How Does the Brain Represent Speech?

Oiwi Parker Jones[1] and Jan W. H. Schnupp[2]

[1] Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, UK
[2]Department of Biomedical Sciences, City University of Hong Kong, Hong Kong

Author Manuscript of Chapter 3 in: The Handbook of Speech Perception 2nd ed. , Pardo, Nygaard, Remez, Pisoni eds., 2021, p.58-96, Wiley

1. Introduction

In this chapter, we hope to provide a brief overview of how the brain's auditory system represents speech. The topic is vast, many decades of research on the subject have generated several book's worth of insight into this fascinating question, and getting close up and personal with this subject matter does necessitate a fair bit of background knowledge about neuroanatomy and physiology, as well as acoustics and linguistic sciences. Providing a reasonably comprehensive overview of the topic which is accessible to a wide readership, all in a short chapter, is a near impossible task, and we apologize in advance for the shortcomings that this chapter will inevitably have. With these caveats out of the way and without further ado, let us jump right in and begin by examining the question: what is there to "represent" in a speech signal?

The word "representation" is quite widely used in sensory neuroscience, but it is rarely clearly defined. A "neural representation" tends to refer to the manner in which neural activity patterns encode or processes some key aspects of the sensory world. Of course, if we want to understand how the brain listens to speech, then grasping how neural activity in early stages of the nervous system encodes speech sounds is really only a very small part of what we would ideally like to understand. It is a necessary first step that leaves many interesting questions unanswered, as you can easily appreciate if you consider that fairly simple technological devices such as telephone lines are able to "represent" speech with patterns of electrical activity, but these devices tell us relatively little about what it means for a brain to hear speech. Phone lines merely have to capture enough of the physical parameters of an acoustic waveform to allow the re-synthesis of a sufficiently similar acoustic waveform to facilitate comprehension by another person at the other end of the line. Brains, in contrast, don't just deliver signals to "a mind at the other end of the line", they have to make the mind at the other end of the line, and to do that they have to try to learn something from the speech signal about who speaks, where they might be, what mood they are in, and, most importantly, about the ideas the speaker is trying to communicate. Consequently it would be nice to know how the brain represents not just the acoustics, but also the phonetic, prosodic and semantic features of the speech it hears.

Readers of this volume are likely to be well aware that extracting such “higher order” features from speech signals is difficult and intricate. Once the physical aspects of the acoustic waveform are encoded, phonetic properties, such as formant frequencies, voicing and voice pitch, must be inferred, interpreted and classified in a context-dependent manner, which in turn facilitates the creation of a semantic representation of speech. In the auditory brain, this occurs along a processing hierarchy, where the lowest levels of the auditory nervous system, the inner ear, auditory nerve fibres and brainstem, encode the physical attributes of the sound and compute what might be described as “low level features”, which are then passed on via the midbrain and the thalamus toward an extensive network of auditory and multisensory cortical areas, whose task it is to form phonetic and semantic representations. As this chapter progresses, we will look in some detail at this progressive transformation of an initially largely “acoustic” representation of speech sounds in the auditory nerve, brainstem, midbrain and primary cortex to an increasingly “linguistic” feature representation in a part of the brain called the superior temporal gyrus, and finally to semantic representations in brain areas stretching well beyond those classically thought of as “auditory” structures.

While it is apt to think of this neural speech processing stream as a hierarchical process, it would nevertheless be wrong to think of this process as entirely feed-forward. It is well known that, for each set of ascending nerve fibres carrying auditory signals from the inner ear to the brainstem, from brainstem to midbrain, from midbrain to thalamus, and from thalamus to cortex, there is a parallel, descending pathway going from cortex back to thalamus, midbrain, brainstem and all the way back to the ear. This is thought to allow feedback signals to be sent in order to focus attention, and to make use of the fact that the rules of language make the temporal evolution of speech sounds partly predictable, and such predictions can facilitate hearing speech in noise, or to “tune the ear” to the voice or dialect of a particular speaker.

To orient the readers who are unfamiliar with the neuroanatomy of the auditory pathway we include a sketch in Figure 1, which shows the approximate location of some of the key stages of the early parts of this pathway, from the ear to the primary auditory cortex, projected onto a drawing of a frontal section through the brain running vertically roughly through the middle of the ear canals.

