HHS Public Access

Author manuscript

Annu Rev Psychol. Author manuscript; available in PMC 2022 September 06.

Published in final edited form as:

Annu Rev Psychol. 2022 January 04; 73: 79–102. doi:10.1146/annurev-psych-022321-035256.

Speech Computations of the Human Superior Temporal Gyrus

Ilina Bhaya-Grossman[1,2], **Edward F. Chang**[1]

[1]Department of Neurological Surgery, University of California, San Francisco, California 94143, USA

[2]Joint Graduate Program in Bioengineering, University of California, Berkeley and San Francisco, California 94720, USA

Abstract

Human speech perception results from neural computations that transform external acoustic speech signals into internal representations of words. The superior temporal gyrus (STG) contains the nonprimary auditory cortex and is a critical locus for phonological processing. Here, we describe how speech sound representation in the STG relies on fundamentally nonlinear and dynamical processes, such as categorization, normalization, contextual restoration, and the extraction of temporal structure. A spatial mosaic of local cortical sites on the STG exhibits complex auditory encoding for distinct acoustic-phonetic and prosodic features. We propose that as a population ensemble, these distributed patterns of neural activity give rise to abstract, higher-order phonemic and syllabic representations that support speech perception. This review presents a multi-scale, recurrent model of phonological processing in the STG, highlighting the critical interface between auditory and language systems.

Keywords

superior temporal gyrus; phonological processing; categorization; contextual restoration; temporal landmarks

INTRODUCTION

Speech perception relies on a set of transformations that convert a complex acoustic signal 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 temporal and phonological context, speaker identity, speech rate, and more (Liberman et al. 1952, 1967; Ladefoged & Johnson 2014). Determining the mechanisms by which the human brain overcomes these challenges to comprehend the speech signal reliably and flexibly has been the subject of spirited debate over the past century.

edward.chang@ucsf.edu .

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

Speech sounds can be described by acoustic properties of the physical sound wave, or articulatory properties that relate to the way in which the speech sound was formed in the vocal tract (acoustic and articulatory representations are illustrated in Figure 1). We use the term "phonological computation" to refer 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 the acoustic input. They are also dynamic in that speech representations vary as a function of time, depending heavily on adjacent words and sentence-level context. These key qualities of phonological computation give rise to a distinct form of auditory perception that facilitates speech processing.

Sound wave:

the variations in physical air pressure over time that characterize a sound

Classic anatomical lesion studies have long implicated the posterior region of the superior temporal gyrus (STG) as a critical locus for speech perception and language comprehension (Wernicke 1874, Geschwind 1970). The STG contains part of the nonprimary 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-hemisphere STG (usually the left one) (Hillis et al. 2017). Transient functional lesions of the same area, using electrocortical stimulation, demonstrate acute behavioral impairments in phonemic and word perception (Boatman et al. 2000, Boatman 2004, Roux et al. 2015, Leonard et al. 2019) and in sentence comprehension (Matsumoto et al. 2011), without clear effects on acoustic pure tone discrimination. Converging evidence from neuroimaging studies also suggests that the anterior and posterior regions of the STG play a special role in the neural analysis of speech (Figure 2a). The STG is responsive to speech over nonspeech sounds (Zatorre et al. 1992, Binder et al. 2000, Benson et al. 2001, Humphries et al. 2014) as well as intelligible over nonintelligible 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).

Although 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 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 the 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 review, we address the recent discoveries enabled by high-resolution cortical electrophysiology that contribute to our mechanistic understanding of speech processing. Direct cortical neurophysiological recordings allow for high spatial and temporal resolution at the millimeter and millisecond scales. These parameters are critical, because adjacent recording electrodes even a few millimeters apart show highly distinct tuning (Figure 2b), 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 the 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 (HGA; 50–200 Hz), 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, Steinschneider et al. 2008, Towle et al. 2008, Ray & Maunsell 2011, Leszczyński et al. 2020). The term "local neural populations" refers in this review to the activity recorded at a single electrode contact.

We begin by defining the role of phonological units in speech processing and review recent work that describes how these units may be represented by STG cortical activity. We then consider evidence for nonlinear neural transformations of speech input, including categorization of speech segments and the normalization of speaker-dependent acoustic properties. Next, we focus on the dynamic neural computations that enable the extraction of key temporal landmarks for syllable processing and context-sensitive phonological predictions. We argue that distinct patterns of neural activity, defined by nonlinear and dynamic tuning, give rise to emergent and higher-order phonological representations. In prevailing theories of speech processing, sounds are converted into higher-order speech units along a serial, feedforward cortical pathway. The empirical findings discussed in this review suggest an alternative. In the last section, we present a distributed model of speech processing in which multiple levels of auditory and phonological representation exist concurrently across the STG at both the local and the population spatial scales.

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 articulatory or acoustic properties and are typically binary: A given sample of speech can be described as a set of features that are either present or absent. 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, or phonemes (Halle & Chomsky 1968). For instance, a /b/ is a sound produced through voicing (vibrating vocal cords), bilabial (movement of the two lips), and plosive (the burst of air released after an articulatory closure) features. Any of these phonetic features taken alone is not linguistically meaningful. In combination, however, they give rise to all consonant and vowel sounds.

Phoneme:

any of the perceptually distinct units of sound in a language that distinguish one word from another (e.g., *b*, *m*, *d*, and *t* in *bad*, *bat*, *mad*, and *mat*)

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 the 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 the production of the vowel sound.

In natural speech, consonant and vowel sounds occur in highly organized repeated motifs that define syllables. The main component of a syllable is 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), and many perceptual and acoustically informed candidates have been proposed. Psycholinguistic studies attempting to address this question have yielded largely inconsistent results due in part to their use of task paradigms that impose distinct cognitive and linguistic demands (Savin & Bever 1970, Massaro 1974, Healy & Cutting 1976). 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). Ultimately, this body of research demonstrates that listeners focus on different perceptual units depending on the task context. The perceptual strategies employed by the listener can be optimized to accomplish situation-specific listening goals (Sendlmeier 1995, Goldinger & Azuma 2003).

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 no single phonemes or syllables (Mesgarani et al. 2014). In other words, STG-evoked neural responses are tuned to high-order auditory cues that correspond to the acoustic and articulatory features found across speech sounds. Neural populations in the STG are particularly sensitive to the manner of articulation, an articulatory feature that correlates with the greatest acoustic difference between vowels and consonants and between consonant categories (e.g., fricative versus plosive). Tuning to acoustic-phonetic features rather than phonemes is observed at the level of both single ECoG electrode contacts (thousands of neurons) and single neurons in the STG (Chan et al. 2014, Lakertz et al. 2021).

Acoustic-phonetic:

the acoustic and articulatory parameters of physical sound properties that are important for the realization of consonants and vowels

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 to acoustic phonetic features throughout the STG contains the information necessary to decode abstract phonemic categories, suggesting that spatial patterns of neural activity in the STG may reflect higher-order perceptual content. Different phonological units, such as the acoustic-phonetic feature and the phoneme, are thus realized at two different spatial scales (the local and the population ensemble, respectively). Of note, no cortical area has yet been shown to selectively encode single invariant phonemes or syllables at local sites.

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 phonological class, a process called categorization (Holt & Lotto 2010). These mappings are not only many-to-one but 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 (Mann 1980, Mann & Repp 1980, Johnson 2008). 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 whereby 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 used to describe nonlinearities in speech perception that allows researchers to quantify the extent to which incremental changes to an acoustic signal alter the 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 neurophysiologically 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 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 consonants, vowels, and lexical tone categories.

Consonants

In a classic example of categorical perception, Liberman et al. (1957) showed that when participants are asked to identify synthesized speech sounds that contain smoothly and continuously changing spectral content among the voiced plosive sound categories in English (i.e., /ba/, /ga/, and /da/), their perception of phoneme category changes abruptly rather than gradually. 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), the participants reported categorical perception of three distinct phonemes, suggesting nonlinear 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 categories. The underlying dimensions driving the neural activity were the onset frequency of the second spectral peak (labeled F2 in Figure 3) and the magnitude of the spectral transition, which corresponds to lingual articulatory movements during speech production (Lindblom & Sussman 2012). Interestingly, a decoding analysis revealed that several neighboring cortical sites discriminated between different category pairs, underscoring the functional heterogeneity across the posterior STG. This result suggests that a distributed neural representation of acoustic spectral cues can encode a nonlinearity matching the subjective experience of the listener. Although it remains unclear whether the neural responses at single cortical sites show strongly nonlinear tuning to these specific spectral cues, 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 (Figure 3).

