge, J. F. (2009). Temporal envelope of time-compressed speech represented in the human auditory cortex. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *29*(49), 15564–15574.

Nourski, K. V., Steinschneider, M., Rhone, A. E., Kovach, C. K., Kawasaki, H., & Howard, M. A., 3rd. (2019). Differential responses to spectrally degraded speech within human auditory cortex: An intracranial electrophysiology study. *Hearing Research*, *371*, 53–65.

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

Obleser, J., & Eisner, F. (2009). Pre-lexical abstraction of speech in the auditory cortex. *Trends in Cognitive Sciences*, *13*(1), 14–19.

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

Obleser, J., Leaver, A. M., Vanmeter, J., & Rauschecker, J. P. (2010). Segregation of vowels and consonants in human auditory cortex: evidence for distributed hierarchical organization. *Frontiers in Psychology*, *1*, 232.

Oganian, Y., Bhaya-Grossman, I., Johnson, K., & Chang, E. F. (2023). Vowel and formant representation in the human auditory speech cortex. *Neuron*, *111*(13), 2105–2118.e4.

Oganian, Y., & Chang, E. F. (2019). A speech envelope landmark for syllable encoding in human superior temporal gyrus. *Science Advances*, *5*(11), eaay6279.

Ohala, J. J. (1975). In G. Fant & M. A. A. Tatham (Eds.), *Auditory analysis and perception of speech* (pp. 431–453). Academic Press.

Ohl, F. W., & Scheich, H. (1997). Orderly cortical representation of vowels based on formant interaction. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(17), 9440–9444.

Okada, K., Rong, F., Venezia, J., Matchin, W., Hsieh, I.-H., Saberi, K., Serences, J. T., & Hickok, G. (2010). Hierarchical organization of human auditory cortex: evidence from acoustic invariance in the response to intelligible speech. *Cerebral Cortex* , *20*(10), 2486–2495.

Overath, T., McDermott, J. H., Zarate, J. M., & Poeppel, D. (2015). The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nature Neuroscience*, *18*(6), 903–911.

Ozker, M., Schepers, I. M., Magnotti, J. F., Yoshor, D., & Beauchamp, M. S. (2017). A Double Dissociation between Anterior and Posterior Superior Temporal Gyrus for Processing Audiovisual Speech Demonstrated by Electrocorticography. *Journal of Cognitive Neuroscience*, *29*(6), 1044–1060.

Parvizi, J., & Kastner, S. (2018). Promises and limitations of human intracranial electroencephalography. *Nature Neuroscience*, *21*(4), 474–483.

Peelle, J. E., & Davis, M. H. (2012). Neural Oscillations Carry Speech Rhythm through to Comprehension. *Frontiers in Psychology*, *3*, 320.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442.

Pérez-González, D., & Malmierca, M. S. (2014). Adaptation in the auditory system: an overview. *Frontiers in Integrative Neuroscience*, *8*, 19.

Peterson, G. E. (1961). Parameters of vowel quality. *Journal of Speech and Hearing Research*, *4*, 10–29.

Pineda, L. A., Castellanos, H., Cuétara, J., Galescu, L., Juárez, J., Llisterri, J., Pérez, P., & Villaseñor, L. (2010). The Corpus DIMEx100: transcription and evaluation. *Language Resources and Evaluation*,

44(4), 347–370.

Pineda, L. A., Pineda, L. V., Cuétara, J., Castellanos, H., & López, I. (2004). DIMEx100: A New Phonetic and Speech Corpus for Mexican Spanish. *Advances in Artificial Intelligence – IBERAMIA 2004*, 974–983.

Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception & Psychophysics*, *13*(2), 253–260.

Połczyńska, M. M., & Bookheimer, S. Y. (2021). General principles governing the amount of neuroanatomical overlap between languages in bilinguals. *Neuroscience and Biobehavioral Reviews*, *130*, 1–14.

Rabinowitz, N. C., Willmore, B. D. B., Schnupp, J. W. H., & King, A. J. (2011). Contrast gain control in auditory cortex. *Neuron*, *70*(6), 1178–1191.

Rauschecker, J. P., & Scott, S. K. (2009). Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nature Neuroscience*, *12*(6), 718–724.

Ray, S., & Maunsell, J. H. R. (2011). Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. *PLoS Biology*, *9*(4), e1000610.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science*, *212*(4497), 947–949.

