CellPress

Neuron
Review

The Encoding of Speech Sounds in the Superior Temporal Gyrus

Han Gyol Yi,[1,2] Matthew K. Leonard,[1,2] and Edward F. Chang[1,*]
[1]Department of Neurological Surgery, University of California, San Francisco, 675 Nelson Rising Lane, San Francisco, CA 94158, USA
[2]These authors contributed equally
*Correspondence: edward.chang@ucsf.edu
https://doi.org/10.1016/j.neuron.2019.04.023

The human superior temporal gyrus (STG) is critical for extracting meaningful linguistic features from speech input. Local neural populations are tuned to acoustic-phonetic features of all consonants and vowels and to dynamic cues for intonational pitch. These populations are embedded throughout broader functional zones that are sensitive to amplitude-based temporal cues. Beyond speech features, STG representations are strongly modulated by learned knowledge and perceptual goals. Currently, a major challenge is to understand how these features are integrated across space and time in the brain during natural speech comprehension. We present a theory that temporally recurrent connections within STG generate context-dependent phonological representations, spanning longer temporal sequences relevant for coherent percepts of syllables, words, and phrases.

Speech is a unique form of communication that enables humans to convey an unlimited range of thoughts and ideas with a limited set of fundamental elements. Linguists have characterized the units and structures of speech sounds that make up the world's spoken languages through a system known as phonology (Baudouin de Courtenay, 1972; De Saussure, 1879; Sapir, 1925). Although phonology provides a useful description of the sound structure of speech, we have a strikingly incomplete understanding of its implementation in terms of neural computations in the human brain.

Here we examine the nature of speech representation in the human superior temporal gyrus (STG), which sits at a functional and anatomical interface between lower-level auditory structures and higher-level association areas that support abstract aspects of language. Injury to the mid- to posterior part of STG results in an array of profound deficits in speech comprehension (Wernicke, 1874, 1881), an observation that has led to the view that this region is an important locus for speech perception (Geschwind, 1970). However, it remains unclear why these deficits arise when these specific neural structures are damaged.

Converging evidence from non-invasive functional magnetic resonance imaging (fMRI) (Binder et al., 2000; DeWitt and Rauschecker, 2012; Price, 2012; Scott et al., 2000) and electro- and magneto-encephalography (EEG and MEG; Di Liberto et al., 2015; Giraud and Poeppel, 2012; Gwilliams et al., 2018; Sohoglu et al., 2012; Wöstmann et al., 2017) has implicated the STG in various aspects of phonological processing. These studies have helped to shape important theories regarding the localization of speech and language function, and they have also raised fundamental questions about the nature of phonological representation. What sound features are encoded in STG? How do they correspond to both acoustic and linguistic descriptions of speech? What computational principles underlie the higher-order auditory processing that is necessary for extracting relevant structure and information from speech?

In this review, we focus on the emerging role of high-density intracranial neurophysiological recordings in humans to address these questions. The high spatial and temporal resolution of direct recordings has facilitated a deeper investigation of the nature of speech representation in the human cortex at the scale of millimeters and milliseconds. These methods have enabled the estimation of receptive fields at local sites as well as population ensemble activity at the rapid timescale of speech (Berezutskaya et al., 2017; Chan et al., 2013; Holdgraf et al., 2016; Nourski et al., 2014). Furthermore, they have made it possible to describe the selective encoding of speech sounds in STG, accounting for critical phonological representations of consonants and vowels as well as prosodic features such as intonational and syllabic cues.

In the first sections, we review evidence that STG representations demonstrate properties of high-order auditory encoding, including invariance, non-linearity (Steinschneider et al., 1999), and context-dependence. We also describe emerging evidence that STG neural population activity directly reflects the subjective experience of listeners, adjusting for the presence of noisy or ambiguous sounds (Gwilliams et al., 2018; Holdgraf et al., 2016). These computations may be critical for linking acoustic sensory input with deeply learned knowledge about the structure of language to generate meaningful perceptual representations. In the last section, we consider one of the most substantial and important challenges in neurolinguistics: understanding how the brain binds a continuous acoustic signal into discrete and meaningful representations like words and phrases. Drawing on well-established mechanisms from sensory and perceptual neuroscience, we speculate that a simple and neurobiologically plausible computational framework can explain how local, context-dependent representations in STG may be implemented as a function of time. In the context of the existing evidence, we suggest that STG may play a more substantial role in multiple aspects of speech perception than previously understood.

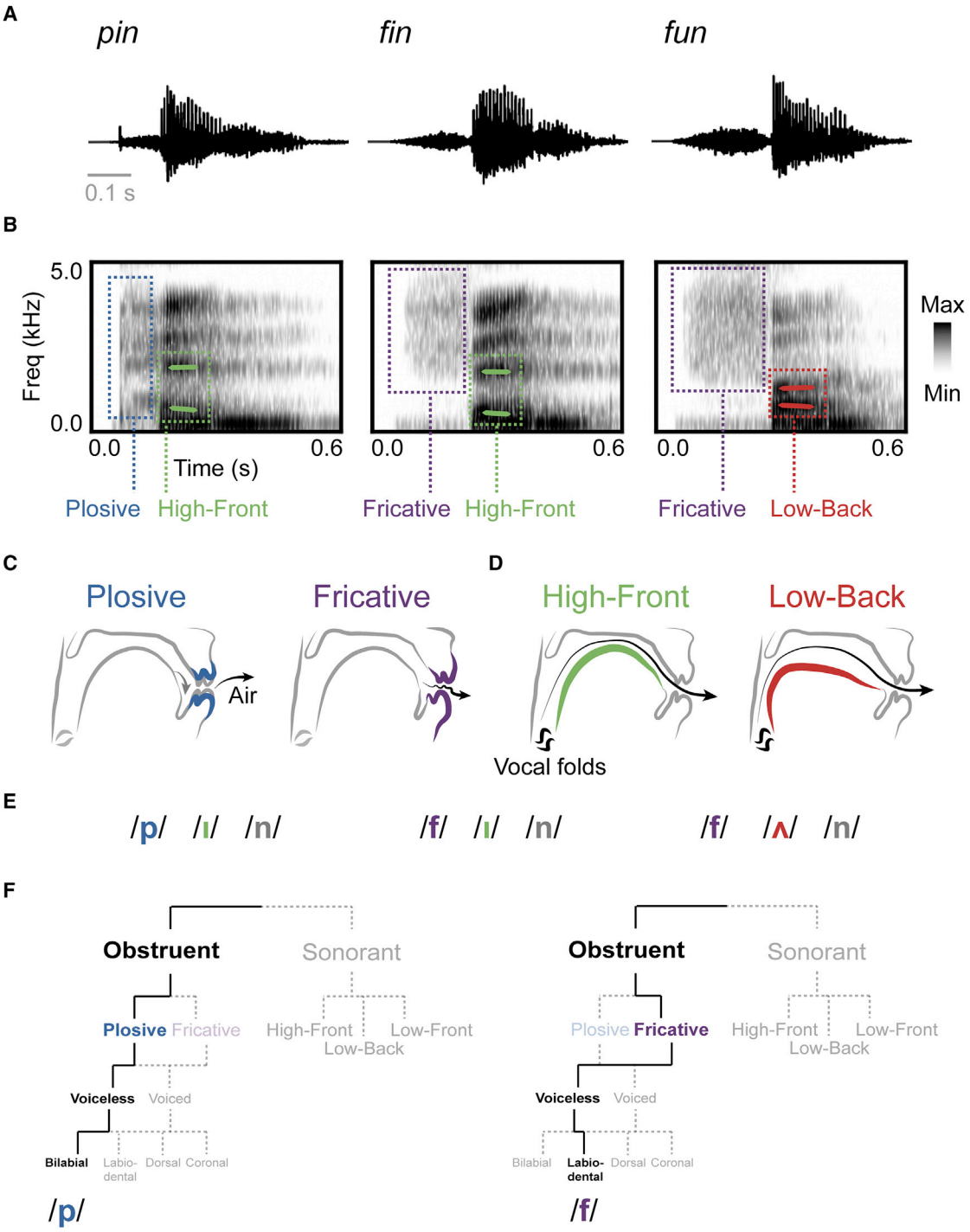

Figure 1. Speech Sounds Can Be Described in Multiple Complementary Ways

For example, the English words "pin," "fin," and "fun" are characterized according to several different but related descriptions, ranging from physical acoustic features to abstract linguistic features.

(A) Acoustic waveforms show a broad distinction between low amplitude/aperiodic features (consonants) and high amplitude/periodic features (vowels).

(B) Spectrogram representations of these words show how each sound is characterized by different spectrotemporal patterns of acoustic energy, including frequency bands that reflect resonance in the vocal tract (formants; green and red lines).

(C) Articulatory descriptions of these sounds characterize acoustic-phonetic features. Plosives are produced by initially blocking the airflow (gray) and then releasing air through the mouth (black), generating a short broadband burst in the spectrogram. Fricatives are produced by partially occluding the passage of air in the mouth, generating a longer-duration, high-frequency broadband noise in the spectrogram. These two features are examples of obstruents.

(legend continued on next page)

Acoustic-Phonetic Features Provide a Framework for Phonological Encoding

The Taxonomy of Speech Sounds

Speech sounds can be described in several different yet complementary ways, ranging from physical characteristics of sound to abstract categories and linguistic features (Figure 1). It is of great interest, therefore, not only which representations truly exist in the brain but how each type of representation is implemented computationally. In this section, we briefly introduce some of these linguistic descriptions and how they relate to the acoustic properties of speech sounds.

At the most basic level, speech, like all sounds, consists of vibrations of air molecules at different amplitudes across time. For simple speech sounds, like the words "pin," "fin," and "fun," the initial portion of each word (i.e., the first consonant) has relatively low amplitude and aperiodic structure, lasting approximately 100 ms (Figure 1A). As these sound waveforms enter the ear, the cochlea decomposes them into time-frequency representations (Delgutte and Kiang, 1984; Shamma, 1985), as shown in the spectrograms in Figure 1B. Here the differences among the initial portions of these words become clear: "pin" begins with a transient broadband noise with rapid onset (Figure 1B) that is produced by the release of a burst of air through the lips when they are opened (Figure 1C), whereas "fin" and "fun" begin with noise with relatively higher spectral frequencies with longer durations (Figure 1B) that is produced by generating a turbulence of aperiodic noise through a partial closure of the mouth (Figure 1C). The middle portions of each example word (i.e., the vowels) are characterized by relatively higher amplitude, periodic structure, and more sustained power in discrete frequency bands (Figures 1A and 1B). These frequency bands, known as formants, are generated by configuring the vocal tract into specific shapes that produce distinct sound resonance patterns. The vowel /ɪ/ in "pin" and "fin" has a larger distance between the first two formants compared with /ʌ/ in "fun" (Figure 1B), which reflects different positions of the tongue (Figure 1D).

