## Title
The Encoding of Speech Sounds in the Superior Temporal Gyrus



## Journal
Neuron, 102(6)

## ISSN
0896-6273

## Authors
Yi, Han Gyol
Leonard, Matthew K
Chang, Edward F

## Publication Date
2019-06-01

## DOI
10.1016/j.neuron.2019.04.023

Peer reviewed

Published in final edited form as:

*Neuron*. 2019 June 19; 102(6): 1096–1110. doi:10.1016/j.neuron.2019.04.023.

# The Encoding of Speech Sounds in the Superior Temporal Gyrus

**Han Gyol Yi**[1,*], **Matthew K. Leonard**[1,*], and **Edward F. Chang**[1,†]

[1]Department of Neurological Surgery, University of California, San Francisco, 675 Nelson Rising Lane, San Francisco, CA 94158, USA

## Summary

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.

### Keywords

Speech processing; superior temporal gyrus; auditory cortex; acoustic-phonetic features; temporal landmarks; context-dependent representation; phonological sequence; temporal integration; temporally-recurrent connections; electrocorticography

## Introduction

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). While 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

†Correspondence: edward.chang@ucsf.edu.
*These authors contributed equally.

structures and higher-level association areas that support abstract aspects of language. Injury of the mid-to-posterior part of the 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 & Rauschecker, 2012; Price, 2012; Scott, Blank, Rosen, & Wise, 2000) and electro- and magneto-encephalography (E/MEG; Di Liberto, O'Sullivan, & Lalor, 2015; Giraud & Poeppel, 2012; Gwilliams, Linzen, Poeppel, & Marantz, 2018; Sohoglu, Peelle, Carlyon, & Davis, 2012; Wöstmann, Fiedler, & Obleser, 2017) has implicated STG in various aspects of phonological processing. While these studies have helped shape important theories on the localization of speech and language function, they have also raised fundamental questions about the nature of phonological representation: What sound features are encoded in the 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 time scale of speech (Berezutskaya, Freudenburg, Güçlü, van Gerven, & Ramsey, 2017; Chan et al., 2013; Holdgraf et al., 2016; Nourski et al., 2012). Furthermore, they have made it possible to describe the selective encoding of speech sounds in the 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, Volkov, Noh, Garell, & Howard III, 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 has been previously understood.

## 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 & 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), which is produced by the release of a burst of air through the lips when they are opened (Figure 1C), while “fin” and “fun” begin with noise with relatively higher spectral frequencies with longer durations (Figure 1B), which 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 (Figure 1A–B). 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 to /ʌ/ in “fun” (Figure 1B), which reflect 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 & Halle, 1968; Jakobson, Fant, & Halle, 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, which 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 & Stevens, 1981; Hillenbrand, Getty, Clark, & Wheeler, 1995; Lisker, 1986; Stevens & Blumstein, 1981). Importantly, these levels of representation are not mutually exclusive from each other. For instance, in noisy listening situations where not all acoustic cues are available, listeners make perceptual errors that reflect the independent nature of acoustic-phonetic features (Miller & 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 & Lindig, 1973). Together, all of these descriptions reflect the physical, relational, and hierarchical structure of speech, which differ across languages in surface characteristics, but describe an intrinsic aspect of speech in all of the world's languages (Clements, 1985; Keyser & Stevens, 1994; Lahiri & 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 the human superior temporal gyrus (STG) supports phonological processing by implementing acoustic-phonetic feature detectors in local neural populations (Figure 2). The STG is generally considered to be a part of the high-order associative auditory cortex in the human brain (Howard et al., 2000; Moerel, De Martino, & Formisano, 2014; Schönwiesner & Zatorre, 2009), encoding sound features that are more complex and heterogeneous compared to earlier regions in the auditory hierarchy (Delgutte & Kiang, 1984; Escabí, Miller, Read, & Schreiner, 2003; Nourski et al., 2012; Shamma, 1985; Steinschneider et al., 2014). Anatomically, STG is homologous to the non-human primate parabelt auditory cortex (Brewer & Barton, 2016; Hackett, Preuss, & Kaas, 2001; Kaas & Hackett, 2000; Petkov, Kayser, Augath, & Logothetis, 2006). Foundational work using electrocorticography (ECoG) has demonstrated that neural activity, particularly in the high-gamma range (~50-200 Hz) reflects evoked activity to sounds including speech in the STG (Crone, Boatman, Gordon, & Hao, 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, David, Fritz, & Shamma, 2008; Steinschneider, Nourski, & Fishman, 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 the 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 electrocorticography (ECoG) in human epilepsy patients (Figure 2A), Mesgarani and colleagues 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, Cheung, Johnson, & Chang, 2014). Notably, the relationships among responses to different speech sounds mirrors the hierarchy of acoustic-phonetic features, with obstruent/sonorant sounds constituting the main distinction, and other features like manner of articulation (e.g., plosive vs. fricative) and voicing showing more fine-grained separability (Clements, 1985; Keyser & Stevens, 1994; Lahiri & Reetz, 2010; Miller & 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 vs. back) and voice-onset time (e.g., /b/ vs. /p/) (Steinschneider et al., 2011).

Encoding of acoustic-phonetic features in the STG has also been observed in recent functional neuroimaging studies using voxel-wise modeling (Arsenault & Buchsbaum, 2015; de Heer, Huth, Griffiths, Gallant, & Theunissen, 2017). It is likely that these results reflect sensitivity to complex spectrotemporal tuning that is characteristic of higher-order sensory/perceptual cortex (King & 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 narrow-band frequencies, but rather appears to exhibit properties of spectral integration, with tuning to specific distributions of peaks of acoustic frequency resonance of formants (Figure 1B, 2D) which distinguish different vowels (Hillenbrand et al., 1995; Peterson & 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 for nonlinear encoding of acoustic input across neural populations population that encode acoustic-phonetic features, which corresponds to categorical phoneme percepts (Chang et al., 2010; Evans & Davis, 2015; Formisano, De Martino, Bonte, & Goebel, 2008; Lee, Turkeltaub, Granger, & Raizada, 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, Dahan, & Van Donselaar, 1997; Shattuck-Hufnagel & 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 the STG, which are spatially intermixed with, but functionally independent from those that encode traditional acoustic-phonetic features for consonants and vowels (Tang, Hamilton, & Chang, 2017). Neural populations that encode absolute pitch were also observed in STG, though 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 human primary auditory cortex (Griffiths et al., 2010) as well as in non-human auditory cortical regions (Bizley, Walker, Silverman, King, & Schnupp, 2009; Steinschneider, Reser, Fishman, Schroeder, & Arezzo, 1998; Walker, Bizley, King, & Schnupp, 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 for 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, Baker, & Goodglass, 1977; Geschwind, 1970; Robson, Keidel, Ralph, & Sage, 2012; Wernicke, 1874, 1881). Similarly, electrical stimulation to left, but not right, posterior STG causes acute interference of speech perception, as well as induces phonological processing deficits, such as paraphasic errors, during verbal repetition (Boatman, 2004; Boatman, Lesser, & Gordon, 1995; Corina et al., 2010; Leonard, Cai, Babiak, Ren, & Chang, 2016; Roux et al., 2015). While 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, Thompson, Weintraub, & Rogalski, 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 (c.f., Arsenault & Buchsbaum, 2015). Even in the few rare opportunities 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, Ojemann, & Lettich, 1989; Engel, Moll, Fried, & Ojemann, 2005). The complexity of STG responses is in stark contrast with those of the tonotopically organized lemniscal auditory pathway, which reflects the spatial gradients for frequency information originating in the cochlea (Delgutte & Kiang, 1984; Escabí et al., 2003; Shamma, 1985), including the human primary auditory cortex, where single neurons show narrow frequency tuning (Bitterman, Mukamel, Malach, Fried, & Nelken, 2008).

## Encoding of Temporal Landmarks Parcellates the 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), while middle-to-anterior STG may track ongoing changes in the amplitude envelope of continuous sound (Figure 3C). This spatial organization across the STG explains the largest proportion of variance in speech responses, and is highly conserved across individuals, unlike for encoding of acoustic-phonetic feature detectors (Hamilton, Edwards, & Chang, 2018), suggesting that the encoding of amplitude-based cues is a critical function of the STG.

The posterior STG is highly responsive to sound onsets 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 observed 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 conversations (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 the anterior STG (Ozker, Schepers, Magnotti, Yoshor, & Beauchamp, 2017; Ozker, Yoshor, & Beauchamp, 2018). Additionally, during speech production, responses in a focal region of posterior STG are suppressed at the onset of speech, compared to passively listening to the same sounds (Chang, Niziolek, Knight, Nagarajan, & Houde, 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 & Larson, 2011; Behroozmand et al., 2016). Thus, these results suggest that sensitivity to temporal context may provide sensorimotor predictions relevant for vocal control, where the posterior STG shows the most pronounced speaking-induced suppression at onset of vocalization, while middle STG shows enhancement for altered feedback pertubations during ongoing vocalizations.

