27 The Cortical Processing of Speech Sounds in the Temporal Lobe

MATTHIAS J. SJERPS AND EDWARD F. CHANG

Speech perception is a complex process that transforms the continuous stream of clicks, hisses, and vibrations that make up speech sounds into meaningful linguistic representations. This process unfolds at a remarkable speed, as naturally spoken speech typically contains around five syllables per second (Ding et al., 2017; Miller, Grosjean, & Lomanto, 1984). The cortical processing of spoken language involves a network of regions in the temporal, parietal, and frontal lobes in which the specific involvement of regions may vary depending on the task demands or goals of the listener (Hickok & Poeppel, 2004). It is widely recognized, however, that the posterior portions of the superior temporal gyrus (STG) and superior temporal sulcus (STS; see figure 27.1) play a pivotal role in early processing of speech sounds (e.g., Belin, Zatorre, Lafaille, Ahad, & Pike, 2000; Hickok & Poeppel, 2004; 2007; 2015; Rauschecker & Scott, 2009).

Indeed, local disruption of neural activity with focal electrical stimulation of the STG leads to sensory errors and/or phonemic errors (see, e.g., Boatman, 2004; Boatman, Hall, Goldstein, Lesser, & Gordon, 1997; Leonard, Cai, Babiak, Ren, & Chang, 2016; Quigg & Fountain, 1999; Roux et al., 2015). Furthermore, damage to the posterior part of the superior temporal lobe (STL, i.e., STS and STG combined) has been repeatedly associated with speech-perception deficits (Buchman, Garron, Trost-Cardamone, Wichter, & Schwartz, 1986; Buchsbaum, Baldo, et al., 2011; Rogalsky et al., 2015; Wilson et al., 2015). The STL is thus thought to play a critical role in the transformation of acoustic information into phonetic and prelexical representations.

One of the major questions that drives current research on early speech sound processing is the actual nature of speech representations in the STL (the STL is defined here as the *lateral parabelt auditory cortex*, including parts of Brodmann areas 41, 42, and 22; Hackett, 2011). Does this region mostly represent acoustic features (i.e., a responsiveness to energy at specific frequencies or perhaps to sounds for which the dominant frequencies change over time)? Or does this region mostly represent linguistic units such as phonemes and/or syllables? And how does the brain arrive at phonetic representations (or another form of prelexical representation) that allows for lexical access independently of how or by whom the speech sound was produced (i.e., abstract representations)? These questions are of particular importance for understanding the processing of spoken language as a whole because the representations in the STL constitute a critical link in processing, receiving direct input from primary input areas as well as interacting with associative auditory areas with higher-level representations (DeWitt & Rauschecker, 2012; Hickok & Poeppel, 2004; 2007; Lerner, Honey, Silbert, & Hasson, 2011; Rauschecker & Scott, 2009; Scott & Johnsrude, 2003; Steinschneider et al., 2011).

The current chapter provides a review of several concepts and recent findings that have informed our understanding of the role of the STL in early speech sound processing. Because this field of research is broad and highly active, we will focus our discussion by especially highlighting research that addresses the nature of speech sound representations in the STL. This approach, focusing on representations as distributed patterns of activation, has been especially informed by noninvasive imaging methods such as functional MRI (fMRI) and magnetoencephalography. In addition, invasive methods such as electrocorticography (ECoG) recordings, the main method used in our work, have also contributed meaningfully to research.

In section 1, we will briefly discuss speech sound processing in the primary auditory cortex (PAC), the main source of input for the STL with regard to acoustic information (chapter 35 by Formisano in this volume provides a more in-depth description of the language-relevant dominant properties of PAC organization). Subsequent sections will discuss the representation of speech sounds as acoustic phonetic features, the emergence of categorical/abstract representations, and how these representations are influenced by visual cues and other "contextual information" such as phoneme sequencing and lexical-semantic representations.

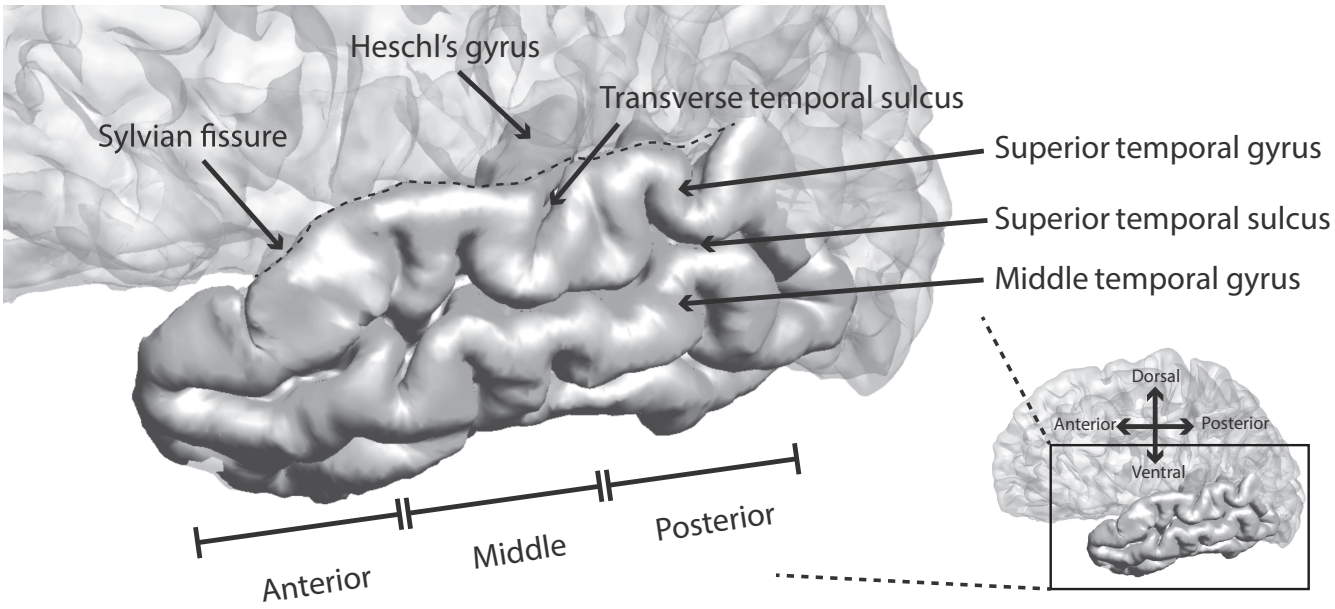

FIGURE 27.1 Anatomical landmarks of the temporal lobe on and around the regions involved in early speech sound processing. Regions outside the temporal lobe are displayed as transparent, allowing for the visualization of Heschl's gyrus, which is located inside the Sylvian fissure.

The research discussed here stresses the role of the STL as a highly versatile auditory association cortex that displays sensitivity to acoustic patterns at multiple levels of granularity (i.e., from acoustic features to phoneme sequences) but is also robustly influenced by concurrent visual information and lexical-semantic context. Moreover, abstraction, the property that allows for categorical and context-invariant mapping, seems to be an emergent but distributed property of processing in the STL.

1. From Acoustics to Prelexical Abstraction

1.1. REPRESENTATIONS IN PAC AND CLOSELY SURROUNDING REGIONS It is important to understand the functional pathway through which key speech auditory regions receive most of their input. The ascending auditory pathway projects to PAC through afferent input from the medial geniculate complex, which is part of the thalamus. Processing at these subcortical levels is subject to important transformations and is already influenced by linguistic and musical exposure (Bidelman, Gandour, & Krishnan, 2011; Krishnan, Gandour, & Bidelman, 2012; Weiss & Bidelman, 2015). Important for the current review, however, is that the representations also largely transmit the time-frequency properties of the sound waveform (Shamma & Lorenzi, 2013; Weiss & Bidelman; Young, 2008). This information is transmitted in a partly nonlinear fashion especially along the frequency axis (i.e., frequency resolution follows the so-called *mel scale*, which is a loglike scale, overrepresenting lower frequencies). PAC in humans is mostly confined to the bilateral transverse temporal gyrus (Heschl's gyrus; see figure 27.1). Its organization is traditionally characterized as having neuronal populations that display very fine frequency tuning, with at least two mirror-symmetric tonotopic frequency gradients (Bauman, Petkov, & Griffiths, 2013; Bitterman, Mukamel, Malach, Fried, & Nelken, 2008; Humphries, Liebenthal, & Binder, 2010; Moerel, De Martino, & Formisano, 2012; Saenz & Langers, 2014). As a result, sound representations in PAC allow for the transmission of acoustic cues that are critical for the perception of speech such as formants, formant transitions and amplitude modulations (e.g., Young, 2008). In addition to tonotopic representations, however, studies in animal models have also demonstrated more complex properties in PAC, such as tuning for temporal and spectral modulations rather than specific frequency representations per se (e.g., Schreiner, Froemke, & Atencio, 2011).

Secondary auditory areas such as the planum temporale (PT; located posterior to Heschl's gyrus) and the lateral STG largely depend on inputs from PAC (Hackett, 2011). This flow of information is facilitated by (bidirectional) functional connections between parts of PAC and its closely surrounding region, as well as direct projections from auditory thalamus. This has been demonstrated, for example, by activity in the laterally exposed STG that is observed at very short latencies after electrical stimulation in the PAC (Brugge,

Volkov, Garell, Reale, & Howard, 2003). Functionally, the regions immediately surrounding PAC, both within the Sylvian fissure and on the lateral part of the STG, display both tuning to narrow frequency ranges and sensitivity to increasingly complex spectrotemporal information. To exemplify, parts of the lateral STL display fairly low-level acoustic response properties. For example, Nourski et al. (2012) observed strong responses to simple pure tone stimuli in a restricted region surrounding the laterally exposed part of the transverse temporal sulcus (see figure 27.1), which runs parallel along the posterior side of Heschl's gyrus. The observation that this region inherits some amount of tonotopic organization is further supported by a body of research (Humphries et al. 2010; Moerel et al., 2012; Nourski et al.; Striem-Amit, Hertz, & Amedi, 2011; Talavage et al., 2004). In addition to these tonotopic characteristics, however, the regions surrounding PAC also display spectral preferences that become more complex, with more widespread tuning at octave intervals and harmonically related frequency intervals (Moerel et al., 2013; Ohl & Scheich, 1997). The characteristics of auditory representations in the PAC, on the one hand, and the lateral STG, on the other, are, thus, partly overlapping. However, the dominant representation in PAC is one of tonotopic distributions, whereas the more dominant forms of representation outside of PAC are of a more complex spectrotemporal nature (Hullett, Hamilton, Mesgarani, Schreiner, & Chang, 2016).

1.2. STL: From Spectrotemporal Receptive Fields to Speech Sound Representations The predominant preference for more complex spectrotemporal patterns shows that large portions of the medial and posterior STL transform relatively basic acoustic properties such as pure tones and sweeps into combined representations (Hickok & Poeppel, 2015; Peelle, Johnsrude, & Davis, 2010; Rauschecker & Scott, 2009). A popular approach in research on the involvement of the STL in speech sound processing has been to compare levels of activation to clear speech sounds with degraded speech sounds or nonspeech sounds (Belin et al., 2000; Binder et al., 2000; Davis & Johnsrude, 2003; Liebenthal, Binder, Spitzer, Possing, & Medler, 2005; Obleser, Eisner, & Kotz, 2008; Obleser, Zimmermann, Van Meter, & Rauschecker, 2007; Rosen, Wise, Chadha, Conway, & Scott, 2011; Scott, Blank, Rosen, & Wise, 2000; Takeichi, et al., 2010; Turkeltaub & Coslett, 2010; Zaehle, Geiser, Alter, Jancke, & Meyer, 2008). The general view that arises from this body of research is a hierarchy of responsiveness to increasingly speech-specific signal characteristics as activation spreads to more anterior and ventral regions (see Obleser & Eisner, 2009; Price, 2012, for general review, and Liebenthal, Desai, Humphries, Sabri, & Desai, 2014; Turkeltaub & Coslett; DeWitt & Rauschecker, 2012, for fMRI- and positron-emission tomography [PET]-based Activation Likelihood Estimation [ALE] meta-analyses). Turkeltaub and Coslett, for example, performed two ALE meta-analyses on studies that compared sublexical speech versus nonspeech signals. In a first analysis, they compared listening to speech with listening to relatively simple nonspeech signals (i.e., listening to isolated vowels or consonant-vowel sequences, compared to a variety of nonspeech signals such as pure tones, band-passed noise, music). Their analysis revealed large clusters in the bilateral STG extending into the STS that respond more strongly to speech, suggesting that these regions are involved in the processing of more complex acoustic properties of speech sounds. A subsequent analysis compared sublexical speech sounds with nonspeech stimuli that were closely matched to speech in terms of their spectrotemporal properties. This second analysis revealed a much smaller region of speech specificity, mostly located in the left STS but extending somewhat into the ventral bank of the left STG (see Desai, Liebenthal, Waldron, & Binder, 2008; DeWitt & Rauschecker; Jäncke, Wüstenberg, Scheich, & Heinze, 2002; Liebenthal et al., 2005, 2010, 2014; and Price, 2012, for very similar results), suggesting that only this more ventral portion of the left STL was involved in speech processing per se.

