Title
Dynamic speech representations in the human temporal lobe

Journal
Trends in Cognitive Sciences, 18(9)

ISSN
1364-6613

Authors
Leonard, Matthew K
Chang, Edward F

Publication Date
2014-09-01

DOI
10.1016/j.tics.2014.05.001

Peer reviewed

Published in final edited form as:
Trends Cogn Sci. 2014 September ; 18(9): 472–479. doi:10.1016/j.tics.2014.05.001.

Dynamic speech representations in the human temporal lobe

Matthew K. Leonard and **Edward F. Chang**
Department of Neurological Surgery, University of California, San Francisco, 675 Nelson Rising Ln., Room 535. San Francisco, CA 94158

Abstract

Speech perception requires rapid integration of acoustic input with context-dependent knowledge. Recent methodological advances have allowed researchers to identify underlying information representations in primary and secondary auditory cortex, and to examine how context modulates these representations. We review recent studies that focus on contextual modulations of neural activity in the superior temporal gyrus (STG), a major hub for spectrotemporal encoding. Recent findings suggest a highly interactive flow of information processing through the auditory ventral stream, including influences of higher-level linguistic and metalinguistic knowledge. Such mechanisms may give rise to more abstract representations, such as those for words. We discuss the importance of characterizing neural representations of context-dependent and dynamic patterns of neural activity in the approach to speech perception research.

Introduction

How does the human brain generate phenomenologically rich representations of words from the complex and noisy acoustic speech signal? This is not a new question, with many of our current theories and observations heavily influenced by those nearly 140 years old [1,2]. In this review, we consider the implications of progress that has been made in redefining the issues central to speech perception. Recent advances have allowed researchers to examine the functioning human brain with an unprecedented level of detail, with particular attention to decoding the representations contained in speech-evoked neural responses [3–5], an important step beyond localizing task-dependent activity. Combined with a growing and productive interaction between linguistics and neuroscience [6], new recording and analysis methods have created a pivotal moment for understanding the neural basis of speech perception.

Organization of the ventral stream

Human neuroimaging and neurophysiology studies support the concept of an information processing hierarchy for speech perception in the temporal lobe. Responses evoked by

Corresponding Author: Chang, E.F., changed@neurosurg.ucsf.edu.

speech sounds, words, and sentences show activity that spreads primarily from posterior to anterior temporal areas [7–14]. This dominant direction of information flow is supported by anatomical connections between the superior temporal plane and anteroventral temporal areas [15], and is commonly known as the 'ventral stream' for speech perception [16,17]. This contrasts with a distinct but related network that connects posterior superior temporal areas to both ventral and dorsal frontal, as well as inferior parietal cortex, known as the 'dorsal stream' [18]. It is clear from numerous imaging studies that dorsal stream areas are active during speech perception, however their exact functional roles are debated [19].

Most current conceptions of the speech perception system view the ventral stream as the primary pathway for transforming acoustic sensory signals into abstract linguistic representations such as phonemes and words. These theories posit a hierarchical flow of information among temporal lobe regions that support largely discrete (albeit strongly connected) aspects of linguistic encoding. Hickok & Poeppel's [16] dominant view argues that frequency and amplitude information from primary auditory cortex (A1) is fed to the posterior superior temporal gyrus (STG), which supports a spectrotemporal encoding of the most fundamental features of the speech signal. Along the ventral stream, STG is connected to the posterior superior temporal sulcus (STS), which encodes phonological-level processes (e.g., phonemes). Finally, STS is directly connected to posterior middle temporal gyrus (MTG) and inferior temporal sulcus (ITS), which are the "lexical interface", where abstract representations of words are stored [20,21]. This network is hypothesized to be largely bilateral, particularly for lower-level aspects of acoustic and phonological processing [16,22,23], although the extent and function of lateralized activity is debated [24].

As we review below, this conception finds strong support from a vast body of neuroimaging studies, with particular convergence in superior temporal areas that are hypothesized to support spectrotemporal and phonological processing. The primary goal of the present review is to describe recent advances that provide important extensions of these findings, particularly in the STG. We will argue that, in large part due to the nature of the processing that occurs in this region and due to methodological advances, our understanding of the role of STG in speech perception exceeds that of most other brain areas. Specifically, the ability to decode neural activity along spectrotemporal, linguistic, and metalinguistic dimensions means that STG is characterized according to its underlying *representations*, rather than its differential responses to stimuli that vary along theoretically interesting dimensions (e.g., clear speech versus acoustic controls that maintain aspects of the spectrotemporal structure of the input, but degrade intelligibility). We will present an argument that this level of specificity is necessary, although difficult to achieve, if we wish to understand more abstract linguistic representations, such as those for words.

Early cortical auditory encoding

To examine the specific roles that STG plays in the speech perception hierarchy, it is important to understand the inputs to this region. A large body of work has established important aspects of sensory processing that occur in the ascending auditory system en route to the primary auditory cortex in several mammalian species [25–27]. A1 in humans, located on the posteromedial portion of Heschl's gyrus, is characterized by at least one major

tonotopic axis [28]. An important aspect of this tonotopic organization is that A1 neuronal populations show narrow spectral tuning [29,30], which combined with selectivity for temporal features of the stimulus give rise to perceptual distinctions such as pitch [31,32]. Heschl's gyrus also encompasses non-core auditory regions, and exhibits frequency-specific response characteristics indicative of both rate and temporal coding of auditory stimuli both within and across neural populations [33]. In secondary auditory areas such as planum temporale (PT), preferences for temporal features are significantly decreased relative to A1, while spectral specificity is more finely tuned [32]. This spectral preference includes complex multi-peak tuning at octave intervals [34], potentially allowing multiple stimulus features to be integrated into distinct auditory objects. In sum, recent advances in optimizing the spatiotemporal and frequency resolutions of human neural recording methodologies have demonstrated in new detail that activity in primary and early secondary auditory areas indicates highly specialized tuning for relevant stimulus features, however this does not appear to be specialized for speech (see [35] for an excellent review comparing direct neurophysiological responses from human and nonhuman primates).