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 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 (~50 ms) 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 (Steinschneider et al. 1999, 2011; Blumstein et al. 2005; Mesgarani et al. 2014; Fox et al. 2017, 2020). Functional magnetic resonance imaging (fMRI) studies of VOT perception reveal that the

activation of the 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/ versus /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 the 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 (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 that are selective for distinct categories of voicing.

Vowels

Vowel sounds are generated by the different shapes of the oral cavity, giving rise to resonances at particular frequencies known as formants. Whereas the fundamental frequency (F0) is determined by the rate at which the vocal cords vibrate during voiced sounds (with pitch as the perceptual correlate), 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) corresponds to resonance at higher frequencies. Formants are roughly correlated with the articulatory dimensions that characterize vowel sounds (i.e., the closeness of the tongue to the roof of the mouth in F1, the degree of tongue front-backness in F2, height, and backness, respectively). Vowels are not perceived as categorically as consonants are (Pisoni 1973); however, the perceptual vowel space is warped (nonlinearly) toward category prototypes, which is 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 sounds 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 those of speakers with shorter vocal tracts, and 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 listeners' representations of relevant linguistic content discard irrelevant speaker-dependent properties (Johnson 2008).

There is some evidence that vowel category information beyond the 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 to a narrowband 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. A majority of single electrode responses in English listeners show strong selectivity for either high-front (low F1, high F2) or low-back (high F1, low F2) vowels. 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 as well as absolute fundamental and formant frequencies (F0–F4). 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 fMRI study comparing STG activation in a speaker and vowel classification task, Bonte et al. (2014) found that different regions within the STG are differentially activated in the two task types, indicating a heterogeneous and task-dependent representation of vowel sounds. This is consistent with the findings of previous fMRI studies in which speaker and vowel identity were simultaneously decoded from interspersed regions of the temporal cortex (Formisano et al. 2008). Although it appears that distinct cortical sites are dedicated to representing speaker and formant information separately, these neuroimaging results do not necessarily preclude the existence of speaker-normalized formant encoding in the STG.

In a recent ECoG study addressing the question of speaker normalization, subjects were asked to categorize synthesized vowel sounds. Carrier sentences that preceded the synthesized vowels were produced by either a low- or a 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. For both speaker types, the target vowel sounds with extreme F1 frequencies corresponded unambiguously to either an /u/ or an /o/. However, the category of vowel sounds with intermediate F1 frequencies depended on the speaker type. Comparing psychometric and corresponding neurometric functions derived from STG activity revealed that both behavioral and neural responses to vowel sounds represented the speaker-normalized formant value (Sjerps et al. 2019). That is, the neurometric function derived from the STG neural responses over the range of F1 values shifted in accordance with the dynamic F1 range of the speaker. Whereas only a small subset of electrode responses showed consistent effects of speaker-dependent F1 tuning, multivariate decoding over regional electrode responses in the STG showed a strong speaker-dependent encoding of F1 frequencies. These findings demonstrate that the neural encoding of vowel sounds in the STG is normalized for speaker type, with reduced sensitivity to the acoustic details that are irrelevant for determining vowel category.

Neurometric function:

computed by relating changes in acoustic signal to changes in neural activity (magnitude, location, or latency) rather than behavior

Intonational Prosody and Lexical Tones

Intonational prosody refers to the changes in speaker pitch that occur over the course of a phrase or sentence and convey linguistic meaning. However, as is the case for vowel categories, intonational pitch is dependent on the speaker, such that linguistic meaning is embedded in the normalized relative pitch patterns of a sentence. For example, in the sentence "Jim likes to *ski*," raising the pitch on the final word may indicate that Jim likes to ski but not to swim. For different speakers, the raised pitch portion of the sentence likely corresponds to different absolute frequencies, as it is the relative pitch change that expresses linguistic meaning. In fMRI studies, blood oxygenation level-dependent (BOLD) activation in the STG depends on pitch change in speech and music (Zatorre et al. 2012, Allen et al. 2017) as well as on speaker identity (Formisano et al. 2008). Consistent with these studies, Tang et al. (2017) found that a subset of local neural populations in the middle STG encodes relative pitch, or the pitch contour normalized for the speaker's baseline pitch, by translating absolute pitch into speaker-invariant intonation patterns. This finding demonstrates that speaker-invariant relative pitch change can be extracted from the speech signal online and represented locally within single neural populations. The authors also report that populations that encode relative pitch change are distinct from, but interspersed with, those that encode speaker-identity (i.e., vocal tract length cued by fundamental frequency) and acoustic-phonetic features (Formisano et al. 2008, Tang et al. 2017). It remains an open question as to how these functionally diverse populations interact with one another in order to generate the listener's perceptual experience.

Intonation:

the changes in pitch that occur over the course of an utterance

Prosody:

the properties of syllables and larger units of speech that convey speech-related meaning, including intonation, tone, stress, and rhythm

In contrast to English, tonal languages use the pitch contour that occurs over a syllable to discriminate between words, a cue known as the lexical tone (Howie 1976). Similar to vowels, pitch contours that map to the same lexical tone category are highly variable across speakers and situational contexts (Ladd 2008) and are perceived categorically. Neuroimaging studies have indicated that the STG is involved in the processing of lexical tone (Zatorre & Gandour 2008, Feng et al. 2018, Liang & Du 2018). In an ECoG study exploring the effect of tonal language experience on STG activity, native English and Mandarin speakers passively listened to natural Mandarin speech stimuli (Figure 4a). In both native English and Mandarin speakers, the neural responses recorded at single electrodes in the STG encoded a language-independent representation of speaker-normalized pitch (i.e., relative pitch and pitch change; Figure 4b). In contrast, the responses pooled across electrode sites revealed a language-dependent representation warped toward tone category in native Mandarin speakers (Li et al. 2021). This could be a result of the relative proportion of local responses tuned to specific pitch changes. Whereas native Mandarin speakers

were found to have local sites tuned to both negative and positive pitch change, English speakers had predominant tuning to positive pitch changes (Figure 4b,c). In Mandarin, a balanced representation of pitch change in both positive and negative directions is critical for tone categorization. In English, because pitch information is primarily used for intonational processing, sensitivity to positive pitch changes may be sufficient for speech comprehension. These data suggest that although the underlying neural representation may correspond to language-invariant speaker-normalized pitch, language experience contributes to a distributed categorical encoding that biases the distribution of tuning parameters for speaker-normalized pitch.

Tonal language:

language in which the pitch contour is used to discriminate between words or grammatical forms

SPEECH DYNAMICS

Up to this point, we have reviewed evidence for nonlinearities in speech encoding in the human STG. However, natural speech generally consists of sequences, and the structure of the speech signal as it unfolds across time is critical to communication. In this section, 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 a series of beads on a string, or linear sequences of context-invariant linguistic units. That is, the perception of a single speech element, like a phoneme, can be considered 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 scales 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 Syllabic and Phrasal 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) (see the sidebar titled Early Dynamical Systems Approach to Speech Analysis). 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 the speech envelope is not sufficient for speech comprehension, especially in the presence of noise (Lorenzi et al. 2006, Hopkins & Moore 2009, Moon & Hong 2014).

Speech envelope:

slow modulation in the amplitude of the acoustic waveform over time

Stress:

the emphasis given to a syllable in a word, or a word in a sentence, with acoustic correlates including duration and peak intensity

Extensive literature shows that brain activity continuously tracks the speech envelope (Ahissar et al. 2001, Liégeois-Chauvel et al. 2004, Nourski et al. 2009, Peelle & Davis 2012, Drennan & Lalor 2019). 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 the nucleus. The peakRate event occurs *within* the syllable, in contrast to previous models that have postulated “chunking” at syllabic boundaries. Further, the magnitude of peakRate events, or the 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 the magnitude of the peakRate event, effectively encoding information about syllable structure, stress, and speech rate (Ohala 1975, Oganian & Chang 2019). These findings suggest that at the local neural population scale, there exists a discrete event-based neural encoding of syllable timing and stress opposed to an encoding that continuously tracks the envelope amplitude. Syllable information—including the timing, stress, and acoustic-phonetic content—can be thought of as distributed across several functionally distinct local neural populations.

The onset of speech after a brief period of silence is another event that elicits a unique neural response in the STG. Onsets play a pivotal role in the segmentation of the speech signal, cueing both phrasal and sentential units. Silent pauses at phrase boundaries are thought to bind auditory sequences, acting as a frame to hold auditory sequences in memory (Frazier et al. 2006). Neural populations in the posterior STG selectively respond to the onset of speech following at least 200 ms of silence (Hamilton et al. 2018). Notably, the posterior STG onset

populations are anatomically separated from middle STG populations that are sensitive to peakRate. Within both of these regions, however, there are neural populations that encode acoustic-phonetic features and relative pitch, some of which jointly encode these temporal cues as well.