To further exemplify, a recent fMRI study demonstrated that cortical regions in the posteromedial STS reveal a preference for speechlike sounds over sounds that are matched by a number of spectrotemporal characteristics (Overath, McDermott, Zarate, & Poeppel, 2015). Overath et al. divided natural speech signals into short sound segments that were then randomly reshuffled but which adhered to local speechlike statistics. They found that the STS is increasingly sensitive to speechlike sounds when those sounds consisted of increasingly longer segments (ranging from 30 ms to ~1 s). The STS thus seemed to prefer speech sound cues, but increasingly so if they appear in a longer sequence. That is, STL prefers sequences that conform to the way that speech input is typically heard in more everyday listening situations. This was despite the fact that all stimuli (both long and short sequences) were equally meaningless to the listeners. Similarly, Canolty et al. (2007) compared ECoG responses in patients listening to both clear speech stimuli and to nonphonemic speechlike sounds (speech sounds for which specific formant details were removed, rendering complex but unintelligible sounds). Cortical responses in the STG were larger for speech than for the nonphonemic

sounds. Interestingly, the differences in activation arose in a serial manner, with differences arising in posterior STG at ~120 ms, in the mid-STG at ~193 ms, and in the mid-STS at ~268 ms. This suggests a similar gradient across the STL (posterior to anterior/medial) of both the temporal progression of information and of increasing speech specificity.

Despite the demonstration of a gradient across the STL displaying selectivity for increasingly speechlike sounds, a fundamental question that remains is what properties of speech sounds are reflected by this neural activity. The dominant representations could be, among others, low-level spectrotemporal parameters, acoustic-phonetic features, or phonemes. In addressing this question, however, it is important to note that receptive fields and response properties of neurons are tightly matched to the statistics of natural input. This property has been demonstrated in both animal- and human-based research (David, Vinje, & Gallant, 2004; Hsu, Woolley, Fremouw, & Theunissen, 2004; Rieke, Bodnar, & Bialek, 1995; Talebi & Baker, 2012; Theunissen, Sen, & Doupe, 2000; Young, 2008). To obtain a detailed picture of how the human STL integrates auditory features into some form of higher-level representations, it is thus important to rely on natural or ecologically valid stimuli.

A powerful approach that has been developed in animal research to investigate auditory representations is *spectrotemporal receptive field* (*STRF*) *estimation*. STRFs are computed by first recording activity from a neural site in response to acoustic input. Then, through a procedure such as reverse correlation (e.g., Klein, Depireux, Simon, & Shamma, 2000; Theunissen et al., 2000; see, e.g., Hullett et al., 2016, who describe another estimation method called maximally informative dimension analysis), properties of the acoustic spectrogram are established that are found to either excite or inhibit neural activity (i.e., specific frequency bands at particular time lags). This method is distinguished from other measures by its broader descriptive power for encompassing both dynamics and spectral selectivity and for not requiring much prior knowledge such as frequency tuning or threshold. Moreover, this method has some advantages for analyzing responses to using natural stimuli.

Recently, STRF models have been used to describe the encoding of specific stimulus features in human STG for both speech and nonspeech input (e.g., Hullett et al., 2016; Mesgarani, Cheung, Johnson, & Chang, 2014; see figure 27.2A). Hullett et al., for example, observed that the human STG displays an anterior to posterior organization of different types of spectrotemporal tuning. Using ECoG, they demonstrated that sites toward the posterior STG were found to be increasingly tuned for speech sounds that have relatively constant energy across the frequency range (low spectral modulation; see figure 27.2B) but which are temporally changing at a fast rate. In contrast, sites toward the anterior STG were found to be increasingly tuned for speech sounds that show a high degree of spectral variation across the frequency range (high spectral modulation) and which are temporally changing at a slow rate (see Santoro et al., 2014 for corroborating findings from fMRI). This sensitivity to two types of modulations seems to be an important property of processing in the auditory processing stream that has been observed in animal models as well (e.g., Woolley, Fremouw, Hsu, & Theunissen, 2005; Nagel & Doupe, 2008). Of the regions in STG that Hullett et al., found to be responsive to basic auditory properties (these were mostly confined to our definition of posterior and medial STG in figure 27.1), on average about 23% of the variance in neural activity at specific sites could be explained by the patterns described in STRFs (Hullett et al., 2016; Pasley et al., 2012). This shows that a significant component of the information represented in the posterior and medial STG is closely related to acoustic features rather than higher-level (e.g., lexical or semantic) ones.

Despite this sensitivity to low-level acoustic properties of sound, it is clear that processing in the STG is strongly related to the behavioral relevance of the input. For example, when pure-tone selectivity is observed in the STG, it is generally confined to the low-frequency portions of the tonotopic map, which are the frequencies that are predominant in human voice sounds and speech, in particular (Moerel et al., 2012). Furthermore, auditory-based predictions of activity in the STG such as those already described (i.e., figure 27.2A) perform best for the ranges of spectrotemporal modulations most critical to speech intelligibility (Chi, Gao, Guyton, Ru, & Shamma, 1999; Elliott & Theunissen, 2009; Pasley et al., 2012). As another example, dynamic ripple stimuli (combinations of fluctuating sine tones) contain the same basic spectrotemporal modulations that are reflected in the speech, yet, because they occur in nonbehaviorally relevant auditory objects, they don't excite regions in the STG the way that speech does (Hullett et al., 2016). These findings suggest that the auditory stimulus preferences for regions on the lateral STL are closely related to the general acoustic properties of speech sounds.

In an investigation of how tuning to spectrotemporal properties in the STG is related to the processing of speech sounds, Mesgarani et al. (2014) presented listeners with a large number of naturally spoken sentences. Their participants were patients undergoing surgical monitoring for medically refractory epilepsy and were

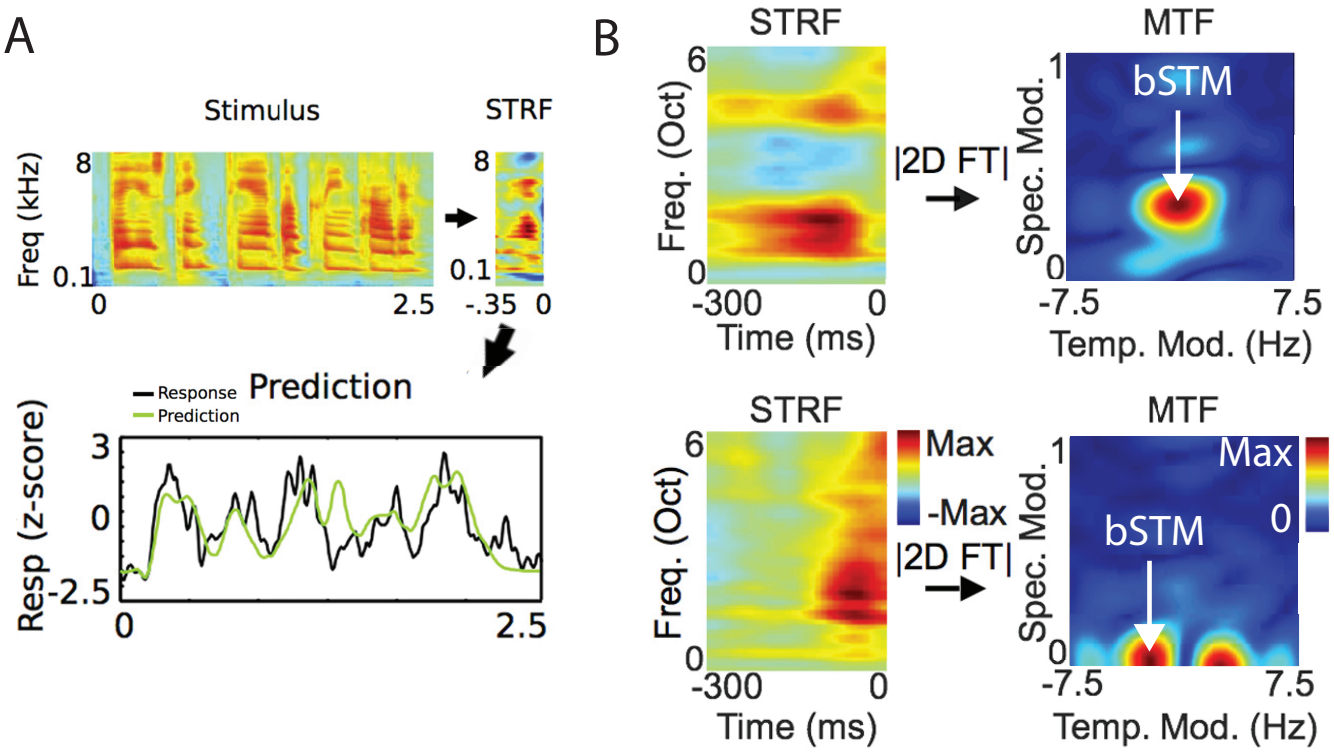

FIGURE 27.2 (A) STRF mapping. The spectrogram of a spoken sentence, an STRF, and the predicted and measured response for the sentence. Predicted responses are obtained by convolving the stimulus spectrogram with the STRF and are proportional to the similarity between the spectrotemporal content in the stimulus and the receptive field. (B) Computation of the modulation transfer function (MTF). The MTF is derived as the magnitude of the two-dimensional Fourier Transform (2D FT) of the STRF. It characterizes spectrotemporal modulation tuning for each site. Like the "best frequency" of a frequency tuning curve, the peak of the MTF defines the "best spectrotemporal modulation" (bSTM). For the site with the STRF shown on the top row, the MTF indicates that high spectral modulations and low temporal modulations drive activity at that site (i.e., prototypical of anterior sites). In contrast, the site on the bottom row has a bSTM at high temporal modulations and low spectral modulations, indicating that the site is driven by changes in temporal and not spectral energy (i.e., prototypical of posterior sites; figure reproduced from Hullet et al., 2016).

implanted with ECoG grids directly over the perisylvian cortex for clinical purposes (figure 27.3A). They listened to speech samples from the TIMIT corpus (figure 27.3B) covering a wide range of different sentences and speakers. Focal patterns of activity on the cortex of individual patients revealed selectivity for phonetic features, but not for individual phonemes. For example, one electrode (e1; figure 27.3C) displayed a reliable response to plosive phonemes /d/, /b/, /g/, /p/, /k/, and /t/. E2 displayed a reliable response to sibilant fricatives: /ʃ/, /z/, and /s/. E3 displayed a reliable response to low-back vowels (e.g., /a/ and /aʊ/). E4 displayed a reliable response to high-front vowels and glides (/i/ and /j/). And e5 was selective for nasals (/n/, /m/, and /ŋ/).