Thus, two aspects of early cortical auditory processing are clear. First, A1 and surrounding areas have been well-characterized, both in their responses to a wide variety of stimuli (including speech), and also in the nature of the information that those responses represent. Second, and perhaps most important for understanding the early cortical stages of speech perception, is the fact that these areas do not show strictly linear responses that can be characterized as faithful representations of the physical stimulus. As we shall see, if the goal of the speech perception hierarchy is to reach abstracted representations of the input, it is critical to understand that abstraction is not a feature that is unique to downstream regions in the anterior and ventral temporal lobes.

Stimulus and linguistic representations in STG

Despite showing stimulus- and context- dependent modulations in neural activity, few would argue that A1 exhibits responses that are specific to speech. In contrast, a major target of primary auditory outputs is the STG, which is one of the best-characterized regions in the speech perception system, and which shows responses that suggest the earliest stages of speech-tuned representation. Like its upstream neighbors, STG is highly sensitive to the spectrotemporal content of the acoustic signal. Recent studies have used electrocorticography (ECoG), an invasive method in which electrodes are implanted directly on the brain surface in humans, to understand how distinct neural populations in STG respond to sound. The majority of these studies examine neural activity in the high-gamma (∼70-200Hz) range [36], which is strongly correlated with multiunit spiking activity [37,38] and the BOLD response in fMRI [39]. There are two major notable characteristics of responses to sound in this region. Distinct STG neuronal populations encode the temporal structure of non-speech acoustic input differently depending on the frequency content of the signal [40]. Likewise, in the spectral domain, populations are selective (although generally broadly-tuned) to ranges of frequencies [41]. Importantly, this selectivity is both amplitude-invariant and malleable over a millisecond time scale, suggesting that spectrotemporally complex stimuli may be encoded through cross-frequency integration mechanisms in relatively local areas of cortex. Thus, studies that have examined STG responses to non-

speech stimuli have demonstrated local selective responses that might give rise to population activity that encodes the perception of spectrotemporally complex input.

Although significantly more complex (both physically and behaviorally) than pure-tones and clicks, it is possible to examine the nature of the spectrotemporal representation of speech in the superior temporal lobe. Stimulus reconstruction methods have demonstrated a strong correspondence between the speech spectrogram and distributed neural activity along the posterior STG [4]. This relationship is particularly strong for the spectrotemporal aspects of the acoustic input that are relevant for speech intelligibility, specifically temporal modulation rates that correspond to syllable onsets and offsets.

A recent ECoG study explored how this spectrotemporal sensitivity relates to local tuning for phonetic features. Mesgarani and colleagues played hundreds of sentences spoken by hundreds of different speakers while neural activity was recorded from high-density grids over STG [42]. The stimuli provided a large number of examples of all English phonemes, which allowed the authors to examine the relative selectivity of each electrode to each phoneme. They found that, rather than being selective to individual phonemes (e.g., /s/, /m/, /k/), STG neural populations are tuned to particular acoustic features, such as fricatives, nasals, and plosives. Vowels showed similar feature-based representations for low-back, low-front, and high-front features, which were directly related to the encoding of formant frequency variability, particularly the difference between F1 and F2. These results are consistent with the view that individual phonemes are not represented by discrete spatial points on the STG, but rather are represented by population neural activity defined by a multidimensional feature space.

Data from multiple recording methods including fMRI, MEG, and ECoG has shown that STG activity is sensitive to phonological manipulations of the speech signal that alter its intelligibility. Numerous studies, particularly those pioneered by Scott [43] and Davis and colleagues [44] have compared neural responses for speech to non-speech sounds that preserve important spectral or temporal aspects of the signal [45–51]. Using a process called noise-vocoding, in which the spectrogram is essentially smoothed in the spectral or temporal axis, it has been demonstrated that left superior temporal areas are more sensitive to the temporal content of speech, while the homologous right hemisphere regions are more sensitive to the frequency content [50]. These results are in line with a provocative theory put forward by Poeppel and colleagues on the mechanistic differences between the cerebral hemispheres during speech perception [52]. These findings also suggest that abstract representations such as phonemes arise from finely-tuned encoding of acoustic features in local regions of the temporal lobe.

Recent advances in recording and multivariate analysis methods have provided more detailed information about how neural activity is tuned in STG. Responses in this region track important contrastive acoustic cues such as voice-onset time (VOT; [53]) and place of articulation [54]. These findings are particularly important because they demonstrate that STG is sensitive to acoustic cues that also reflect important linguistic distinctions. A well-studied phenomenon in speech processing is categorical phoneme perception, in which a linear continuum of speech sounds is perceived non-linearly. Several recent neuroimaging

studies have localized this perceptual phenomenon to the lateral superior temporal cortex [55–58], and one recent ECoG study provided a detailed examination of local STG activity during a categorical phoneme perception task. Participants listened to a continuum of synthesized speech sounds that ranged from /ba/ to /da/ to /ga/, by changing F2 onset in linear stepwise increments [54] (Figure 1a). Across the population, neural activity patterns were spatially distinct for the three stimulus categories, even within the space of only a couple of centimeters (Figure 1b). Using multivariate classification methods, stimulus-specific discriminability was observed in this activity (Figure 1c), suggesting that at the peak of pattern dissimilarity across categories, certain perceptual contrasts arose from specific neural populations. Interestingly, there was also evidence of organization along acoustic sensitivities, as the representations of speech tokens were ordered according to F2 in one dimension (see ordering along x-axis in Figure 1d), but the overall pattern was categorical in two dimensions (Figure 1d), demonstrating that this perceptual phenomenon is encoded non-linearly in the brain. These categorical effects were strongest at ∼110ms, essentially at the same time as the peak response, suggesting that the representation occurs in situ or previous to the STG, rather than through top-down influences of other brain regions.