Neural populations in mid-anterior STG, in contrast, show more heterogeneous and sustained physiological responses to speech compared to posterior STG (Hamilton et al., 2018). While the specific functional roles of this region are less clear, there is extensive evidence from non-invasive methods that neural activity in human auditory regions is correlated with fluctuations in the amplitude envelope of the speech signal (Ahissar et al., 2001; Ding, Chatterjee, & Simon, 2014; Doelling, Arnal, Ghitza, & Poeppel, 2014; Kubanek, Brunner, Gunduz, Poeppel, & Schalk, 2013; Liégeois-Chauvel, Lorenzi, Trébuchon, Régis, & Chauvel, 2004; Nourski et al., 2009; Overath, Zhang, Sanes, & Poeppel, 2012). It is possible that the sustained responses in mid-anterior STG, which is

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 the world's languages, the amplitude envelope provides important syllable-level temporal cues (e.g., syllabic nuclei; Figure 3C) (Blevins & Goldsmith, 1995; Byrd, 1996; Zec, 1995). Perceptually, amplitude envelope plays a critical role in comprehension and intelligibility of spoken sentences (Drullman, Festen, & Plomp, 1994b, 1994a; Rosen, 1992; Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995). Based on long-standing theories that have postulated that amplitude events signal key temporal landmarks for spectral analysis of the speech signal (Chistovich & 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. While there is evidence that the human auditory cortex activity entrains to the continuous amplitude envelope (Gross et al., 2013; Peelle & Davis, 2012), there is 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 cortex that are tuned specifically to the rate of change in the amplitude envelope (Fishbach, Nelken, & Yeshurun, 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 & Chang, 2018).

Although much remains to be characterized regarding the functional parcellation of STG, the broad spatial organization of posterior onset and mid-anterior amplitude cues aligns with previous observations of spatial tuning to different temporal and spectral acoustic modulation rates (Hullett, Hamilton, Mesgarani, Schreiner, & Chang, 2016; Santoro et al., 2014; Schönwiesner & 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 the amplitude associated with onsets of speech sounds (Hamilton et al., 2018). In contrast, the middle-to-anterior STG prefers high spectral modulation (Hullett et al., 2016), which is characteristic of vowel sounds (Elliott & Theunissen, 2009; Versnel & Shamma, 1998), the timing for which is strongly correlated with the temporal envelope in natural speech (Hermes, 1990; Zec, 1995). We propose that the broad spatial organization for 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, the embedding of acoustic-phonetic detectors throughout STG allows for local processing of highly dynamic complex acoustic input (Figure 3D). Together, this organization may suggest that acoustic-phonetic feature representations in the posterior zone are modulated by phrase onsets, and 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 due to the temporal context

provided by the detection of the temporal landmark for the vowel nucleus. If true, 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 the STG in encoding sequences as holistic units. We propose that the 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, Sussillo, Shenoy, & Newsome, 2013; W. Phillips, Clark, & Silverstein, 2015; Sussillo & Abbott, 2009; Wang, Narain, Hosseini, & Jazayeri, 2018) may be implemented in STG through the laminar or cross-cortical organization of the cortex 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 the STG are strongly influenced by multiple forms of context. These include not only the temporal context cues provided by amplitude-based events of the acoustic envelope (Figure 3B,C), 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/ vs. “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 sequences (phonotactics; Furl et al., 2011; Leonard, Bouchard, Tang, & Chang, 2015; Yaron, Hershenhoren, & Nelken, 2012) and the predictability of sub-lexical units based on lexical statistics (e.g., word frequency and cohort density; Cibelli, Leonard, Johnson, & Chang, 2015; Davis, Johnsrude, Hervais-Adelman, Taylor, &

McGettigan, 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 & Simon, 2012; Golumbic et al., 2013; Mesgarani & Chang, 2012) or target detection (Chang et al., 2011; Nourski, Steinschneider, Oya, Kawasaki, & Howard III, 2015; Nourski, Steinschneider, Rhone, & Howard III, 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, such 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 & Davis, 2016; Friston, 2005; Kiebel, Von Kriegstein, Daunizeau, & Friston, 2009; Yildiz, von Kriegstein, & Kiebel, 2013) and Hebbian learning processes for object recognition (Dan & 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 & Cartwright, 1996; McQueen, 1998; Saffran, Newport, & Aslin, 1996), which may not be as readily identifiable using amplitude-based temporal landmarks as are syllables or phrases (Figure 3A–C).

This integrated representation of acoustic-phonetic, temporal landmark, and contextual features allows 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, Spangenberg, & Davis, 2018; Holdgraf et al., 2016; Khoshkhoo, Leonard, Mesgarani, & Chang, 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 (Grossberg & 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 (Figure 4A–C) (Leonard, Baud, et al., 2016). These results strongly suggest that contextual sources of linguistic knowledge and expectation influence up-stream representations of sound (McClelland & Elman, 1986), allowing listeners to recover from noisy environments and interruptions almost instantaneously. Similarly, whereas some neural populations in human primary auditory cortex are not sensitive to the intelligibility of speech as reported by listeners, neural populations throughout the lateral STG show stronger

responses to intelligible sounds (Nourski et al., 2019), further demonstrating that the STG represents perceptual, rather than purely sensory, experience (Figure 4C).

Together, these findings illustrate that computations associated with feature detection and temporal/contextual integration occur at least partially within STG. While 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 & Kotz, 2009; Obleser, Wise, Dresner, & Scott, 2007; Park, Ince, Schyns, Thut, & Gross, 2015; Sohoglu et al., 2012), they nevertheless demonstrate the highly contextual nature of acoustic-phonetic representations themselves in STG. 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

Thus far, we have described evidence which demonstrates that multiple sources of information are encoded in 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 on 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, Meyniel, Wacongne, Wang, & Pallier, 2015)? (2) How are the hierarchical units of phonology encoded as meaningful, perceptual chunks that unfold over longer timescales (Figure 1; Figure 5A)? (3) What is the computational implementation that allows the speech system to parse an acoustic signal that changes rapidly and is ephemeral (Christiansen & 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 the 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 & Poeppel, 2007). According to this view, the 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 (Figure 5C–D). 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 (Figure 5E–F). 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 & Poeppel, 2007). There is evidence for specialized temporal processing circuits in the brain (Paton & Buonomano, 2018), which 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 & 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 & Hitch, 1999; Cogan et al., 2017). While these models theoretically allow the order of each phonological unit to be tagged (e.g., $/ʃ/_1$ - $/ɑ/_2$ - $/p/_3$ vs. $/p/_1$ - $/ɑ/_2$ - $/ʃ/_3$) (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 (W. A. Phillips & Singer, 1997). Some of this binding process may be accounted for by cues which 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, Lotto, & Holt, 2004). However, others exist only in the internal representations of the listener, including temporal and linguistic context (Leonard & 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, as 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 & Martin, 2007; Larkum, 2013; W. Phillips et al., 2015; Xing, Yeh, Burns, & Shapley, 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, which 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, such that $/p/_{/ɑ/}$ is fundamentally different from $/p/_{/l/}$, 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/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 time scales, 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 these 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 & Callaway, 2008; Mitani et al., 1985). These recurrent connections form the foundation of local microcircuits that represent sounds as functional units (Atencio & Schreiner, 2016; Sakata & Harris, 2009; See, Atencio, Sohal, & Schreiner, 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, Elgueda, Englitz, Fritz, & Shamma, 2018; Guo et al., 2012; Li et al., 2014; O’Connell, Barczak, Schroeder, & Lakatos, 2014; Winkowski & Kanold, 2013). While 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, De Martino, Uğurbil, Formisano, & Yacoub, 2018; Moerel, De Martino, Uğurbil, Yacoub, & Formisano, 2019; Wu, Chu, Lin, Kuo, & Lin, 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 structure 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 which 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 in 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 & 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 & Davis, 2016; Gwilliams et al., 2018; Yildiz, Mesgarani, & Deneve, 2016). Neural representation of a given acoustic-phonetic feature cannot be adequately understood separate from the surrounding temporal context, but should rather be considered as 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, Baud, Sjerps, & Chang, 2016; Leonard et al., 2015) reflect different sources of context, which are inextricably 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 the processing of the surrounding acoustic-phonetic feature input to push neural activity to a distinct part of the neural state-space, compared to 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 once speaker, temporal and coarticulatory contexts were averaged out (Mesgarani et al., 2014), or when a limited set of stimuli were used (Arsenault & 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 cortex provides plausible answers to the key questions outlined in the beginning of this section: (1) What has been previously described as instantaneous feature representations are in fact temporally context-dependent representations reflecting 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, Idsardi, & Van Wassenhove, 2008) have allowed significant progress to be made in characterizing the neural computations involved in transforming continuous acoustic signals into language-specific phonological codes. While 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 the STG, but of speech and language networks more broadly.