Mesgarani et al. then used unsupervised hierarchical clustering analyses on an electrode-specific measure of phoneme selectivity to find groups of electrodes with similar response characteristics. The clusters of electrodes that emerged from this procedure revealed very specific and speech-relevant STRFs. For example, a first cluster revealed an STRF (top row of figure 27.3D) displaying tuning for broadband excitation, a spectral property that is indicative of plosives (first panel in bottom row of figure 27.3D). A second cluster revealed an STRF that was tuned to a high-frequency component that is a defining feature of sibilant fricatives. Further clusters displayed STRFs indicative of other classes of speech sounds such as tuning for characteristic formants that define low-back, low-front, and high-front vowels, and a cluster revealed tuning for low acoustic frequencies, conforming to a general property of nasal speech sounds. These observations suggest that, at least on the lateral surface of the STG, speech sounds are most dominantly encoded as acoustic-phonetic features such as manner of articulation (e.g., /ta/ vs. /sa/) and voicing (e.g., /ba/ vs. /pa/). Moreover, features with acoustically very distinct cues such as manner of articulation were very strong determinants of selectivity, while acoustically weaker distinctions such as place of

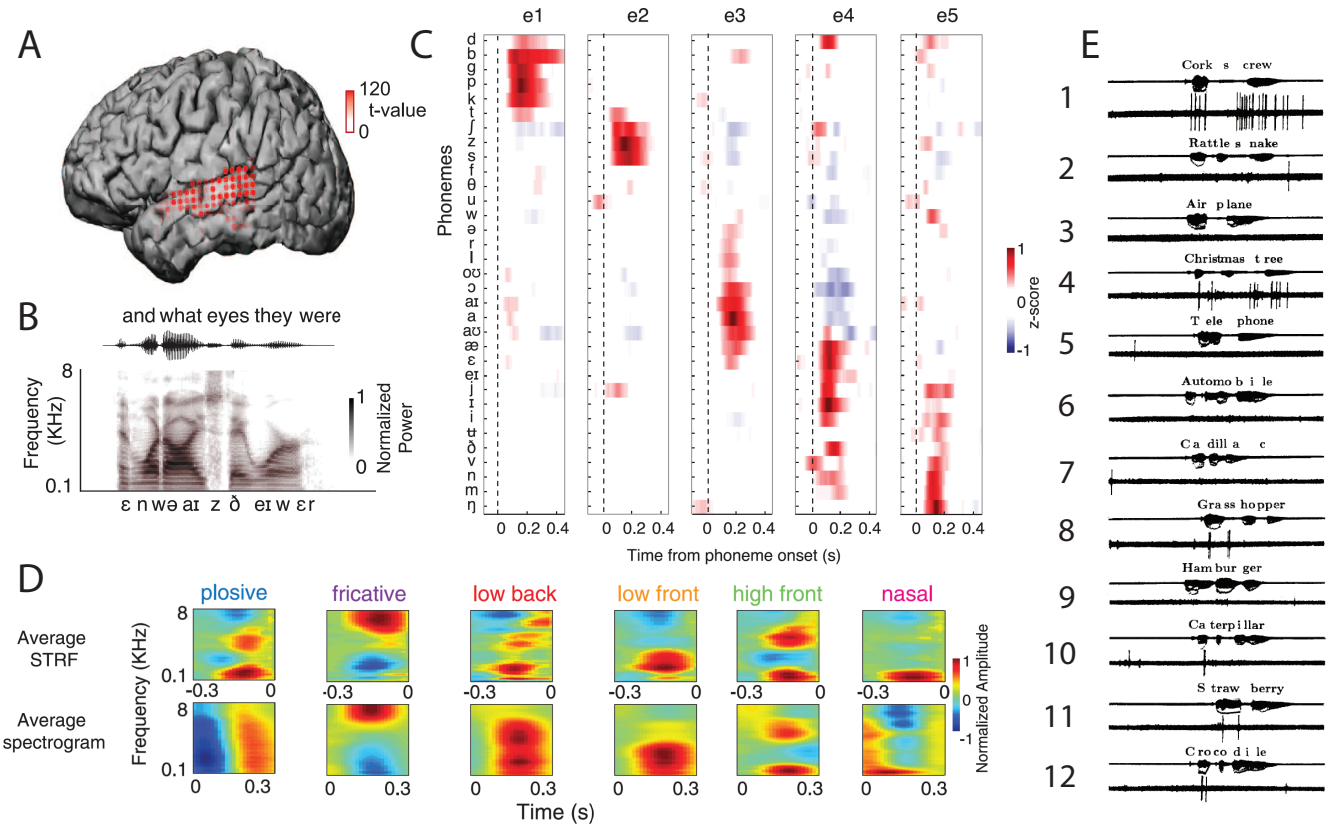

FIGURE 27.3 Human STG cortical selectivity to speech sounds. (A) MRI surface reconstruction of one participant's cerebrum. Auditory-responsive electrodes (red) are plotted with opacity signifying the *t*-test value when comparing responses to silence and speech. (B) Example sentence and its acoustic waveform, spectrogram, and phonetic transcription. (C) Average responses at five example electrodes to all English phonemes. (D, *top row*) Weighted average STRFs of main electrode clusters; (*bottom row*) average acoustic spectrograms for phonemes associated with each population cluster. (E) Oscillograms and corresponding responses of a single neuron in the right STG during passive listening to a list of 12 words. Panels A–D reproduced from Mesgarani et al. (2014); panel E reproduced from Creutzfeldt et al. (1989).

articulation (e.g., /pa/ vs. /ta/) were much less discriminable. It is important to note that single neuron recordings also failed to show selectivity to single phonemes (Chan et al., 2014; Creutzfeldt, Ojeman, & Lettich, 1989), suggesting that this phonetic feature organization is not simply a confound of the meso-scale ECoG sampling of thousands of neurons (see figure 27.3E, which displays activation after obstruent consonants such as /k/ and /t/, especially when these appear in obstruent clusters, e.g., *st* or *sk*). The data presented by Mesgarani et al. reveal that feature-level selectivity is a dominant property for STG processing of speech sounds (see also Arsenault & Buchsbaum, 2015; Steinschneider et al., 2011). This demonstrates that the organization of speech sounds in the STL is tightly linked to acoustic/phonetic cues and not to discrete phonemic or even articulatory ones.

The findings discussed in this section demonstrate that processing across the STG is dedicated to complex spectrotemporal events. The spectral and temporal modulation ranges that are covered by the STG closely align to the spectrotemporal properties that are important for the processing of natural speech sounds. This observation helps to better understand why processing across the STL becomes increasingly speech-specific: only those stimuli that involve longer sequences that adhere to speechlike statistics evoke activity in the more ventral and anterior regions of the STL. Moreover, it is clear that the dominant form of speech sound representations is one that reflects speech sound features rather than specific phonemes per se. That is, representations in the STL are closely related to the acoustic properties of the natural classes of speech sounds, such as fricatives, vowels, plosives.

1.3. Abstract Representations and Contextual Invariance

One of the main challenges of speech perception is that there is no one-to-one mapping between sounds and words or even between sounds and some form of prelexical representation such as

features, phonemes, or syllables. This lack of a one-to-one mapping has many origins, such as differences in listening conditions, speakers with differently sized vocal tracts and speaker accents. To uniquely select lexicosemantic representations, however, there must exist a form of abstraction in the neural hierarchy that allows listeners to rely on some form of contextually invariant code. One way in which such abstraction has been demonstrated behaviorally is through categorical perception experiments where tokens on an acoustically linear continuum (e.g., a linear continuum spanning from the speech sound *ba* to *pa*, a distinction primarily cued by Voice Onset Time) show a nonlinear, sigmoidal, pattern of categorization by listeners (Harnad, 1987; Liberman, Harris, Kinney, & Lane, 1961). Abstraction in this case, then, requires some form of "warping" of neural space in the sense that ambiguous tokens (which can be considered as nonoptimal instances of prototypical speech sound representations) are assimilated to one's native category structure (Kuhl, 1993; Kuhl et al., 2008). In the following, we will describe how such forms of abstraction have indeed been observed to be an emergent property of processing across the STL.

Abstraction is a property that is useful for all behaviorally relevant sound classes, not only speech sounds. In fact, in some cases, meaningless nonspeech sounds may be an interesting alternative to investigate the acquisition of abstract representations because their representation has not yet been adjusted by lifelong exposure. Ley et al. (2012) presented participants with a set of complex (nonspeech) sound categories, both before and after training participants to distinguish them into two sound classes (based on pitch). An analysis of the distributed activation patterns during listening both before and after a training session demonstrated that sound category could only be decoded from the distributed blood oxygenation level-dependent (BOLD) response patterns after training. This suggests that some form of functional cortical reorganization had taken place as a result of learning that allowed for the enhanced processing of those aspects of the stimuli that are relevant for behavioral classification. Furthermore, the patterns of neural activity across the testing continuum after training revealed a nonlinear (i.e., sigmoidal) pattern of similarity between items on the pitch continuum thereby showing tight correspondence to participants' categorical behavior (Harnad, 1987). Interestingly, the regions that most strongly contributed to decoding involved a wide region spanning Heschl's gyrus, the PT, and portions of both posteromedial STG and STS. This suggests that this rather distributed region contributed to the emergence of more abstract representations of sound classes.

The distributed nature of the emergence of categorical representations has also been observed for the encoding of speech sounds, in particular for vowel and speaker representations. Formisano, De Martino, Bonte, and Goebel (2008) presented participants with speech sounds consisting of three vowels recorded from three different speakers. Statistical models were then trained to label the associated multivoxel BOLD responses according to vowel identity or according to speaker identity. The models could accurately classify novel trials (i.e., trials that were not in the training data) for both features. However, these features relied on separate, distributed regions in the STL. Importantly, the representations of these features were independent of specific acoustics and generalized across speakers and vowels. Correct classification of vowel identity was based on regions that included large portions of bilateral posteromedial STG and STS, along with a left-sided part of the PT. Speaker identity was mostly decoded based on a portion of the right STS. Similarly, Chang et al. (2010) presented ECoG patients with sounds on a place-of-articulation continuum spanning from /ba/ to /da/ to /ga/, with intermediate steps between the unambiguous speech sounds as well (a distinction that is cued by the F2 onset trajectory; cf. Liberman, Harris, Hoffman, & Griffith, 1957). It was observed that distinct spatiotemporal patterns of activity occurred across the posterior STG when patients listened to the different speech sounds. Importantly, however, they observed a nonlinearity in the neural encoding of the acoustically linearly spaced sound continuum. Tokens that were close to the unambiguous speech sounds (e.g., clear /da/) gave rise to patterns of activity that were very similar to those of the clear ones themselves. These signatures of abstraction arose within a latency range as short as 110–150 ms (see Tsunada, Lee, & Cohen, 2011, for closely related findings in macaque auditory belt regions; and Okada et al., 2010; and Altmann et al., 2014, for speech sound abstraction, especially in the STS).

Finally, converging evidence for the distributed effects of categorical representations have come from fMRI adaptation, or repetition suppression, paradigms (Grill-Spector, Henson & Martin, 2006; Grill-Spector & Malach, 2001). Adaptation can be used to reveal cortical regions that are sensitive to a particular characteristic that remains constant across a set of repeated items but differs in an oddball stimulus. For example, using speech sound stimuli on a continuum from /ga/ to /da/, Joanisse, Zevin, and McCandliss (2007) compared adaptation responses to pairs of stimuli that lay on either the same side of the phoneme-category boundary or stimuli that straddled the category boundary.

They found adaptation effects that were specific to phonemic (as opposed to acoustic) content in left mid-STG, suggesting abstraction in relatively early speech sound processing (but see Chevillet, Jiang, Rauschecker, & Riesenhuber, 2013, for conflicting results). Furthermore, in a similar approach, Leaver and Rauschecker (2010) showed adaptation effects for phonetic categories in left mid-STS (see also Humphries, Sabri, Lewis, & Liebenthal, 2014, for similar results).