Together, these studies demonstrate that linguistic phenomena such as categorical phoneme perception arise from neural sensitivity to acoustic features, primarily within the lateral superior temporal cortex. Furthermore, these new approaches have extended what was previously known about phonemic representations by showing that neural activity based on low-level feature selectivity is modulated by higher-level linguistic knowledge and experience to represent complex and increasingly abstract information.

Cognitive and linguistic modulation

The studies described thus far provide compelling evidence that STG is a major hub for sublexical processing in the speech perception hierarchy. Like many other brain regions, responses in STG are non-linear not only along physical stimulus dimensions (such as categorical phoneme perception), but also according to complex cognitive contexts and task demands. For example, several recent studies have demonstrated that STG activity is powerfully modulated by the attentional constraints of the task [59–63]. One study showed that the relatively fine-scale representation of both spectral and temporal acoustic information in STG is highly dependent on whether the listener is attending to the content of the speech stream [64] (Figure 2). In this study, participants listened to two speakers simultaneously while ECoG was recorded (Figure 2a-b). The participants were asked to report the content of just one of the speakers, thus attending to only part of the acoustic input. Humans are known to be experts at solving this so-called ‘cocktail party problem’, which is much more difficult for artificial speech recognition systems. The authors found that, consistent with previous work, STG population activity encoded the fine spectrotemporal details of the stimulus. However, while the attended speech stream was robustly represented, it was as if the ignored speaker had not been heard at all (Figure 2c). This striking result is another example of STG activity representing behaviorally-relevant aspects of the stimulus through contextually modulatory activity.

There is also extensive evidence that higher-level linguistic knowledge affects lower-level speech processes. Neural responses along the temporal lobe, including posterior superior temporal regions, are related to speech intelligibility [51,65–67], listener attention to sublexical features [46], number of words (but not pseudowords) in a sentence [68], congruency with a preceding semantic context [66,69], whether the stimulus is a known word or a phonotactically legal pseudoword [70], and familiarity with the specific language being heard [71,72]. In addition, the surrounding acoustic context (including coarticulation and the size of the temporal integration window necessary to understand the input) can impact perception [73,74]. This collection of work is consistent with models positing that low- and high-level representations interact in real-time [75], although this is still a contentious claim [76,77].

To summarize, a key feature of neural activity in STG that has been described in detail is that it reflects a *context-dependent* spectrotemporal representation of speech. In this sense, an area that is typically thought of as having a relatively low-level acoustic processing function [16] actually encodes linguistically and behaviorally meaningful information. This raises a series of important questions that we believe reflect a crucial turning point for the neuroscience of speech perception.

Lexical representations in the ventral stream

The studies reviewed above suggest that activity in STG during speech perception is non-deterministic. That is, it is not possible to predict activity with a high degree of precision simply based on the physical characteristics of the stimulus. This principle is a defining feature of abstract representations, and historically has made it rather difficult to study the underlying representations of neural systems beyond early sensory cortices. It also makes it potentially even more difficult to answer a fundamental question in speech perception: How are sublexical representations combined across time and brain regions to form or access abstract lexical representations? In the following section, we propose that recent methodological and theoretical advances may allow us to tackle this question, which has both scientific and clinically relevant applications.

Perhaps the most difficult aspect of this question is the fact that there is no agreed-upon definition of a word. Recent work [78,79] has revived an old debate on what information should be attributed to words in the mental lexicon. The problem is that there is simply too much knowledge (semantic, syntactic, morphological, and broadly contextual) that is attached to a given word in a given context. These attributions have even been extended to include speaker identity [80], which is typically thought of as a paradigmatic example of metalinguistic information particularly because the spectrotemporal features that define speakers are less abstract even than a phoneme [81] (although even phoneme representations have been shown to include speaker identity [82,83]).

What, then, is the nature of the underlying representation of stored lexical information in the brain? Nearly all models of speech perception include a lexical level of representation, which is the ultimate target of lower-level acoustic-phonetic and phonemic inputs [75,84–

86]. However, it is unclear what information is attributed to the lexicon, and if such information should include the contextual content described above.

To date, the neural data on the lexicon have not resolved this question. As previously discussed, most current theories posit a “lexical interface” where abstractword representations are stored [16,20]. This includes distinct areas that are hypothesized to play different roles in representing discrete, yet interconnected aspects of lexical items (e.g., lexical-semantics, lexical phonology, articulatory representations). Numerous studies (reviewed in [20,21]) have compared responses to known real words and phonotactically legal but unfamiliar pseudowords to derive a network centered primarily in posterior MTG and temporal-parietal areas that respond more strongly to familiar word forms. It has been shown that these responses show specificity for certain lexical characteristics, such as being modulated by lexical frequency [87]. Indeed, lexical access itself is strongly modulated by word frequency [88], suggesting a distributed representation of word-level knowledge. Morphological complexity and grammatical category also appear to be integral aspects of word representations, as they dynamically modulate neural activity along the left hemisphere ventral stream [89].