These descriptions of speech sounds are based entirely on the acoustic properties of the signal, which are perceived by listeners to be language-specific categories. To formalize these properties, linguists have developed the system of phonology, which describes both abstract categorical linguistic units, called phonemes, and a taxonomy of features that make up phonemes. Specifically, the words "pin," "fin," and "fun" are each made up of three phonemes, which are the minimally contrastive units of meaning in speech (Figure 1E; Chomsky and Halle, 1968; Jakobson et al., 1951). This means that changing the phoneme /p/ to /f/ changes the meaning of the word in English (Figure 1E; Baudouin de Courtenay, 1972; De Saussure, 1879; Sapir, 1925).

In phonology, phonemes can be decomposed into smaller, more elemental acoustic-phonetic features that link abstract categorical phoneme representations to the underlying acoustic properties and articulatory gestures that generate them (Figure 1F). Acoustic-phonetic features are related to each other hierarchically, where different combinations of features compose a unique phoneme. Whereas each phoneme is mutually exclusive, and only one can exist at a given time as an abstract unit, features are combined in specific ways, overlapping in time. Each acoustic-phonetic feature describes a particular aspect of how the sound is produced; for example, occluding airflow through the mouth (obstruent) for a relatively short time (plosive) without vibrating the vocal folds (voiceless) and having the place of occlusion be at the lips (bilabial). These features [obstruent + plosive + voiceless + bilabial] together describe the English phoneme /p/. Changing the plosive feature to the fricative feature and the bilabial feature to the labio-dental feature changes the description to the phoneme /f/ (Figure 1F), demonstrating how relationships among acoustic-phonetic features create a flexible system for phonological representation.

For listeners, each representation from acoustic to linguistic features provides flexibility that allows for rapid and robust analysis of speech at multiple levels (Blumstein and Stevens, 1981; Hillenbrand et al., 1995; Lisker, 1986; Stevens and Blumstein, 1981). Importantly, these levels of representation are not mutually exclusive of each other. For instance, in noisy listening situations where not all acoustic cues are available, listeners make perceptual errors that reflect the hierarchical nature of acoustic-phonetic features (Miller and Nicely, 1955). At the same time, when the perceptual task involves making phoneme-level decisions, listeners clearly have access to the more abstract level of representation (McNeill and Lindig, 1973). Together, all of these descriptions reflect the physical, relational, and hierarchical structure of speech, which differs across languages in surface characteristics but describes an intrinsic aspect of speech in all of the world's languages (Clements, 1985; Keyser and Stevens, 1994; Lahiri and Reetz, 2010). Here we argue that an important goal in speech neuroscience is to understand how the human brain supports each of these units across the auditory and speech hierarchy and how those units are bound together into perceptually and cognitively relevant entities such as words and phrases.

The Encoding of Acoustic-Phonetic Features in STG

In this section, we describe how human STG supports phonological processing by implementing acoustic-phonetic feature detectors in local neural populations (Figure 2). The STG is generally considered to be part of the high-order associative auditory cortex in the human brain (Howard et al., 2000; Moerel et al., 2014; Schönwiesner and Zatorre, 2009), encoding sound features that are more complex and heterogeneous compared with earlier regions in the auditory hierarchy (Delgutte and Kiang,

(D) High-front vowels are produced by moving the tongue to the top and front of the mouth, creating a resonance cavity that generates relatively low first-formant and high second-formant values (green lines on B). In contrast, low-back vowels show the reverse pattern (red lines on B). These two features are examples of sonorants.
(E) Each of the example words can also be characterized as a set of abstract phonemes: /pɪn/, /fɪn/, and /fʌn/.
(F) Multiple acoustic-phonetic features are combined to describe unique phonemes. Here, obstruent, plosive, unvoiced, and labial features are combined to describe the English phoneme /p/. Changing the plosive feature to fricative and the bilabial feature to labio-dental describes the phoneme /f/ (not all possible features are shown for simplicity).

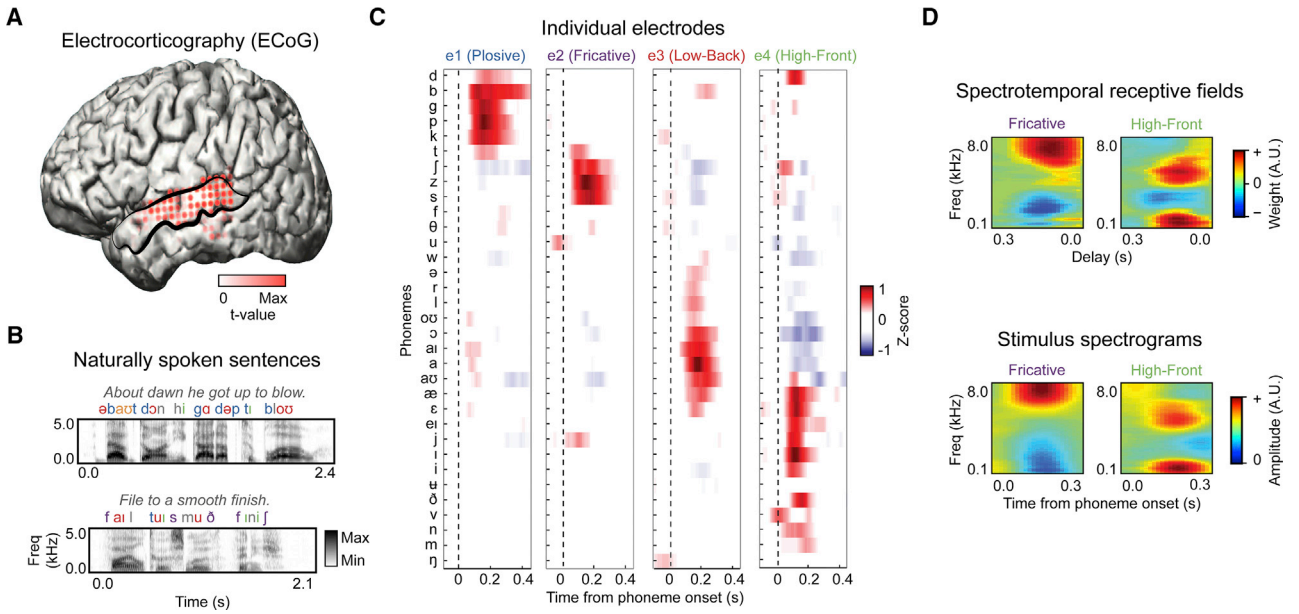

Figure 2. Local Encoding of Acoustic-Phonetic Features in the Human STG

Using direct electrocorticography (ECoG), neural responses to speech can be measured with concurrently high spatial and temporal resolution. These data reveal the encoding of acoustic-phonetic features in local populations during speech perception.

(A and B) ECoG electrodes over human STG (A, outlined in black; opacity signifies t-value for speech versus silence comparison) show robust evoked responses to distinct sounds during listening to (B) naturally spoken sentences.

(C) Each electrode shows selective responses to groups of phonemes, corresponding to acoustic-phonetic features. *Z*-score indicates normalized response.

(D) Electrodes sensitive to specific acoustic-phonetic features (e.g., fricative or low-back vowels) have spectrotemporal receptive fields that strongly resemble the average acoustic spectrograms of sounds characterized by those features (adapted from Mesgarani et al., 2014).

1984; Escabí et al., 2003; Nourski et al., 2014; Shamma, 1985; Steinschneider et al., 2014). Anatomically, STG is homologous to the non-human primate parabelt auditory cortex (Brewer and Barton, 2016; Hackett et al., 2001; Kaas and Hackett, 2000; Petkov et al., 2006). Foundational work using electrocorticography (ECoG) has demonstrated that cortical neural activity, particularly in the high-gamma range (~50–200 Hz), reflects evoked activity to sounds, including speech, in STG (Crone et al., 2001; Towle et al., 2008).

Below, primarily based on insights from ECoG recordings, we argue that the encoding of acoustic-phonetic features arises from the cortical infrastructure for auditory processing that is neither entirely specific nor selective to speech (Mesgarani et al., 2008; Steinschneider et al., 2013) but is nevertheless heavily specialized and causal for speech perception. Rather than attempting to adjudicate the existence of speech-specific and abstract levels of linguistic representations in the brain, we focus on the nature of relevant computations performed on the acoustic speech signal within STG.

Recent work has examined evoked neural responses to natural, continuous speech (Figure 2) and found activity that reflects the local encoding of acoustic-phonetic features in STG. Using ECoG in human epilepsy patients (Figure 2A), Mesgarani et al. (2014) showed that, when neural activity is time-aligned to every individual phoneme in English (Figure 2B), there is clear selectivity for groups of phonemes at the scale of single electrodes. These groups correspond to acoustic-phonetic features such as plosives, fricatives, and vowels (Figure 2C; Mesgarani et al., 2014). Notably, the relationships among responses to different speech sounds mirror the hierarchy of acoustic-phonetic features, with obstruent versus sonorant sounds constituting the main distinction and other features, like manner of articulation (e.g., plosive versus fricative) and voicing, showing more fine-grained separability (Clements, 1985; Keyser and Stevens, 1994; Lahiri and Reetz, 2010; Miller and Nicely, 1955). This work extends previous intracranial recording studies that found local encoding of English phonemes that were distinguished by both place of articulation (e.g., front versus back) and voice-onset time (e.g., /b/ versus /p/) (Steinschneider et al., 2011).

Encoding of acoustic-phonetic features in STG has also been observed in recent functional neuroimaging studies using voxel-wise modeling (Arsenault and Buchsbaum, 2015; de Heer et al., 2017). It is likely that these results reflect sensitivity to complex spectrotemporal tuning, which is characteristic of the higher-order sensory-perceptual cortex (King and Nelken, 2009; Sharpee, 2016). Spectrotemporal receptive fields for ECoG electrodes tuned to specific acoustic-phonetic features closely mirror the acoustic properties of their preferred speech sounds, including relatively complex multi-peak spectral tuning (Figure 2D). For vowels in particular, STG does not show encoding of narrowband frequencies but, rather, appears to exhibit properties of spectral integration, including tuning to specific distributions of acoustic frequency resonance peaks of formants (Figures 1B and 2D) that distinguish different vowels (Hillenbrand et al., 1995; Peterson and Barney, 1952).

At a linguistic level, individual phonemes are described by combinations of acoustic-phonetic features, reflecting different aspects of the same underlying acoustic signal (Figure 1). Indeed, there is evidence of non-linear encoding of acoustic input across neural populations that encode acoustic-phonetic features, which corresponds to categorical phoneme percepts (Chang et al., 2010; Evans and Davis, 2015; Formisano et al., 2008; Lee et al., 2012). Thus, we do not consider these different and complementary descriptions of speech to be mutually exclusive; neural populations that encode one description (e.g., acoustic-phonetic features at local sites) may also contribute to neural codes for other descriptions (e.g., phonemes at the population level).