Roux, F.-E., Miskin, K., Durand, J.-B., Sacko, O., Réhault, E., Tanova, R., & Démonet, J.-F. (2015). Electrostimulation mapping of comprehension of auditory and visual words. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *71*, 398–408.

Saarinen, T., Jalava, A., Kujala, J., Stevenson, C., & Salmelin, R. (2015). Task-sensitive reconfiguration of corticocortical 6-20 Hz oscillatory coherence in naturalistic human performance. *Human Brain Mapping*, *36*(7), 2455–2469.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word Segmentation: The Role of Distributional

Cues. *Journal of Memory and Language*, *35*(4), 606–621.

Saffran, J. R., Werker, J. F., & Werner, L. A. (1042). The Infant's Auditory World: Hearing, Speech, and the Beginnings of Language. In D. Kuhn (Ed.), *Handbook of child psychology: Cognition, perception, and language, Vol* (Vol. 2, pp. 58–108). John Wiley & Sons, Inc., xxvii.

Samuel, A. G. (1987). Lexical uniqueness effects on phonemic restoration. *Journal of Memory and Language*, *26*(1), 36–56.

Sankaran, N., Leonard, M. K., Theunissen, F., & Chang, E. F. (2024). Encoding of melody in the human auditory cortex. *Science Advances*, *10*(7), eadk0010.

Santoro, R., Moerel, M., De Martino, F., Valente, G., Ugurbil, K., Yacoub, E., & Formisano, E. (2017). Reconstructing the spectrotemporal modulations of real-life sounds from fMRI response patterns. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(18), 4799–4804.

Savin, H. B., & Bever, T. G. (1970). The nonperceptual reality of the phoneme. *Journal of Verbal Learning and Verbal Behavior*, *9*(3), 295–302.

Scharinger, M., Idsardi, W. J., & Poe, S. (2011). A Comprehensive Three-dimensional Cortical Map of Vowel Space. In *Journal of Cognitive Neuroscience* (Vol. 23, Issue 12, pp. 3972–3982). https://doi.org/10.1162/jocn_a_00056

Schlosser, M. J., Aoyagi, N., Fulbright, R. K., Gore, J. C., & McCarthy, G. (1998). Functional MRI studies of auditory comprehension. *Human Brain Mapping*, *6*(1), 1–13.

Schönwiesner, M., Rübsamen, R., & Von Cramon, D. Y. (2005). Hemispheric asymmetry for spectral and temporal processing in the human antero-lateral auditory belt cortex. In *European Journal of Neuroscience* (Vol. 22, Issue 6, pp. 1521–1528). https://doi.org/10.1111/j.1460-9568.2005.04315.x

Schönwiesner, M., & Zatorre, R. J. (2009). Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(34), 14611–14616.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with

primarily temporal cues. *Science*, *270*(5234), 303–304.

Shatzman, K. B., & McQueen, J. M. (2006). Segment duration as a cue to word boundaries in spoken-word recognition. *Perception & Psychophysics*, *68*(1), 1–16.

Shestakova, A., Brattico, E., Soloviev, A., Klucharev, V., & Huotilainen, M. (2004). Orderly cortical representation of vowel categories presented by multiple exemplars. *Brain Research. Cognitive Brain Research*, *21*(3), 342–350.

Sjerps, M. J., Fox, N. P., Johnson, K., & Chang, E. F. (2019). Speaker-normalized sound representations in the human auditory cortex. *Nature Communications*, *10*(1), 2465.

Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature*, *388*(6640), 381–382.

Steinschneider, M., Fishman, Y. I., & Arezzo, J. C. (2008). Spectrotemporal analysis of evoked and induced electroencephalographic responses in primary auditory cortex (A1) of the awake monkey. *Cerebral Cortex* , *18*(3), 610–625.

Steinschneider, M., Nourski, K. V., Kawasaki, H., Oya, H., Brugge, J. F., & Howard, M. A., 3rd. (2011). Intracranial study of speech-elicited activity on the human posterolateral superior temporal gyrus. *Cerebral Cortex* , *21*(10), 2332–2347.

Steinschneider, M., Volkov, I. O., Noh, M. D., Garell, P. C., & Howard, M. A., 3rd. (1999). Temporal encoding of the voice onset time phonetic parameter by field potentials recorded directly from human auditory cortex. *Journal of Neurophysiology*, *82*(5), 2346–2357.