Collectively, these studies demonstrate that there exists an efficient neural code for extracting key temporal landmarks in the speech signal such as peakRate and speech onset. Neural populations that encode the peakRate events are localized in the middle STG, whereas onset encoding is primarily found in a region of the posterior STG. The neural representation of these temporal cues is embedded in overlapping subdivisions of the STG, contributing to the heterogeneity of neural response types in the region. We speculate that these response types provide the temporal context information that is critical for speech processing.

Word-Level Contextual Dynamics

A highly illustrative example of the brain's ability to use linguistic knowledge, and specifically word-level context, to understand noisy acoustic input is the case of perceptual restoration, in which speech segments are entirely unavailable in the input (Warren 1970, Warren & Sherman 1974, Remez et al. 1981). 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 a 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 in which 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/ versus /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 they 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 occurring 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 ~300 ms 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 as well as possible 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 the 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 are 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), in which 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 show higher neural activation when SWS is intelligible compared to when it is not (Dehaene-Lambertz et al. 2005, Benson et al. 2006, Möttönen et al. 2006). Whereas 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 listening to the corresponding unfiltered speech reported an increase in intelligibility after hearing 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 becomes 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 exposed to SWS or vocoded speech did not find enhanced STG responses dependent on intelligibility, although other areas, including the 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?

Word-level knowledge seems to be integrated into the phonological representation of speech in the STG. This property illustrates the dynamic nature of phonological computations in this higher-order auditory area. Although it is difficult to parcel out the effects of prior linguistic knowledge on neural responses to natural speech corpora, controlled experimental paradigms that exploit masked and degraded speech signals expose the mechanisms by which word-level information is combined with phonological representation. Because the degree of degradation in natural speech is continuous, it is unlikely that this neurophysiological process of integration only occurs during extreme adverse listening conditions. Rather, the perceptual and corresponding neural systems for processing speech sounds may strongly rely on linguistic knowledge in any situations in which it is relevant and useful to do so.

FUNCTIONAL SPECIFICITY IN THE STG

To what extent are the mechanisms of phonological processing in the STG exclusive to speech stimuli? That is, are the nonlinear and context-sensitive features of phonological processing exhibited in the STG dependent on an explicit speech input? Whereas some have argued that a functionally and anatomically distinct speech processor in the brain readily performs the computations necessary to facilitate speech perception (Liberman et al. 1967, Liberman & Mattingly 1985), others have argued that the mechanisms for processing speech are identical to the auditory mechanisms in the brain that exist to process all sounds (for a review, see Diehl et al. 2004). Here, we attempt to reconcile these contrasting perspectives, arguing that STG processing is neither speech specific nor general auditory. Rather, auditory processing in the STG is tuned to the statistical properties of speech sounds and engages key nonlinear and dynamic neural mechanisms. The functional difference between the types of processing that occur in the primary and nonprimary auditory cortices (STG) is in part a consequence of anatomical circuitry (see the sidebar titled The Auditory Pathway).

Certain nonspeech sounds elicit neural activity in speech-responsive regions of the STG (Binder et al. 2000, Liebenthal et al. 2005, Leech et al. 2009, Hamilton et al. 2018). The ECoG studies described previously also used nonspeech controls for pitch contours and speech envelope amplitude ramps (Tang et al. 2017, Oganian & Chang 2019) that elicit activity in the same populations that encode intonation and peakRate cues in speech. Further, several studies have demonstrated that the cognitive operations supporting the categorical perception of speech sounds can be explained by a learned nonlinear biasing of underlying auditory receptive field processing due to extended experience and expertise (Holt & Lotto 2010, Liebenthal et al. 2010, Liu & Holt 2011). This may suggest that the difference between neural activities in response to speech and to simple nonspeech sounds (pure tones, unstructured noise) is in part due to a difference in acoustic signal complexity and familiarity.

The context-dependent predictive capabilities of the STG also are not necessarily speech specific. It has been shown that neural populations in the STG encode language-specific phoneme transition probabilities when presented with speech sequences, and this effect is modulated by the lexical status of the sequence (Leonard et al. 2015). However, the encoding of learned expectation in the auditory cortex when presented with sequences of

tones and other nonspeech stimuli is well documented (Heilbron & Chait 2018, Furl et al. 2011). It appears that the dynamic prediction mechanisms of nonprimary auditory cortices can be invoked by many different types of auditory sequences, but in the presence of speech they are modulated by linguistic knowledge and experience.

A RECURRENT MODEL OF PHONOLOGICAL PROCESSING IN THE STG

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 (see, for example, Blumstein 2009) (Figure 5a). For instance, two dual-stream models (Hickok & Poeppel 2004, 2007; 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 Hickok & Poeppel's (2004, 2007) model, a ventral processing stream begins with spectrotemporal analysis in the bilateral STG, followed by phonological processing in the bilateral STS and lexical processing in the middle and inferior temporal gyri. In contrast, in Rauschecker & Scott's (2009) model, spectrotemporal analysis is thought to occur in the posteromedial primary auditory cortex, and auditory word-form recognition is achieved as activity flows toward the anterior and lateral regions of the STG. Rauschecker & Scott's ventral processing stream runs parallel to the visual ventral stream in the infratemporal cortex (Rauschecker & Scott 2009, Jasmin et al. 2019). 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.

Neural activity in spatially distinct regions of the STG has been shown to be sensitive to specific spectrotemporal fluctuations (Schönwiesner & Zatorre 2009, Hullett et al. 2016, Santoro et al. 2017). Nonetheless, we have also established 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 syllabic segmentation. In Hickok & Poeppel's (2004, 2007) model, phonological processes are assumed to take place largely in the bilateral STS. In Rauschecker & Scott's (2009) model, phonological-relevant encoding is observed through the middle and anterior STG. Whereas in these models information is assumed to travel from the primary auditory to the nonprimary 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 the STG pose significant challenges to the 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 the evidence presented above, the functional organization of brain computations need not align with the levels of representation assumed in linguistics and psychology. Although there is little evidence for a dedicated cortical processing stream from phonemes to syllables to words, there exist important functional subdivisions of the STG, for example, that correspond to the temporal landmarks at different timescales described previously (Yi et al. 2019).

As an alternative to the models described above, we seek one that is explanatory but does not have strong commitments to predefined linguistic units (Figure 5c). To account for the neurophysiological evidence presented in this review, this alternative model of auditory word recognition must include the neural processing units characterized in previous neuroimaging and ECoG work. 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, onset), and prosodic features. It has yet to be determined how these local processors may interact with one another. Contextual restoration effects, which require a dynamic and time-dependent representation of speech sound sequences, can occur if the top-down word-level information can modulate phonological analysis in real time. 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, this proposed model emphasizes a distributed, dynamic representation of speech sounds that changes with time. 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 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 neurophysiologically implemented through the recurrent connections that link layers within and across cortical columns (Douglas & Martin 2007, Yi et al. 2019).

EMERGING PRINCIPLES AND OPEN QUESTIONS

Here, we synthesize the findings summarized above and review the key principles that guide phonological computations in the STG. We contend that neural populations in the STG exhibit nonlinear and dynamic representations of speech sounds. STG neurons are not passive sound filters or feature detectors but rather exert an interpretive function by integrating prior linguistic knowledge and recent temporal context in real time (Figure 5c).

Local Acoustic-Phonetic and Pitch Tuning Exhibit Key Nonlinearities

STG responses to speech are often nonlinear, resulting from a set of transformations that enable the perceptual system to extract linguistic content. In categorization, neural responses show stronger sensitivities to acoustic changes that occur across phonetically relevant

category boundaries rather than within the same category. In speaker normalization, neural responses to vowel sounds and pitch become invariant to speaker characteristics that do not contribute meaningfully to phonological content. Further detailed characterization of these nonlinearities is critical for understanding the diversity of local neural responses in the STG. Additional research questions may be posed such as, To what extent does the category information integrated into phonological analysis adapt to changing language contexts in the case of multilingual speech perception?

The Process by Which Linguistic Information Is Integrated into the Representation of Phonological Units Is Dynamic

STG cortical activity not only represents the instantaneous spectral content of the speech signal but also is highly time-dependent and influenced by the phonological context. Temporal landmarks such as peakRate and the onset of sound from silence cue critical syllabic and phrasal events that facilitate speech comprehension. It is unsurprising, then, that these key features are also represented by distinct local neuronal populations in the STG (Hamilton et al. 2018). In the cases of degraded or partially masked speech, listeners often perceive a restored version of the speech signal. STG neural populations respond in real time to context-sensitive segmental phonetic features when single speech sounds are selectively masked by noise. Together, these studies indicate that the neural implementation of speech processing in the STG is modulated by prior linguistic and contextual information. Further development of dynamic encoding models applied to neural activity may open the door to an exciting and largely unexplored line of research (Keshishian et al. 2020).