These studies approach the problem of identifying the characteristics of lexical representation through a clever and unique conceit: If it looks like a word, and acts like a word, it is probably a word. Recent work by Gaskell and colleagues has shown that the acquisition of novel spoken word forms is influenced by existing knowledge of language-specific morphological rules [90]. Word learning paradigms also offer the opportunity to measure changes in neural responses to newly acquired forms over time. Scalp EEG responses to unfamiliar words rapidly become more like those of known words, even with a passive listening task [91]. Still, in such cases, it is unclear what aspects of lexical processing are represented in the temporal lobe regions that show such changes.

One of the most compelling signatures of lexical processing can be found in changes to the structure of the entire mental lexicon when new items are added. A recent study by Gagnepain *et al*. found that the introduction of a novel form (‘formubo’) changes the activity evoked by familiar words (‘formula’) in left superior temporal cortex [86]. These changes occur at the level of the phoneme sequence representations, where the authors propose that a temporal predictive coding scheme compares the real-time phonemic input to a likelihood density function derived from stored word representations. The difference between the bottom-up and top-down representations is the prediction error, which is reflected in the neural signal recorded at the scalp. The ability to adapt the structure of the mental lexicon so rapidly likely not only underlies humans' uncanny ability to learn new words throughout the lifespan, but also reflects the distributed, multilevel hierarchical organization of lexical information in the brain. We believe it will prove useful to apply decoding methods that have successfully uncovered speech representations in STG [3,4,54] to these aspects of neural activity that reflect the characteristics of words at both sub-lexical and abstract levels. While this approach may be viewed as looking for indirect signs of words in the brain, as opposed to direct encoding of acoustic features in STG, lexical representations may be sufficiently complex, abstract, and context-dependent that such an analogous signal simply may not exist.

Gagnepain *et al*. interpret their findings in a Bayesian predictive coding framework, which has become a very popular (if controversial) mechanistic explanation for neural processing of sequential input [92], including robust recognition of spoken words [93,94]. In general, evidence is accumulating to suggest that the basic principles of neural computations are statistical in nature [95,96]. In the auditory domain, low-level responses are highly dependent on the predictability of both local [97] and longer-term event probabilities [98,99]. Even if the primary currency of neural computation turns out not to be prediction error, as advocated by the strong predictive coding argument, the notion that neural representations (including abstract lexical responses) are emergent from contextually-driven integration of low-level input and higher-level predictions is attractive because it allows a large amount of knowledge to reside in the mental lexicon. This is consistent with the non-deterministic nature of representations in higher order auditory areas like STG, and points to fundamental neural processing mechanisms that integrate bottom-up input with knowledge about the world that is stored as statistical distributions (rather than static objects).

Conclusions

We have discussed evidence that representations of speech information cannot be understood in a strictly linear or deterministic hierarchical framework, even for spectrotemporal representations in STG. This presents a challenge for understanding more complex and abstract forms of representation such as words (Box 1), but it also potentially provides a means for major advances in neurolinguistics that parallel those in sensory neuroscience. We believe that machine learning and dynamical systems approaches, combined with high spatial and temporal resolution neuroimaging and neurophysiological recordings will facilitate these advances, since they allow researchers to gain insights into the neural codes that generate the representations we are ultimately interested in. These approaches have proved useful for decoding neural activity in sensory and spectrotemporal brain areas, and their application to higher-order processes like lexical encoding may be achievable if we attempt to decode the characteristics and properties of words and the mental lexicon, rather than the specific signatures of static representations.

Nearly two decades of human brain mapping have led to an unprecedentedly detailed view of the brain bases of speech and language. Now, as we begin to understand some of the more fundamental principles of neural computation, both at the single neuron and network levels, it is becoming possible to move beyond attempts to flesh out the Wernicke-Geschwind model, and instead reconsider some of its basic assumptions. We believe the next five years will be an exciting and productive time for speech neuroscience, which will begin to provide both intellectual and practical benefits beyond what has been possible in the past.

Acknowledgments

M.K.L. was funded by NIH National Research Service Award F32-DC013486. E.F.C. was funded by the US National Institutes of Health grants R00-NS065120, DP2-OD00862 and R01-DC012379, and the Ester A. and Joseph Klingenstein Foundation.