Other higher-order spectral features of the speech signal that convey important aspects of meaning are also encoded locally in STG. For instance, all spoken languages utilize intonational prosody, in which vocal pitch is varied to indicate a question or a statement or to emphasize words (Cutler et al., 1997; Shattuck-Hufnagel and Turk, 1996). Intonational prosody thus communicates meaning along a distinct information channel, and recent work has found that it is encoded in STG neural populations that are sensitive to speaker-normalized pitch. This encoding for pitch-related prosody appears at discrete sites in STG that are spatially intermixed with but functionally independent from those that encode traditional acoustic-phonetic features for consonants and vowels (Tang et al., 2017). Neural populations that encode absolute pitch were also observed in STG, although they were substantially less common than speaker-normalized pitch populations and did not appear to contribute to intonational prosody. Absolute pitch encoding has been observed in other studies of the human primary auditory cortex (Griffiths et al., 2010) as well as in non-human auditory cortical regions (Bizley et al., 2009; Steinschneider et al., 1998; Walker et al., 2011), but it remains unclear how these neural codes contribute to speech processing beyond encoding information like speaker identity.

Lesion and direct electrical stimulation studies have established a causal role of STG neural populations in speech perception. Damage to the left superior temporal area gray matter results in a striking "receptive" language disorder, known as Wernicke's aphasia (Bates et al., 2003; Blumstein et al., 1977; Geschwind, 1970; Robson et al., 2012; Wernicke, 1874, 1881). Similarly, electrical stimulation to the left but not right posterior STG causes acute interference with speech perception and induces phonological processing deficits, such as paraphasic errors, during tasks like verbal repetition (Boatman, 2004; Boatman et al., 1995; Corina et al., 2010; Leonard et al., 2016b; Roux et al., 2015). Although these results are consistent with the notion that STG has an important role in speech perception at a phonological level, they also raise interesting and unresolved questions about whether there are distinct roles of neural activity in left versus right STG. In addition, these studies do not address the functional connectivity of speech and language networks, which may also explain some of these deficits (Mesulam et al., 2015).

Even though the above findings demonstrate precisely localized encoding for distinct acoustic-phonetic features in STG, there is no apparent spatial clustering for acoustic-phonetic feature categories within individuals, nor is there a conserved map across individuals (cf. Arsenault and Buchsbaum, 2015). Even in the few rare instances where it has been possible to study single neurons in human STG in response to speech, firing rates were consistent with tuning to complex spectrotemporal patterns and acoustic-phonetic features but were highly diverse across neighboring cells (Chan et al., 2013; Creutzfeldt et al., 1989; Engel et al., 2005). The spatial complexity of STG speech encoding is in stark contrast with the tonotopically organized lemniscal auditory pathway, which reflects spatial gradients for frequency information originating in the cochlea (Delgutte and Kiang, 1984; Escabí et al., 2003; Shamma, 1985), including the human primary auditory cortex, where single neurons show narrow frequency tuning (Bitterman et al., 2008).

Encoding of Temporal Landmarks Parcellates STG

Recent work has demonstrated that STG is parcellated into broader distinct regions that encode important temporal landmarks in the speech signal. Specifically, posterior STG is sensitive to speech onset following a period of silence (Figure 3B), whereas middle-to-anterior STG may track ongoing changes in the amplitude envelope of continuous sound (Figure 3C). This spatial organization across STG explains the largest proportion of variance in speech responses and is highly conserved across individuals, unlike acoustic-phonetic feature detectors (Hamilton et al., 2018), suggesting that encoding of amplitude-based cues is a critical function of STG.

Posterior STG is highly responsive to sound onset following at least 200 ms of silence, which contrasts with anterior-middle STG, which is more responsive during ongoing speech (Hamilton et al., 2018; Figure 3D). This organization has been discovered using data-driven, unsupervised clustering approaches without explicit constraints on spatial organization. Notably, onset responses are found not only for intelligible speech but also for unintelligible speech as well as for non-speech sounds, suggesting that they may reflect a fundamental auditory computation. In the context of speech, onset responses provide a robust way to detect important acoustic temporal landmarks in continuous speech, like phrase and sentence boundaries, which often mark meaningful changes in topics, speakers, and tone shifts in natural conversation (Figure 3B).

In addition to studies focused on auditory perception, this functional parcellation between posterior and anterior STG has also been observed in other contexts. Recent ECoG work has shown that posterior STG integrates multimodal input from audiovisual speech in a distinct manner from anterior STG (Ozker et al., 2017, 2018). Additionally, during speech production, responses in a focal region of posterior STG are suppressed at the onset of speech compared with passively listening to the same sounds (Chang et al., 2013). This phenomenon is distinguished from neural activity directly associated with auditory feedback used for control of ongoing vocalization, which is observed throughout the middle STG (Chang et al., 2013) and in Heschl's gyrus (Behroozmand and Larson, 2011; Behroozmand et al., 2016). Thus, these results suggest that sensitivity to temporal context may provide sensorimotor predictions relevant for vocal control, where posterior STG shows the most pronounced speaking-induced suppression at onset of vocalization,

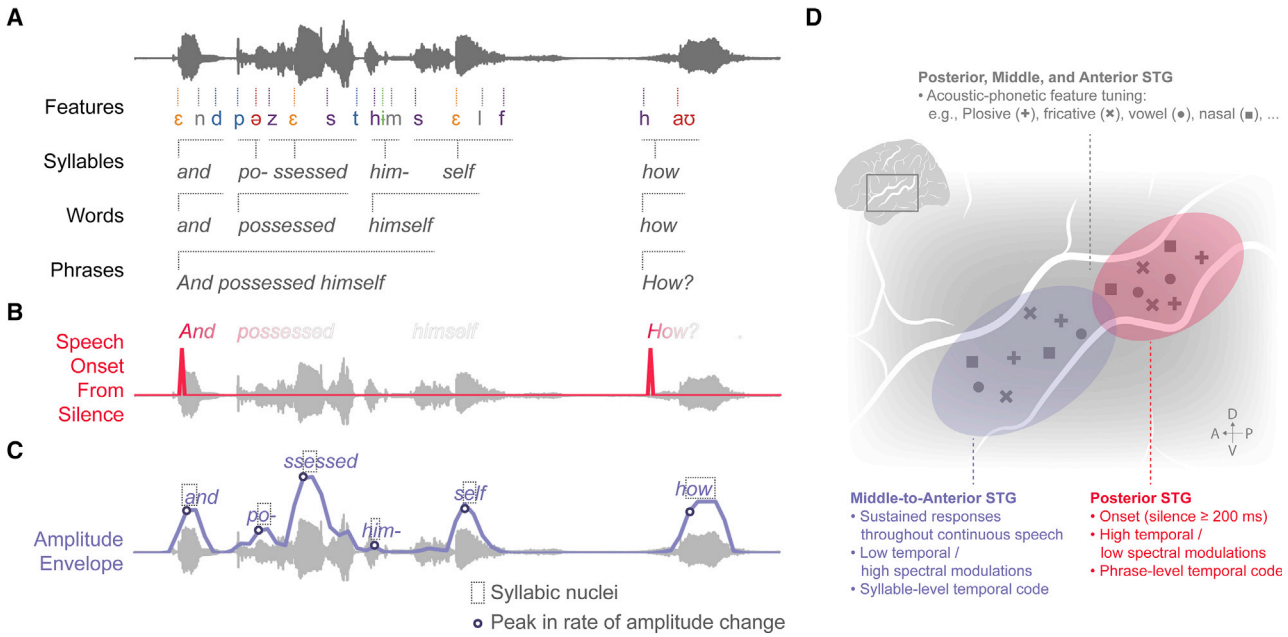

Figure 3. STG Is Parcellated into Two Major Zones that Track Temporal Landmarks Relevant for Speech Processing

Broad regions encoding temporal landmarks have acoustic-phonetic feature detectors embedded in them, facilitating temporal context-dependent speech representations.

(A) Speech can be characterized by multiple temporal and linguistic scales ranging from features to syllables to words to phrases.

(B) Onsets from silence cue prosodic phrase boundaries.

(C) Amplitude envelope change dynamics are a major source of acoustic variability, and peaks in the rate of change correspond to syllabic nuclei.

(D) STG is characterized by a global spatial organization for temporal landmarks. Posterior STG tracks onsets following a period of silence that is 200 ms or longer, whereas middle-to-anterior STG has more sustained responses that may track peaks in the rate of amplitude envelope change. Neural populations in both regions are tuned to acoustic-phonetic features, suggesting that STG integrates temporal landmarks and instantaneous phonetic units.

whereas middle STG shows enhancement for altered feedback perturbations during ongoing vocalizations.

Neural populations in middle-to-anterior STG, in contrast, show more heterogeneous and sustained physiological responses to speech compared with posterior STG (Hamilton et al., 2018). Although the specific functional roles of this region are less clear, there is extensive evidence from non-invasive methods that neural activity in human auditory cortex is correlated with fluctuations in the amplitude envelope of the speech signal (Ahissar et al., 2001; Ding et al., 2014; Doelling et al., 2014; Kubanek et al., 2013; Liégeois-Chauvel et al., 2004; Nourski et al., 2009; Overath et al., 2012). It is possible that the sustained responses in middle-to-anterior STG, which are observed when averaging activity across many sentences (Hamilton et al., 2018), may reflect the encoding of envelope-based cues at the single-trial level. Across nearly all of the world's languages, the amplitude envelope provides important syllable-level temporal cues (e.g., syllabic nuclei; Figure 3C; Blevins, 1995; Byrd, 1996; Zec, 1995). Perceptually, the amplitude envelope plays a critical role in comprehension and intelligibility of spoken sentences (Drullman et al., 1994a, 1994b; Rosen, 1992; Shannon et al., 1995). Based on long-standing theories that have postulated that amplitude events signal key temporal landmarks for spectral analysis of the speech signal (Chistovich and Lublinskaya, 1979; Stevens, 2002), we suggest that the amplitude envelope may be encoded as a discrete landmark feature. Neural populations that are tuned to detect this feature provide a temporal frame for organizing the rapid stream of alternating consonants and vowels in natural speech, which are analyzed in local STG populations that are tuned to specific spectral acoustic-phonetic features (Figure 3D).

Currently, the specific neural code for amplitude envelope information has not been firmly established. Although there is evidence that the human auditory cortex activity entrains to the continuous amplitude envelope (Gross et al., 2013; Peelle and Davis, 2012), there are also data suggesting that encoding is based on a sparser cue (Doelling et al., 2014). In particular, animal neurophysiology has found neurons throughout the mammalian auditory pathway, including the cortex, that are tuned specifically to the rate of change in the amplitude envelope (Fishbach et al., 2001; Heil, 1997, 2004). This topic remains under active investigation, including ongoing efforts using ECoG in humans to examine the extent to which neural populations in human STG respond to sparse, amplitude-based temporal cues of continuous speech (Oganian and Chang, 2018).