Acoustic-Phonetic and Higher-Order Linguistic Representations Coexist in the STG at Distinct Cortical Scales

Whereas the neural responses recorded from a single electrode or neuronal site represent acoustic-phonetic features, the spatial pattern of local responses across the STG are relevant for representing phonemes and other linguistic category-level information. 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 (Figure 5c). However, it remains unclear how perceptual units emerge from this distributed neural code: How are the local neural populations that encode distinct acoustic properties ultimately integrated into a cohesive and experience-based reflection of speech input?

CONCLUSION

Decades of neuroscientific research have shown that the human STG plays an essential role in speech perception, interfacing with both auditory and language-processing systems. Several important principles govern the patterns of cortical activity in the STG and suggest that this brain region may be performing more complex computations than originally assumed. Technological and experimental advances have enabled researchers to document the characters of functionally heterogeneous neural populations in the nonprimary auditory cortex. Despite this significant progress, many important challenges still 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.

ACKNOWLEDGMENTS

The authors would like to thank Matthew Leonard, Yulia Oganian, and the rest of the Chang lab for helpful comments on the manuscript and figures. This work was supported by National Institute of Health (NIH) grants R01-DC012379 and U01-NS117765 and by the National Science Foundation Graduate Research Fellowship Program.

LITERATURE CITED

Ahissar E, Nagarajan S, Ahissar M, Protopapas A, Mahncke H, Merzenich MM. 2001. Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. PNAS 98(23):13367–72 [PubMed: 11698688]

Allen EJ, Burton PC, Olman CA, Oxenham AJ. 2017. Representations of pitch and timbre variation in human auditory cortex. J. Neurosci 37(5):1284–93 [PubMed: 28025255]

Anderson LA, Linden JF. 2011. Physiological differences between histologically defined subdivisions in the mouse auditory thalamus. Hear. Res 274(1–2):48–60 [PubMed: 21185928]

Bartlett EL. 2013. The organization and physiology of the auditory thalamus and its role in processing acoustic features important for speech perception. Brain Lang 126(1):29–48 [PubMed: 23725661]

Benson RR, Richardson M, Whalen DH, Lai S. 2006. Phonetic processing areas revealed by sinewave speech and acoustically similar non-speech. NeuroImage 31(1):342–53 [PubMed: 16530428]

Benson RR, Whalen DH, Richardson M, Swainson B, Clark VP, et al. 2001. Parametrically dissociating speech and nonspeech perception in the brain using fMRI. Brain Lang 78(3):364–96 [PubMed: 11703063]

Binder JR, Frost JA, Hammeke TA, Bellgowan PS, Springer JA, et al. 2000. Human temporal lobe activation by speech and nonspeech sounds. Cereb. Cortex 10(5):512–28 [PubMed: 10847601]

Blumstein SE. 2009. Auditory word recognition: evidence from aphasia and functional neuroimaging. Lang. Linguist. Compass 3(4):824–38 [PubMed: 19915692]

Blumstein SE, Myers EB, Rissman J. 2005. The perception of voice onset time: an fMRI investigation of phonetic category structure. J. Cogn. Neurosci 17(9):1353–66 [PubMed: 16197689]

Boatman D 2004. Cortical bases of speech perception: evidence from functional lesion studies. Cognition 92(1–2):47–65 [PubMed: 15037126]

Boatman D, Gordon B, Hart J, Selnes O, Miglioretti D, Lenz F. 2000. Transcortical sensory aphasia: revisited and revised. Brain 123(Pt. 8):1634–42 [PubMed: 10908193]

Bonte M, Hausfeld L, Scharke W, Valente G, Formisano E. 2014. Task-dependent decoding of speaker and vowel identity from auditory cortical response patterns. J. Neurosci 34(13):4548–57 [PubMed: 24672000]

Chan AM, Dykstra AR, Jayaram V, Leonard MK, Travis KE, et al. 2014. Speech-specific tuning of neurons in human superior temporal gyrus. Cereb. Cortex 24(10):2679–93 [PubMed: 23680841]

Chang EF. 2015. Towards large-scale, human-based, mesoscopic neurotechnologies. Neuron 86(1):68–78 [PubMed: 25856487]

Chang EF, Rieger JW, Johnson K, Berger MS, Barbaro NM, et al. 2010. Categorical speech representation in human superior temporal gyrus. Nat. Neurosci 13(11):1428–32 [PubMed: 20890293]

Chistovich LA. 1980. Auditory processing of speech. Lang. Speech 23(1):67–73 [PubMed: 7421370]

Crone NE, Boatman D, Gordon B, Hao L. 2001. Induced electrocorticographic gamma activity during auditory perception. Clin. Neurophysiol 112(4):565–82 [PubMed: 11275528]

Cutler A, Mehler J, Norris D, Segui J. 1986. The syllable's differing role in the segmentation of French and English. J. Mem. Lang 25(4):385–400

Cutler A, Otake T. 1994. Mora or phonemes? Further evidence for language-specific listening. J. Mem. Lang 33(6):824–44

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 [PubMed: 15588593]

Diehl RL, Lotto AJ, Holt LL. 2004. Speech perception. Annu. Rev. Psychol 55:149–79 [PubMed: 14744213]

Douglas RJ, Martin KAC. 2007. Recurrent neuronal circuits in the neocortex. Curr. Biol 17(13):R496–500 [PubMed: 17610826]

Drennan DP, Lalor EC. 2019. Cortical tracking of complex sound envelopes: modeling the changes in response with intensity. eNeuro 6(3). 10.1523/ENEURO.0082-19.2019

Drullman R 1995. Temporal envelope and fine structure cues for speech intelligibility. J. Acoust. Soc. Am 97(1):585–92 [PubMed: 7860835]

Elman JL. 1990. Finding structure in time. Cogn. Sci 14(2):179–211

Feng G, Gan Z, Wang S, Wong PCM, Chandrasekaran B. 2018. Task-general and acoustic-invariant neural representation of speech categories in the human brain. Cereb. Cortex 28(9):3241–54 [PubMed: 28968658]

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):970–73 [PubMed: 18988858]

Fox NP, Leonard MK, Sjerps MJ, Chang EF. 2020. Transformation of a temporal speech cue to a spatial neural code in human auditory cortex. eLife 9:e53051 [PubMed: 32840483]

Fox NP, Sjerps MJ, Chang EF. 2017. Dynamic emergence of categorical perception of voice-onset time in human speech cortex. J. Acoust. Soc. Am 141(5):3571–71

Frazier L, Carlson K, Clifton C Jr. 2006. Prosodic phrasing is central to language comprehension. Trends Cogn. Sci 10(6):244–49 [PubMed: 16651019]

Furl N, Kumar S, Alter K, Durrant S, Shawe-Taylor J, Griffiths TD. 2011. Neural prediction of higher-order auditory sequence statistics. NeuroImage 54(3):2267–77 [PubMed: 20970510]

Garofolo JS, Lamel LF, Fisher WM, Fiscus JG, Pallett DS. 1993. DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1–1.1. CD-ROM https://ui.adsabs.harvard.edu/abs/1993STIN…9327403G/abstract

Geschwind N 1970. The organization of language and the brain. Science 170(3961):940–44 [PubMed: 5475022]

Goldinger SD, Azuma T. 2003. Puzzle-solving science: the quixotic quest for units in speech perception. J. Phonet 31(3–4):305–20

Halle M, Chomsky N. 1968. The Sound Pattern of English New York: Harper & Row

Hamilton LS, Edwards E, Chang EF. 2018. A spatial map of onset and sustained responses to speech in the human superior temporal gyrus. Curr. Biol 28(12):1860–71.e4 [PubMed: 29861132]

Hamilton LS, Oganian Y, Chang EF. 2020. Topography of speech-related acoustic and phonological feature encoding throughout the human core and parabelt auditory cortex. bioRxiv 121624. 10.1101/2020.06.08.121624

Healy AF, Cutting JE. 1976. Units of speech perception: phoneme and syllable. J. Verb. Learn. Verb. Behav 15(1):73–83

Heil P, Neubauer H. 2001. Temporal integration of sound pressure determines thresholds of auditory-nerve fibers. J. Neurosci 21(18):7404–15 [PubMed: 11549751]

Heilbron M, Chait M. 2018. Great expectations: Is there evidence for predictive coding in auditory cortex? Neuroscience 389:54–73 [PubMed: 28782642]

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 [PubMed: 15037127]