Strange, W. (1989a). Dynamic specification of coarticulated vowels spoken in sentence context. In *The Journal of the Acoustical Society of America* (Vol. 85, Issue 5, pp. 2135–2153). https://doi.org/10.1121/1.397863

Strange, W. (1989b). Evolving theories of vowel perception. *The Journal of the Acoustical Society of America*, *85*(5), 2081–2087.

Strange, W., & Shafer, V. L. (2008). 6. Speech perception in second language learners: The re-education of selective perception. In *Studies in Bilingualism* (pp. 153–191). John Benjamins Publishing

Company.

Strange, W., Verbrugge, R. R., Shankweiler, D. P., & Edman, T. R. (1976). Consonant environment specifies vowel identity. *The Journal of the Acoustical Society of America*, *60*(1), 213–224.

SUBTLEXus. (2015, March 25). Universiteit Gent. https://www.ugent.be/pp/experimentele-psychologie/en/research/documents/subtlexus

Sulpizio, S., Del Maschio, N., Fedeli, D., & Abutalebi, J. (2020). Bilingual language processing: A meta-analysis of functional neuroimaging studies. *Neuroscience and Biobehavioral Reviews*, *108*, 834–853.

Syrdal, A. K., & Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. In *The Journal of the Acoustical Society of America* (Vol. 79, Issue 4, pp. 1086–1100). https://doi.org/10.1121/1.393381

Theunissen, F. E., David, S. V., Singh, N. C., Hsu, A., Vinje, W. E., & Gallant, J. L. (2001). Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *Network*, *12*(3), 289–316.

Towle, V. L., Yoon, H.-A., Castelle, M., Edgar, J. C., Biassou, N. M., Frim, D. M., Spire, J.-P., & Kohrman, M. H. (2008). ECoG gamma activity during a language task: differentiating expressive and receptive speech areas. *Brain: A Journal of Neurology*, *131*(Pt 8), 2013–2027.

Tuller, B., Nguyen, N., Lancia, L., & Vallabha, G. K. (2011). Nonlinear Dynamics in Speech Perception. In R. Huys & V. K. Jirsa (Eds.), *Nonlinear Dynamics in Human Behavior* (pp. 135–150). Springer Berlin Heidelberg.

Turk, A. E., & Shattuck-Hufnagel, S. (2000). Word-boundary-related duration patterns in English. *Journal of Phonetics*, *28*(4), 397–440.

Tyler, M. D., & Cutler, A. (2009). Cross-language differences in cue use for speech segmentation. *The Journal of the Acoustical Society of America*, *126*(1), 367–376.

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: a new and improved word frequency database for British English. *Quarterly Journal of Experimental*

Psychology (2006), *67*(6), 1176–1190.

Vannest, J. J., Karunanayaka, P. R., Altaye, M., Schmithorst, V. J., Plante, E. M., Eaton, K. J., Rasmussen, J. M., & Holland, S. K. (2009). Comparison of fMRI data from passive listening and active-response story processing tasks in children. In *Journal of Magnetic Resonance Imaging* (Vol. 29, Issue 4, pp. 971–976). https://doi.org/10.1002/jmri.21694

von Kriegstein, K., Eger, E., Kleinschmidt, A., & Giraud, A. L. (2003). Modulation of neural responses to speech by directing attention to voices or verbal content. *Brain Research. Cognitive Brain Research*, *17*(1), 48–55.

Wang, Y., Spence, M. M., Jongman, A., & Sereno, J. A. (1999). Training American listeners to perceive Mandarin tones: Transfer to production. *The Journal of the Acoustical Society of America*, *105*(2_Supplement), 1095–1095.

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science*, *167*(3917), 392–393.

Warren, R. M., & Sherman, G. L. (1974). Phonemic restorations based on subsequent context. *Perception & Psychophysics*, *16*(1), 150–156.

Werker, J. F., & Lalonde, C. E. (1988). Cross-language speech perception: Initial capabilities and developmental change. *Developmental Psychology*, *24*(5), 672–683.

Werker, J. F., & Tees, R. C. (1999). Influences on infant speech processing: toward a new synthesis. *Annual Review of Psychology*, *50*(1), 509–535.

Wernicke, C. (1874). *Der aphasische Symptomencomplex: Eine psychologische Studie auf anatomischer Basis*. Cohn.