As demonstrated, signatures of categorical representations seemed to be widely distributed (e.g., Chang et al., 2010; Formisano et al., 2008), including regions such as PT and posterior STG, regions typically thought to perform more basic acoustic integration. This observation demonstrates that variable levels of representation show a fair amount of overlap. In addition, one may speculate that abstraction can be a processing characteristic that is not restricted to phonemic and postphonemic levels of processing (Mitterer, Scharenborg, & McQueen, 2013), but may occur at the level of speech sound features as well. Abstraction at prelexical levels of representation is important because it allows listeners to understand the meaning of speech spoken by different speakers, despite their differences in pronunciation. Furthermore, when a listener comes across a speaker with an idiosyncratic pronunciation of, say, the phoneme /s/ (perhaps a speaker who lisps), the abstract nature of prelexical units allow the listener to apply their knowledge of this speaker's /s/ also to words that they have not heard this speaker produce before (e.g., McQueen, Cutler, & Norris, 2006; Sjerps & McQueen, 2010). An important addition to these observations, however, is the fact that neural abstraction is often not complete (e.g., Chang et al., 2010). That is, although these representations enhance between-category distinctiveness, "behaviorally irrelevant" auditory detail is not completely removed from the neural representations: within class items do not become neurally identical, just more similar. This property is important because it allows listeners to have access to fine phonetic detail as well when necessary, which can be extremely useful in reanalysis in ambiguous sentences and in the integration of speech cues over longer stretches of time (e.g., Andruski, Blumstein, & Burton, 1994; McMurray, Tanenhaus, Aslin, & Spivey, 2003; McMurray, Tanenhaus, & Aslin, 2009).

2. Integrating Speech Sound Representations with Context

Bottom-up processing of auditory information in speech perception is heavily influenced by various forms of context (e.g., Leonard & Chang, 2014). Such context includes factors such as concurrent visual information, phoneme sequence probabilities, and lexicosemantic properties. In the following, we will discuss recent findings that demonstrate how processing in the posteromedial STG and STS is affected by these factors during speech perception. In a final section, we will briefly describe the main regions in the human cortex that receive information from the STL as it forms part of the larger perisylvian language network.

2.1. Multimodal Integration

Despite the focus on auditory processing of speech so far in this review, a typical setting for speech perception is a situation where the listener achieves comprehension through the incorporation of both auditory and visual signals. In fact, in people with extreme hearing loss, lip-reading alone can sometimes provide them with enough information to understand speech (Bernstein, Tucker, & Demorest, 2000; Suh, Lee, Kim, Chung, & Oh, 2009). Typically, however, Audiovisual (AV) perception will involve the integration of convergent signals, where listeners can utilize the relative importance of both signals by weighting their contribution depending on how informative they are (e.g., Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007). In the following, we will describe how the visual and auditory flows of information interact in the STL and how their relative contributions can be adjusted to situation-specific demands.

Visual information has been shown to influence auditory processing throughout most of its cortical processing. For example, fMRI-based research has shown that both primary and secondary auditory cortices can be activated by visual (lip-read) information alone (Paulesu et al., 2003; Calvert et al., 1997). In addition, interactions between auditory and visual information occur across a number of these regions (Okada, Venezia, Matchin, Saberi, & Hickock, 2013; Skipper, van Wassenhove, Nusbaum, & Small, 2007; Miller & D'Esposito, 2005). In specific situations, auditory and visual information may conflict, as demonstrated in the well-known McGurk effect (e.g., auditory /ba/ presented with visual /ga/ often merge to a /da/ percept; McGurk & MacDonald, 1976). In an ALE meta-analysis of fMRI research on AV speech perception, Erickson, Heeg, Rauschecker, & Turkeltaub (2014) demonstrated the involvement of a large set of regions in resolving conflicting AV information. In addition, however, there was a (much smaller) set of regions involved when auditory and visual signals were in agreement. In the temporal lobe, Erickson et al. found that the posteromedial STL, especially the STS, was involved in both "validating" and conflicting AV situations, with a larger, more ventrally/posteriorly located region involved in

situations of conflict. To further investigate the role of the STS in AV integration, Beauchamp, Nath, and Pasalar (2010) presented participants with McGurk stimuli and, on some trials, applied Transcranial Magnetic Stimulation (TMS) during presentation. Without TMS, most trials elicited the well-known McGurk fusion. In contrast, when TMS was applied to the STS, the proportion of reported fusions was significantly lower. In a further study, Nath and Beauchamp (2011) investigated functional connectivity between the AV integration area on the posterior STS (extending onto the lateral STG), with two regions: (i) a region on and around PAC and(ii) a primarily visual region. They then manipulated the reliability of the visual or the auditory signal (a clear auditory signal with a blurred visual signal, or a clear visual signal with an auditory signal in noise) and observed that when the auditory signal was more reliable, functional connectivity increased between the auditory region and the STS. When the visual signal was more reliable, functional connectivity increased between the visual region and the STS. This observation provides a crucial insight into the mechanism behind listeners' ability to flexibly change their dependence on one or the other channel to the most informative one (Ma, Zhou, Ross, Foxe, & Parra, 2009; Nath & Beauchamp, 2011).

To gain an insight into the time course of AV interactions, Rhone et al. (2016) presented ECoG patients with AV stimuli that consisted of the possible combinations of a speech/ nonspeech sound and a speech/ nonspeech lip movement. Rhone et al. recorded cortical signals from the STG, Heschl's gyrus (and premotor cortex, not discussed here). They observed that initial processing of sound in Heschl's gyrus was mostly unaffected by initial visual information (although they do report a small modulation of low-frequency oscillations), while STG activity was influenced by visual input. That is, in STG, stronger (additive) cortical responses were observed when both the visual and auditory signals consisted of speech than when they consisted of nonspeech. Their findings support a model of audiovisual processing in which visual information is integrated with auditory information in the STG, mostly beyond the PAC (but see Pekkola et al., 2005, for fMRI findings in favor of PAC integration). In a related approach, also using ECoG, Besle et al. (2008) reported AV interactions predominantly in secondary auditory cortex, almost immediately (~30 ms) after sound onset (see Reale et al., 2007, for further evidence of the involvement of STG in AV integration).

The findings reviewed here suggest the posterior STS, extending into the lateral STG, is a major site for the integration of auditory and visual speech signals. Recently, Peelle and Sommers (2015) proposed a multistage integration process where, at a first stage, visual information may aid early auditory processing (as early as primary auditory cortex) by predicting the timing of upcoming acoustic events, potentially by resetting ongoing oscillations (see, e.g., Schroeder, Lakatos, Kajikawa, Partan, & Puce, 2008; Stekelenburg & Vroomen, 2007; see Arnal, Poeppel, & Giraud, 2015; and Gross & Poeppel, chapter 29 of this volume, for recent discussions of the broader role of oscillations in speech perception). It should be noted, however, that the extent to which visual information may influence PAC processing beyond providing information about timing remains an active and hotly debated issue. In addition to influences in PAC, however, in regions involving mostly STS (and other higher-level regions outside the temporal lobe) visual information could help to constrain lexical processing, for example in noisy environments. That is, visual information could potentially contribute in distinguishing contrasts based on place of articulation (e.g., *bet* vs. *get*), which, as also described by Mesgarani et al. (2014), are not very clearly represented in the auditory signal and subsequent processing in the STG. Such integration, especially for auditorily weak signals, would thus involve the relative weighting of auditory and visual inputs (Peelle & Sommers; Ross, et al., 2007) through some of the mechanisms described here. Future research should focus on the representational form of the visually elicited activations on the posterior STL to further elucidate what the dominant code is for visual information that is integrated with auditory information on the STL.

2.2. STG Sensitivity to Phoneme Sequences In natural languages, words are not formed by random concatenations of phonemes or syllables. Instead, speech sounds are sequentially organized based on language-specific constraints (termed *phonotactics*). One such statistical regularity is the probability of a specific sound (e.g., *po*) being followed by another (e.g., *te*), called *transition probability* (or, p(B|A) in a sequence AB). Syllable transition probabilities tend to be higher within words than across word boundaries. Listeners can therefore use these probabilities, for example to segment continuous speech into words (e.g., McQueen, 1998), despite the fact that word boundaries are rarely marked by silence. Indeed, listeners, from a very young age, have been shown to be highly sensitive to such regularities (e.g., Pelucchi, Hay, & Saffran, 2009; Saffran, Aslin, & Newport, 1996; Tremblay, Baroni, & Hasson, 2013), despite typically not being consciously aware of them.

To understand at what stages of perceptual analysis transition probabilities affect ongoing cortical processing

of speech sounds, McNealy, Mazziotta, and Dapretto (2006) examined the functional neuroanatomical correlates of speech processing while listeners became familiar with recurring patterns of syllables (these sequences, however, were all nonwords). During fMRI recording, the McNealy et al. presented listeners with syllables that appeared in either sequences with statistical regularities (i.e., some regularly recurring syllable triplets) or sequences with no statistical regularities (a third condition, involving stressed syllables, is not discussed here). They found that the bilateral (but left dominant) posterior STG was more active when listeners were presented with one of the statistically regular sequences than those which contained no regularities. Interestingly, this pattern displayed a buildup across the duration of the experiment as participants became more familiar with the sequences (see McNealy, Mazziotta, & Dapretto, 2010, for a replication in children; and Karuza et al., 2013, for a related study, again implicating bilateral STG). Sound processing in the STG is thus sensitive to local statistical relations, and it appears to be able to learn such relations even over the duration of a single experiment.

In addition to regularities at the syllabic level, similar constraints exist at the phoneme level (i.e., within syllables). For example, in English, hearing the sound /k/ followed by /uw/ (*koo*) is more common than hearing /k/ followed by /iy/ *kee* (we will refer to this relation as *forward probabilities* [*Pfwd*]; see figure 27.4A). Native speakers of English are behaviorally sensitive to these probabilities (e.g., Vitevitch & Luce, 1999), and recent work has begun to characterize the neural basis of these effects. Leonard, Bouchard, Tang, and Chang (2015) presented ECoG patients with a range of consonant-vowel-consonant (CVC) sounds of which the transitions between the CV and the VC parts had variable forward probabilities (and backward probabilities, not discussed here for brevity) based on patterns in spoken English. Firstly, in line with the literature discussed in section 1.3, Leonard et al. found that some electrodes displayed clear phoneme selectivity: Figure 27.4B displays the response of an example electrode, revealing that across the stimulus set this particular electrode had a clear preference for syllable-initial /n/ (blue lines in the left panel). This preference was related to the electrode's STRF which revealed sensitivity to low-frequency components that are characteristic of /n/ (figure 27.4C; STRF is estimated on independent data).

After controlling for the portion of the neural response explained by its acoustic sensitivity (the STRF), the electrode displays a strong effect of forward probability (figure 27.4D). That is, when the vowel was predictable based on the initial consonant, this electrode's response was significantly attenuated. Across electrodes, Leonard et al. observed both attenuating and facilitating effects of forward probability on processing of both the vowel and the final consonant. Time courses of the linear weights for STRF-based models and transition probability models (figure 27.4E) show that transition probabilities affect speech sound processing at a slight delay compared to auditory influences, as expected. These findings demonstrate that even when listening to individual CVC syllables, spoken English-based local transition probabilities have a strong effect on speech sound processing.

The two studies described here thus demonstrate that processing in the STG is highly sensitive to local statistical probabilities of sound patterns at a level beyond individual features of phonemes. These influences were observed both as emerging across the duration of an experiment (McNealy et al., 2006) and as a result of lifelong exposure on short CVC syllables (Leonard et al., 2015; see also Tremblay, Deschamps, Baroni, & Hasson, 2016). Furthermore, effects of local sequence probabilities at the word level have been argued to additionally influence cortical responses in the STG such that words that are statistically unlikely (given the just preceding speech) give rise to stronger activation (Willems, Frank, Nijhof, Hagoort, & Van den Bosch, 2015). Furthermore, a recent study has related listeners' abilities to learn syllable-wise statistical regularities to cortical thickness in a number of regions among which is the bilateral STG (Deschamps, Hasson, & Tremblay, 2016). These combined observations reveal that the human STG functions as an acoustic phonetic pattern recognizer that operates over a range of levels of granularity, from feature sequences to syllable sequences and possibly even word sequences.