Hickok G, Poeppel D. 2007. The cortical organization of speech processing. Nat. Rev. Neurosci 8(5):393–402 [PubMed: 17431404]

Hillis AE, Rorden C, Fridriksson J. 2017. Brain regions essential for word comprehension: drawing inferences from patients. Ann. Neurol 81(6):759–68 [PubMed: 28445916]

Holdgraf CR, de Heer W, Pasley B, Rieger J, Crone N, et al. 2016. Rapid tuning shifts in human auditory cortex enhance speech intelligibility. Nat. Commun 7:13654 [PubMed: 27996965]

Holt LL, Lotto AJ. 2010. Speech perception as categorization. Attent. Percept. Psychophys 72(5):1218–27

Hopkins K, Moore BCJ. 2009. The contribution of temporal fine structure to the intelligibility of speech in steady and modulated noise. J. Acoust. Soc. Am 125(1):442–46 [PubMed: 19173429]

Howie JM. 1976. Acoustical Studies of Mandarin Vowels and Tones Cambridge, UK: Cambridge Univ. Press

Hullett PW, Hamilton LS, Mesgarani N, Schreiner CE, Chang EF. 2016. Human superior temporal gyrus organization of spectrotemporal modulation tuning derived from speech stimuli. J. Neurosci 36(6):2014–26 [PubMed: 26865624]

Humphries C, Sabri M, Lewis K, Liebenthal E. 2014. Hierarchical organization of speech perception in human auditory cortex. Front. Neurosci 8:406 [PubMed: 25565939]

Hutchison ER, Blumstein SE, Myers EB. 2008. An event-related fMRI investigation of voice-onset time discrimination. NeuroImage 40(1):342–52 [PubMed: 18248740]

Jasmin K, Lima CF, Scott SK. 2019. Understanding rostral-caudal auditory cortex contributions to auditory perception. Nat. Rev. Neurosci 20(7):425–34 [PubMed: 30918365]

Johnson K 2008. Speaker normalization in speech perception. In The Handbook of Speech Perception, ed. Pisoni DB, Remez RE, pp. 363–89. Malden, MA: Blackwell

Jordan MI. 1997. Serial order: a parallel distributed processing approach. In Advances in Psychology, ed. Donahoe JW, Packard Dorsel V, pp. 471–95. Amsterdam: North-Holland

Keshishian M, Akbari H, Khalighinejad B, Herrero JL, Mehta AD, Mesgarani N. 2020. Estimating and interpreting nonlinear receptive field of sensory neural responses with deep neural network models. eLife 9:e53445 [PubMed: 32589140]

Khoshkhoo S, Leonard MK, Mesgarani N, Chang EF. 2018. Neural correlates of sine-wave speech intelligibility in human frontal and temporal cortex. Brain Lang 187:83–91 [PubMed: 29397190]

Kuhl PK. 1991. Human adults and human infants show a “perceptual magnet effect” for the prototypes of speech categories, monkeys do not. Percept. Psychophys 50(2):93–107 [PubMed: 1945741]

Kuhl PK, Williams KA, Lacerda F, Stevens KN, Lindblom B. 1992. Linguistic experience alters phonetic perception in infants by 6 months of age. Science 255(5044):606–8 [PubMed: 1736364]

Ladd R 2008. Intonational Phonology Cambridge, UK: Cambridge Univ. Press

Ladefoged P, Broadbent DE. 1957. Information conveyed by vowels. J. Acoust. Soc. Am 29(1):98–104

Ladefoged P, Johnson K. 2014. A Course in Phonetics Toronto: Nelson Educ.

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 [PubMed: 33186717]

Leech R, Holt LL, Devlin JT, Dick F. 2009. Expertise with artificial nonspeech sounds recruits speech-sensitive cortical regions. J. Neurosci 29(16):5234–39 [PubMed: 19386919]

Leonard MK, Baud MO, Sjerps MJ, Chang EF. 2016. Perceptual restoration of masked speech in human cortex. Nat. Commun 7:13619 [PubMed: 27996973]

Leonard MK, Bouchard KE, Tang C, Chang EF. 2015. Dynamic encoding of speech sequence probability in human temporal cortex. J. Neurosci 35(18):7203–14 [PubMed: 25948269]

Leonard MK, Cai R, Babiak MC, Ren A, Chang EF. 2019. The peri-Sylvian cortical network underlying single word repetition revealed by electrocortical stimulation and direct neural recordings. Brain Lang 193:58–72 [PubMed: 27450996]

Lesogor LV, Chistovich LA. 1978. Detection of consonant in two-component complex sounds and interpretation of stimulus as a sequence of elements. Fiziol. Cheloveka 4:213–19

Leszczyński M, Barczak A, Kajikawa Y, Ulbert I, Falchier AY, et al. 2020. Dissociation of broadband high-frequency activity and neuronal firing in the neocortex. Sci. Adv 6(33):eabb0977

Li Y, Tang C, Lu J, Wu J, Chang EF. 2021. Human cortical encoding of pitch in tonal and non-tonal languages. Nat. Commun 12:1161 [PubMed: 33608548]

Liang B, Du Y. 2018. The functional neuroanatomy of lexical tone perception: an activation likelihood estimation meta-analysis. Front. Neurosci 12:495 [PubMed: 30087589]

Liberman AM, Cooper FS, Shankweiler DP, Studdert-Kennedy M. 1967. Perception of the speech code. Psychol. Rev 74(6):431–61 [PubMed: 4170865]

Liberman AM, Delattre PC, Cooper FS. 1952. The role of selected stimulus-variables in the perception of the unvoiced stop consonants. Am. J. Psychol 65(4):497–516 [PubMed: 12996688]

Liberman AM, Delattre PC, Cooper FS. 1958. Some cues for the distinction between voiced and voiceless stops in initial position. Lang. Speech 1(3):153–67

Liberman AM, Harris KS, Hoffman HS, Griffith BC. 1957. The discrimination of speech sounds within and across phoneme boundaries. J. Exp. Psychol 54(5):358–68 [PubMed: 13481283]

Liberman AM, Mattingly IG. 1985. The motor theory of speech perception revised. Cognition 21(1):1–36 [PubMed: 4075760]

Liebenthal E, Binder JR, Spitzer SM, Possing ET, Medler DA. 2005. Neural substrates of phonemic perception. Cereb. Cortex 15(10):1621–31 [PubMed: 15703256]

Liebenthal E, Desai R, Ellingson MM, Ramachandran B, Desai A, Binder JR. 2010. Specialization along the left superior temporal sulcus for auditory categorization. Cereb. Cortex 20(12):2958–70 [PubMed: 20382643]

Liégeois-Chauvel C, Lorenzi C, Trébuchon A, Régis J, Chauvel P. 2004. Temporal envelope processing in the human left and right auditory cortices. Cereb. Cortex 14(7):731–40 [PubMed: 15054052]

Lindblom B, Sussman HM. 2012. Dissecting coarticulation: how locus equations happen. J. Phonet 40(1):1–19

Lisker L, Abramson AS. 1964. A cross-language study of voicing in initial stops: acoustical measurements. Word World 20(3):384–422

Liu R, Holt LL. 2011. Neural changes associated with nonspeech auditory category learning parallel those of speech category acquisition. J. Cogn. Neurosci 23(3):683–98 [PubMed: 19929331]

Lorenzi C, Gilbert G, Carn H, Garnier S, Moore BCJ. 2006. Speech perception problems of the hearing impaired reflect inability to use temporal fine structure. PNAS 103(49):18866–69 [PubMed: 17116863]

Lublinskaja V, Ross J, Ogorodnikova EV. 2006. Auditory perception and processing of amplitude modulation in speech-like signals: legacy of the Chistovich-Kozhevnikov group. In Dynamics of Speech Production and Perception, ed. Divenyi PL, Greenberg S, Meyer G, pp. 87–101. Clifton, VA: IOS Press

Mann VA. 1980. Influence of preceding liquid on stop-consonant perception. Percept. Psychophys 28(5):407–12 [PubMed: 7208250]

Mann VA, Repp BH. 1980. Influence of vocalic context on perception of the [[ʃ]]-[s] distinction. Percept. Psychophys 28(3):213–28 [PubMed: 7432999]

Marslen-Wilson WD, Welsh A. 1978. Processing interactions and lexical access during word recognition in continuous speech. Cogn. Psychol 10(1):29–63

Massaro DW. 1974. Perceptual units in speech recognition. J. Exp. Psychol 102(2):199–208 [PubMed: 4811941]

Matsumoto R, Imamura H, Inouchi M, Nakagawa T, Yokoyama Y, et al. 2011. Left anterior temporal cortex actively engages in speech perception: a direct cortical stimulation study. Neuropsychologia 49(5):1350–54 [PubMed: 21251921]