Although much remains to be characterized regarding the functional parcellation of STG, the broad spatial organization of posterior onset and middle-to-anterior amplitude cues aligns with previous observations of spatial tuning to different temporal and spectral acoustic modulation rates (Hullett et al., 2016; Santoro et al., 2014; Schönwiesner and Zatorre, 2009). Specifically, posterior STG has been shown to prefer high temporal modulation (Hullett et al., 2016), which is consistent with the rapid increase in amplitude associated with onset of speech sounds

(Hamilton et al., 2018). In contrast, middle-to-anterior STG prefers high spectral modulation (Hullett et al., 2016), which is characteristic of vowels (Elliott and Theunissen, 2009; Versnel and Shamma, 1998), the timing of which is strongly correlated with the temporal envelope in natural speech (Hermes, 1990; Zec, 1995). We propose that the broad spatial organization of temporal cues may be crucial for the encoding of phonological information in STG, where these cues serve as temporal landmarks for organizing and binding spectral content across time, such as into syllables, words, or phrases (Figure 3A). Furthermore, embedding of acoustic-phonetic detectors throughout STG allows local processing of highly dynamic complex acoustic input (Figure 3D). This organization may suggest that acoustic-phonetic feature representations in the posterior zone are modulated by phrase onsets and that acoustic-phonetic feature representations in the middle-to-anterior zone are modulated by the syllabic context. Thus, temporal landmarks can provide an intrinsic mechanism for tracking time and, therefore, the order of phonological units. For example, /m/ at the beginning of the word "mom" could be differentiated from the final /m/ in part because of the temporal context provided by detection of the temporal landmark for the vowel nucleus. If true, then this would suggest that STG represents context-dependent speech input across multiple perceptually relevant timescales.

Temporal Binding and Lexical Representation

Up to this point, we have described how STG neural populations encode instantaneous representations of spectral and temporal features. A crucial question is the extent to which these acoustic representations are integrated on longer timescales to reflect more abstract linguistic information, like words and sequences of words that make up phrases (Figure 3A). More generally, how does STG contribute to the representation of phonological sequences? In this section, we describe evidence for a computational role of STG in encoding sequences as holistic units. We propose that STG integrates representations of acoustic-phonetic features (e.g., /ʃ/ - /ɑ/ - /p/), taking into account the temporal context provided by amplitude-based prosodic cues and learned knowledge about the statistics and structure of the language, into a more abstract, holistic unit of a word (e.g., "shop"). Specifically, we hypothesize that the types of recurrent computations that have been observed throughout the brain for other perceptual and cognitive tasks (Mante et al., 2013; Phillips et al., 2015; Sussillo and Abbott, 2009; Wang et al., 2018) may be implemented in STG through laminar or cortico-cortical organization to generate context-sensitive representations of both lexical and sub-lexical information.

STG Computes Representations of Perceptual Experience

STG plays a crucial role in interpreting auditory input to generate perceptual representations. Mounting evidence has shown that acoustic-phonetic representations in STG are strongly influenced by multiple forms of context. These include not only the temporal context cues provided by amplitude-based landmarks in the acoustic envelope (Figures 3B and 3C) but also those that are not physically part of the sound. Broadly, this means that activity in these neural populations reflects information beyond an instantaneous sensory representation of acoustics. Rather, speech encoding in STG reflects multiple sources of knowledge about speech and language, ultimately generating representations of the listener's subjective perception.

For instance, when the input to STG is a set of words that differ in a single sound (e.g., "faster" /fæstr/ versus "factor" /fæktr/; Figure 4A), neural populations that are tuned to the specific acoustic-phonetic features encode this difference (Figure 4B). However, neural networks within STG—and possibly in other brain regions—also contain information other than the signal acoustics that guide perception. There are many sources of context that modulate speech-evoked STG responses, including both learned knowledge about language structure and task-related goals like attention. For example, learned language-specific statistics such as phoneme sequence probabilities (phonotactics; Furl et al., 2011; Leonard et al., 2015; Yaron et al., 2012) and the predictability of sub-lexical units based on lexical statistics (e.g., word frequency and cohort density; Cibelli et al., 2015; Davis et al., 2005) exert strong effects on STG neural populations that show tuning to acoustic-phonetic features.

In addition, domain-general cognitive factors like selective attention in the context of multiple concurrent speakers (cocktail party phenomenon; Ding and Simon, 2012; Zion Golumbic et al., 2013; Mesgarani and Chang, 2012) or target detection (Chang et al., 2011; Nourski et al., 2015, 2017) can have effects on neural activity, like changing overall gain or signal-to-noise of evoked responses (Figure 4B). Moreover, sources of contextual modulation for speech in STG extend beyond the auditory modality, as in the case of multisensory integration (Ozker et al., 2017, 2018; Rhone et al., 2016).

Computationally, integration of many of these sources of context can be implemented by mechanisms that facilitate the rapid transformation of sensory input into perceptual representations, including predictive coding (e.g., forward transition probabilities for prediction) (Blank and Davis, 2016; Friston, 2005; Kiebel et al., 2009; Yildiz et al., 2013) and Hebbian learning processes for object recognition (Dan and Poo, 2004). We hypothesize that these mechanisms constitute a fundamental part of the neural circuitry involved in high-level auditory processing, embedded in the networks that process the physical properties of speech and, therefore, resulting in an integrated representation (Figure 4B). Crucially, this kind of learned information about the statistical structure of speech and language can provide a strong foundation for binding input into perceptually coherent and meaningful units like words (Figure 4C; Brent and Cartwright, 1996; McQueen, 1998; Saffran et al., 1996), which may not be as readily identifiable using amplitude-based temporal landmarks such as syllables or phrases (Figures 3A–3C).

This integrated representation of acoustic-phonetic, temporal landmark, and contextual features allows for some remarkable capabilities. Recent studies indicate that these neural populations rapidly and dynamically change their activity depending on the listener's perceptual experience, influenced by the predictability of longer-timescale phonological, lexical, and semantic knowledge (Blank et al., 2018; Holdgraf et al., 2016; Khoshkhoo et al., 2018). For example, when part of a word is completely masked by noise (Figure 4A, bottom), listeners report hearing the full word as if the missing sound were present

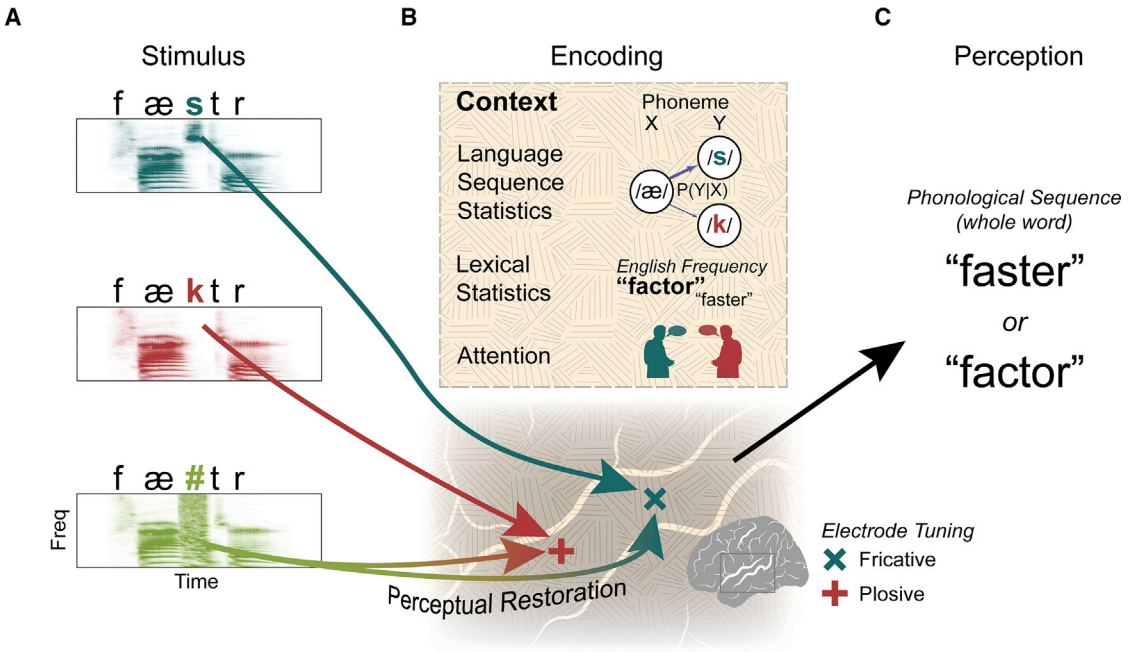

Figure 4. STG Combines Acoustic-Phonetic Tuning with Various Sources of Context to Compute Perceptual Representations of Speech

(A) Example words and their acoustic spectrograms that differ in a single phoneme or acoustic-phonetic feature (/s/ versus /k/) and a stimulus with masking noise (/#/) completely replacing the middle sound.

(B) Stimulus encoding involves detecting acoustic-phonetic features with tuned neural populations (e.g., fricative populations respond to /s/, and plosive populations respond to /k/). This response is embedded in both local and distributed representations of context (orange texture), including sensitivity to language-level sequence statistics (phonotactics), lexical statistics like word frequency, and attention to particular speakers. In the case of the ambiguous sound, STG neural populations "restore" the missing phoneme by activating the appropriate acoustic-phonetic tuned population in real time, possibly using a combination of these multiple sources of context.

(C) The output of STG population activity reflects the perceptual experience of the listener. Specifically, STG activity encodes the percept of the phonological sequence, in this case the whole words "faster" or "factor." In the case of ambiguous input (A, bottom), these percepts do not directly correspond to the input acoustic signal.

(Grossberg and Kazerounian, 2011; Warren, 1970). Even when told that a sound is missing, listeners have trouble reporting the identity and timing of the noise, suggesting that its percept was "restored" (Samuel, 1987). A recent ECoG study demonstrated that this ambiguous input is rapidly transformed to generate the listener's perceptual experience by activating the appropriately tuned STG neural populations in real time (Figures 4A–4C; Leonard et al., 2016a). These results strongly suggest that contextual sources of linguistic knowledge and expectation influence up-stream representations of sound (McClelland and Elman, 1986), allowing listeners to recover from noisy environments and interruptions almost instantaneously. Similarly, although some neural populations in the human primary auditory cortex are not sensitive to the intelligibility of speech as reported by listeners, neural populations throughout lateral STG show stronger responses to intelligible sounds (Nourski et al., 2019), further demonstrating that STG represents perceptual rather than purely sensory experience (Figure 4C).

Together, these findings illustrate that computations associated with feature detection and temporal or contextual integration occur at least partially within STG. Although there may also be a role for top-down modulation from other regions in the speech and language network for many of these findings (Cope et al., 2017; Obleser and Kotz, 2010; Obleser et al., 2007; Park et al., 2015; Sohoglu et al., 2012), they nevertheless demonstrate the highly contextual nature of acoustic-phonetic representations themselves. Here we argue that evoked STG activity closely reflects subjective perceptual experience, resulting from the integration of sensory inputs and the internal dynamics governed by the task demands.