## Acknowledgements

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, and by 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, & Merzenich MM (2001). Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. Proceedings of the National Academy of Sciences, 98(23), 13367–13372.

Arsenault JS, & Buchsbaum BR (2015). Distributed neural representations of phonological features during speech perception. Journal of Neuroscience, 35(2), 634–642. [PubMed: 25589757]

Atencio CA, & Schreiner CE (2016). Functional congruity in local auditory cortical microcircuits. Neuroscience, 316, 402–419. [PubMed: 26768399]

Baddeley A (1992). Working memory. Science, 255(5044), 556–559. [PubMed: 1736359]

Barbour DL, & Callaway EM (2008). Excitatory local connections of superficial neurons in rat auditory cortex. Journal of Neuroscience, 28(44), 11174–11185. [PubMed: 18971460]

Bates E, Wilson SM, Saygin AP, Dick F, Sereno MI, Knight RT, & Dronkers NF (2003). Voxel-based lesion-symptom mapping. Nature Neuroscience, 6(5), 448. [PubMed: 12704393]

Baudouin de Courtenay J (1972). An attempt at a theory of phonetic alternations. Stankiewicz Edward *(Ed. and Trans.)* A Baudouin de Courtenay Anthology: The Beginnings of Structural Linguistics, 144–212.

Behroozmand R, & Larson CR (2011). Error-dependent modulation of speech-induced auditory suppression for pitch-shifted voice feedback. BMC Neuroscience, 12(1), 54. [PubMed: 21645406]

Behroozmand R, Oya H, Nourski KV, Kawasaki H, Larson CR, Brugge JF, … Greenlee JD (2016). Neural correlates of vocal production and motor control in human Heschl's gyrus. Journal of Neuroscience, 36(7), 2302–2315. [PubMed: 26888939]

Berezutskaya J, Freudenburg ZV, Güçlü U, van Gerven MA, & Ramsey NF (2017). Neural tuning to low-level features of speech throughout the perisylvian cortex. Journal of Neuroscience, 0238–17.

Binder JR, Frost JA, Hammeke TA, Bellgowan PS, Springer JA, Kaufman JN, & Possing ET (2000). Human temporal lobe activation by speech and nonspeech sounds. Cerebral Cortex, 10(5), 512–528. [PubMed: 10847601]

Bitterman Y, Mukamel R, Malach R, Fried I, & Nelken I (2008). Ultra-fine frequency tuning revealed in single neurons of human auditory cortex. Nature, 451(7175), 197. [PubMed: 18185589]

Bizley JK, Walker KM, Silverman BW, King AJ, & Schnupp JW (2009). Interdependent encoding of pitch, timbre, and spatial location in auditory cortex. Journal of Neuroscience, 29(7), 2064–2075. [PubMed: 19228960]

Blank H, & Davis MH (2016). Prediction errors but not sharpened signals simulate multivoxel fMRI patterns during speech perception. PLoS Biology, 14(11), e1002577. [PubMed: 27846209]

Blank H, Spangenberg M, & Davis MH (2018). Neural Prediction Errors Distinguish Perception and Misperception of Speech. Journal of Neuroscience, 38(27), 6076–6089. [PubMed: 29891730]

Blevins J, & Goldsmith J (1995). The syllable in phonological theory. 1995, 206–244.

Blumstein SE, Baker E, & Goodglass H (1977). Phonological factors in auditory comprehension in aphasia. Neuropsychologia, 15(1), 19–30. [PubMed: 831150]

Blumstein SE, & Stevens KN (1981). Phonetic features and acoustic invariance in speech. Cognition, 10(1–3), 25–32. [PubMed: 7198546]

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

Boatman D, Lesser RP, & Gordon B (1995). Auditory speech processing in the left temporal lobe: an electrical interference study. Brain and Language, 51(2), 269–290. [PubMed: 8564472]

Brent MR, & Cartwright TA (1996). Distributional regularity and phonotactic constraints are useful for segmentation. Cognition, 61, 93–125. [PubMed: 8990969]

Brewer AA, & Barton B (2016). Maps of the auditory cortex. Annual Review of Neuroscience, 39, 385–407.

Burgess N, & Hitch GJ (1999). Memory for serial order: a network model of the phonological loop and its timing. Psychological Review, 106(3), 551.

Byrd D (1996). Influences on articulatory timing in consonant sequences. Journal of Phonetics, 24(2), 209–244.

Chan AM, Dykstra AR, Jayaram V, Leonard MK, Travis KE, Gygi B, … others. (2013). Speech-specific tuning of neurons in human superior temporal gyrus. Cerebral Cortex, 24(10), 2679–2693. [PubMed: 23680841]

Chang EF, Edwards E, Nagarajan SS, Fogelson N, Dalal SS, Canolty RT, … Knight RT (2011). Cortical spatio-temporal dynamics underlying phonological target detection in humans. Journal of Cognitive Neuroscience, 23(6), 1437–1446. [PubMed: 20465359]

Chang EF, Niziolek CA, Knight RT, Nagarajan SS, & Houde JF (2013). Human cortical sensorimotor network underlying feedback control of vocal pitch. Proceedings of the National Academy of Sciences, 110(7), 2653–2658.

Chang EF, Rieger JW, Johnson K, Berger MS, Barbaro NM, & Knight RT (2010). Categorical speech representation in human superior temporal gyrus. Nature Neuroscience, 13(11), 1428. [PubMed: 20890293]

Chistovich LA, & Lublinskaya VV (1979). The 'center of gravity'effect in vowel spectra and critical distance between the formants: Psychoacoustical study of the perception of vowel-like stimuli. Hearing Research, 1(3), 185–195.

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

Christiansen MH, & Chater N (2016). The Now-or-Never bottleneck: A fundamental constraint on language. Behavioral and Brain Sciences, 39.

Cibelli ES, Leonard MK, Johnson K, & Chang EF (2015). The influence of lexical statistics on temporal lobe cortical dynamics during spoken word listening. Brain and Language, 147, 66–75. [PubMed: 26072003]

Clements GN (1985). The geometry of phonological features. Phonology, 2(1), 225–252.

Cogan GB, Iyer A, Melloni L, Thesen T, Friedman D, Doyle W, … Pesaran B (2017). Manipulating stored phonological input during verbal working memory. Nature Neuroscience, 20(2), 279. [PubMed: 27941789]

Cope TE, Sohoglu E, Sedley W, Patterson K, Jones PS, Wiggins J, … others. (2017). Evidence for causal top-down frontal contributions to predictive processes in speech perception. Nature Communications, S(1), 2154.

Corina DP, Loudermilk BC, Detwiler L, Martin RF, Brinkley JF, & Ojemann G (2010). Analysis of naming errors during cortical stimulation mapping: implications for models of language representation. Brain and Language, 115(2), 101–112. [PubMed: 20452661]

Creutzfeldt O, Ojemann G, & Lettich E (1989). Neuronal activity in the human lateral temporal lobe. Experimental Brain Research, 77(3), 451–475. [PubMed: 2806441]

Crone NE, Boatman D, Gordon B, & Hao L (2001). Induced electrocorticographic gamma activity during auditory perception. Clinical Neurophysiology, 112(4), 565–582. [PubMed: 11275528]

Cutler A, Dahan D, & Van Donselaar W (1997). Prosody in the comprehension of spoken language: A literature review. Language and Speech, 40(2), 141–201. [PubMed: 9509577]

Dan Y, & Poo M (2004). Spike timing-dependent plasticity of neural circuits. Neuron, 44(1), 23–30. [PubMed: 15450157]

Davis MH, Johnsrude IS, Hervais-Adelman A, Taylor K, & McGettigan C (2005). Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences. Journal of Experimental Psychology: General, 134(2), 222. [PubMed: 15869347]

de Heer WA, Huth AG, Griffiths TL, Gallant JL, & Theunissen FE (2017). The hierarchical cortical organization of human speech processing. Journal of Neuroscience, 3267–16.

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, & Pallier C (2015). The neural representation of sequences: from transition probabilities to algebraic patterns and linguistic trees. Neuron, 88(1), 2–19. [PubMed: 26447569]

Delgutte B, & Kiang NY (1984). Speech coding in the auditory nerve: I. Vowel-like sounds. The Journal of the Acoustical Society of America, 75(3), 866–878. [PubMed: 6707316]

DeWitt I, & Rauschecker JP (2012). Phoneme and word recognition in the auditory ventral stream. Proceedings of the National Academy of Sciences, 109(8), E505–E514.