References

1. Wernicke, C. Der aphasische Symptomencomplex: eine psychologische Studie auf anatomischer Basis. Cohn; 1874.
2. Geschwind, N. Disconnexion syndromes in animals and man. Springer; 1974.
3. Formisano E, et al. “ Who” Is Saying“ What”? Brain-Based Decoding of Human Voice and Speech Science. 2008; 322:970–973.
4. Pasley BN, et al. Reconstructing speech from human auditory cortex. PLoS Biol. 2012; 10:e1001251. [PubMed: 22303281]
5. Bouchard KE, et al. Functional organization of human sensorimotor cortex for speech articulation. Nature. 2013; 495:327–332. [PubMed: 23426266]
6. Poeppel D, et al. Speech perception at the interface of neurobiology and linguistics. Philos Trans R Soc B Biol Sci. 2008; 363:1071–1086.
7. Lerner Y, et al. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. J Neurosci. 2011; 31:2906–2915. [PubMed: 21414912]
8. DeWitt I, Rauschecker JP. Phoneme and word recognition in the auditory ventral stream. Proc Natl Acad Sci. 2012; 109:E505–E514. [PubMed: 22308358]
9. Leff AP, et al. The cortical dynamics of intelligible speech. J Neurosci. 2008; 28:13209–13215. [PubMed: 19052212]
10. Okada K, et al. Hierarchical organization of human auditory cortex: evidence from acoustic invariance in the response to intelligible speech. Cereb Cortex. 2010; 20:2486–2495. [PubMed: 20100898]
11. Price CJ. The anatomy of language: a review of 100 fMRI studies published in 2009. Ann N Y Acad Sci. 2010; 1191:62–88. [PubMed: 20392276]
12. Scott BH, et al. Transformation of temporal processing across auditory cortex of awake macaques. J Neurophysiol. 2011; 105:712–730. [PubMed: 21106896]
13. Steinschneider, M. Neural Correlates of Auditory Cognition. Springer; 2013. Phonemic Representations and Categories; p. 151-191.
14. Chevillet M, et al. Functional correlates of the anterolateral processing hierarchy in human auditory cortex. J Neurosci. 2011; 31:9345–9352. [PubMed: 21697384]
15. Hackett TA. Information flow in the auditory cortical network. Hear Res. 2011; 271:133–146. [PubMed: 20116421]
16. Hickok G, Poeppel D. The cortical organization of speech processing. Nat Rev Neurosci. 2007; 8:393–402. [PubMed: 17431404]
17. Rauschecker JP, Scott SK. Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nat Neurosci. 2009; 12:718–724. [PubMed: 19471271]
18. Saur D, et al. Ventral and dorsal pathways for language. Proc Natl Acad Sci. 2008; 105:18035–18040. [PubMed: 19004769]
19. Hickok G, et al. Sensorimotor integration in speech processing: computational basis and neural organization. Neuron. 2011; 69:407–422. [PubMed: 21315253]
20. Gow DW Jr. The cortical organization of lexical knowledge: A dual lexicon model of spoken language processing. Brain Lang. 2012; 121:273–288. [PubMed: 22498237]
21. Davis MH, Gaskell MG. A complementary systems account of word learning: neural and behavioural evidence. Philos Trans R Soc B Biol Sci. 2009; 364:3773–3800.
22. Bozic M, et al. Bihemispheric foundations for human speech comprehension. Proc Natl Acad Sci. 2010; 107:17439–17444. [PubMed: 20855587]
23. McGettigan C, Scott SK. Cortical asymmetries in speech perception: what's wrong, what's right and what's left? Trends Cogn Sci. 2012; 16:269–276. [PubMed: 22521208]
24. Evans S, et al. The Pathways for Intelligible Speech: Multivariate and Univariate Perspectives. Cereb Cortex. 2013
25. Sharpee TO, et al. Hierarchical representations in the auditory cortex. Curr Opin Neurobiol. 2011; 21:761–767. [PubMed: 21704508]

26. Schreiner CE, Winer JA. Auditory cortex mapmaking: principles, projections, and plasticity. Neuron. 2007; 56:356–365. [PubMed: 17964251]
27. Nelken I. Processing of complex sounds in the auditory system. Curr Opin Neurobiol. 2008; 18:413–417. [PubMed: 18805485]
28. Baumann S, et al. A unified framework for the organization of the primate auditory cortex. Front Syst Neurosci. 2013; 7
29. Moerel M, et al. Processing of natural sounds in human auditory cortex: Tonotopy, spectral tuning, and relation to voice sensitivity. J Neurosci. 2012; 32:14205–14216. [PubMed: 23055490]
30. Bitterman Y, et al. Ultra-fine frequency tuning revealed in single neurons of human auditory cortex. Nature. 2008; 451:197–201. [PubMed: 18185589]
31. Griffiths TD, et al. Direct recordings of pitch responses from human auditory cortex. Curr Biol. 2010; 20:1128–1132. [PubMed: 20605456]
32. Schönwiesner M, Zatorre RJ. Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. Proc Natl Acad Sci. 2009; 106:14611–14616. [PubMed: 19667199]
33. Brugge JF, et al. Coding of repetitive transients by auditory cortex on Heschl's gyrus. J Neurophysiol. 2009; 102:2358–2374. [PubMed: 19675285]
34. Moerel M, et al. Processing of Natural Sounds: Characterization of Multipeak Spectral Tuning in Human Auditory Cortex. J Neurosci. 2013; 33:11888–11898. [PubMed: 23864678]
35. Steinschneider M, et al. Representation of speech in human auditory cortex: Is it special? Hear Res. 2013
36. Crone NE, et al. Induced electrocorticographic gamma activity during auditory perception. Clin Neurophysiol. 2001; 112:565–582. [PubMed: 11275528]
37. Steinschneider M, et al. Spectrotemporal analysis of evoked and induced electroencephalographic responses in primary auditory cortex (A1) of the awake monkey. Cereb Cortex. 2008; 18:610–625. [PubMed: 17586604]
38. Ray S, Maunsell JH. Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. PLoS Biol. 2011; 9:e1000610. [PubMed: 21532743]
39. Mukamel R, et al. Coupling between neuronal firing, field potentials, and FMRI in human auditory cortex. Science. 2005; 309:951–954. [PubMed: 16081741]
40. Nourski KV, et al. Coding of repetitive transients by auditory cortex on posterolateral superior temporal gyrus in humans: an intracranial electrophysiology study. J Neurophysiol. 2013; 109:1283–1295. [PubMed: 23236002]
41. Nourski KV, et al. Spectral organization of the human lateral superior temporal gyrus revealed by intracranial recordings. Cereb Cortex. 2012
42. Mesgarani N, et al. Phonetic feature encoding in human superior temporal gyrus. Science. 2014
43. Scott SK, et al. Identification of a pathway for intelligible speech in the left temporal lobe. Brain. 2000; 123:2400–2406. [PubMed: 11099443]
44. Davis MH, Johnsrude IS. Hierarchical processing in spoken language comprehension. J Neurosci. 2003; 23:3423–3431. [PubMed: 12716950]
45. Zaehle T, et al. Segmental processing in the human auditory dorsal stream. Brain Res. 2008; 1220:179–190. [PubMed: 18096139]
46. Turkeltaub PE, Branch Coslett H. Localization of sublexical speech perception components. Brain Lang. 2010; 114:1–15. [PubMed: 20413149]
47. Rosen S, et al. Hemispheric asymmetries in speech perception: sense, nonsense and modulations. PloS One. 2011; 6:e24672. [PubMed: 21980349]
48. Takeichi H, et al. Comprehension of degraded speech sounds with m-sequence modulation: An fMRI study. Neuroimage. 2010; 49:2697–2706. [PubMed: 19878726]
49. Travis KE, et al. Independence of early speech processing from word meaning. Cereb Cortex. 2012
50. Obleser J, et al. Bilateral speech comprehension reflects differential sensitivity to spectral and temporal features. J Neurosci. 2008; 28:8116–8123. [PubMed: 18685036]
51. Obleser J, et al. Functional integration across brain regions improves speech perception under adverse listening conditions. J Neurosci. 2007; 27:2283–2289. [PubMed: 17329425]