Implementational Challenges for Existing Models of Phonological Sequence Encoding

So far, we have described evidence that demonstrates that multiple sources of information are encoded in the STG, often within the same local population. In particular, the presence of context-dependent perceptual representations suggests that various speech features may be dynamically integrated across time. Below, we describe how the existing neuroanatomical models account for temporal integration and binding processes that are central to speech perception. We then speculate how simple and commonly used recurrent computations can alternatively provide a parsimonious explanation for several lines of existing research. The primary goal of describing these hypotheses is to address three key questions that have remained unanswered for decades: (1) do instantaneous responses to acoustic-phonetic features also contain information about sequential order (Dehaene et al., 2015); (2) how are the hierarchical units of phonology encoded as meaningful, perceptual chunks that

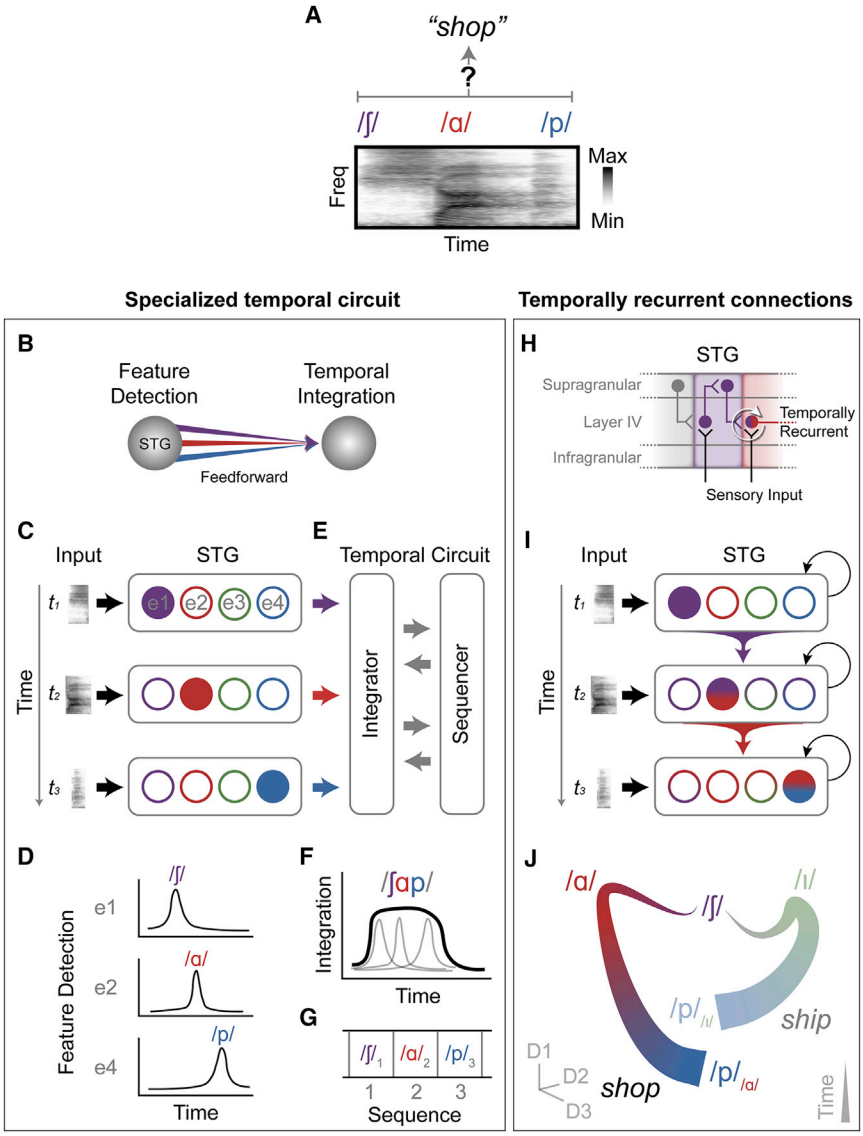

Figure 5. Computational Implementations of Temporal Sequencing and Binding in Speech Cortex

(A) How does the brain bind instantaneous acoustic-phonetic features (e.g., /ʃ/, /ɑ/, and /p/) into perceptually coherent sequences (e.g., "shop")?

(B) A dedicated temporal integrator receives feature representations from STG.

(C and D) Distinct STG populations (recorded with different electrodes; e1, e2, etc.) detect acoustic-phonetic features from the acoustic input by generating spatially (C) and temporally independent (D) neural responses.

(E) Detected features are passed to a separate mechanism that tracks temporal order and is capable of temporal integration.

(F) The temporal integrator and sequencer have a relatively long temporal window and are able to bind multiple feature inputs across time.

(G) The sequence representation contains markers of temporal order (e.g., $/ʃ/_1$, $/ɑ/_2$, and $/p/_3$).

(H–J) An alternative framework has context-dependent acoustic-phonetic feature representations that arise from temporally recurrent connections.

(H) The laminar organization of the human cortex provides a means for input and output connections across layers and columns to implement temporal recurrence, where input to layer IV is contextually modulated by prior output from supragranular layers and thalamic inputs.

(I) Unfolded across time, the neural representation of the input is a function of the past state of the network via temporally recurrent connections among feature detectors.

(J) At the population level, the representation across time of the sequence "shop" (/ʃɑp/) is distinguishable from that of "ship" (/ʃɪp/), not only based on the instantaneous responses to the vowels (/ɑ/ versus /ɪ/) but also from the context-modulated responses to the final consonants ($/p/_{/ɑ/}$ versus $/p/_{/ɪ/}$; i.e., /p/ does not occupy a single point in the state-space).

unfold over longer timescales (Figures 1 and 5A); and (3) what is the computational implementation that allows the speech system to parse an acoustic signal that changes rapidly and is ephemeral (Christiansen and Chater, 2016)? In the absence of work that has directly tackled this set of questions in speech neuroscience, we draw from diverse fields of research to propose a model of temporal sequencing and binding for speech in the brain (Figure 5).

The classical neuroanatomical model of speech processing posits a hierarchical organization where acoustic-phonetic features detected in STG are combined by a separate brain region that tracks the specific order of acoustic-phonetic activity (Figure 5B) to give rise to longer units such as words (Hickok and Poeppel, 2007). According to this view, STG acts as a spectrotemporal feature detector with relatively short temporal integration windows, where its individual neural populations respond preferentially to their preferred combination of acoustic features in specific temporal contexts defined by temporal landmark events (Figures 5C and 5D). For example, groups of STG populations that prefer unvoiced fricatives (/ʃ/), low-back vowels (/ɑ/), and bilabial unvoiced plosives (/p/) could be bound together into the word "shop" by a neural population with a longer temporal integration window, which also tags the activity of each feature detector with sequential order information (Figures 5E and 5F). This sequential neural population activity across time is therefore what determines the larger unit of representation, such as words, which can then be associated with meaning.

Under this hypothesis, the processing of time and sequential order is both computationally and spatially independent from the processing of the constituent features of the input (Dehaene et al., 2015; Hickok and Poeppel, 2007). There is evidence for specialized temporal processing circuits in the brain (Paton and Buonomano, 2018) that can precisely track cues that are important for some phonological category distinctions, like the relative time between the closure of the lips and the subsequent release of air that distinguishes /b/ and /p/, referred to as voice-onset time (Klatt, 1975). It is unclear, however, what

computations are used to combine the separate input feature detection and temporal sequence information into perceptually relevant representations, particularly in a highly dynamic stimulus like speech.

Moreover, for a given sequence to be properly understood, the order of its elements must be tracked. For example, the instantaneous acoustic-phonetic elements of the words "shop" and "posh" (/ʃ/ - /ɑ/ - /p/ and /p/ - /ɑ/ - /ʃ/) are nearly identical and are primarily distinguished by the order of the elements. Classic models of speech perception attempt to solve this problem in part by relying on reduplication of network states across each time step of processing (McClelland and Elman, 1986), which is biologically implausible. Computational models of phonological working memory have further proposed specialized time-context units that represent the abstract sequence order (Baddeley, 1992; Burgess and Hitch, 1999; Cogan et al., 2017). Although these models theoretically allow the order of each phonological unit to be tagged (e.g., $/ʃ/_1$ - $/ɑ/_2$ - $/p/_3$ versus $/p/_1$ - $/ɑ/_2$ - $/ʃ/_3$) (Figure 5G), they have primarily been evaluated at the level of word or non-word sequences or lists rather than sub-lexical sequences. Furthermore, these implementations are generally unable to explain listeners' ability to understand sequences with highly variable sequence lengths, including contextual cues necessitated by common phenomena such as homophony.

In our view, a model of speech perception requires accounting for these computational issues associated with temporal integration and sequencing. At its core, such a model must address the fact that speech is not a purely linear, feedforward process of sequential phonetic, phonemic, or lexical identification. Specifically, perceiving and comprehending speech requires binding multiple sources of information into a coherent representation (Phillips and Singer, 1997). Some of this binding process may be accounted for by cues that are present in the acoustic signal, such as coarticulation, where speech sounds are produced differently depending on the sounds that precede and follow them (Diehl et al., 2004). However, others exist only in the internal representations of the listener, including temporal and linguistic context (Leonard and Chang, 2014). In the final section, we speculate that understanding the neural basis of speech perception requires a computational framework that incorporates all of these different sources because none of them alone can explain the perceptual experience of comprehending spoken input.

Recurrence as a Potential Mechanism for Sequencing and Binding in Speech

In this section, we hypothesize that temporally recurrent computations within high-order auditory cortex may provide a neurobiological basis for temporal binding and integration of speech. Based on extensive work from other sensory domains (Douglas and Martin, 2007; Larkum, 2013; Phillips et al., 2015; Xing et al., 2012), we hypothesize that recurrent connections across auditory cortical layers allow cortical columns to respond to incoming sensory input in a manner that is modulated by preceding activity from other columns that have different stimulus tuning properties (Figure 5H). Computationally, recurrent connections provide a mechanism for representing temporally dependent sequences, where the representation at time t is inextricably a function of past representations of input at times $t-1$, $t-2$, …, $t-n$ (Jordan, 1986; Figure 5I). This principle has been implemented in multilayer neural network models where the hidden layers contain representations of the input that are influenced by the identity, predictability, and temporal separation of preceding input (Elman, 1990).

Following this principle, a neural population that responds preferentially to acoustic features that define unvoiced bilabial plosives (e.g., /p/) is represented differently depending on the content of preceding speech, which may simultaneously provide acoustic, phonetic, lexical, semantic, prosodic, and many other sources of context. This means that the representation of speech sounds in the network is intrinsically context-dependent so that $/p/_{/ɑ/}$ is fundamentally different from $/p/_{/ɪ/}$, where the subscript denotes the input or sequence of inputs that preceded the sound currently being heard. Thus, the way in which the speech system represents /p/ is fundamentally distinct depending on whether it occurs in the word "shop" or "ship" (Figure 5J). This is true at an acoustic level (Diehl et al., 2004), but it is also true at an algorithmic and representational level (Marr, 1982) and is based on the experience and statistical structure of the input training data to the network. Thus, at their core, recurrent computations provide a means for compact, efficient, and local representations of sequences at multiple behaviorally relevant timescales, which constitute a central trait of time-evolving signals like speech.