2.3. Integration with Lexicosemantic Representations in STG and STS Going beyond sequences of features, a crucial step in language comprehension is the activation of lexical representations. Lexical representations allow for the linking of incoming information to stored semantic representations, which are themselves thought to be widely distributed across the cortex (e.g., Huth, de Heer, Griffiths, Theunissen & Gallant, 2016; Nastase et al., 2017; Ralph, Jefferies, Patterson, & Rogers, 2017, and references therein). Adult native speakers of English have access to somewhere between 25,000 and 75,000 lexical entries (e.g., Altmann, 1997; McMurray, 2007). These words are highly variable in terms of their frequency of use and both their semantic and phonological similarities among each other. To build up this massive lexical inventory,

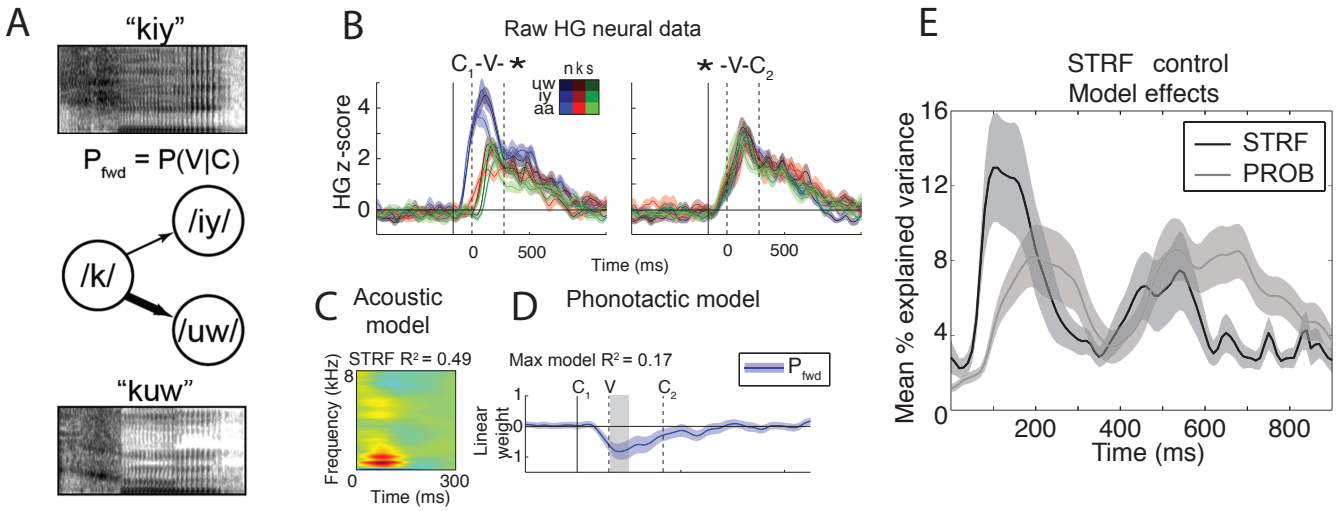

FIGURE 27.4 Phonotactic transition probabilities affect processing in the STG. (A) Two sequences, /k/-/iy/ (*top spectrogram*) and /k/-/uw/ (*bottom spectrogram*), have different transition probabilities between the consonant and the vowel (normalized by the marginal probabilities of the individual elements). The probability that /k/ is followed by /iy/ or /uw/ in English (Pfwd) is indicated by the thickness of the arrows in the *left diagram*. (B) Mean (± standard error of the mean [SEM]) High-Gamma (HG) cortical responses to all stimulus-CVC combinations in an example STG electrode. The electrode shows preferential responses to specific speech sounds in specific positions (see color grid; the electrode is most active when /n/ is the first phoneme). Dashed vertical lines indicate average V and C2 phoneme onsets. (C) The linear STRF of the example electrode in B demonstrates that selectivity is partly driven by phoneme acoustics. (D) Transition probabilities can modulate neural activity not explained by the linear STRF for the example electrode in B: time courses of linear weights show a significant effect during the vowel for Pfwd ($P < 0.05$; gray shading). (E) Mean (±SEM) R^2 time courses across all significant electrodes show that auditory (STRF-model) effects generally precede phonotactic probability (PROB-model) effects. Figure reproduced from Leonard et al. (2015).

between birth and adulthood, people are thought to learn up to 10 new words a day on average. Even in adulthood, learning does not stop, as wordlike forms start to be processed as potential real words after only limited exposure (De Vaan, Schreuder, & Baayen, 2007; Lindsay, Sedin, & Gaskell, 2012), allowing the ongoing introduction of new words into a language's repertoire (e.g., *blog*, *selfie*, or *emoji*). In the following section, we will discuss some of the findings that suggest that speech representations in the STG are strongly affected by lexical- and semantic-level linguistic representations.

One influential approach to study lexical processing has been to compare processing of words and nonwords and especially where in the processing stream the two types of stimuli evoke different responses. Recently, Cibelli, Leonard, Johnson, and Chang (2015) presented ECoG patients with a list of auditory words and nonwords in an overt repetition task. Cibelli et al. observed that processing of both words and nonwords involved a temporal progression of peak latency high-gamma activity from more posterior-dorsal to more anterior-ventral temporal lobe sites, consistent with previous findings. In addition, they observed stronger responses to nonwords over words across the length of STG, an effect that was increasingly larger for more anterior sites. This finding aligns with a recent ALE meta-analysis of PET and fMRI research comparing activation between words and nonwords (Davis & Gaskell, 2009). In that meta-analysis, more activation was found for nonwords than words in a large region of the STG but, interestingly, the opposite was found for a number of downstream regions among which were the middle temporal gyrus (MTG) and a large area covering the supramarginal gyrus and adjacent regions of the most posterior portions of the STG. The increased activation for nonwords in the STG appears to reflect additional processing that is necessary while no lexical item has been selected. The repeated presentation of these nonwords, however, can lead to rapid changes in nonword processing. Davis, Di Betta, Macdonald, and Gaskell (2009) have shown, for example, that BOLD responses in the STG after hearing words and nonwords become increasingly similar once listeners have become familiar with the nonwords (and consolidated learning through sleep). The relative dominance of activation for words over nonwords for regions outside the STG probably reflects more semantic-level processing that fails to activate for nonwords.

A considerable behavioral literature has demonstrated that the ease with which a word is recognized is influenced by the number of words that are phonologically similar to it in a person's lexicon (Luce & Large,

2001; Luce & Pisoni, 1998; Vitevitch & Luce, 1999). That is, words that have many "phonological neighbors" are relatively hard to access at later stages of speech perception because of increased competition (although note that early on in processing having many frequent neighbors may be facilitatory). This pattern has been demonstrated, for example, with increased reaction times in lexical decision tasks or picture-naming latencies (Luce & Pisoni, 1998). To further understand the nature of lexical representations in the cortex, a number of researchers have manipulated these more subtle lexical properties. Although reports of regions that are affected by these manipulations appear to be somewhat variable, lexical statistics have been observed to affect processing in both the STG and STS. Cibelli et al. (2015) observed that for the processing of words, small and low-frequency cohorts (i.e., the number of words matching the phonetic input at each time point) led to increased activity in more anterior sites (also, see Zhuang, Randall, Stamatakis, Marslen-Wilson, & Tyler, 2011 for further effects of cohort on STG processing). Others have observed stronger activation in posterior STS for words that have a high phonological neighborhood density (Okada & Hickok, 2006). In a number of reports, however, researchers have failed to observe robust effects in STG/STS, but they have observed greater BOLD response for high-density neighborhood words than to low-density words in the left supramarginal gyrus (see also Righi, Blumstein, Mertus, & Worden, 2010, for the involvement of this region in phonological-lexical competition), and greater activation for high-frequency words in both anterior and posterior left MTG (Prabhakaran, Blumstein, Myers, Hutchison, & Britton, 2006). These findings suggest that among words, the ease of lexical access as governed by cohort size and neighborhood density impact processing in the STG/STS, but the variability in observed topography suggests that these effects may be relatively dependent on specific task requirements.

Further in the hierarchy from sound to meaning are lexicosemantic relations. The influence of lexicosemantic relations on speech processing has often been investigated with semantic priming paradigms (i.e., how does the prime *nurse* affect the processing of the subsequent target word *hospital*). Although effects of semantic relations are typically observed for MTG (see, e.g., Binder, Desai, Graves, & Conant, 2009, for review; Copland et al., 2003; Giesbrecht, Camblin, & Swaab, 2004; Guediche, Reilly, Santiago, Laurent, & Blumstein, 2016; Rissman, Eliassen, & Blumstein, 2003; Wible et al., 2006; and see Dronkers, Wilkins, Van Valin, Redfern, & Jaeger, 2004, for related research on lesions), a number of fMRI studies have observed similar modulations of STG activation (Matsumoto et al., 2005; Minicucci, Guediche, & Blumstein, 2013; Rissman, Eliassen, & Blumstein, 2003; Wible et al., 2006; but see Guediche et al.). For example, Wible et al. presented participants with (i) highly semantically related, (ii) mildly related, or (iii) unrelated prime-target pairs. They observed that activity in a large region in the posterior STG (along with a portion of the MTG) was strongly dependent on the semantic relation between the prime and target. That is, levels of activation in the STG are reduced when a particular target word has a semantic relation to a just-preceding prime stimulus (see also Rissman, Eliassen, & Blumstein, 2003, for corroborating evidence).

To investigate the temporal properties of the influence of lexicosemantic factors on STG processing, Travis et al. (2013) presented ECoG patients with a picture, followed by a spoken word or a noise sound. The spoken word was either congruent or incongruent with the picture. They observed early (within ~60 ms) differences between the processing of the speech sounds and the noise sounds, demonstrating the type of speech preference for the STG electrodes that was discussed in section 1.3. At a later time window (after ~217 ms), however, in the same region and in some cases the same electrodes, they observed differences between the congruent and incongruent prime-target relations. Their results suggest that early auditory processing in the posterior STG was unaffected by semantic-based expectations, whereas activity in the same region revealed lexicosemantic-dependent processing in a later time window (with, typically, larger responses to the incongruous words). Similar results were observed in a combined EEG-magnetoencephalography experiment (Sohoglu, Peelle, Carlyon, & Davis, 2012), suggesting that semantic-level information can robustly influence information processing in the STL as part of a top-down flow of information. These two studies suggested an initial window of information processing in the STG that was unaffected by semantic information, and a later window where it was. The extent and the time course over which higher-level representations such as semantics can affect early speech sound processing are of particular interest because of the important role that online lexical-phonetic interactions have played in the "autonomous" versus "feedback" debate in formal models of speech perception (see Fox & Blumstein, 2016; McClelland & Elman, 1986; Norris, McQueen, & Cutler, 2000, 2016, and references therein). A recent contribution to this debate was the demonstration that the phenomenon of *phoneme restoration* (the perceptual filling in of occluded phonemes) involves a neural

reconstruction of the acoustic/phonetic events of the missing phoneme. This finding thus suggests that higher-level information may indeed feedback during online processing to activate acoustic/phonemic representations. A crucial next question, however, is what the role of such patterns of activation is (i.e., a role in online perception or in learning; Norris et al., 2016). Further research using the high-spatial and -temporal resolution as offered by ECoG is likely to provide important contributions to this ongoing debate.

The findings presented in this section demonstrate that both lexical- and semantic-level factors have a strong influence on processing in the posteromedial STL. These findings reveal that information typically thought to be represented in regions outside the STL can affect STL processing over short timescales.

2.4. The Role of the STL in the Larger Speech Perception Hierarchy The current chapter has discussed some of the core processing characteristics of the human STL in speech sound processing. The STL is just one part of a vastly interconnected language network in the broader perisylvian region, of which each of the components display processing characteristics that are too complex and detailed to describe here. However, some brief description of the regions that receive information from the STL and the functions that have been ascribed to them is in order.