Di Liberto GM, O'Sullivan JA, & Lalor EC (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. Current Biology, 25(19), 2457–2465. [PubMed: 26412129]

Diehl RL, Lotto AJ, & Holt LL (2004). Speech perception. Annu. Rev. Psychol, 55, 149–179. [PubMed: 14744213]

Ding N, Chatterjee M, & Simon JZ (2014). Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. Neuroimage, 88, 41–46. [PubMed: 24188816]

Ding N, & Simon JZ (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. Proceedings of the National Academy of Sciences, 109(29), 11854–11859.

Doelling KB, Arnal LH, Ghitza O, & Poeppel D (2014). Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing. Neuroimage, 85, 761–768. [PubMed: 23791839]

Douglas RJ, & Martin KA (2007). Recurrent neuronal circuits in the neocortex. Current Biology, 17(13), R496–R500. [PubMed: 17610826]

Drullman R, Festen JM, & Plomp R (1994a). Effect of reducing slow temporal modulations on speech reception. The Journal of the Acoustical Society of America, 95(5), 2670–2680. [PubMed: 8207140]

Drullman R, Festen JM, & Plomp R (1994b). Effect of temporal envelope smearing on speech reception. The Journal of the Acoustical Society of America, 95(2), 1053–1064. [PubMed: 8132899]

Elliott TM, & Theunissen FE (2009). The modulation transfer function for speech intelligibility. PLoS Computational Biology, 5(3), e1000302. [PubMed: 19266016]

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

Engel AK, Moll CK, Fried I, & Ojemann GA (2005). Invasive recordings from the human brain: clinical insights and beyond. Nature Reviews Neuroscience, 6(1), 35. [PubMed: 15611725]

Escabí MA, Miller LM, Read HL, & Schreiner CE (2003). Naturalistic auditory contrast improves spectrotemporal coding in the cat inferior colliculus. Journal of Neuroscience, 23(37), 11489–11504. [PubMed: 14684853]

Evans S, & Davis MH (2015). Hierarchical organization of auditory and motor representations in speech perception: evidence from searchlight similarity analysis. Cerebral Cortex, 25(12), 4772–4788. [PubMed: 26157026]

Fishbach A, Nelken I, & Yeshurun Y (2001). Auditory edge detection: a neural model for physiological and psychoacoustical responses to amplitude transients. Journal of Neurophysiology, 85(6), 2303–2323. [PubMed: 11387378]

Formisano E, De Martino F, Bonte M, & Goebel R (2008). " Who" is saying" what"? Brain-based decoding of human voice and speech. Science, 322(5903), 970–973. [PubMed: 18988858]

Francis NA, Elgueda D, Englitz B, Fritz JB, & Shamma SA (2018). Laminar profile of task-related plasticity in ferret primary auditory cortex. Scientific Reports, 8(1), 16375. [PubMed: 30401927]

Friston K (2005). A theory of cortical responses. Philosophical Transactions of the Royal Society of London B: Biological Sciences, 360(1456), 815–836. [PubMed: 15937014]

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), 2267–2277. [PubMed: 20970510]

Geschwind N (1970). The organization of language and the brain. Science, 170(3961), 940–944. [PubMed: 5475022]

Giraud A-L, & Poeppel D (2012). Cortical oscillations and speech processing: emerging computational principles and operations. Nature Neuroscience, 15(4), 511. [PubMed: 22426255]

Golumbic EMZ, Ding N, Bickel S, Lakatos P, Schevon CA, McKhann GM, … others. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party." Neuron, 77(5), 980–991. [PubMed: 23473326]

Griffiths TD, Kumar S, Sedley W, Nourski KV, Kawasaki H, Oya H, … Howard MA (2010). Direct recordings of pitch responses from human auditory cortex. Current Biology, 20(12), 1128–1132. [PubMed: 20605456]

Gross J, Hoogenboom N, Thut G, Schyns P, Panzeri S, Belin P, & Garrod S (2013). Speech rhythms and multiplexed oscillatory sensory coding in the human brain. PLoS Biology, 11(12), e1001752. [PubMed: 24391472]

Grossberg S, & Kazerounian S (2011). Laminar cortical dynamics of conscious speech perception: Neural model of phonemic restoration using subsequent context in noise. The Journal of the Acoustical Society of America, 130(1), 440–460. [PubMed: 21786911]

Guo W, Chambers AR, Darrow KN, Hancock KE, Shinn-Cunningham BG, & Polley DB (2012). Robustness of cortical topography across fields, laminae, anesthetic states, and neurophysiological signal types. Journal of Neuroscience, 32(27), 9159–9172. [PubMed: 22764225]

Gwilliams L, Linzen T, Poeppel D, & Marantz A (2018). In Spoken Word Recognition, the Future Predicts the Past. Journal of Neuroscience, 38(35), 7585–7599. [PubMed: 30012695]

Hackett TA, Preuss TM, & Kaas JH (2001). Architectonic identification of the core region in auditory cortex of macaques, chimpanzees, and humans. Journal of Comparative Neurology, 441(3), 197–222. [PubMed: 11745645]

Hamilton LS, Edwards E, & Chang EF (2018). A Spatial Map of Onset and Sustained Responses to Speech in the Human Superior Temporal Gyrus. Current Biology.

Heil P (1997). Auditory cortical onset responses revisited. I. First-spike timing. Journal of Neurophysiology, 77(5), 2616–2641. [PubMed: 9163380]

Heil P (2004). First-spike latency of auditory neurons revisited. Current Opinion in Neurobiology, 14(4), 461–467. [PubMed: 15321067]

Hermes DJ (1990). Vowel-onset detection. The Journal of the Acoustical Society of America, 87(2), 866–873. [PubMed: 2307780]

Hickok G, & Poeppel D (2007). The cortical organization of speech processing. Nature Reviews Neuroscience, 8(5), 393. [PubMed: 17431404]

Hillenbrand J, Getty LA, Clark MJ, & Wheeler K (1995). Acoustic characteristics of American English vowels. The Journal of the Acoustical Society of America, 97(5), 3099–3111. [PubMed: 7759650]

Holdgraf CR, De Heer W, Pasley B, Rieger J, Crone N, Lin JJ, … Theunissen FE (2016). Rapid tuning shifts in human auditory cortex enhance speech intelligibility. Nature Communications, 7, 13654.

Howard MA, Volkov I, Mirsky R, Garell P, Noh M, Granner M, … others. (2000). Auditory cortex on the human posterior superior temporal gyrus. Journal of Comparative Neurology, 416(1), 79–92. [PubMed: 10578103]

Hullett PW, Hamilton LS, Mesgarani N, Schreiner CE, & Chang EF (2016). Human superior temporal gyrus organization of spectrotemporal modulation tuning derived from speech stimuli. Journal of Neuroscience, 36(6), 2014–2026. [PubMed: 26865624]

Jakobson R, Fant CG, & Halle M (1951). Preliminaries to speech analysis: The distinctive features and their correlates.

Jordan M (1986). Serial order: a parallel distributed approach (ICS Report 8604). San Diego: University of California. Institute for Cognitive Science.

Kaas JH, & Hackett TA (2000). Subdivisions of auditory cortex and processing streams in primates. Proceedings of the National Academy of Sciences, 97(22), 11793–11799.

Keyser SJ, & Stevens KN (1994). Feature geometry and the vocal tract. Phonology, 11(2), 207–236.

Khoshkhoo S, Leonard MK, Mesgarani N, & Chang EF (2018). Neural correlates of sine-wave speech intelligibility in human frontal and temporal cortex. Brain and Language.

Kiebel SJ, Von Kriegstein K, Daunizeau J, & Friston KJ (2009). Recognizing sequences of sequences. PLoS Computational Biology, 5(8), e1000464. [PubMed: 19680429]

King AJ, & Nelken I (2009). Unraveling the principles of auditory cortical processing: can we learn from the visual system? Nature Neuroscience, 12(6), 698. [PubMed: 19471268]

Klatt DH (1975). Voice onset time, frication, and aspiration in word-initial consonant clusters. Journal of Speech, Language, and Hearing Research, 18(4), 686–706.

Kubanek J, Brunner P, Gunduz A, Poeppel D, & Schalk G (2013). The tracking of speech envelope in the human cortex. PloS One, 8(1), e53398. [PubMed: 23408924]

Lahiri A, & Reetz H (2010). Distinctive features: Phonological underspecification in representation and processing. Journal of Phonetics, 58(1), 44–59.