It is well-established that the laminar structure of the cortex provides the structural capacity for implementing precisely the kinds of recurrent computations that may be central to temporal binding and integration. In the auditory cortex, anatomical connectivity is characterized by extensive recurrent connections across the superficial and deep cortical layers (Barbour and Callaway, 2008; Mitani et al., 1985). These recurrent connections form the foundation of local microcircuits that represent sounds as functional units (Atencio and Schreiner, 2016; Sakata and Harris, 2009; See et al., 2018), in which superficial layers exhibit substantially more fine-tuned, flexible, and complex receptive field properties relative to their deeper counterparts, which receive direct input from the thalamus (Francis et al., 2018; Guo et al., 2012; Li et al., 2014; O'Connell et al., 2014; Winkowski and Kanold, 2013). Although little is currently known about the functional implications of such circuitry for speech processing in human STG, recent advances in high-resolution, non-invasive neuroimaging have begun to allow characterization of interlaminar variability in acoustic feature representation in the human auditory cortex (Moerel et al., 2018, 2019; Wu et al., 2018).

Crucially, recurrence allows this context dependence to exert effects over arbitrarily long timescales, allowing this basic computation to explain temporal binding at multiple levels of linguistic representation, including syllables, words, and phrases (Figure 3A). By representing the current input as a function of preceding input with an unknown but measurable temporal decay, neural populations that explicitly represent context-dependent acoustic-phonetic information may provide a mechanism for local representation of phonological sequences.

This mechanism also provides a way for multiple sources of context to influence and define the phonological sequence.

Recurrent connections are capable of creating a context-dependent representation via many types of structures in the training data. This includes sequence statistics like phonotactic probability, syllable sequence probability, and lexical cohort statistics, for which there is evidence of neural encoding (Cibelli et al., 2015; Leonard et al., 2015). It may also encompass important physical dependencies in the input, like coarticulation (Diehl et al., 2004), which provides smooth trajectories through acoustic space across time and may be important cues in the neural encoding of speech feature sequences. Notably, many of these sources of context are also used to predict upcoming input, generating neural representations of speech that have context dependence for both past and future input. Recurrent models have also provided useful insights into other domains related to cortical processing, demonstrating that contextual dependence is a fundamental part of neuronal function during complex cognitive behaviors (Mante et al., 2013; Sussillo and Abbott, 2009; Wang et al., 2018).

At the neuronal population level, temporal context itself is an integral part of the representation. The activity of each neural unit does not represent an output from a static response function (Figure 5C) but a dynamic response to the past and present activity of other neural units (Figure 5I; Blank and Davis, 2016; Gwilliams et al., 2018; Yildiz et al., 2016). Neural representation of a given acoustic-phonetic feature cannot be adequately understood separate from the surrounding temporal context but, rather, should be considered a reflection of an ongoing process to parse the continuous speech input (Yildiz et al., 2016). In this sense, STG sensitivity to temporal landmarks (e.g., sound onsets and amplitude envelope; Hamilton et al., 2018) and learned linguistic knowledge (Leonard et al., 2015, 2016a) reflect different sources of context that are linked to the sequence of temporally evolving featural representations. For instance, detecting a temporal landmark like a phrase or vowel onset (the syllabic nucleus in most languages) could initiate processing of the surrounding acoustic-phonetic feature input to push neural activity to a distinct part of the neural state space compared with when the same acoustic-phonetic feature occurs in a different temporal context (Figure 5J). Indeed, local encoding of acoustic-phonetic features in STG was observed most clearly when speaker, temporal, and coarticulatory contexts were averaged out (Mesgarani et al., 2014) or when a limited set of stimuli was used (Arsenault and Buchsbaum, 2015). In this view, acoustic-phonetic representations devoid of any context may not have any meaning or realization in speech perception.

In summary, we hypothesize that recurrence, as implemented by the laminar structure of the cortex, provides plausible answers to the key questions outlined in the beginning of this section: (1) what has been described previously as instantaneous feature representations are, in fact, temporally context-dependent representations reflecting properties of the longer phonological sequences in which they occur; (2) different putative phonological units like syllables, words, and phrases emerge from the binding and integration of input across time and differently tuned neural populations, possibly locally within STG; and (3) temporal recurrence provides a simple and local computational mechanism for these kinds of context-dependent representations.

Concluding Remarks

For nearly 150 years, the STG has been viewed as an important hub for speech and language in the brain (Geschwind, 1970; Wernicke, 1874, 1881). Modern advances in neuroscience and linguistics (Poeppel et al., 2008) have allowed significant progress to be made in characterizing the neural computations involved in transforming continuous acoustic signals into language-specific phonological codes. Although the various contributions of STG to speech processing have been largely characterized separately, it is possible that their combined function is what gives rise to the ultimate perceptual experience of comprehending speech. The next several years will yield a more comprehensive and cohesive view, not only of STG but, more broadly, of speech and language networks in the human brain.

ACKNOWLEDGMENTS

We are grateful to Yulia Oganian, Neal Fox, and the rest of the Chang lab for helpful discussions and comments on the manuscript. This work was supported by NIH grants R01-DC012379 and R01-DC015504 the New York Stem Cell Foundation, the Howard Hughes Medical Institute, the McKnight Foundation, the Shurl and Kay Curci Foundation, and the William K. Bowes Foundation.

REFERENCES

Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., and Merzenich, M.M. (2001). Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. Proc. Natl. Acad. Sci. USA *98*, 13367–13372.

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

Atencio, C.A., and Schreiner, C.E. (2016). Functional congruity in local auditory cortical microcircuits. Neuroscience *316*, 402–419.

Baddeley, A. (1992). Working memory. Science *255*, 556–559.

Barbour, D.L., and Callaway, E.M. (2008). Excitatory local connections of superficial neurons in rat auditory cortex. J. Neurosci. *28*, 11174–11185.

Bates, E., Wilson, S.M., Saygin, A.P., Dick, F., Sereno, M.I., Knight, R.T., and Dronkers, N.F. (2003). Voxel-based lesion-symptom mapping. Nat. Neurosci. *6*, 448–450.

Baudouin de Courtenay, J. (1972). An attempt at a theory of phonetic alternations. In A Baudouin de Courtenay Anthology: The Beginnings of Structural Linguistics, E. Stankiewicz, ed. (Indiana University Press), pp. 144–212.

Behroozmand, R., and Larson, C.R. (2011). Error-dependent modulation of speech-induced auditory suppression for pitch-shifted voice feedback. BMC Neurosci. *12*, 54.

Behroozmand, R., Oya, H., Nourski, K.V., Kawasaki, H., Larson, C.R., Brugge, J.F., Howard, M.A., 3rd, and Greenlee, J.D. (2016). Neural correlates of vocal production and motor control in human Heschl's gyrus. J. Neurosci. *36*, 2302–2315.

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

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

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

Bizley, J.K., Walker, K.M., Silverman, B.W., King, A.J., and Schnupp, J.W. (2009). Interdependent encoding of pitch, timbre, and spatial location in auditory cortex. J. Neurosci. *29*, 2064–2075.

Blank, H., and Davis, M.H. (2016). Prediction errors but not sharpened signals simulate multivoxel fMRI patterns during speech perception. PLoS Biol. *14*, e1002577.

Blank, H., Spangenberg, M., and Davis, M.H. (2018). Neural prediction errors distinguish perception and misperception of speech. J. Neurosci. *38*, 6076–6089.

Blevins, J. (1995). The syllable in phonological theory. In The Handbook of Phonological Theory, J.A. Goldsmith, ed. (Blackwell Publishers), pp. 206–244.

Blumstein, S.E., and Stevens, K.N. (1981). Phonetic features and acoustic invariance in speech. Cognition *10*, 25–32.

Blumstein, S.E., Baker, E., and Goodglass, H. (1977). Phonological factors in auditory comprehension in aphasia. Neuropsychologia *15*, 19–30.

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

Boatman, D., Lesser, R.P., and Gordon, B. (1995). Auditory speech processing in the left temporal lobe: an electrical interference study. Brain Lang. *51*, 269–290.

Brent, M.R., and Cartwright, T.A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. Cognition *61*, 93–125.

Brewer, A.A., and Barton, B. (2016). Maps of the auditory cortex. Annu. Rev. Neurosci. *39*, 385–407.

Burgess, N., and Hitch, G.J. (1999). Memory for serial order: a network model of the phonological loop and its timing. Psychol. Rev. *106*, 551.

Byrd, D. (1996). Influences on articulatory timing in consonant sequences. J. Phonetics *24*, 209–244.

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., and Cash, S.S. (2013). Speech-specific tuning of neurons in human superior temporal gyrus. Cereb. Cortex *24*, 2679–2693.

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

Chang, E.F., Edwards, E., Nagarajan, S.S., Fogelson, N., Dalal, S.S., Canolty, R.T., Kirsch, H.E., Barbaro, N.M., and Knight, R.T. (2011). Cortical spatio-temporal dynamics underlying phonological target detection in humans. J. Cogn. Neurosci. *23*, 1437–1446.

Chang, E.F., Niziolek, C.A., Knight, R.T., Nagarajan, S.S., and Houde, J.F. (2013). Human cortical sensorimotor network underlying feedback control of vocal pitch. Proc. Natl. Acad. Sci. USA *110*, 2653–2658.

Chistovich, L.A., and Lublinskaya, V.V. (1979). The 'center of gravity' effect in vowel spectra and critical distance between the formants: Psychoacoustical study of the perception of vowel-like stimuli. Hear. Res. *1*, 185–195.

Chomsky, N., and Halle, M. (1968). The Sound Pattern of English, First Edition (Harper and Row).

Christiansen, M.H., and Chater, N. (2016). The Now-or-Never bottleneck: a fundamental constraint on language. Behav. Brain Sci. *39*, e62.

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

Clements, G.N. (1985). The geometry of phonological features. Phonology *2*, 225–252.

Cogan, G.B., Iyer, A., Melloni, L., Thesen, T., Friedman, D., Doyle, W., Devinsky, O., and Pesaran, B. (2017). Manipulating stored phonological input during verbal working memory. Nat. Neurosci. *20*, 279–286.

Cope, T.E., Sohoglu, E., Sedley, W., Patterson, K., Jones, P.S., Wiggins, J., Dawson, C., Grube, M., Carlyon, R.P., Griffiths, T.D., et al. (2017). Evidence for causal top-down frontal contributions to predictive processes in speech perception. Nat. Commun. *8*, 2154.

Corina, D.P., Loudermilk, B.C., Detwiler, L., Martin, R.F., Brinkley, J.F., and Ojemann, G. (2010). Analysis of naming errors during cortical stimulation mapping: implications for models of language representation. Brain Lang. *115*, 101–112.

Creutzfeldt, O., Ojemann, G., and Lettich, E. (1989). Neuronal activity in the human lateral temporal lobe. I. Responses to speech. Exp. Brain Res. *77*, 451–475.