52. Poeppel D. The analysis of speech in different temporal integration windows: cerebral lateralization as “asymmetric sampling in time”. Speech Commun. 2003; 41:245–255.
53. Steinschneider M, et al. Intracranial study of speech-elicited activity on the human posterolateral superior temporal gyrus. Cereb Cortex. 2011; 21:2332–2347. [PubMed: 21368087]
54. Chang EF, et al. Categorical speech representation in human superior temporal gyrus. Nat Neurosci. 2010; 13:1428–1432. [PubMed: 20890293]
55. Liebenthal E, et al. Specialization along the left superior temporal sulcus for auditory categorization. Cereb Cortex. 2010; 20:2958–2970. [PubMed: 20382643]
56. Tsunada J, et al. Representation of speech categories in the primate auditory cortex. J Neurophysiol. 2011; 105:2634–2646. [PubMed: 21346209]
57. Zhang L, et al. Cortical dynamics of acoustic and phonological processing in speech perception. PLoS One. 2011; 6:e20963. [PubMed: 21695133]
58. Joanisse MF, et al. Brain mechanisms implicated in the preattentive categorization of speech sounds revealed using fMRI and a short-interval habituation trial paradigm. Cereb Cortex. 2007; 17:2084–2093. [PubMed: 17138597]
59. Zion Golumbic EM, et al. Mechanisms Underlying Selective Neuronal Tracking of Attended Speech at a “Cocktail Party”. Neuron. 2013; 77:980–991. [PubMed: 23473326]
60. Ding N, Simon JZ. Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. J Neurophysiol. 2012; 107:78–89. [PubMed: 21975452]
61. Ding N, Simon JZ. Emergence of neural encoding of auditory objects while listening to competing speakers. Proc Natl Acad Sci. 2012; 109:11854–11859. [PubMed: 22753470]
62. Sabri M, et al. Attentional and linguistic interactions in speech perception. Neuroimage. 2008; 39:1444–1456. [PubMed: 17996463]
63. Kerlin JR, et al. Attentional gain control of ongoing cortical speech representations in a “cocktail party”. J Neurosci. 2010; 30:620–628. [PubMed: 20071526]
64. Mesgarani N, Chang EF. Selective cortical representation of attended speaker in multi-talker speech perception. Nature. 2012; 485:233–236. [PubMed: 22522927]
65. Peelle JE, et al. Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. Cereb Cortex. 2013; 23:1378–1387. [PubMed: 22610394]
66. Wild CJ, et al. Human auditory cortex is sensitive to the perceived clarity of speech. Neuroimage. 2012; 60:1490–1502. [PubMed: 22248574]
67. Davis MH, et al. Does semantic context benefit speech understanding through “top–down” processes?. Evidence from time-resolved sparse fMRI. J Cogn Neurosci. 2011; 23:3914–3932. [PubMed: 21745006]
68. Pallier C, et al. Cortical representation of the constituent structure of sentences. Proc Natl Acad Sci. 2011; 108:2522–2527. [PubMed: 21224415]
69. Sohoglu E, et al. Predictive top-down integration of prior knowledge during speech perception. J Neurosci. 2012; 32:8443–8453. [PubMed: 22723684]
70. Wolmetz M, et al. What does the right hemisphere know about phoneme categories? J Cogn Neurosci. 2011; 23:552–569. [PubMed: 20350179]
71. Leonard MK, et al. Language proficiency modulates the recruitment of non-classical language areas in bilinguals. PloS One. 2011; 6:e18240. [PubMed: 21455315]
72. Ylinen S, et al. Training the brain to weight speech cues differently: a study of Finnish second-language users of english. J Cogn Neurosci. 2010; 22:1319–1332. [PubMed: 19445609]
73. Holt LL, Lotto AJ. Speech perception within an auditory cognitive science framework. Curr Dir Psychol Sci. 2008; 17:42–46. [PubMed: 19060961]
74. Chait M, et al. Processing asymmetry of transitions between order and disorder in human auditory cortex. J Neurosci. 2007; 27:5207–5214. [PubMed: 17494707]
75. McClelland JL, Elman JL. The TRACE model of speech perception. Cognit Psychol. 1986; 18:1–86. [PubMed: 3753912]
76. McQueen JM, et al. Are there really interactive processes in speech perception? Trends Cogn Sci. 2006; 10:533. [PubMed: 17067845]