Larkum M (2013). A cellular mechanism for cortical associations: an organizing principle for the cerebral cortex. Trends in Neurosciences, 36(3), 141–151. [PubMed: 23273272]

Lee Y-S, Turkeltaub P, Granger R, & Raizada RD (2012). Categorical speech processing in Broca's area: an fMRI study using multivariate pattern-based analysis. Journal of Neuroscience, 32(11), 3942–3948. [PubMed: 22423114]

Leonard MK, Baud MO, Sjerps MJ, & Chang EF (2016). Perceptual restoration of masked speech in human cortex. Nature Communications, 7, 13619.

Leonard MK, Bouchard KE, Tang C, & Chang EF (2015). Dynamic encoding of speech sequence probability in human temporal cortex. Journal of Neuroscience, 35(18), 7203–7214. [PubMed: 25948269]

Leonard MK, Cai R, Babiak MC, Ren A, & Chang EF (2016). The peri-Sylvian cortical network underlying single word repetition revealed by electrocortical stimulation and direct neural recordings. Brain and Language.

Leonard MK, & Chang EF (2014). Dynamic speech representations in the human temporal lobe. Trends in Cognitive Sciences, 18(9), 472–479. [PubMed: 24906217]

Li L, Ji X, Liang F, Li Y, Xiao Z, Tao HW, & Zhang LI (2014). A feedforward inhibitory circuit mediates lateral refinement of sensory representation in upper layer 2/3 of mouse primary auditory cortex. Journal of Neuroscience, 34(41), 13670–13683. [PubMed: 25297094]

Liégeois-Chauvel C, Lorenzi C, Trébuchon A, Régis J, & Chauvel P (2004). Temporal envelope processing in the human left and right auditory cortices. Cerebral Cortex, 14(7), 731–740. [PubMed: 15054052]

Lisker L (1986). "Voicing" in English: A catalogue of acoustic features signaling/b/versus/p/in trochees. Language and Speech, 29(1), 3–11. [PubMed: 3657346]

Mante V, Sussillo D, Shenoy KV, & Newsome WT (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. Nature, 503(7474), 78. [PubMed: 24201281]

Marr D (1982). Vision: A computational investigation into the human representation and processing of visual information. MIT Press Cambridge, Massachusetts.

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

McNeill D, & Lindig K (1973). The perceptual reality of phonemes, syllables, words, and sentences. Journal of Memory and Language, 12(4), 419.

McQueen JM (1998). Segmentation of continuous speech using phonotactics. Journal of Memory and Language, 39(1), 21–46.

Mesgarani N, & Chang EF (2012). Selective cortical representation of attended speaker in multi-talker speech perception. Nature, 485(7397), 233. [PubMed: 22522927]

Mesgarani N, Cheung C, Johnson K, & Chang EF (2014). Phonetic feature encoding in human superior temporal gyrus. Science, 1245994.

Mesgarani N, David SV, Fritz JB, & Shamma SA (2008). Phoneme representation and classification in primary auditory cortex. The Journal of the Acoustical Society of America, 123(2), 899–909. [PubMed: 18247893]

Mesulam M-M, Thompson CK, Weintraub S, & Rogalski EJ (2015). The Wernicke conundrum and the anatomy of language comprehension in primary progressive aphasia. Brain, 138(8), 2423–2437. [PubMed: 26112340]

Miller GA, & Nicely PE (1955). An analysis of perceptual confusions among some English consonants. The Journal of the Acoustical Society of America, 27(2), 338–352.

Mitani A, Shimokouchi M, Itoh K, Nomura S, Kudo M, & Mizuno N (1985). Morphology and laminar organization of electrophysiologically identified neurons in the primary auditory cortex in the cat. Journal of Comparative Neurology, 235(4), 430–447. [PubMed: 3998218]

Moerel M, De Martino F, & Formisano E (2014). An anatomical and functional topography of human auditory cortical areas. Frontiers in Neuroscience, 8, 225. [PubMed: 25120426]

Moerel M, De Martino F, Uğurbil K, Formisano E, & Yacoub E (2018). Evaluating the columnar stability of acoustic processing in the human auditory cortex. Journal of Neuroscience, 38(36), 7822–7832. [PubMed: 30185539]

Moerel M, De Martino F, Uğurbil K, Yacoub E, & Formisano E (2019). Processing complexity increases in superficial layers of human primary auditory cortex. Scientific Reports, 9(1), 5502 10.1038/s41598-019-41965-w [PubMed: 30940888]

Nourski KV, Reale RA, Oya H, Kawasaki H, Kovach CK, Chen H, … Brugge JF (2009). Temporal envelope of time-compressed speech represented in the human auditory cortex. Journal of Neuroscience, 29(49), 15564–15574. [PubMed: 20007480]

Nourski KV, Steinschneider M, Oya H, Kawasaki H, & Howard MA III (2015). Modulation of response patterns in human auditory cortex during a target detection task: an intracranial electrophysiology study. International Journal of Psychophysiology, 95(2), 191–201. [PubMed: 24681353]

Nourski KV, Steinschneider M, Oya H, Kawasaki H, Jones RD, & Howard MA (2012). Spectral organization of the human lateral superior temporal gyrus revealed by intracranial recordings. Cerebral Cortex, 24(2), 340–352. [PubMed: 23048019]

Nourski KV, Steinschneider M, Rhone AE, & Howard MA III (2017). Intracranial electrophysiology of auditory selective attention associated with speech classification tasks. Frontiers in Human Neuroscience, 10, 691. [PubMed: 28119593]

Nourski KV, Steinschneider M, Rhone AE, Kovach CK, Kawasaki H, & Howard MA III (2019). Differential responses to spectrally degraded speech within human auditory cortex: An intracranial electrophysiology study. Hearing Research, 371, 53–65. [PubMed: 30500619]

Obleser J, & Kotz SA (2009). Expectancy constraints in degraded speech modulate the language comprehension network. Cerebral Cortex, 20(3), 633–640. [PubMed: 19561061]

Obleser J, Wise RJ, Dresner MA, & Scott SK (2007). Functional integration across brain regions improves speech perception under adverse listening conditions. Journal of Neuroscience, 27(9), 2283–2289. [PubMed: 17329425]

O'Connell MN, Barczak A, Schroeder CE, & Lakatos P (2014). Layer specific sharpening of frequency tuning by selective attention in primary auditory cortex. Journal of Neuroscience, 34(49), 16496–16508. [PubMed: 25471586]

Oganian Y, & Chang EF (2018). A speech envelope landmark for syllable encoding in human superior temporal gyrus. BioRxiv, 388280.

Overath T, Zhang Y, Sanes DH, & Poeppel D (2012). Sensitivity to temporal modulation rate and spectral bandwidth in the human auditory system: fMRI evidence. Journal of Neurophysiology, 107(8), 2042–2056. [PubMed: 22298830]

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. Journal of Cognitive Neuroscience, 29(6), 1044–1060. [PubMed: 28253074]

Ozker M, Yoshor D, & Beauchamp MS (2018). Converging Evidence From Electrocorticography and BOLD fMRI for a Sharp Functional Boundary in Superior Temporal Gyrus Related to Multisensory Speech Processing. Frontiers in Human Neuroscience, 12.

Park H, Ince RA, Schyns PG, Thut G, & Gross J (2015). Frontal top-down signals increase coupling of auditory low-frequency oscillations to continuous speech in human listeners. Current Biology, 25(12), 1649–1653. [PubMed: 26028433]

Paton JJ, & Buonomano DV (2018). The Neural Basis of Timing: Distributed Mechanisms for Diverse Functions. Neuron, 98(4), 687–705. [PubMed: 29772201]

Peelle JE, & Davis MH (2012). Neural oscillations carry speech rhythm through to comprehension. Frontiers in Psychology, 3, 320. [PubMed: 22973251]

Peterson GE, & Barney HL (1952). Control methods used in a study of the vowels. The Journal of the Acoustical Society of America, 24(2), 175–184.

Petkov CI, Kayser C, Augath M, & Logothetis NK (2006). Functional imaging reveals numerous fields in the monkey auditory cortex. PloS Biology, 4(7), e215. [PubMed: 16774452]

Phillips WA, & Singer W (1997). In search of common foundations for cortical computation. Behavioral and Brain Sciences, 20(4), 657–683. [PubMed: 10097008]

Phillips W, Clark A, & Silverstein SM (2015). On the functions, mechanisms, and malfunctions of intracortical contextual modulation. Neuroscience & Biobehavioral Reviews, 52, 1–20. [PubMed: 25721105]

Poeppel D, Idsardi WJ, & Van Wassenhove V (2008). Speech perception at the interface of neurobiology and linguistics. Philosophical Transactions of the Royal Society of London B: Biological Sciences, 353(1493), 1071–1086.