Crone, N.E., Boatman, D., Gordon, B., and Hao, L. (2001). Induced electrocorticographic gamma activity during auditory perception. Brazier Award-winning article, 2001. Clin. Neurophysiol. *112*, 565–582.

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

Dan, Y., and Poo, M.M. (2004). Spike timing-dependent plasticity of neural circuits. Neuron *44*, 23–30.

Davis, M.H., Johnsrude, I.S., Hervais-Adelman, A., Taylor, K., and McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences. J. Exp. Psychol. Gen. *134*, 222–241.

de Heer, W.A., Huth, A.G., Griffiths, T.L., Gallant, J.L., and Theunissen, F.E. (2017). The hierarchical cortical organization of human speech processing. J. Neurosci. *37*, 6539–6557.

De Saussure, F. (1879). Mémoire sur le système primitif des voyelles dans les langues indo-européennes (BG Teubner).

Dehaene, S., Meyniel, F., Wacongne, C., Wang, L., and Pallier, C. (2015). The neural representation of sequences: from transition probabilities to algebraic patterns and linguistic trees. Neuron *88*, 2–19.

Delgutte, B., and Kiang, N.Y. (1984). Speech coding in the auditory nerve: I. Vowel-like sounds. J. Acoust. Soc. Am. *75*, 866–878.

DeWitt, I., and Rauschecker, J.P. (2012). Phoneme and word recognition in the auditory ventral stream. Proc. Natl. Acad. Sci. USA *109*, E505–E514.

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

Diehl, R.L., Lotto, A.J., and Holt, L.L. (2004). Speech perception. Annu. Rev. Psychol. *55*, 149–179.

Ding, N., and Simon, J.Z. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. Proc. Natl. Acad. Sci. USA *109*, 11854–11859.

Ding, N., Chatterjee, M., and Simon, J.Z. (2014). Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. Neuroimage *88*, 41–46.

Doelling, K.B., Arnal, L.H., Ghitza, O., and Poeppel, D. (2014). Acoustic landmarks drive delta-theta oscillations to enable speech comprehension by facilitating perceptual parsing. Neuroimage *85*, 761–768.

Douglas, R.J., and Martin, K.A. (2007). Recurrent neuronal circuits in the neocortex. Curr. Biol. *17*, R496–R500.

Drullman, R., Festen, J.M., and Plomp, R. (1994a). Effect of reducing slow temporal modulations on speech reception. J. Acoust. Soc. Am. *95*, 2670–2680.

Drullman, R., Festen, J.M., and Plomp, R. (1994b). Effect of temporal envelope smearing on speech reception. J. Acoust. Soc. Am. *95*, 1053–1064.

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

Elman, J.L. (1990). Finding structure in time. Cogn. Sci. *14*, 179–211.

Engel, A.K., Moll, C.K., Fried, I., and Ojemann, G.A. (2005). Invasive recordings from the human brain: clinical insights and beyond. Nat. Rev. Neurosci. *6*, 35–47.

Escabí, M.A., Miller, L.M., Read, H.L., and Schreiner, C.E. (2003). Naturalistic auditory contrast improves spectrotemporal coding in the cat inferior colliculus. J. Neurosci. *23*, 11489–11504.

Evans, S., and Davis, M.H. (2015). Hierarchical organization of auditory and motor representations in speech perception: evidence from searchlight similarity analysis. Cereb. Cortex *25*, 4772–4788.

Fishbach, A., Nelken, I., and Yeshurun, Y. (2001). Auditory edge detection: a neural model for physiological and psychoacoustical responses to amplitude transients. J. Neurophysiol. *85*, 2303–2323.

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

Francis, N.A., Elgueda, D., Englitz, B., Fritz, J.B., and Shamma, S.A. (2018). Laminar profile of task-related plasticity in ferret primary auditory cortex. Sci. Rep. *8*, 16375.

Friston, K. (2005). A theory of cortical responses. Philos. Trans. R. Soc. Lond. B Biol. Sci. *360*, 815–836.

Furl, N., Kumar, S., Alter, K., Durrant, S., Shawe-Taylor, J., and Griffiths, T.D. (2011). Neural prediction of higher-order auditory sequence statistics. Neuroimage *54*, 2267–2277.

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

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

Griffiths, T.D., Kumar, S., Sedley, W., Nourski, K.V., Kawasaki, H., Oya, H., Patterson, R.D., Brugge, J.F., and Howard, M.A. (2010). Direct recordings of pitch responses from human auditory cortex. Curr. Biol. *20*, 1128–1132.

Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., and Garrod, S. (2013). Speech rhythms and multiplexed oscillatory sensory coding in the human brain. PLoS Biol. *11*, e1001752.

Grossberg, S., and Kazerounian, S. (2011). Laminar cortical dynamics of conscious speech perception: neural model of phonemic restoration using subsequent context in noise. J. Acoust. Soc. Am. *130*, 440–460.

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

Gwilliams, L., Linzen, T., Poeppel, D., and Marantz, A. (2018). In spoken word recognition, the future predicts the past. J. Neurosci. *38*, 7585–7599.

Hackett, T.A., Preuss, T.M., and Kaas, J.H. (2001). Architectonic identification of the core region in auditory cortex of macaques, chimpanzees, and humans. J. Comp. Neurol. *441*, 197–222.

Hamilton, L.S., Edwards, E., and Chang, E.F. (2018). A spatial map of onset and sustained responses to speech in the human superior temporal gyrus. Curr. Biol. *28*, 1860–1871.e4.

Heil, P. (1997). Auditory cortical onset responses revisited. I. First-spike timing. J. Neurophysiol. *77*, 2616–2641.

Heil, P. (2004). First-spike latency of auditory neurons revisited. Curr. Opin. Neurobiol. *14*, 461–467.

Hermes, D.J. (1990). Vowel-onset detection. J. Acoust. Soc. Am. *87*, 866–873.

Hickok, G., and Poeppel, D. (2007). The cortical organization of speech processing. Nat. Rev. Neurosci. *8*, 393–402.

Hillenbrand, J., Getty, L.A., Clark, M.J., and Wheeler, K. (1995). Acoustic characteristics of American English vowels. J. Acoust. Soc. Am. *97*, 3099–3111.

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

Howard, M.A., Volkov, I.O., Mirsky, R., Garell, P.C., Noh, M.D., Granner, M., Damasio, H., Steinschneider, M., Reale, R.A., Hind, J.E., and Brugge, J.F. (2000). Auditory cortex on the human posterior superior temporal gyrus. J. Comp. Neurol. *416*, 79–92.

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

Jakobson, R., Fant, C.G., and Halle, M. (1951). Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates, First Edition (The MIT Press).

Jordan, M. (1986). Serial order: a parallel distributed approach. Technical Report of the University of California Institute for Cognitive Science, May 1, 1986, ICS Report 8604. https://www.osti.gov/biblio/6910294-serial-order-parallel-distributed-processing-approach-technical-report-june-march.

Kaas, J.H., and Hackett, T.A. (2000). Subdivisions of auditory cortex and processing streams in primates. Proc. Natl. Acad. Sci. USA *97*, 11793–11799.

Keyser, S.J., and Stevens, K.N. (1994). Feature geometry and the vocal tract. Phonology *11*, 207–236.

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

Kiebel, S.J., von Kriegstein, K., Daunizeau, J., and Friston, K.J. (2009). Recognizing sequences of sequences. PLoS Comput. Biol. *5*, e1000464.

King, A.J., and Nelken, I. (2009). Unraveling the principles of auditory cortical processing: can we learn from the visual system? Nat. Neurosci. *12*, 698–701.

Klatt, D.H. (1975). Voice onset time, frication, and aspiration in word-initial consonant clusters. J. Speech Hear. Res. *18*, 686–706.

Kubanek, J., Brunner, P., Gunduz, A., Poeppel, D., and Schalk, G. (2013). The tracking of speech envelope in the human cortex. PLoS ONE *8*, e53398.

Lahiri, A., and Reetz, H. (2010). Distinctive features: phonological underspecification in representation and processing. J. Phonetics *38*, 44–59.

Larkum, M. (2013). A cellular mechanism for cortical associations: an organizing principle for the cerebral cortex. Trends Neurosci. *36*, 141–151.

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

Leonard, M.K., and Chang, E.F. (2014). Dynamic speech representations in the human temporal lobe. Trends Cogn. Sci. *18*, 472–479.

Leonard, M.K., Bouchard, K.E., Tang, C., and Chang, E.F. (2015). Dynamic encoding of speech sequence probability in human temporal cortex. J. Neurosci. *35*, 7203–7214.

Leonard, M.K., Baud, M.O., Sjerps, M.J., and Chang, E.F. (2016a). Perceptual restoration of masked speech in human cortex. Nat. Commun. *7*, 13619.

Leonard, M.K., Cai, R., Babiak, M.C., Ren, A., and Chang, E.F. (2016b). The peri-Sylvian cortical network underlying single word repetition revealed by electrocortical stimulation and direct neural recordings. Brain Lang., S0093-934X(15)30194-2.

Li, L.Y., Ji, X.Y., Liang, F., Li, Y.T., Xiao, Z., Tao, H.W., and Zhang, L.I. (2014). A feedforward inhibitory circuit mediates lateral refinement of sensory representation in upper layer 2/3 of mouse primary auditory cortex. J. Neurosci. *34*, 13670–13683.

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

Lisker, L. (1986). "Voicing" in English: a catalogue of acoustic features signaling /b/ versus /p/ in trochees. Lang. Speech *29*, 3–11.

Mante, V., Sussillo, D., Shenoy, K.V., and Newsome, W.T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. Nature *503*, 78–84.

Marr, D. (1982). Vision: A Computational Investigation into the Human Representation and Processing of Visual Information (MIT Press).

McClelland, J.L., and Elman, J.L. (1986). The TRACE model of speech perception. Cognit. Psychol. *18*, 1–86.

McNeill, D., and Lindig, K. (1973). The perceptual reality of phonemes, syllables, words, and sentences. J. Mem. Lang. *12*, 419.

McQueen, J.M. (1998). Segmentation of continuous speech using phonotactics. J. Mem. Lang. *39*, 21–46.

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

Mesgarani, N., David, S.V., Fritz, J.B., and Shamma, S.A. (2008). Phoneme representation and classification in primary auditory cortex. J. Acoust. Soc. Am. *123*, 899–909.

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

Mesulam, M.-M., Thompson, C.K., Weintraub, S., and Rogalski, E.J. (2015). The Wernicke conundrum and the anatomy of language comprehension in primary progressive aphasia. Brain *138*, 2423–2437.

Miller, G.A., and Nicely, P.E. (1955). An analysis of perceptual confusions among some English consonants. J. Acoust. Soc. Am. *27*, 338–352.

Mitani, A., Shimokouchi, M., Itoh, K., Nomura, S., Kudo, M., and Mizuno, N. (1985). Morphology and laminar organization of electrophysiologically identified neurons in the primary auditory cortex in the cat. J. Comp. Neurol. *235*, 430–447.