As mentioned throughout this chapter, it is often suggested that speech sound processing becomes more speech-specific as activity spreads toward regions further away from PAC, especially on the posterior/dorsal to anterior/ventral axis. It appears that lexical and semantic levels of representations become activated in the anterior temporal lobe, the MTG, and the inferior temporal lobe (Binder et al., 1997; DeWitt & Rauschecker, 2012; Mesulam, Thompson, Weintraub, & Rogalski, 2015; Patterson & Johnsrude, 2008; Rissman et al., 2003; Rodd, Davis, & Johnsrude, 2005; Turken & Dronkers, 2011). In addition to this stream of processing in the temporal lobe, however, information is thought to relay from posteromedial STL to the supramarginal gyrus (Obleser & Eisner, 2009; Turkeltaub & Coslett, 2010), a region presumed to be involved in phonological working memory (see also Buchsbaum, Padmanabhan, & Berman, 2011, for the involvement of closely-situated portions of the posterior STS), but which may also play a role in the activation of lexical representations, especially in situations of phonological/lexical competition (Blumstein, 2009). There are also strong connections from the STL to premotor regions, especially as auditory information is important for self-monitoring in speech production (e.g., Chang, Niziolek, Knight, Nagarajan, & Houde, 2013). It has also been suggested that the motor region may play a role in the perception of speech produced by others, although this remains a hotly debated topic (see, e.g., Cheung, Hamilton, Johnson, & Chang, 2016; Galantucci, Fowler, & Turvey, 2006; Pulvermüller & Fadiga, 2010, and references therein). Finally, one of the main regions involved in speech sound perception and language tasks in general has been the inferior frontal gyrus (IFG; including Broca's area). The IFG is heavily involved in language processing more generally (e.g., Hagoort, Baggio, & Willems, 2009), but also in speech perception tasks, for example in resolving competition between task-relevant alternatives, especially when listeners make decisions about noisy or underspecified signals (e.g., Prabhakaran et al., 2006; Snyder, Feigenson, & Thompson-Schill, 2007; Swaab, Brown, & Hagoort, 1998; Utman, Blumstein, & Sullivan, 2001), and in tasks like phonological target detection (Chang et al., 2011). The higher-level role of IFG in phonetic speech perception tasks is supported by the observation that it is strongly dependent on attention (Alho et al., 2016).

The role of this network in speech perception has been conceptualized as following two parallel streams specialized for analyzing different aspects of the speech signal, which both originate from initial processing in parts of the STL. Although specific interpretations differ, it has been broadly suggested that information follows a ventral stream (involving the STL, MTG, anterior TL, [anterior] IFG) of which the dominant function involves lexicosemantic access and/or comprehension, and a dorsal stream (involving STL; supramarginal gyrus; sensory-motor cortex; [posterior] IFG) of which the primary function involves sensory–motor integration and phonological working memory (Hickok & Poeppel, 2007; Rauschecker & Scott, 2009; Scott & Johnsrude, 2003). Other chapters in part V of this volume provide a more detailed description of the several parts of this network.

3. Conclusion

The findings and concepts discussed in this chapter provide some important descriptions of the processing characteristics of the STL and its role in speech sound processing. Processing in the STL is partly characterized by sensitivity to relatively basic auditory properties, especially in regions close to PAC, but the dominant preference is one for more complex spectrotemporal stimulus properties, especially for those features that occur in speech (e.g., Hullett et al., 2016). For example, large portions of the STG were shown to be sensitive to

specific spectrotemporal patterns that are associated with natural classes of speech sounds such as vowels, plosives, or fricatives. This observation shows that the representation of phonetic features is an important processing property of the STL (e.g., Arsenault & Buchsbaum, 2015; Mesgarani et al., 2014; Steinschneider et al., 2011). In addition, categorical representations were observed as an emergent property of processing in the STL. That is, acoustically continuous sequences of sounds become neurally "warped" depending on which parts of these continua are most behaviorally relevant (Chang et al., 2010; Ley et al., 2012). This property of the STL is fundamental to speech processing. It allows for the selection of representations (e.g., lexical and semantic) that are independent of idiosyncrasies in how particular speech sounds are uttered on different occasions or by different speakers (i.e., also known as the invariance problem).

In addition to processing related to immediate auditory stimulus properties, a number of observations demonstrate that processing in the STL is strongly influenced by contextual factors. Visual and auditory information were found to be rapidly integrated in the STL (especially for the posterior parts, e.g., Beauchamp et al., 2010; Erickson et al., 2014), in a way that allows for the selective weighting of visual and auditory input, depending on their reliability (e.g., Nath & Beauchamp, 2011). In addition, processing in the STL is affected by the local probabilities of perceived acoustic events given preceding and subsequent cues at multiple levels of granularity (i.e., sequence probabilities from acoustic features to phoneme sequences; Leonard et al., 2015; McNealy et al., 2006). Furthermore, evidence was presented for semantic influences affecting processing in the STL (Rissman et al., 2003; Wible et al., 2006), albeit potentially with a slight temporal delay relative to initial bottom-up processes (e.g., Sohoglu et al., 2012; Travis et al., 2013). These observations stress the role of the STL as a highly versatile auditory association cortex. Future research should be aimed at further investigating how representations in the STL are influenced by factors such as visual and semantic information. That is, what aspects of these featural representations are affected by visual and semantic representations? And what forms do these "contextual" representations have themselves that allow them to directly influence feature-based representations in the STL? We believe that such questions are likely to be informed by approaches that focus on speech sounds represented as distributed patterns of activity. That is, representation as distinctive patterns of activation, even in absence of changes in overall changes in signal level in broadly defined regions of interest. We think that this approach in combination with increased availability of invasive and noninvasive imaging techniques that allow for the measurement of cortical activity at high-spatial and -temporal resolution ushers in an exciting period for research on the neural processing of speech.

Acknowledgments

The first author received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7 2007–2013 under REA grant agreement nr. 623072. The authors are grateful to Matthew Leonard, Neal Fox, Nina Dronkers and David Poeppel for their comments on earlier versions of the manuscript.

REFERENCES

Alho, J., Green, B. M., May, P. J., Sams, M., Tiitinen, H., Rauschecker, J. P., & Jääskeläinen, I. P. (2016). Early-latency categorical speech sound representations in the left inferior frontal gyrus. *NeuroImage*, *129*, 214–223.

Altmann, C. F., Uesaki, M., Ono, K., Matsuhashi, M., Mima, T., & Fukuyama, H. (2014). Categorical speech perception during active discrimination of consonants and vowels. *Neuropsychologia*, *64*, 13–23.

Altmann, G. T. M. (1997). *The ascent of Babel: An exploration of language, mind, and understanding*. Oxford: Oxford University Press.

Andruski, J. E., Blumstein, S. E., & Burton, M. (1994). The effect of subphonetic differences on lexical access. *Cognition*, *52*(3), 163–187.

Arnal, L. H., Poeppel, D., & Giraud, A. L. (2015). Temporal coding in the auditory cortex. *Handbook of Clinical Neurology*, *129*, 85–98.

Arsenault, J. S., & Buchsbaum, B. R. (2015). Distributed neural representations of phonological features during speech perception. *Journal of Neuroscience*, *35*(2), 634–642.

Baumann, S., Petkov, C. I., & Griffiths, T. D. (2013). A unified framework for the organization of the primate auditory cortex. *Frontiers in Systems Neuroscience*, *7*, 11.

Beauchamp, M. S., Nath, A. R., & Pasalar, S. (2010). fMRI-guided transcranial magnetic stimulation reveals that the superior temporal sulcus is a cortical locus of the McGurk effect. *Journal of Neuroscience*, *30*(7), 2414–2417.

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature*, *403*(6767), 309–312.

Bernstein, L. E., Tucker, P. E., & Demorest, M. E. (2000). Speech perception without hearing. *Perception and Psychophysics*, *62*(2), 233–252.

Besle, J., Fischer, C., Bidet-Caulet, A., Lecaignard, F., Bertrand, O., & Giard, M. H. (2008). Visual activation and audiovisual interactions in the auditory cortex during speech perception: Intracranial recordings in humans. *Journal of Neuroscience*, *28*(52), 14301–14310.

Bidelman, G. M., Gandour, J. T., & Krishnan, A. (2011). Cross-domain effects of music and language experience on the representation of pitch in the human auditory brainstem. *Journal of Cognitive Neuroscience*, *23*(2), 425–434.

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, *19*(12), 2767–2796.

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

Binder, J. R., Frost, J. A., Hammeke, T. A., Cox, R. W., Rao, S. M., & Prieto, T. (1997). Human brain language areas identified by functional magnetic resonance imaging. *Journal of Neuroscience*, *17*(1), 353–362.

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

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

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

Boatman, D., Hall, C., Goldstein, M. H., Lesser, R., & Gordon, B. (1997). Neuroperceptual differences in consonant and vowel discrimination: As revealed by direct cortical electrical interference. *Cortex*, *33*(1), 83–98.

Brugge, J. F., Volkov, I. O., Garell, P. C., Reale, R. A., & Howard, M. A. (2003). Functional connections between auditory cortex on Heschl's gyrus and on the lateral superior temporal gyrus in humans. *Journal of Neurophysiology*, *90*(6), 3750–3763.

Buchman, A. S., Garron, D. C., Trost-Cardamone, J. E., Wichter, M. D., & Schwartz, M. (1986). Word deafness: One hundred years later. *Journal of Neurology, Neurosurgery and Psychiatry*, *49*(5), 489–499.

Buchsbaum, B. R., Baldo, J., Okada, K., Berman, K. F., Dronkers, N., D'Esposito, M., & Hickok, G. (2011). Conduction aphasia, sensory-motor integration, and phonological short term memory an aggregate analysis of lesion and fMRI data. *Brain and Language*, *119*(3), 119–128.

Buchsbaum, B. R., Padmanabhan, A., & Berman, K. F. (2011). The neural substrates of recognition memory for verbal information: Spanning the divide between short and long-term memory. *Journal of Cognitive Neuroscience*, *23*, 978–991.

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C., McGuire, P. K., … David, A. S. (1997). Activation of auditory cortex during silent lipreading. *Science*, *276*(5312), 593–596.

Canolty, R. T., Soltani, M., Dalal, S. S., Edwards, E., Dronkers, N. F., Nagarajan, S. S., … Knight, R. T. (2007). Spatiotemporal dynamics of word processing in the human brain. *Frontiers in Neuroscience*, *1*, 14.

Chan, A. M., Dykstra, A. R., Jayaram, V., Leonard, M. K., Travis, K. E., Gygi, B., … Cash, S. S. (2014). Speech-specific tuning of neurons in human superior temporal gyrus. *Cerebral Cortex*, *24*(10), 2679–2693.

Chang, E. F., Edwards, E., Nagarajan, S. S., Fogelson, N., Dalal, S. S., Canolty, R. T., … Knight, R. T. (2011). Cortical spatio-temporal dynamics underlying phonological target detection in humans. *Journal of Cognitive Neuroscience*, *23*(6), 1437–1446.

Chang, E. F., Niziolek, C. A., Knight, R. T., Nagarajan, S. S., & Houde, J. F. (2013). Human cortical sensorimotor network underlying feedback control of vocal pitch. *Proceedings of the National Academy of Sciences*, *110*(7), 2653–2658.

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

Cheung, C., Hamilton, L. S., Johnson, K., & Chang, E. F. (2016). The auditory representation of speech sounds in human motor cortex. *eLife*, *5*, e12577.

Chevillet, M. A., Jiang, X., Rauschecker, J. P., & Riesenhuber, M. (2013). Automatic phoneme category selectivity in the dorsal auditory stream. *Journal of Neuroscience*, *33*(12), 5208–5215.

Chi, T., Gao, Y., Guyton, M. C., Ru, P., & Shamma, S. (1999). Spectro-temporal modulation transfer functions and speech intelligibility. *Journal of the Acoustical Society of America*, *106*(5), 2719–2732.

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

Copland, D. A., De Zubicaray, G. I., McMahon, K., Wilson, S. J., Eastburn, M., & Chenery, H. J. (2003). Brain activity during automatic semantic priming revealed by event-related functional magnetic resonance imaging. *NeuroImage*, *20*(1), 302–310.

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

David, S. V., Vinje, W. E., & Gallant, J. L. (2004). Natural stimulus statistics alter the receptive field structure of v1 neurons. *Journal of Neuroscience*, *24*(31), 6991–7006.