Price CJ (2012). A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. Neuroimage, 62(2), 816–847. [PubMed: 22584224]

Rhone AE, Nourski KV, Oya H, Kawasaki H, Howard MA III, & McMurray B (2016). Can you hear me yet? An intracranial investigation of speech and non-speech audiovisual interactions in human cortex. Language, Cognition and Neuroscience, 31(2), 284–302.

Robson H, Keidel JL, Ralph MAL, & Sage K (2012). Revealing and quantifying the impaired phonological analysis underpinning impaired comprehension in Wernicke's aphasia. Neuropsychologia, 50(2), 276–288. [PubMed: 22172546]

Rosen S (1992). Temporal information in speech: acoustic, auditory and linguistic aspects. Phil. Trans. R. Soc. Lond. B, 336(1278), 367–373. [PubMed: 1354376]

Roux F-E, Minkin K, Durand J-B, Sacko O, Réhault E, Tanova R, & Démonet J-F (2015). Electrostimulation mapping of comprehension of auditory and visual words. Cortex, 71, 398–408. [PubMed: 26332785]

Saffran JR, Newport EL, & Aslin RN (1996). Word segmentation: The role of distributional cues. Journal of Memory and Language, 35(4), 606–621.

Sakata S, & Harris KD (2009). Laminar structure of spontaneous and sensory-evoked population activity in auditory cortex. Neuron, 64(3), 404–418. [PubMed: 19914188]

Samuel AG (1987). Lexical uniqueness effects on phonemic restoration. Journal of Memory and Language, 26(1), 36.

Santoro R, Moerel M, De Martino F, Goebel R, Ugurbil K, Yacoub E, & Formisano E (2014). Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. PLoS Computational Biology, 10(1), e1003412. [PubMed: 24391486]

Sapir E (1925). Sound patterns in language. Language, 1(2), 37–51.

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

Scott SK, Blank CC, Rosen S, & Wise RJ (2000). Identification of a pathway for intelligible speech in the left temporal lobe. Brain, 123(12), 2400–2406. [PubMed: 11099443]

See JZ, Atencio CA, Sohal VS, & Schreiner CE (2018). Coordinated neuronal ensembles in primary auditory cortical columns. Elife, 7, e35587. [PubMed: 29869986]

Shamma SA (1985). Speech processing in the auditory system I: The representation of speech sounds in the responses of the auditory nerve. The Journal of the Acoustical Society of America, 78(5), 1612–1621. [PubMed: 4067077]

Shannon RV, Zeng F-G, Kamath V, Wygonski J, & Ekelid M (1995). Speech recognition with primarily temporal cues. Science, 270(5234), 303–304. [PubMed: 7569981]

Sharpee TO (2016). How invariant feature selectivity is achieved in cortex. Frontiers in Synaptic Neuroscience, 8, 26. [PubMed: 27601991]

Shattuck-Hufnagel S, & Turk AE (1996). A prosody tutorial for investigators of auditory sentence processing. Journal of Psycholinguistic Research, 25(2), 193–247. [PubMed: 8667297]

Sohoglu E, Peelle JE, Carlyon RP, & Davis MH (2012). Predictive top-down integration of prior knowledge during speech perception. Journal of Neuroscience, 32(25), 8443–8453. [PubMed: 22723684]

Steinschneider M, Nourski KV, & Fishman YI (2013). Representation of speech in human auditory cortex: is it special? Hearing Research, 305, 57–73. [PubMed: 23792076]

Steinschneider M, Nourski KV, Kawasaki H, Oya H, Brugge JF, & Howard MA III (2011). Intracranial study of speech-elicited activity on the human posterolateral superior temporal gyrus. Cerebral Cortex, 21(10), 2332–2347. [PubMed: 21368087]

Steinschneider M, Nourski KV, Rhone AE, Kawasaki H, Oya H, & Howard MA III (2014). Differential activation of human core, non-core and auditory-related cortex during speech categorization tasks as revealed by intracranial recordings. Frontiers in Neuroscience, 8, 240. [PubMed: 25157216]

Steinschneider M, Reser DH, Fishman YI, Schroeder CE, & Arezzo JC (1998). Click train encoding in primary auditory cortex of the awake monkey: evidence for two mechanisms subserving pitch perception. The Journal of the Acoustical Society of America, 104(5), 2935–2955. [PubMed: 9821339]

Steinschneider M, Volkov IO, Noh MD, Garell PC, & Howard MA III (1999). Temporal encoding of the voice onset time phonetic parameter by field potentials recorded directly from human auditory cortex. Journal of Neurophysiology, 82(5), 2346–2357. [PubMed: 10561410]

Stevens KN (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. The Journal of the Acoustical Society of America, 777(4), 1872–1891.

Stevens KN, & Blumstein SE (1981). The search for invariant acoustic correlates of phonetic features. Perspectives on the Study of Speech, 1–38.

Sussillo D, & Abbott LF (2009). Generating coherent patterns of activity from chaotic neural networks. Neuron, 63(4), 544–557. [PubMed: 19709635]

Tang C, Hamilton L, & Chang E (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, Frim DM, … Kohrman MH (2008). ECoG gamma activity during a language task: differentiating expressive and receptive speech areas. Brain, 737(8), 2013–2027.

Versnel H, & Shamma SA (1998). Spectral-ripple representation of steady-state vowels in primary auditory cortex. The Journal of the Acoustical Society of America, 703(5), 2502–2514.

Walker KM, Bizley JK, King AJ, & Schnupp JW (2011). Multiplexed and robust representations of sound features in auditory cortex. Journal of Neuroscience, 37(41), 14565–14576.

Wang J, Narain D, Hosseini EA, & Jazayeri M (2018). Flexible timing by temporal scaling of cortical responses. Nature Neuroscience, 27(1), 102.

Warren RM (1970). Perceptual restoration of missing speech sounds. Science, 767(3917), 392–393.

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

Wernicke C (1881). Lehrbuch der gehirnkrankheiten für aerzte undstudirende (Vol. 2). Fischer.

Winkowski DE, & Kanold PO (2013). Laminar transformation of frequency organization in auditory cortex. Journal of Neuroscience, 33(4), 1498–1508. [PubMed: 23345224]

Wöstmann M, Fiedler L, & Obleser J (2017). Tracking the signal, cracking the code: Speech and speech comprehension in non-invasive human electrophysiology. Language, Cognition and Neuroscience, 32(7), 855–869.

Wu P-Y, Chu Y-H, Lin J-FL, Kuo W-J, & Lin F-H (2018). Feature-dependent intrinsic functional connectivity across cortical depths in the human auditory cortex. Scientific Reports, 8(1), 13287. [PubMed: 30185951]

Xing D, Yeh C-I, Burns S, & Shapley RM (2012). Laminar analysis of visually evoked activity in the primary visual cortex. Proceedings of the National Academy of Sciences, 709(34), 13871–13876.

Yaron A, Hershenhoren I, & Nelken I (2012). Sensitivity to complex statistical regularities in rat auditory cortex. Neuron, 76(3), 603–615. [PubMed: 23141071]

Yildiz IB, Mesgarani N, & Deneve S (2016). Predictive ensemble decoding of acoustical features explains context-dependent receptive fields. Journal of Neuroscience, 36(49), 12338–12350. [PubMed: 27927954]

Yildiz IB, von Kriegstein K, & Kiebel SJ (2013). From birdsong to human speech recognition: Bayesian inference on a hierarchy of nonlinear dynamical systems. PLoS Computational Biology, 9(9), e1003219. [PubMed: 24068902]

Zec D (1995). Sonority constraints on syllable structure. Phonology, 72(1), 85–129.