Moerel, M., De Martino, F., and Formisano, E. (2014). An anatomical and functional topography of human auditory cortical areas. Front. Neurosci. *8*, 225.

Moerel, M., De Martino, F., Uğurbil, K., Formisano, E., and Yacoub, E. (2018). Evaluating the columnar stability of acoustic processing in the human auditory cortex. J. Neurosci. *38*, 7822–7832.

Moerel, M., De Martino, F., Uğurbil, K., Yacoub, E., and Formisano, E. (2019). Processing complexity increases in superficial layers of human primary auditory cortex. Sci. Rep. *9*, 5502.

Nourski, K.V., Reale, R.A., Oya, H., Kawasaki, H., Kovach, C.K., Chen, H., Howard, M.A., 3rd, and Brugge, J.F. (2009). Temporal envelope of time-compressed speech represented in the human auditory cortex. J. Neurosci. *29*, 15564–15574.

Nourski, K.V., Steinschneider, M., Oya, H., Kawasaki, H., Jones, R.D., and Howard, M.A. (2014). Spectral organization of the human lateral superior temporal gyrus revealed by intracranial recordings. Cereb. Cortex *24*, 340–352.

Nourski, K.V., Steinschneider, M., Oya, H., Kawasaki, H., and Howard, M.A., 3rd (2015). Modulation of response patterns in human auditory cortex during a target detection task: an intracranial electrophysiology study. Int. J. Psychophysiol. *95*, 191–201.

Nourski, K.V., Steinschneider, M., Rhone, A.E., and Howard Iii, M.A. (2017). Intracranial electrophysiology of auditory selective attention associated with speech classification tasks. Front. Hum. Neurosci. *10*, 691.

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

O'Connell, M.N., Barczak, A., Schroeder, C.E., and Lakatos, P. (2014). Layer specific sharpening of frequency tuning by selective attention in primary auditory cortex. J. Neurosci. *34*, 16496–16508.

Obleser, J., and Kotz, S.A. (2010). Expectancy constraints in degraded speech modulate the language comprehension network. Cereb. Cortex *20*, 633–640.

Obleser, J., Wise, R.J., Dresner, M.A., and Scott, S.K. (2007). Functional integration across brain regions improves speech perception under adverse listening conditions. J. Neurosci. *27*, 2283–2289.

Oganian, Y., and Chang, E.F. (2018). A speech envelope landmark for syllable encoding in human superior temporal gyrus. bioRxiv. https://doi.org/10.1101/388280.

Overath, T., Zhang, Y., Sanes, D.H., and Poeppel, D. (2012). Sensitivity to temporal modulation rate and spectral bandwidth in the human auditory system: fMRI evidence. J. Neurophysiol. *107*, 2042–2056.

Ozker, M., Schepers, I.M., Magnotti, J.F., Yoshor, D., and Beauchamp, M.S. (2017). A double dissociation between anterior and posterior superior temporal gyrus for processing audiovisual speech demonstrated by electrocorticography. J. Cogn. Neurosci. *29*, 1044–1060.

Ozker, M., Yoshor, D., and Beauchamp, M.S. (2018). Converging evidence from electrocorticography and BOLD fMRI for a sharp functional boundary in superior temporal gyrus related to multisensory speech processing. Front. Hum. Neurosci. *12*, 141.

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

Paton, J.J., and Buonomano, D.V. (2018). The neural basis of timing: distributed mechanisms for diverse functions. Neuron *98*, 687–705.

Peelle, J.E., and Davis, M.H. (2012). Neural oscillations carry speech rhythm through to comprehension. Front. Psychol. *3*, 320.

Peterson, G.E., and Barney, H.L. (1952). Control methods used in a study of the vowels. J. Acoust. Soc. Am. *24*, 175–184.

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

Phillips, W.A., and Singer, W. (1997). In search of common foundations for cortical computation. Behav. Brain Sci. *20*, 657–683, discussion 683–722.

Phillips, W.A., Clark, A., and Silverstein, S.M. (2015). On the functions, mechanisms, and malfunctions of intracortical contextual modulation. Neurosci. Biobehav. Rev. *52*, 1–20.

Poeppel, D., Idsardi, W.J., and van Wassenhove, V. (2008). Speech perception at the interface of neurobiology and linguistics. Philos. Trans. R. Soc. Lond. B Biol. Sci. *363*, 1071–1086.

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

Rhone, A.E., Nourski, K.V., Oya, H., Kawasaki, H., Howard, M.A., 3rd, and McMurray, B. (2016). Can you hear me yet? An intracranial investigation of speech and non-speech audiovisual interactions in human cortex. Lang. Cogn. Neurosci. *31*, 284–302.

Robson, H., Keidel, J.L., Ralph, M.A.L., and Sage, K. (2012). Revealing and quantifying the impaired phonological analysis underpinning impaired comprehension in Wernicke's aphasia. Neuropsychologia *50*, 276–288.

Rosen, S. (1992). Temporal information in speech: acoustic, auditory and linguistic aspects. Philos. Trans. R. Soc. Lond. B Biol. Sci. *336*, 367–373.

Roux, F.-E., Miskin, K., Durand, J.-B., Sacko, O., Réhault, E., Tanova, R., and Démonet, J.-F. (2015). Electrostimulation mapping of comprehension of auditory and visual words. Cortex *71*, 398–408.

Saffran, J.R., Newport, E.L., and Aslin, R.N. (1996). Word segmentation: the role of distributional cues. J. Mem. Lang. *35*, 606–621.

Sakata, S., and Harris, K.D. (2009). Laminar structure of spontaneous and sensory-evoked population activity in auditory cortex. Neuron *64*, 404–418.

Samuel, A.G. (1987). Lexical uniqueness effects on phonemic restoration. J. Mem. Lang. *26*, 36.

Santoro, R., Moerel, M., De Martino, F., Goebel, R., Ugurbil, K., Yacoub, E., and Formisano, E. (2014). Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. PLoS Comput. Biol. *10*, e1003412.

Sapir, E. (1925). Sound patterns in language. Language *1*, 37–51.

Schönwiesner, M., and Zatorre, R.J. (2009). Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. Proc. Natl. Acad. Sci. USA *106*, 14611–14616.

Scott, S.K., Blank, C.C., Rosen, S., and Wise, R.J. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. Brain *123*, 2400–2406.

See, J.Z., Atencio, C.A., Sohal, V.S., and Schreiner, C.E. (2018). Coordinated neuronal ensembles in primary auditory cortical columns. eLife *7*, e35587.

Shamma, S.A. (1985). Speech processing in the auditory system. I: the representation of speech sounds in the responses of the auditory nerve. J. Acoust. Soc. Am. *78*, 1612–1621.

Shannon, R.V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (1995). Speech recognition with primarily temporal cues. Science *270*, 303–304.

Sharpee, T.O. (2016). How invariant feature selectivity is achieved in cortex. Front. Synaptic Neurosci. *8*, 26.

Shattuck-Hufnagel, S., and Turk, A.E. (1996). A prosody tutorial for investigators of auditory sentence processing. J. Psycholinguist. Res. *25*, 193–247.

Sohoglu, E., Peelle, J.E., Carlyon, R.P., and Davis, M.H. (2012). Predictive top-down integration of prior knowledge during speech perception. J. Neurosci. *32*, 8443–8453.

Steinschneider, M., Reser, D.H., Fishman, Y.I., Schroeder, C.E., and Arezzo, J.C. (1998). Click train encoding in primary auditory cortex of the awake monkey: evidence for two mechanisms subserving pitch perception. J. Acoust. Soc. Am. *104*, 2935–2955.

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

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

Steinschneider, M., Nourski, K.V., and Fishman, Y.I. (2013). Representation of speech in human auditory cortex: is it special? Hear. Res. *305*, 57–73.

Steinschneider, M., Nourski, K.V., Rhone, A.E., Kawasaki, H., Oya, H., and Howard, M.A., 3rd (2014). Differential activation of human core, non-core and auditory-related cortex during speech categorization tasks as revealed by intracranial recordings. Front. Neurosci. *8*, 240.

Stevens, K.N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. J. Acoust. Soc. Am. *111*, 1872–1891.

Stevens, K.N., and Blumstein, S.E. (1981). The search for invariant acoustic correlates of phonetic features. In Perspectives on the Study of Speech (L. Erlbaum Associates), pp. 1–38.

Sussillo, D., and Abbott, L.F. (2009). Generating coherent patterns of activity from chaotic neural networks. Neuron *63*, 544–557.

Tang, C., Hamilton, L.S., and Chang, E.F. (2017). Intonational speech prosody encoding in the human auditory cortex. Science *357*, 797–801.

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

Versnel, H., and Shamma, S.A. (1998). Spectral-ripple representation of steady-state vowels in primary auditory cortex. J. Acoust. Soc. Am. *103*, 2502–2514.

Walker, K.M., Bizley, J.K., King, A.J., and Schnupp, J.W. (2011). Multiplexed and robust representations of sound features in auditory cortex. J. Neurosci. *31*, 14565–14576.

Wang, J., Narain, D., Hosseini, E.A., and Jazayeri, M. (2018). Flexible timing by temporal scaling of cortical responses. Nat. Neurosci. *21*, 102–110.

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

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

Wernicke, C. (1881). Lehrbuch der gehirnkrankheiten für aerzte und studirende*Volume 2* (Fischer).

Winkowski, D.E., and Kanold, P.O. (2013). Laminar transformation of frequency organization in auditory cortex. J. Neurosci. *33*, 1498–1508.

Wöstmann, M., Fiedler, L., and Obleser, J. (2017). Tracking the signal, cracking the code: Speech and speech comprehension in non-invasive human electrophysiology. Lang. Cogn. Neurosci. *32*, 855–869.

Wu, P.-Y., Chu, Y.-H., Lin, J.L., Kuo, W.-J., and Lin, F.-H. (2018). Feature-dependent intrinsic functional connectivity across cortical depths in the human auditory cortex. Sci. Rep. *8*, 13287.

Xing, D., Yeh, C.-I., Burns, S., and Shapley, R.M. (2012). Laminar analysis of visually evoked activity in the primary visual cortex. Proc. Natl. Acad. Sci. USA *109*, 13871–13876.

Yaron, A., Hershenhoren, I., and Nelken, I. (2012). Sensitivity to complex statistical regularities in rat auditory cortex. Neuron *76*, 603–615.

Yildiz, I.B., von Kriegstein, K., and Kiebel, S.J. (2013). From birdsong to human speech recognition: bayesian inference on a hierarchy of nonlinear dynamical systems. PLoS Comput. Biol. *9*, e1003219.

Yildiz, I.B., Mesgarani, N., and Deneve, S. (2016). Predictive ensemble decoding of acoustical features explains context-dependent receptive fields. J. Neurosci. *36*, 12338–12350.

Zec, D. (1995). Sonority constraints on syllable structure. Phonology *12*, 85–129.

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