Wilson, S. M., Bautista, A., & McCarron, A. (2018). Convergence of spoken and written language processing in the superior temporal sulcus. *NeuroImage*, *171*, 62–74.

Yi, H. G., Chandrasekaran, B., Nourski, K. V., Rhone, A. E., Schuerman, W. L., Howard, M. A., 3rd, Chang, E. F., & Leonard, M. K. (2021). Learning nonnative speech sounds changes local encoding in the adult human cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(36), e2101777118.

Yi, H. G., Leonard, M. K., & Chang, E. F. (2019a). The Encoding of Speech Sounds in the Superior Temporal Gyrus. *Neuron*, *102*(6), 1096–1110.

Yi, H. G., Leonard, M. K., & Chang, E. F. (2019b). The encoding of speech sounds in the superior temporal gyrus. *Neuron*, *102*(6), 1096–1110.

Zatorre, R. J. (2001). Spectral and Temporal Processing in Human Auditory Cortex. In *Cerebral Cortex* (Vol. 11, Issue 10, pp. 946–953). https://doi.org/10.1093/cercor/11.10.946

Zatorre, R. J., Evans, A. C., Meyer, E., & Gjedde, A. (1992). Lateralization of phonetic and pitch discrimination in speech processing. *Science*, *256*(5058), 846–849.

Zeng, F.-G., Nie, K., Stickney, G. S., Kong, Y.-Y., Vongphoe, M., Bhargave, A., Wei, C., & Cao, K. (2005). Speech recognition with amplitude and frequency modulations. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(7), 2293–2298.

Zhang, Y., Leonard, M. K., Bhaya-Grossman, I., Gwilliams, L., & Chang, E. F. (2025). Human cortical dynamics of auditory word form encoding. *Neuron*, *0*(0). https://doi.org/10.1016/j.neuron.2025.10.011

Zhao, T. C., & Kuhl, P. K. (2018). Linguistic effect on speech perception observed at the brainstem. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(35), 8716–8721.

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:
Ilina Bhaya-Grossman
11AAAC31C6184D7...
Author Signature

12/08/2025 | 11:02 PM PST
Date

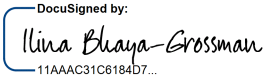
Certificate Of Completion

Envelope Id: 21CBC240-7731-49CB-A07D-1546A85CD96C
Status: Completed
Subject: Publishing Agreement for Ilina Bhaya-Grossman
Source Envelope:
Document Pages: 1
Signatures: 1
Certificate Pages: 1
Initials: 0
AutoNav: Enabled
EnvelopeId Stamping: Disabled
Time Zone: (UTC-08:00) Pacific Time (US & Canada)

Envelope Originator:
UCSF Graduate Division
1855 Folsom St
Suite 601
San Francisco, CA 94103
graduate.division@ucsf.edu
IP Address: 67.161.30.218

Record Tracking

Status: Original
12/8/2025 11:00:00 PM
Holder: UCSF Graduate Division
graduate.division@ucsf.edu
Location: DocuSign

Signer Events	Signature	Timestamp
Ilina Bhaya-Grossman ilina.bhaya-grossman@ucsf.edu Graduate Student University of California, San Francisco Security Level: Email, Account Authentication (Optional)	DocuSigned by:  11AAAC31C6184D7... Signature Adoption: Pre-selected Style Using IP Address: 67.161.30.218	Sent: 12/8/2025 11:00:03 PM Viewed: 12/8/2025 11:02:09 PM Signed: 12/8/2025 11:02:54 PM

Electronic Record and Signature Disclosure:
Not Offered via Docusign

In Person Signer Events	Signature	Timestamp

Editor Delivery Events	Status	Timestamp

Agent Delivery Events	Status	Timestamp

Intermediary Delivery Events	Status	Timestamp

Certified Delivery Events	Status	Timestamp

Carbon Copy Events	Status	Timestamp

Witness Events	Signature	Timestamp

Notary Events	Signature	Timestamp

Envelope Summary Events	Status	Timestamps
Envelope Sent	Hashed/Encrypted	12/8/2025 11:00:03 PM
Certified Delivered	Security Checked	12/8/2025 11:02:09 PM
Signing Complete	Security Checked	12/8/2025 11:02:54 PM
Completed	Security Checked	12/8/2025 11:02:54 PM

Payment Events	Status	Timestamps