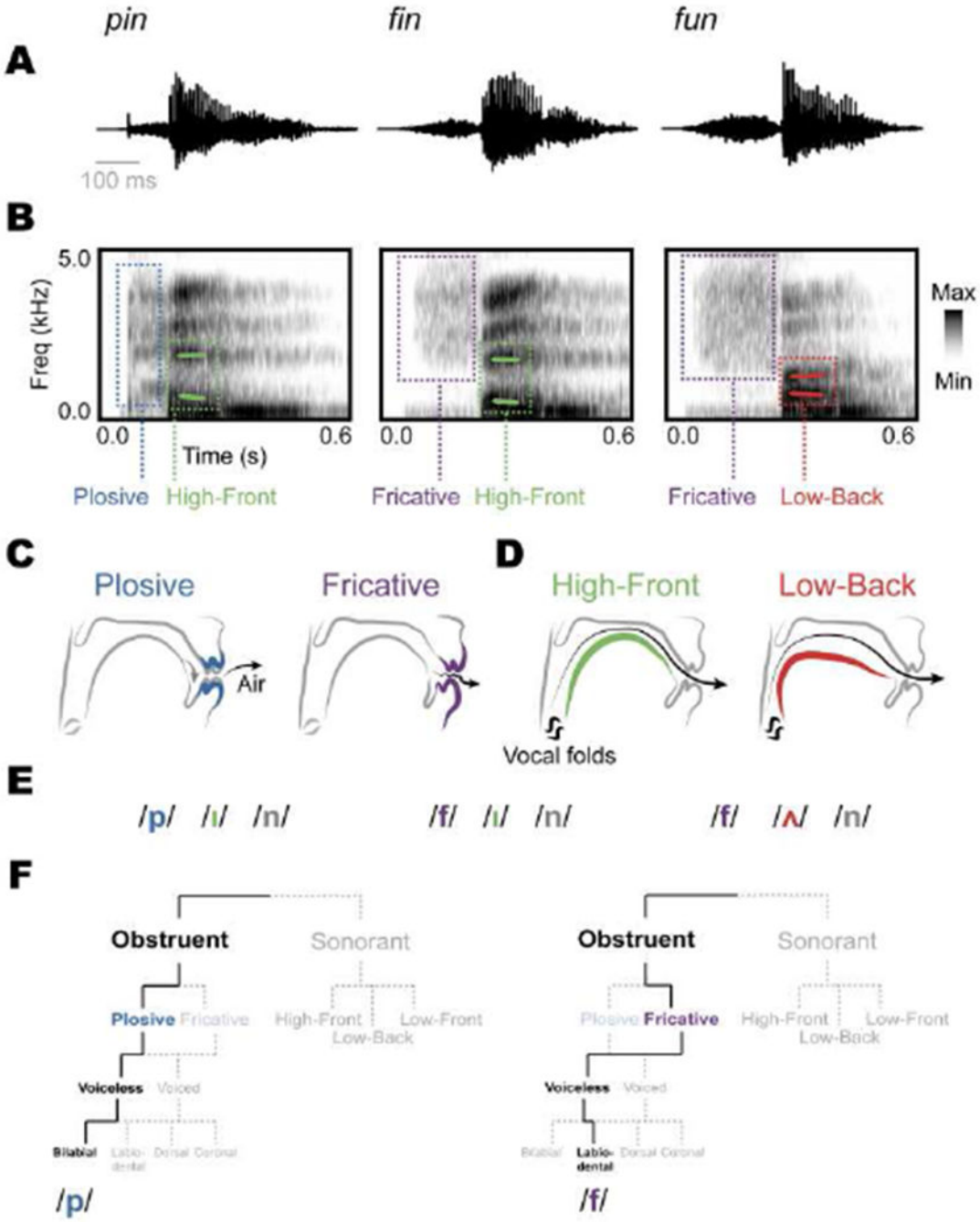


**Figure 1. Speech sounds can be described in multiple complementary ways.**

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**) The acoustic waveforms of show a broad distinction between low amplitude and aperiodic features (consonants), and high amplitude and strong periodic features (vowels). (**B**) Spectrogram representations of these words show how each sound is characterized by different spectrotemporal patterns of acoustic energy. (**C**) Articulatory descriptions of these sounds characterize acoustic-phonetic features. *Plosives* are produced

by initially blocking the airflow (gray), 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. (**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. In contrast, *low-back* vowels show the reverse pattern. These two features are examples of sonorants. (**E**) Each of the example words can also be characterized as a set of successive abstract phonemes: /pɪn/, /fɪn/, and /fʌn/. (**F**) Multiple 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).

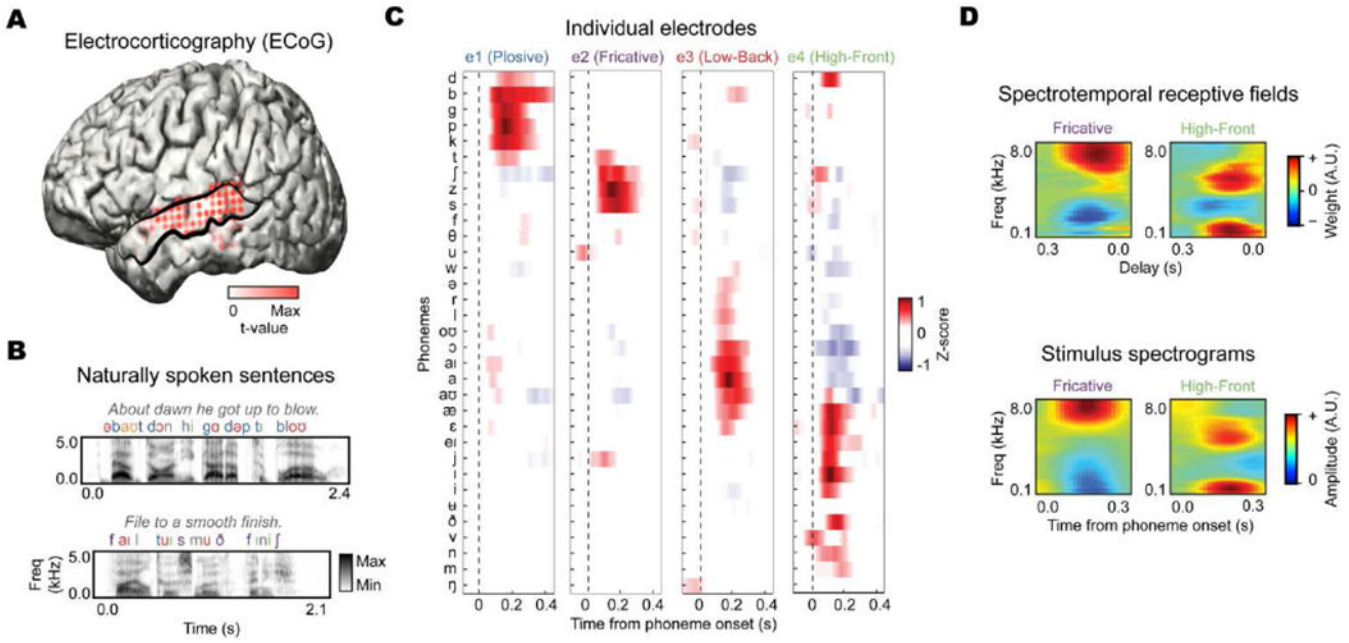


**Figure 2. Local encoding of acoustic-phonetic features in human superior temporal gyrus (STG).**

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**) ECoG electrodes over human STG (outlined in black) 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. (**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).

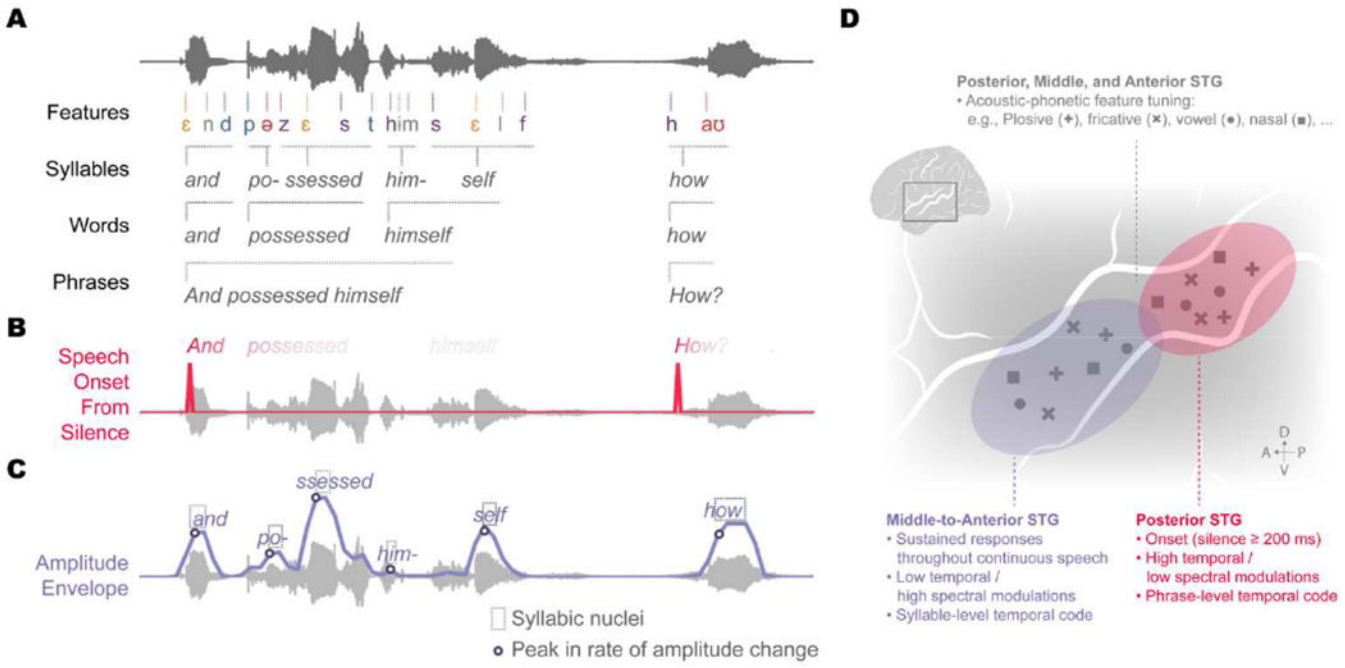


**Figure 3. STG is parcellated into two major zones that track temporal landmarks relevant for speech processing.**

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/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, while 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.