Davis, M. H., Di Betta, A. M., Macdonald, M. J., & Gaskell, M. G. (2009). Learning and consolidation of novel spoken words. *Journal of Cognitive Neuroscience*, *21*(4), 803–820.

Davis, M. H., & Gaskell, M. G. (2009). A complementary systems account of word learning: Neural and behavioural evidence. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *364*(1536), 3773–3800.

Davis, M. H., & Johnsrude, I. S. (2003). Hierarchical processing in spoken language comprehension. *Journal of Neuroscience*, *23*(8), 3423–3431.

Desai, R., Liebenthal, E., Waldron, E., & Binder, J. R. (2008). Left posterior temporal regions are sensitive to auditory categorization. *Journal of Cognitive Neuroscience*, *20*(7), 1174–1188.

Deschamps, I., Hasson, U., & Tremblay, P. (2016). The structural correlates of statistical information processing during speech perception. *PLOS ONE*, *11*(2), e0149375.

De Vaan, L., Schreuder, R., & Baayen, R. H. (2007). Regular morphologically complex neologisms leave detectable traces in the mental lexicon. *Mental Lexicon*, *2*(1), 1–24.

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

Ding, N., Patel, A. D., Chen, L., Butler, H., Luo, C., & Poeppel, D. (2017). Temporal modulations in speech and music. *Neuroscience and Biobehavioral Reviews 81*(Pt. B), 181–187.

Dronkers, N. F., Wilkins, D. P., Van Valin, R. D., Redfern, B. B., & Jaeger, J. J. (2004). Lesion analysis of the brain areas involved in language comprehension. *Cognition*, *92*(1), 145–177.

Elliott, T. M., & Theunissen, F. E. (2009). The modulation transfer function for speech intelligibility. *PLOS Computational Biology*, *5*(3), e1000302.

Erickson, L. C., Heeg, E., Rauschecker, J. P., & Turkeltaub, P. E. (2014). An ALE meta-analysis on the audiovisual integration of speech signals. *Human Brain Mapping*, *35*(11), 5587–5605.

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

Fox, N. P., & Blumstein, S. E. (2016). Top-down effects of syntactic sentential context on phonetic processing. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(5), 730.

Galantucci, B., Fowler, C. A., & Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin and Review*, *13*(3), 361–377.

Giesbrecht, B., Camblin, C. C., & Swaab, T. Y. (2004). Separable effects of semantic priming and imageability on word processing in human cortex. *Cerebral Cortex*, *14*(5), 521–529.

Grill-Spector, K., Henson, R. N., & Martin, A. (2006). Repetition and the brain: Neural models of stimulus specific effects. *Trends in Cognitive Science*, *10*, 14–25.

Grill-Spector, K., & Malach, R. (2001). fMR-adaptation: A tool for studying the functional properties of human cortical neurons. *Acta Psychologica*, *107*, 293–321.

Guediche, S., Reilly, M., Santiago, C., Laurent, P., & Blumstein, S. E. (2016). An fMRI study investigating effects of conceptually related sentences on the perception of degraded speech. *Cortex*, *79*, 57–74.

Hackett, T. A. (2011). Information flow in the auditory cortical network. *Hearing Research*, *271*(1), 133–146.

Hagoort, P., Baggio, G., & Willems, R. M. (2009). Semantic unification. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (4th ed., pp. 819–836). Cambridge, MA: MIT Press.

Harnad, S. (Ed.) (1987). *Categorical perception: The groundwork of cognition*. Cambridge: Cambridge University Press.

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

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

Hickok, G., & Poeppel, D. (2015). Neural basis of speech perception. *Human Auditory System: Fundamental Organization and Clinical Disorders*, *129*, 149.

Hsu, A., Woolley, S. M., Fremouw, T. E., & Theunissen, F. E. (2004). Modulation power and phase spectrum of natural sounds enhance neural encoding performed by single auditory neurons. *Journal of Neuroscience*, *24*(41), 9201–9211.

Hullett, P. W., Hamilton, L. S., Mesgarani, N., Schreiner, C. E., & Chang, E. F. (2016). Human superior temporal gyrus organization of spectrotemporal modulation tuning derived from speech stimuli. *Journal of Neuroscience*, *36*(6), 2014–2026.

Humphries, C., Liebenthal, E., & Binder, J. R. (2010). Tonotopic organization of human auditory cortex. *NeuroImage*. *50*, 1202–1211.

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

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453–458.

Jäncke, L., Wüstenberg, T., Scheich, H., & Heinze, H. J. (2002). Phonetic perception and the temporal cortex. *NeuroImage*, *15*(4), 733–746.

Joanisse, M. F., Zevin, J. D., & McCandliss, B. D. (2007). Brain mechanisms implicated in the preattentive categorization of speech sounds revealed using fMRI and a short-interval habituation trial paradigm. *Cerebral Cortex*, *17*(9), 2084–2093.

Karuza, E. A., Newport, E. L., Aslin, R. N., Starling, S. J., Tivarus, M. E., & Bavelier, D. (2013). The neural correlates of statistical learning in a word segmentation task: An fMRI study. *Brain and Language*, *127*(1), 46–54.

Klein, D. J., Depireux, D. A., Simon, J. Z., & Shamma, S. A. (2000). Robust spectrotemporal reverse correlation for the auditory system: Optimizing stimulus design. *Journal of Computational Neuroscience*, *9*(1), 85–111.

Krishnan, A., Gandour, J. T., & Bidelman, G. M. (2012). Experience-dependent plasticity in pitch encoding: From brainstem to auditory cortex. *NeuroReport*, *23*(8), 498.

Kuhl, P. K. (1993). Early linguistic experience and phonetic perception: Implications for theories of developmental speech perception. *Journal of Phonetics*, *21*, 125–139.

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2008). Phonetic learning as a pathway to language: New data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *363*(1493), 979–1000.

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

Leonard, M. K., Bouchard, K. E., Tang, C., & Chang, E. F. (2015). Dynamic encoding of speech sequence probability in human temporal cortex. *Journal of Neuroscience*, *35*(18), 7203–7214.

Leonard, M. K., Cai, R., Babiak, M. C., Ren, A., & Chang, E. F. (2016). The peri-Sylvian cortical network underlying single word repetition revealed by electrocortical stimulation and direct neural recordings. *Brain and Language,* S0093-934X(15)30194-2. Advance online publication. doi:10.1016/j.bandl.2016.06.001

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

Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *Journal of Neuroscience*, *31*(8), 2906–2915.

Ley, A., Vroomen, J., Hausfeld, L., Valente, G., De Weerd, P., & Formisano, E. (2012). Learning of new sound categories shapes neural response patterns in human auditory cortex. *Journal of Neuroscience*, *32*(38), 13273–13280.

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

Liberman, A. M., Harris, K. S., Kinney, J. A., & Lane, H. (1961). The discrimination of relative onset-time of the

components of certain speech and non-speech patterns. *Journal of Experimental Psychology*, *61*(5), 379.
Liebenthal, E., Binder, J. R., Spitzer, S. M., Possing, E. T., & Medler, D. A. (2005). Neural substrates of phonemic perception. *Cerebral Cortex*, *15*, 1621–1631.
Liebenthal, E., Desai, R., Ellingson, M. M., Ramachandran, B., Desai, A., & Binder, J. R. (2010). Specialization along the left superior temporal sulcus for auditory categorization. *Cerebral Cortex*, *20*(12), 2958–2970.
Liebenthal, E., Desai, R., Humphries, C., Sabri, M., & Desai, A. (2014). The functional organization of the left STS: A large scale meta-analysis of PET and fMRI studies of healthy adults. *Frontiers in Neuroscience*, *8*, 289.
Lindsay, S., Sedin, L. M., & Gaskell, M. G. (2012). Acquiring novel words and their past tenses: Evidence from lexical effects on phonetic categorisation. *Journal of Memory and Language*, *66*(1), 210–225.
Luce, P. A., & Large, N. R. (2001). Phonotactics, density, and entropy in spoken word recognition. *Language and Cognitive Processes*, *16*(5–6), 565–581.
Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, *19*(1), 1.
Ma, W. J., Zhou, X., Ross, L. A., Foxe, J. J., & Parra, L. C. (2009). Lip-reading aids word recognition most in moderate noise: A Bayesian explanation using high-dimensional feature space. *PLOS ONE*, *4*(3), e4638.
Matsumoto A., Iidaka T., Haneda K., Okada T., & Sadato N (2005). Linking semantic priming effect in functional MRI and event-related potentials. *Neuro-image, 24*(3), 624–634.
McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1–86.
McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–748.
McMurray, B. (2007). Defusing the childhood vocabulary explosion. *Science*, *317*(5838), 631.
McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2009). Within-category VOT affects recovery from "lexical" garden-paths: Evidence against phoneme-level inhibition. *Journal of Memory and Language*, *60*(1), 65–91.
McMurray, B., Tanenhaus, M. K., Aslin, R. N., & Spivey, M. J. (2003). Probabilistic constraint satisfaction at the lexical/phonetic interface: Evidence for gradient effects of within-category VOT on lexical access. *Journal of Psycholinguistic Research*, *32*(1), 77–97.
McNealy, K., Mazziotta, J. C., & Dapretto, M. (2006). Cracking the language code: Neural mechanisms underlying speech parsing. *Journal of Neuroscience*, *26*(29), 7629–7639.
McNealy, K., Mazziotta, J. C., & Dapretto, M. (2010). The neural basis of speech parsing in children and adults. *Developmental Science*, *13*(2), 385–406.
McQueen, J. M. (1998). Segmentation of continuous speech using phonotactics. *Journal of Memory and Language*, *39*(1), 21–46.
McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science*, *30*(6), 1113–1126.
Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*, *343*(6174), 1006–1010.
Mesulam, M. M., Thompson, C. K., Weintraub, S., & Rogalski, E. J. (2015). The Wernicke conundrum and the anatomy of language comprehension in primary progressive aphasia. *Brain*, *138*(8), 2423–2437.
Miller, J. L., Grosjean, F., & Lomanto, C. (1984). Articulation rate and its variability in spontaneous speech: A reanalysis and some implications. *Phonetica*, *41*(4), 215–225.
Miller, L. M., & D'Esposito, M. (2005). Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *Journal of Neuroscience*, *25*(25), 5884–5893.
Minicucci, D., Guediche, S., & Blumstein, S. E. (2013). An fMRI examination of the effects of acoustic-phonetic and lexical competition on access to the lexical-semantic network. *Neuropsychologia*, *51*(10), 1980–1988.
Mitterer, H., Scharenborg, O., & McQueen, J. M. (2013). Phonological abstraction without phonemes in speech perception. *Cognition*, *129*(2), 356–361.
Moerel, M., De Martino, F., & Formisano, E. (2012). Processing of natural sounds in human auditory cortex: Tonotopy, spectral tuning, and relation to voice sensitivity. *Journal of Neuroscience*, *32*(41), 14205–14216.
Moerel, M., De Martino, F., Santoro, R., Ugurbil, K., Goebel, R., Yacoub, E., & Formisano, E. (2013). Processing of natural sounds: Characterization of multipeak spectral tuning in human auditory cortex. *Journal of Neuroscience*, *33*(29), 11888–11898.
Nagel, K. I., & Doupe, A. J. (2008). Organizing principles of spectro-temporal encoding in the avian primary auditory area field L. *Neuron*, *58*(6), 938–955.
Nastase, S. A., Connolly, A. C., Oosterhof, N. N., Halchenko, Y. O., Guntupalli, J. S., Visconti di Oleggio Castello, M., … Haxby, J. V. (2017). Attention selectively reshapes the geometry of distributed semantic representation. *Cerebral Cortex*, *27*(8), 4277–4291.
Nath, A. R., & Beauchamp, M. S. (2011). Dynamic changes in superior temporal sulcus connectivity during perception of noisy audiovisual speech. *Journal of Neuroscience*, *31*(5), 1704–1714.
Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences*, *23*(03), 299–325.
Norris, D., McQueen, J. M., & Cutler, A. (2016). Prediction, Bayesian inference and feedback in speech recognition. *Language, Cognition and Neuroscience*, *31*(1), 4–18.
Nourski, K. V., Steinschneider, M., Oya, H., Kawasaki, H., Jones, R. D., & Howard, M. A. (2012). Spectral organization of the human lateral superior temporal gyrus revealed by intracranial recordings. *Cerebral Cortex*, *24*(2), 340–352.
Obleser, J., & Eisner, F. (2009). Pre-lexical abstraction of speech in the auditory cortex. *Trends in Cognitive Sciences*, *13*(1), 14–19.
Obleser, J., Eisner, F., & Kotz, S. A. (2008). Bilateral speech comprehension reflects differential sensitivity to spectral and temporal features. *Journal of Neuroscience*, *28*(32), 8116–8123.
Obleser, J., Zimmermann, J., Van Meter, J., & Rauschecker, J. P. (2007). Multiple stages of auditory speech perception reflected in event-related FMRI. *Cerebral Cortex*, *17*(10), 2251–2257.
Ohl, F. W., & Scheich, H. (1997). Orderly cortical representation of vowels based on formant interaction. *Proceedings of the National Academy of Sciences*, *94*(17), 9440–9444.
Okada, K., & Hickok, G. (2006). Identification of lexical-phonological networks in the superior temporal sulcus using functional magnetic resonance imaging. *NeuroReport*, *17*(12), 1293–1296.