77. McClelland JL, et al. Are there interactive processes in speech perception? Trends Cogn Sci. 2006; 10:363–369. [PubMed: 16843037]

78. Elman JL. An alternative view of the mental lexicon. Trends Cogn Sci. 2004; 8:301–306. [PubMed: 15242689]

79. Elman JL. On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. Cogn Sci. 2009; 33:547–582. [PubMed: 19662108]

80. Creel SC, et al. Heeding the voice of experience: The role of talker variation in lexical access. Cognition. 2008; 106:633–664. [PubMed: 17507006]

81. Creel SC, Bregman MR. How talker identity relates to language processing. Lang Linguist Compass. 2011; 5:190–204.

82. Kraljic T, Samuel AG. Perceptual learning for speech: Is there a return to normal? Cognit Psychol. 2005; 51:141–178. [PubMed: 16095588]

83. Von Kriegstein K, et al. How the human brain recognizes speech in the context of changing speakers. J Neurosci. 2010; 30:629–638. [PubMed: 20071527]

84. Marslen-Wilson WD. Functional parallelism in spoken word-recognition. Cognition. 1987; 25:71–102. [PubMed: 3581730]

85. Norris D, et al. Merging information in speech recognition: Feedback is never necessary. Behav Brain Sci. 2000; 23:299–325. [PubMed: 11301575]

86. Gagnepain P, et al. Temporal predictive codes for spoken words in auditory cortex. Curr Biol. 2012; 22:615–621. [PubMed: 22425155]

87. Prabhakaran R, et al. An event-related fMRI investigation of phonological– lexical competition. Neuropsychologia. 2006; 44:2209–2221. [PubMed: 16842827]

88. Dahan D, et al. Time course of frequency effects in spoken-word recognition: Evidence from eye movements. Cognit Psychol. 2001; 42:317–367. [PubMed: 11368527]

89. Whiting CM, et al. Neural dynamics of inflectional and derivational processing in spoken word comprehension: laterality and automaticity. Front Hum Neurosci. 2013; 7

90. Lindsay S, et al. Acquiring novel words and their past tenses: Evidence from lexical effects on phonetic categorisation. J Mem Lang. 2012; 66:210–225.

91. Shtyrov Y, et al. Rapid cortical plasticity underlying novel word learning. J Neurosci. 2010; 30:16864–16867. [PubMed: 21159957]

92. Clark A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behav Brain Sci. 2012

93. Yildiz IB, et al. From Birdsong to Human Speech Recognition: Bayesian Inference on a Hierarchy of Nonlinear Dynamical Systems. PLoS Comput Biol. 2013; 9:e1003219. [PubMed: 24068902]

94. Norris D, McQueen JM. Shortlist B: a Bayesian model of continuous speech recognition. Psychol Rev. 2008; 115:357. [PubMed: 18426294]

95. Pouget, A., et al. Nat Neurosci. Advance Online Publication; 2013. Probabilistic brains: Knowns and unknowns.

96. Friston K. The free-energy principle: a unified brain theory? Nat Rev Neurosci. 2010; 11:127–138. [PubMed: 20068583]

97. Ulanovsky N, et al. Processing of low-probability sounds by cortical neurons. Nat Neurosci. 2003; 6:391–398. [PubMed: 12652303]

98. Yaron A, et al. Sensitivity to complex statistical regularities in rat auditory cortex. Neuron. 2012; 76:603–615. [PubMed: 23141071]

99. McQueen JM, Huettig F. Changing only the probability that spoken words will be distorted changes how they are recognized. J Acoust Soc Am. 2012; 131:509. [PubMed: 22280612]

Box 1

Outstanding Questions

- How is sublexical information integrated over time to allow lexical access to occur?
- How are dynamic, context-dependent representations encoded for abstract stimulus categories such as words?
- How is the structure of individual words and the mental lexicon in general encoded in local and network-level neural activity?
- How does the auditory ventral stream for speech allow these local networks to communicate with one another?
- What level of resolution (spatial, temporal, spectral) is necessary to be able to observe the dynamics of these networks?

Highlights

- Recent methodological advances reveal underlying information representations
- Spectrotemporal regions such as STG show strong context-dependent responses
- Contextual modulation occurs both *in situ* and through interactive connectivity between regions
- Context-dependent representations may give rise to abstract representations of words
- Multivariate and machine learning statistics will help uncover how sounds transforms into words