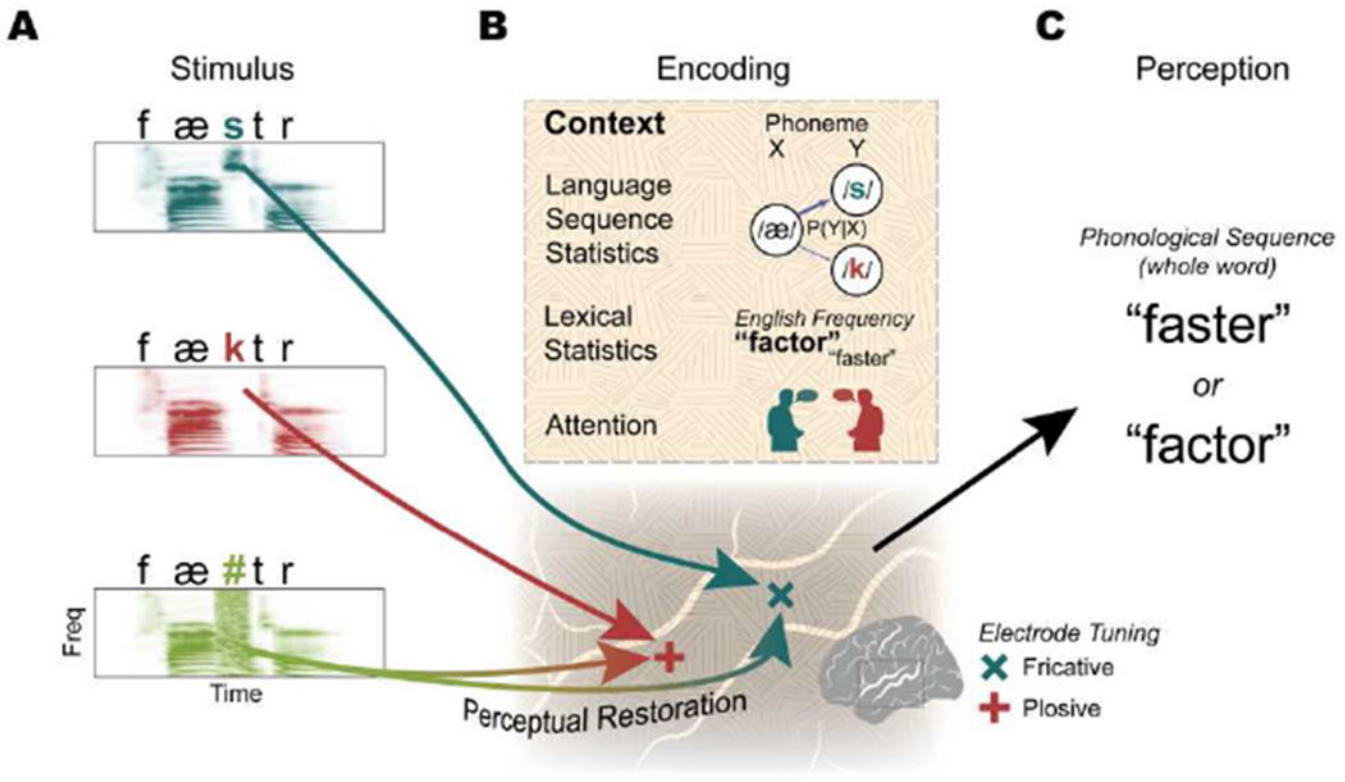


**Figure 4. STG combines acoustic-phonetic tuning with various sources of context to compute perceptual representations of speech.**

(**A**) Example words and their acoustic spectrograms that differ in a single phoneme/acoustic-phonetic feature (/s/ vs. /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.

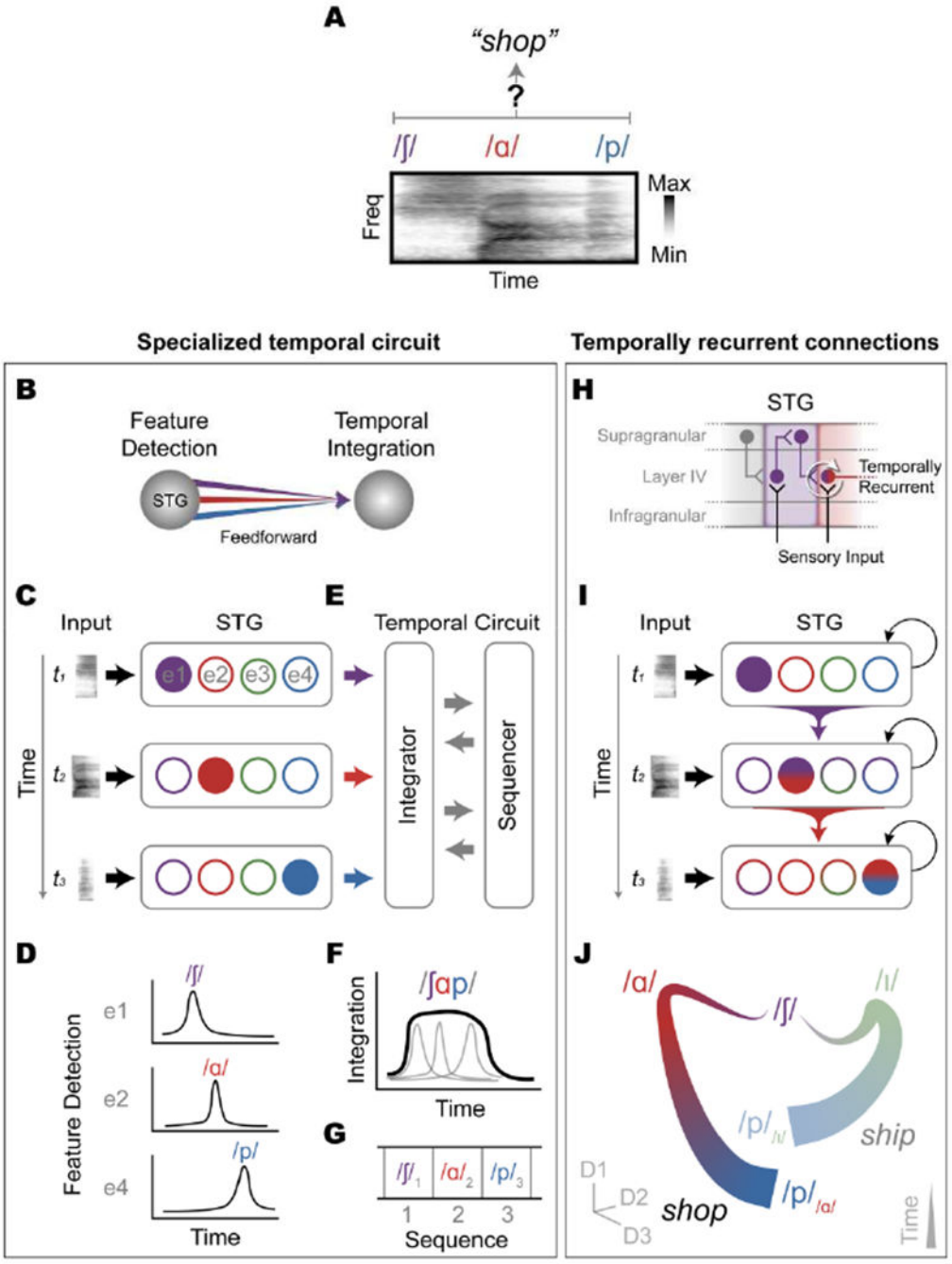


**Figure 5. Computational implementations of temporal sequencing and binding in speech cortex.**

(**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**) Distinct STG populations (recorded with different electrodes: e1, e2, etc.) detect acoustic-phonetic features from the acoustic input (**D**) by generating spatially and temporally independent 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/sequencer has a relatively long temporal

window, and is thus 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 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/ vs. /ɪ/), but also from the context-modulated responses to the final consonants (/p/$_{/ɑ/}$ vs. /p/$_{/ɪ/}$; i.e., /p/ does not occupy a single point in the state-space).