McClelland JL, Elman JL. 1986. The TRACE model of speech perception. Cogn. Psychol 18(1):1–86 [PubMed: 3753912]

Mesgarani N, Cheung C, Johnson K, Chang EF. 2014. Phonetic feature encoding in human superior temporal gyrus. Science 343(6174):1006–10 [PubMed: 24482117]

Moon IJ, Hong SH. 2014. What is temporal fine structure and why is it important? Korean J. Audiol 18(1):1–7 [PubMed: 24782944]

Möttönen R, Calvert GA, Jääskeläinen IP, Matthews PM, Thesen T, et al. 2006. Perceiving identical sounds as speech or non-speech modulates activity in the left posterior superior temporal sulcus. NeuroImage 30(2):563–69 [PubMed: 16275021]

Nittrouer S, Miller ME, Crowther CS, Manhart MJ. 2000. The effect of segmental order on fricative labeling by children and adults. Percept. Psychophys 62(2):266–84 [PubMed: 10723207]

Norris D, McQueen JM, Cutler A. 2000. Merging information in speech recognition: Feedback is never necessary. Behav. Brain Sci 23(3):299–325; discuss. 325–70 [PubMed: 11301575]

Nourski KV, Reale RA, Oya H, Kawasaki H, Kovach CK, et al. 2009. Temporal envelope of time-compressed speech represented in the human auditory cortex. J. Neurosci 29(49):15564–74 [PubMed: 20007480]

Nourski KV, Steinschneider M, Rhone AE, Kovach CK, Kawasaki H, Howard MA. 2019. Differential responses to spectrally degraded speech within human auditory cortex: an intracranial electrophysiology study. Hear. Res 371:53–65 [PubMed: 30500619]

Oganian Y, Chang EF. 2019. A speech envelope landmark for syllable encoding in human superior temporal gyrus. Sci. Adv 5(11):eaay6279

Ohala JJ. 1975. The temporal regulation of speech. In Auditory Analysis and Perception of Speech, ed. Fant G, Tatham MAA, pp. 431–53. San Diego, CA: Academic

Overath T, McDermott JH, Zarate JM, Poeppel D. 2015. The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. Nat. Neurosci 18(6):903–11 [PubMed: 25984889]

Ozker M, Schepers IM, Magnotti JF, Yoshor D, Beauchamp MS. 2017. A double dissociation between anterior and posterior superior temporal gyrus for processing audiovisual speech demonstrated by electrocorticography. J. Cogn. Neurosci 29(6):1044–60 [PubMed: 28253074]

Parvizi J, Kastner S. 2018. Promises and limitations of human intracranial electroencephalography. Nat. Neurosci 21(4):474–83 [PubMed: 29507407]

Peelle JE, Davis MH. 2012. Neural oscillations carry speech rhythm through to comprehension. Front. Psychol 3:320 [PubMed: 22973251]

Pisoni DB. 1973. Auditory and phonetic memory codes in the discrimination of consonants and vowels. Percept. Psychophys 13(2):253–60 [PubMed: 23226880]

Rauschecker JP, Scott SK. 2009. Maps and streams in the auditory cortex: Nonhuman primates illuminate human speech processing. Nat. Neurosci 12(6):718–24 [PubMed: 19471271]

Ray S, Maunsell JHR. 2011. Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. PLOS Biol 9(4):e1000610 [PubMed: 21532743]

Remez RE, Rubin PE, Pisoni DB, Carrell TD. 1981. Speech perception without traditional speech cues. Science 212(4497):947–49 [PubMed: 7233191]

Roux F-E, Miskin K, Durand J-B, Sacko O, Réhault E, et al. 2015. Electrostimulation mapping of comprehension of auditory and visual words. Cortex 71:398–408 [PubMed: 26332785]

Samuel AG. 1987. Lexical uniqueness effects on phonemic restoration. J. Mem. Lang 26(1):36–56

Santoro R, Moerel M, De Martino F, Valente G, Ugurbil K, et al. 2017. Reconstructing the spectrotemporal modulations of real-life sounds from fMRI response patterns. PNAS 114(18):4799–804 [PubMed: 28420788]

Savin HB, Bever TG. 1970. The nonperceptual reality of the phoneme. J. Verb. Learn. Verb. Behav 9(3):295–302

Schönwiesner M, Zatorre RJ. 2009. Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. PNAS 106(34):14611–16 [PubMed: 19667199]

Sendlmeier WF. 1995. Feature, phoneme, syllable or word: How is speech mentally represented? Phonetica 52(3):131–43 [PubMed: 7568391]

Shannon RV, Zeng FG, Kamath V, Wygonski J, Ekelid M. 1995. Speech recognition with primarily temporal cues. Science 270(5234):303–4 [PubMed: 7569981]

Sjerps MJ, Fox NP, Johnson K, Chang EF. 2019. Speaker-normalized sound representations in the human auditory cortex. Nat. Commun 10(1):2465 [PubMed: 31165733]

Steinschneider M, Fishman YI, Arezzo JC. 2008. Spectrotemporal analysis of evoked and induced electroencephalographic responses in primary auditory cortex (A1) of the awake monkey. Cereb. Cortex 18(3):610–25 [PubMed: 17586604]

Steinschneider M, Nourski KV, Kawasaki H, Oya H, Brugge JF, Howard MA. 2011. Intracranial study of speech-elicited activity on the human posterolateral superior temporal gyrus. Cereb. Cortex 21(10):2332–47 [PubMed: 21368087]

Steinschneider M, Volkov IO, Noh MD, Garrell PC, Howard MA. 1999. Temporal encoding of the voice onset time phonetic parameter by field potentials recorded directly from human auditory cortex. J. Neurophysiol 82(5):2346–57 [PubMed: 10561410]

Tang C, Hamilton LS, Chang EF. 2017. Intonational speech prosody encoding in the human auditory cortex. Science 357(6353):797–801 [PubMed: 28839071]

Towle VL, Yoon H-A, Castelle M, Edgar JC, Biassou NM, et al. 2008. ECoG gamma activity during a language task: differentiating expressive and receptive speech areas. Brain 131(Pt. 8):2013–27 [PubMed: 18669510]

Tuller B, Nguyen N, Lancia L, Vallabha GK. 2011. Nonlinear dynamics in speech perception. In Nonlinear Dynamics in Human Behavior, ed. Huys R, Jirsa VK, pp. 135–50. Berlin: Springer

Warren RM. 1970. Perceptual restoration of missing speech sounds. Science 167(3917):392–93 [PubMed: 5409744]

Warren RM, Sherman GL. 1974. Phonemic restorations based on subsequent context. Percept. Psychophys 16(1):150–56

Wernicke C 1874. Der aphasische Symptomencomplex: Eine psychologische Studie auf anatomischer Basis Breslau, Ger.: Max Cohn & Weigert

Yi HG, Leonard MK, Chang EF. 2019. The encoding of speech sounds in the superior temporal gyrus. Neuron 102(6):1096–110 [PubMed: 31220442]

Zatorre RJ, Delhommeau K, Zarate JM. 2012. Modulation of auditory cortex response to pitch variation following training with microtonal melodies. Front. Psychol 3:544 [PubMed: 23227019]

Zatorre RJ, Evans AC, Meyer E, Gjedde A. 1992. Lateralization of phonetic and pitch discrimination in speech processing. Science 256(5058):846–49 [PubMed: 1589767]

Zatorre RJ, Gandour JT. 2008. Neural specializations for speech and pitch: moving beyond the dichotomies. Philos. Trans. R. Soc. B 363(1493):1087–104

Zeng F-G, Nie K, Stickney GS, Kong Y-Y, Vongphoe M, et al. 2005. Speech recognition with amplitude and frequency modulations. PNAS 102(7):2293–98 [PubMed: 15677723]

EARLY DYNAMICAL SYSTEMS APPROACH TO SPEECH ANALYSIS

In the mid-twentieth century, researchers treated the speech signal as a sequence of spectrally detailed time slices mapping to static vocal tract shapes. However, Ludmilla Chistovich (1924–2006), a pioneering Soviet speech scientist, recognized the importance of acoustic dynamics originating from vocal tract movement. Chistovich and colleagues discovered that the speech envelope rise time influences the perceived syllabic structure (Lesogor & Chistovich 1978). By manipulating only the rise time duration, they found that participants' perception of syllable structure changed predictably, revealing that rise time events are critical for temporal order judgement. Chistovich proposed a perceptual system with two parallel auditory analyzers to account for these experimental results. One detects rapid increases in acoustic energy regardless of where in the spectrum changes occur, whereas the other performs spectral analysis. In combination, these analyses give rise to event-based speech perception. Neurally, Chistovich proposed this would correspond to tonic responses, which sustain activity over the course of the signal, and phasic responses, which act as temporal markers for the onset and offset of specific speech sounds (Lublinskaja et al. 2006). This seminal work was an important prelude to later dynamical systems approaches to speech, and it offers a theoretical interpretation of the peakRate detection neural responses observed in the STG.