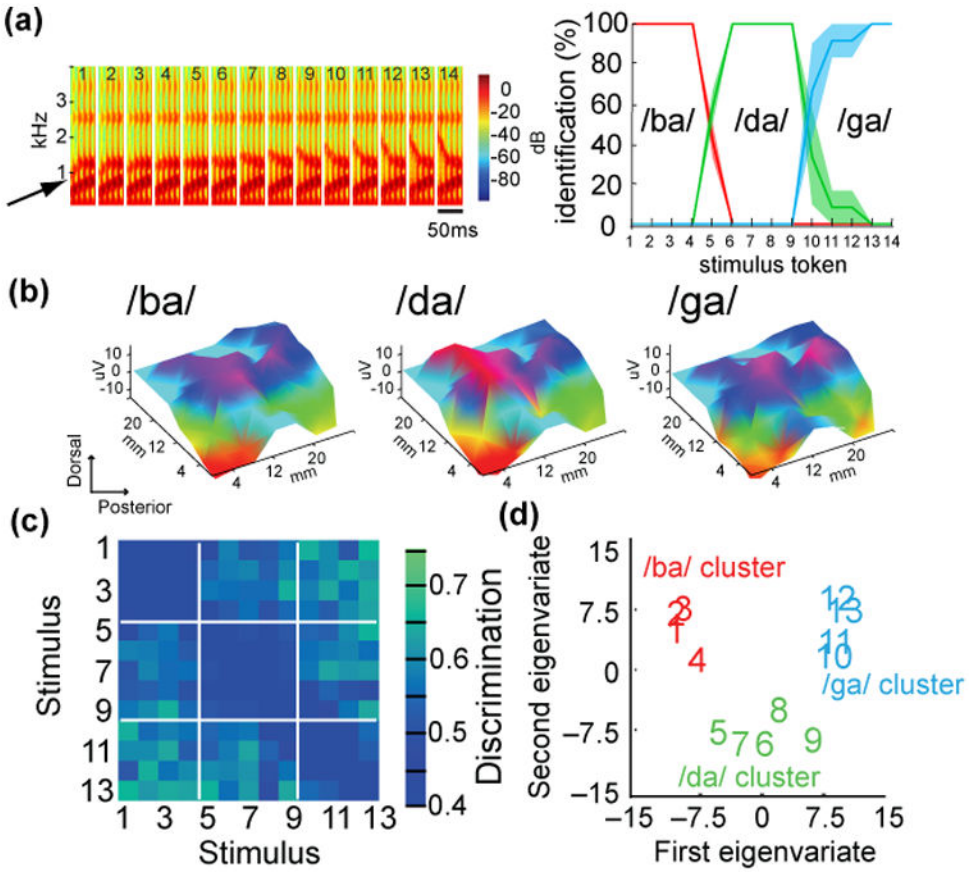

Figure 1. Speech Representation in Human STG

(**a**) Speech sounds synthesized along an acoustic continuum (stimuli 1-14, small increases in the second formant [arrow]) are perceived categorically, and not linearly with stimulus change in behavioral testing (identification). (**b**) Spatial topography of evoked potentials recorded directly from the cortical surface of STG using ECoG for each sound class is highly distributed and complex. (**c**) A neural confusion matrix plots the pair-wise dissimilarity of neural patterns using a multivariate classifier. (**d**) Multidimensional scaling shows that these response patterns are organized in discrete clustered categories along both acoustic and perceptual sensitivities. Adapted from [54].

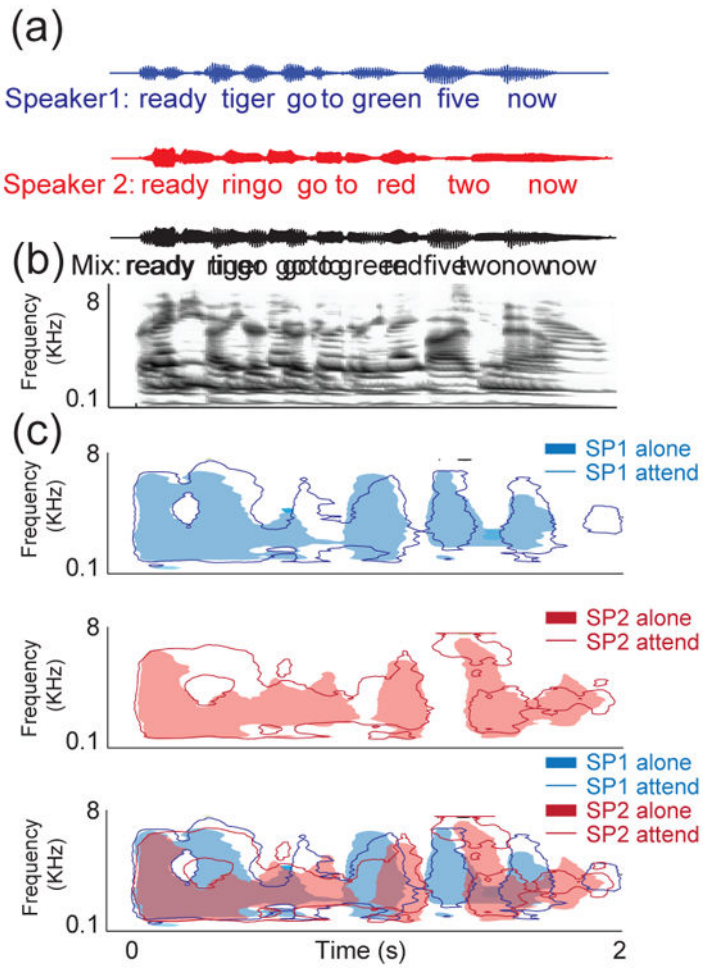

Figure 2. Attention strongly modulates STG representations of spectral and temporal speech content

(**a**) During high-density ECoG recording, participants listened to two speech streams either alone or simultaneously and were cued to focus on a particular call sign (‘tiger’ or ‘ringo’) and to report the color/number combination (e.g., ‘green five’) associated with that speaker. (**b**) The acoustic spectrogram of the mixed speech streams shows highly overlapping energy distributions across time. (**c**) Neural population-based reconstruction of the spectrograms for speaker 1 (blue) and speaker 2 (red), when participants heard each speaker alone (shaded area) or in the mixed condition (outline). Results demonstrate that in the mixed condition, attention to a particular speaker results in a spectrotemporal representation in STG as if that speaker were heard alone. Adapted from [64].