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

Okada, K., Venezia, J. H., Matchin, W., Saberi, K., & Hickok, G. (2013). An fMRI study of audiovisual speech perception reveals multisensory interactions in auditory cortex. *PLOS ONE*, *8*(6), e68959.

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

Pasley, B. N., David, S.V., Mesgarani, N., Flinker, A., Shamma, S.A., Crone, N. E., … Chang, E. F. (2012). Reconstructing speech from human auditory cortex. *PLOS Biology*, *10*(1), e1001251.

Patterson, R. D., & Johnsrude, I. S. (2008) Functional imaging of the auditory processing applied to speech sounds. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1493), 1023–1035.

Paulesu, E., Perani, D., Blasi, V., Silani, G., Borghese, N. A., De Giovanni, U., … Fazio, F. (2003). A functional-anatomical model for lipreading. *Journal of Neurophysiology*, *90*(3), 2005–2013.

Prabhakaran, R., Blumstein, S. E., Myers, E. B., Hutchison, E., & Britton, B. (2006). An event-related fMRI investigation of phonological-lexical competition. *Neuropsychologia*, *44*(12), 2209–2221.

Peelle, J. E., Johnsrude, I., & Davis, M. H. (2010). Hierarchical processing for speech in human auditory cortex and beyond. *Frontiers in Human Neuroscience*, *4*, 51.

Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex*, *68*, 169–181.

Pekkola, J., Ojanen, V., Autti, T., Jääskeläinen, I. P., Möttönen, R., Tarkiainen, A., & Sams, M. (2005). Primary auditory cortex activation by visual speech: An fMRI study at 3 T. *NeuroReport*, *16*(2), 125–128.

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009). Statistical learning in a natural language by 8-month-old infants. *Child Development*, *80*(3), 674–685.

Price, C. J. (2012). A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. *NeuroImage*, *62*(2), 816–847.

Pulvermüller, F., & Fadiga, L. (2010). Active perception: Sensorimotor circuits as a cortical basis for language. *Nature Reviews Neuroscience*, *11*(5), 351–360.

Quigg, M., & Fountain, N. (1999). Conduction aphasia elicited by stimulation of the left posterior superior temporal gyrus. *Journal of Neurology, Neurosurgery, and Psychiatry*, *66*(3), 393.

Ralph, M. L., Jefferies, E., Patterson, K., & Rogers, T. T. (2017). The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, *18*, 42–55.

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

Reale, R. A., Calvert, G. A., Thesen, T., Jenison, R. L., Kawasaki, H., Oya, H., … Brugge, J. F. (2007). Auditory-visual processing represented in the human superior temporal gyrus. *Neuroscience*, *145*(1), 162–184.

Rhone, A. E., Nourski, K. V., Oya, H., Kawasaki, H., Howard III, M. A., & 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.

Rieke, F., Bodnar, D. A., & Bialek, W. (1995). Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents. *Proceedings of the Royal Society of London B: Biological Sciences*, *262*(1365), 259–265.

Righi, G., Blumstein, S. E., Mertus, J., & Worden, M. S. (2010). Neural systems underlying lexical competition: An eye tracking and fMRI study. *Journal of Cognitive Neuroscience*, *22*(2), 213–224.

Rissman, J., Eliassen, J. C., & Blumstein, S. E. (2003). An event related FMRI investigation of implicit semantic priming. *Journal of Cognitive Neuroscience*, *15*(8), 1160–1175.

Rodd, J. M., Davis, M. H., & Johnsrude, I. S. (2005). The neural mechanisms of speech comprehension: FMRI studies of semantic ambiguity. *Cerebral Cortex*, *15*, 1261–1269.

Rogalsky, C., Poppa, T., Chen, K. H., Anderson, S. W., Damasio, H., Love, T., & Hickok, G. (2015). Speech repetition as a window on the neurobiology of auditory–motor integration for speech: A voxel-based lesion symptom mapping study. *Neuropsychologia*, *71*, 18–27.

Rosen, S., Wise, R. J., Chadha, S., Conway, E. J., & Scott, S. K. (2011). Hemispheric asymmetries in speech perception: Sense, nonsense and modulations. *PLOS ONE*, *6*(9), e24672.

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex*, *17*(5), 1147–1153.

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.

Saenz, M., & Langers, D. R. (2014). Tonotopic mapping of human auditory cortex. *Hearing Research*, *307*, 42–52.

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

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.

Schreiner, C. E., Froemke, R. C., & Atencio, C. A. (2011). Spectral processing in auditory cortex. In J. A. Winer & C. E. Schreiner (Eds.), *The auditory cortex* (pp. 275–308). New York: Springer.

Schroeder, C. E., Lakatos, P., Kajikawa, Y., Partan, S., & Puce, A. (2008). Neuronal oscillations and visual amplification of speech. *Trends in Cognitive Sciences*, *12*(3), 106–113.

Scott, S. K., Blank, C. C., Rosen, S., & Wise, R. J. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain*, *123*(12), 2400–2406.

Scott, S. K., & Johnsrude, I. S. (2003). The neuroanatomical and functional organization of speech perception. *Trends in Neurosciences*, *26*(2), 100–107.

Shamma, S., & Lorenzi, C. (2013). On the balance of envelope and temporal fine structure in the encoding of speech in the early auditory system. *Journal of the Acoustical Society of America*, *133*(5), 2818–2833.

Sjerps, M. J., & McQueen, J. M. (2010). The bounds on flexibility in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(1), 195.

Skipper, J. I., van Wassenhove, V., Nusbaum, H. C., & Small, S. L. (2007). Hearing lips and seeing voices: How cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral Cortex*, *17*(10), 2387–2399.

Snyder, H. R., Feigenson, K., & Thompson-Schill, S. L. (2007). Prefrontal cortical response to conflict during semantic and phonological tasks. *Journal of Cognitive Neuroscience*, *19*(5), 761–775.

Sohoglu, E., Peelle, J. E., Carlyon, R. P., & Davis, M. H. (2012). Predictive top-down integration of prior knowledge during speech perception. *Journal of Neuroscience*, *32*(25), 8443–8453.

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

Stekelenburg, J. J., & Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. *Journal of Cognitive Neuroscience*, *19*(12), 1964–1973.

Striem-Amit, E., Hertz, U., & Amedi, A. (2011). Extensive cochleotopic mapping of human auditory cortical fields obtained with phase-encoding fMRI. *PLOS ONE*, *6*(3), e17832.

Suh, M. W., Lee, H. J., Kim, J. S., Chung, C. K., & Oh, S. H. (2009). Speech experience shapes the speechreading network and subsequent deafness facilitates it. *Brain*, *132*(10), 2761–2771.

Swaab, T. Y., Brown, C., & Hagoort, P. (1998). Understanding ambiguous words in sentence contexts: Electrophysiological evidence for delayed contextual selection in Broca's aphasia. *Neuropsychologia*, *36*(8), 737–761.

Takeichi, H., Koyama, S., Terao, A., Takeuchi, F., Toyosawa, Y., & Murohashi, H. (2010). Comprehension of degraded speech sounds with m-sequence modulation: An fMRI study. *NeuroImage*, *49*(3), 2697–2706.

Talavage, T. M., Sereno, M. I., Melcher, J. R., Ledden, P. J., Rosen, B. R., & Dale, A. M. (2004). Tonotopic organization in human auditory cortex revealed by progressions of frequency sensitivity. *Journal of Neurophysiology*, *91*, 1282–1296.

Talebi, V., & Baker, C. L. (2012). Natural versus synthetic stimuli for estimating receptive field models: A comparison of predictive robustness. *Journal of Neuroscience, 32*(5), 1560–1576.

Theunissen, F. E., Sen, K., & Doupe, A. J. (2000). Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. *Journal of Neuroscience*, *20*(6), 2315–2331.

Travis, K. E., Leonard, M. K., Chan, A. M., Torres, C., Sizemore, M. L., Qu, Z., … Halgren, E. (2013). Independence of early speech processing from word meaning. *Cerebral Cortex*, *23*(10), 2370–2379.

Tremblay, P., Baroni, M., & Hasson, U. (2013). Processing of speech and non-speech sounds in the supratemporal plane: Auditory input preference does not predict sensitivity to statistical structure. *NeuroImage*, *66*, 318–332.

Tremblay, P., Deschamps, I., Baroni, M., & Hasson, U. (2016). Neural sensitivity to syllable frequency and mutual information in speech perception and production. *NeuroImage*, *136*, 106–121.

Tsunada, J., Lee, J. H., & Cohen, Y. E. (2011). Representation of speech categories in the primate auditory cortex. *Journal of Neurophysiology*, *105*(6), 2634–2646.

Turkeltaub, P. E., & Coslett, H. (2010). Localization of sublexical speech perception components. *Brain and Language*, *114*(1), 1–15.

Turken, A. U., & Dronkers, N. F. (2011). The neural architecture of the language comprehension network: Converging evidence from lesion and connectivity analyses. *Frontiers in Systems Neuroscience*, 5, 1.

Utman, J. A., Blumstein, S. E., & Sullivan, K. (2001). Mapping from sound to meaning: Reduced lexical activation in Broca's aphasics. *Brain and Language*, *79*(3), 444–472.

Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language*, *40*(3), 374–408.

Weiss, M. W., & Bidelman, G. M. (2015). Listening to the brainstem: Musicianship enhances intelligibility of subcortical representations for speech. *Journal of Neuroscience*, *35*(4), 1687–1691.

Wible, C. G., Han, S. D., Spencer, M. H., Kubicki, M., Niznikiewicz, M. H., Jolesz, F. A., … Nestor, P. (2006). Connectivity among semantic associates: An fMRI study of semantic priming. *Brain and Language*, *97*(3), 294–305.

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., & Van den Bosch, A. (2015). Prediction during natural language comprehension. *Cerebral Cortex*, *26*(6), 2506–2516.

Wilson, S. M., Lam, D., Babiak, M. C., Perry, D. W., Shih, T., Hess, C.P., … Chang, E. F. (2015). Transient aphasias after left hemisphere resective surgery. *Journal of Neurosurgery*, *123*(3), 581–593.

Woolley, S. M., Fremouw, T. E., Hsu, A., & Theunissen, F. E. (2005). Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds. *Nature Neuroscience*, *8*(10), 1371–1379.

Young, E. D. (2008). Neural representation of spectral and temporal information in speech. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *363*(1493), 923–945.

Zaehle, T., Geiser, E., Alter, K., Jancke, L., & Meyer, M. (2008). Segmental processing in the human auditory dorsal stream. *Brain Research*, *1220*, 179–190.

Zhuang, J., Randall, B., Stamatakis, E. A., Marslen-Wilson, W. D., & Tyler, L. K. (2011). The interaction of lexical semantics and cohort competition in spoken word recognition: An fMRI study. *Journal of Cognitive Neuroscience*, *23*(12), 3778–3790.