THE AUDITORY PATHWAY

The primary auditory pathway extending from the cochlea to the cortex is composed of several subcortical synapses. A bundle of axons from the cochlea, known as the auditory nerve, innervates the cochlear and olivary nuclei in the brain stem. From the brainstem, axons that are transmitting auditory information project to the inferior colliculus (IC) in the midbrain. Subdivisions of IC project to distinct subdivisions of the medial geniculate nuclei (MGN) in the auditory thalamus. The lemniscal pathway refers to the projections from the IC central nucleus to the ventral MGN and cortical layers 3/4 of the primary auditory cortex. In contrast, the nonlemniscal pathway consists of the projections from the dorsal IC to the dorsal MGN and cortical layers 3/4 of the secondary auditory cortex, including STG (Bartlett 2013). The anatomical differences between the two pathways correspond to important functional and representational distinctions. Whereas the lemniscal pathway is known to transmit a “high-fidelity, primary-like representation of sound features” (Anderson & Linden 2011, p. 48), the nonlemniscal pathway relays context-dependent information and includes highly adaptive neuronal components that are capable of detecting change (Anderson & Linden 2011).

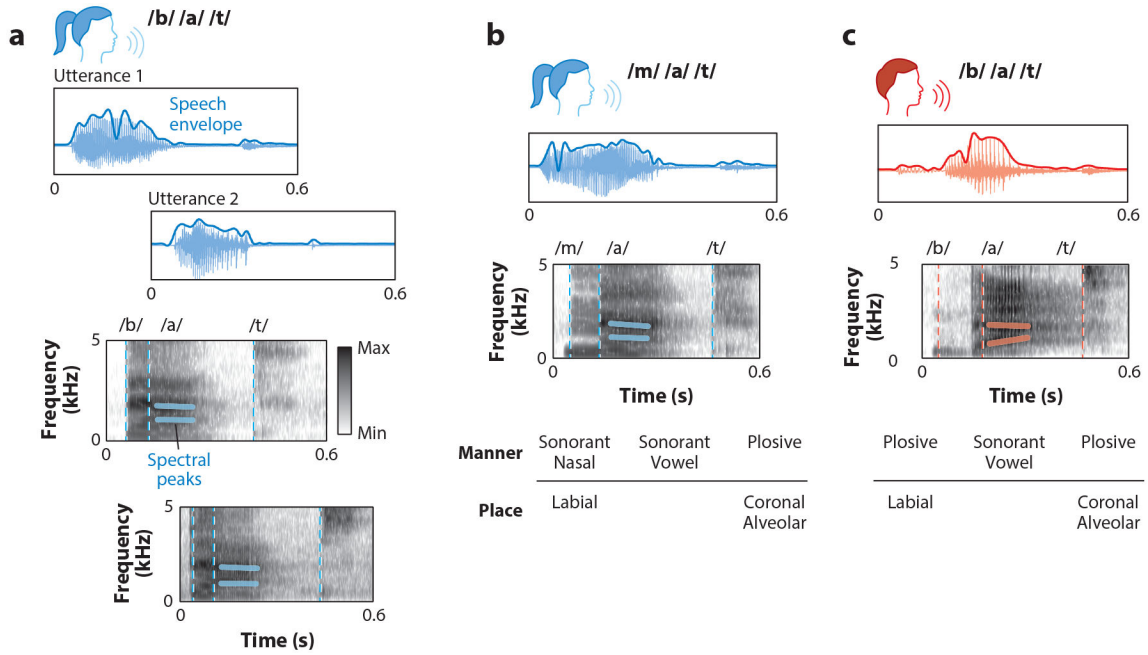

Figure 1.

Within- and between-speaker variability pose a challenge to speech comprehension. (*a*) A higher-pitch speaker produces two instances of “bat” slightly differently (labeled as utterance 1 and utterance 2), but both speech sequences map onto the same linguistic content. Key within-speaker differences in the speech waveform and spectrogram representation of the acoustic signal include changes in the amplitude of the speech envelope, shifted spectral peaks, and different final phoneme durations. (*b*) The same speaker as in panel *a* produces the word “mat.” Corresponding acoustic-phonetic features are shown in the lowest panel, indicating the manner and place of the articulatory gesture that produces the corresponding sound. (*c*) A different, lower-pitch speaker than the speakers in panels *a* and *b* produces the word “bat.” Key between-speaker differences in the speech waveform and spectrogram representation of the acoustic signal include changes in the amplitude of the speech envelope and shifted spectral peaks. Between-speaker variability can be due to several specific speaker characteristics, such as the length of the speaker’s vocal tract, speaker rate, and accent.

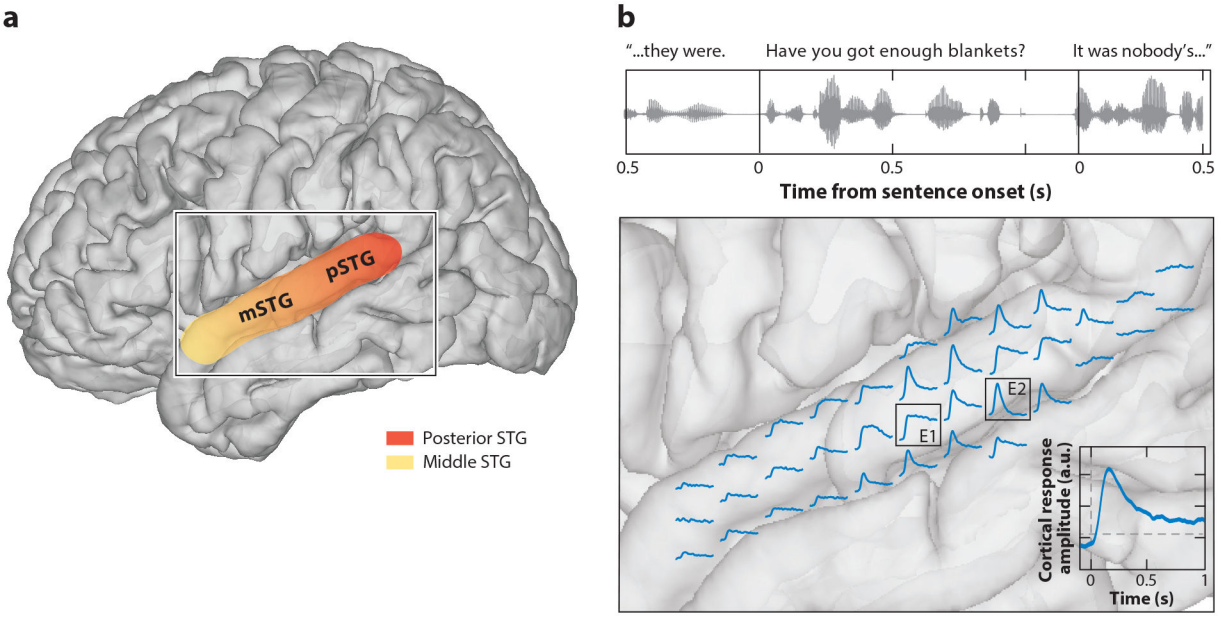

Figure 2.

ECoG enables high-resolution recording of neural activity in the nonprimary auditory cortex. (*a*) This panel illustrates the anatomical boundary of the STG. The color gradient represents the functionally differentiated posterior and middle regions of the STG (Ozker et al. 2017, Yi et al. 2019, Hamilton et al. 2020). (*b*) Example sentences from the TIMIT corpus are shown at the top, 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 (showing slow sustained cortical response for the electrode labeled E1 and rapid response to sentence onset for the electrode labeled E2). Abbreviations: ECoG, electrocorticogram; mSTG, middle superior temporal gyrus; pSTG, posterior superior temporal gyrus; STG, superior temporal gyrus.

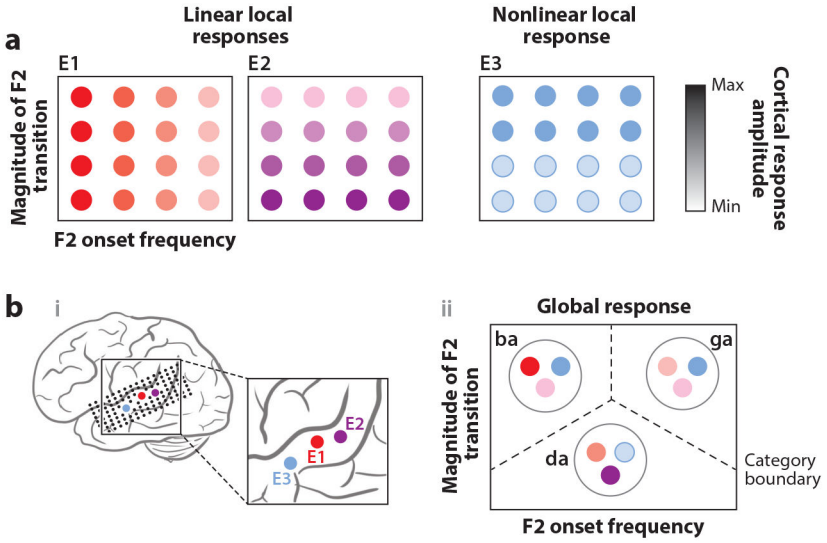

Figure 3.

Patterns of activity across the STG allow for the categorization of phonological units. (*a*) Single electrodes (selected electrodes are shown as colored circles labeled E1, E2, and E3) respond to incremental acoustic change, showing graded linear (E1 and E2) or abrupt nonlinear (E3) 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 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 the electrodes (selected electrodes illustrated in subpanel *i*) can be used to determine the phonemic category of presented speech sounds (e.g., /ba/, /da/, /ga/) (subpanel *ii*) and reflects the perceptual experience of the listener. Further, overlapping functionality in the neural code (*blue* and *purple circles* in the rightmost diagram) may be important for retaining within-category sensitivities. Abbreviations: F2, second spectral peak; STG, superior temporal gyrus.

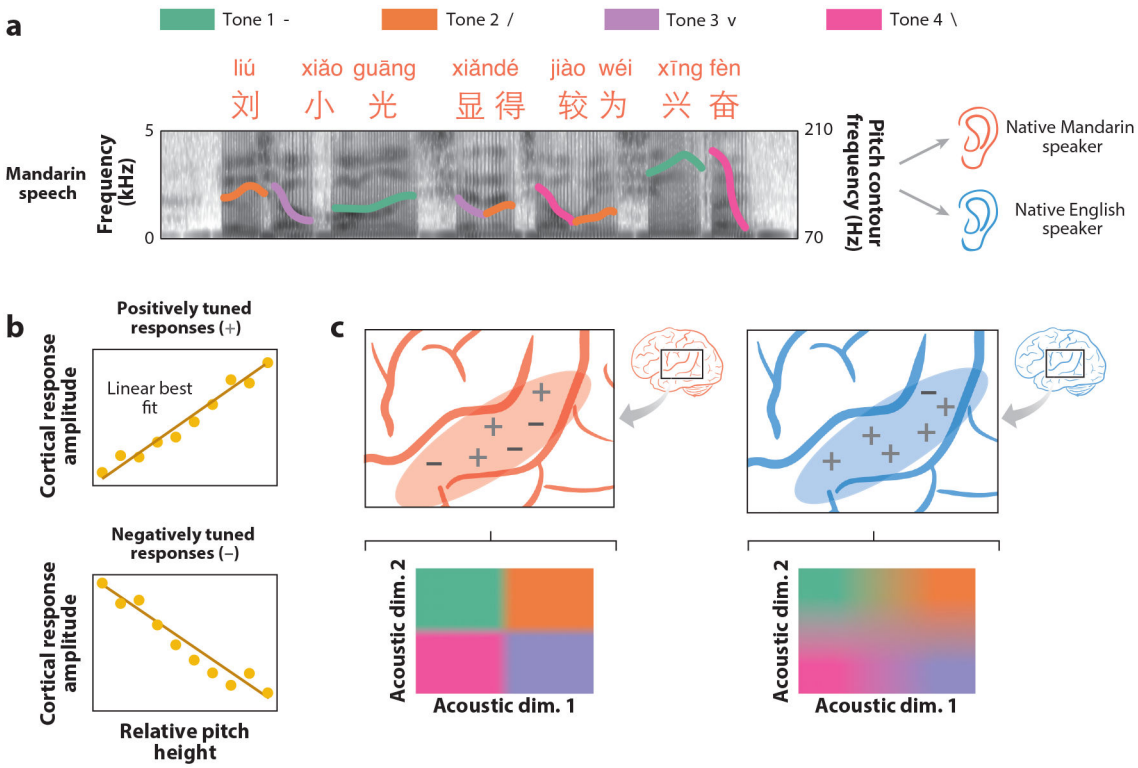

Figure 4.

Language-dependent neural tuning supports the categorization of lexical tone. (*a*) Four distinct tone categories (high, rising, dipping, and falling) were included in this experiment, in which native Mandarin and English speakers were presented with naturally produced Mandarin speech (Li et al. 2021). This panel shows an example Mandarin sentence with the extracted pitch contour overlaid on a spectrogram representation of the speech sequence (color of the contour indicates corresponding tone category). (*b*) Single electrode responses to relative pitch height can be categorized based on the positive or negative relationship between relative pitch height (*x*-axis) and cortical response amplitude (*y*-axis). (*c*) Analysis of electrode pitch encoding reveals a balanced distribution of STG electrodes in native Mandarin speakers that are either negatively or positively tuned to relative pitch (−, +). In native English speakers, STG electrodes show primarily positive relative pitch tuning (+). Whereas lexical tone category can be decoded from the population-level neural response in native Mandarin speakers, the decodability of lexical tone is significantly reduced in English native speakers. These results indicate that the distribution of STG pitch tuning is biased depending on the language experience of the listener.

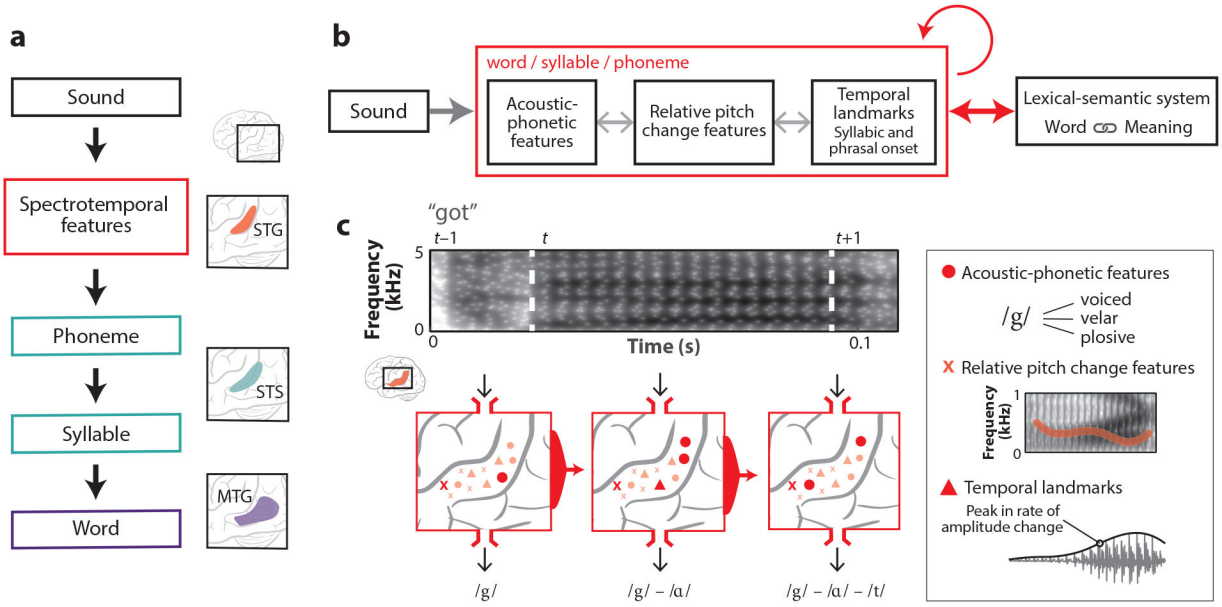

Figure 5.

A new model of phonological analysis. (*a*) Classical model of auditory word recognition in which primarily serial, feedforward, hierarchical processing takes place. The first processing step is spectrotemporal analysis, through which relevant features are extracted. Spectrotemporal features are grouped into phonemic segments that are then sequentially assembled into syllables. Finally, the lexical interface maps phonological sequences onto word-level representations. In classic models of auditory word recognition, each processing step is assigned to an approximate anatomical location (the schematic to the right shows an example of these assignments). The neural representation of speech becomes of increasingly higher order as it moves 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) (see Figure 2). 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. Abbreviations: MTG, middle temporal gyrus, STG, superior temporal gyrus; STS, superior temporal sulcus.