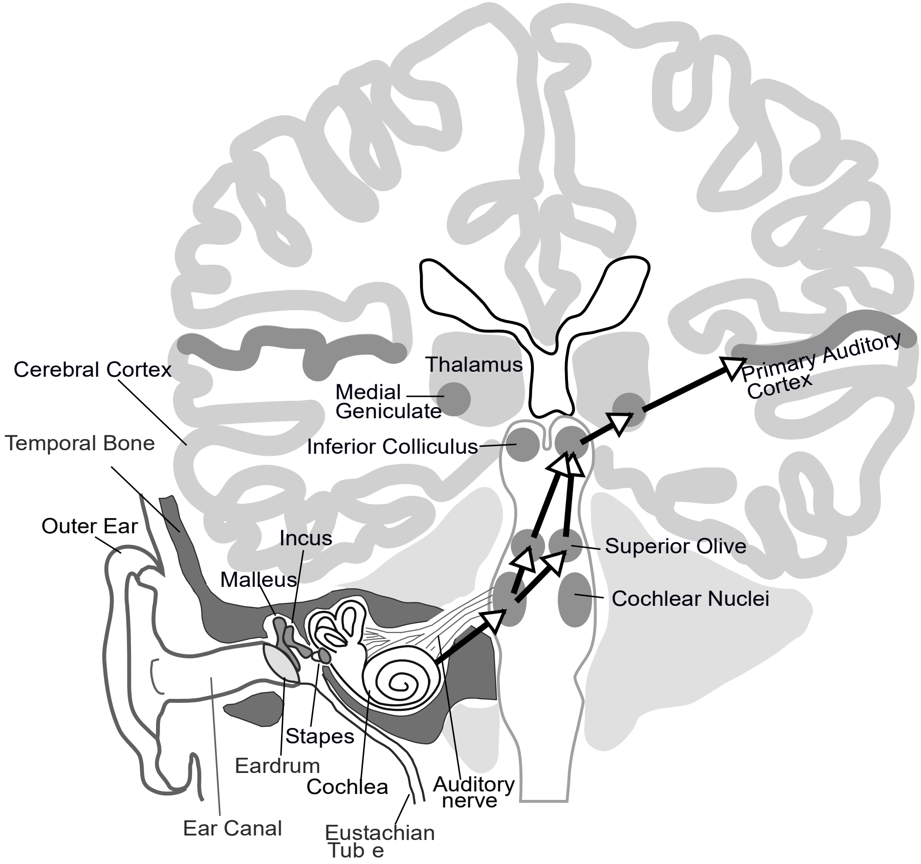

Figure 1: Illustration of the ear and early stages of the ascending auditory pathway.

The arrows in Figure 1 show the principal connections along the main, the so called ‘lemniscal’ ascending auditory pathway. Note, however, that it is impossible to overstate the extent to which Figure 1 oversimplifies the richness and complexity of the brain’s auditory pathways. For example, the cochlear nuclei, the first auditory relay station receiving input from the ear, has no less than three anatomical subdivisions, each comprising many tens to a few hundred thousand neurons of different cell types and with different onward connections. Here we show the output neurons of the cochlear nuclei as projecting to the superior olive bilaterally, which is essentially correct, but for simplicity we omit the fact that the superior olive itself is comprised of around half a dozen intricately interconnected subnuclei, and that there are also connections from the cochlear nuclei which bypass the superior olive and connect straight to the inferior colliculus, the major auditory processing center of the midbrain. The inferior colliculus too comprises several subdivisions, as does the next station on the ascending pathway, the medial geniculate body of the thalamus. And even primary auditory cortex is thought to have two or three distinct subfields, depending on which mammalian species one looks at and which anatomist one asks. In order not to clutter the figure we show none of the descending connections, but to think of each of the arrows here as going in both directions would not be a bad start.

The complexity of the anatomy is quite bewildering, and much remains unknown about the detailed structure and function of its many subdivisions. But we have nevertheless learned a huge amount about these structures and the physiological mechanisms that are at work within them and which underpin our ability to hear speech. Animal experiments have been invaluable in elucidating basic physiological mechanisms of sound encoding, auditory learning and pattern classification in the mammalian brain. Clinical studies on patients with various forms of hearing impairment or aphasia have also helped identify key cortical structures. More recently, functional brain imaging on normal volunteers, as well as invasive electrophysiological recordings from the brains of patients who are undergoing brain surgery for epilepsy have further refined our knowledge of speech representations particularly in higher-order cortical structures.

In the following sections we shall try to highlight some of the insights that have been gained from these types of studies, and we shall try to structure this chapter like a journey, accompanying speech sounds as they leave the vocal tract of a speaker, enter the listener’s ear, become encoded as trains of nerve impulses in the cochlea and auditory nerve, and then travel along the pathways just described and spread out across a phenomenally intricate network of hundreds of millions of neurons whose concerted action underpins our ability to perform the everyday magic of communicating abstract thoughts across space and time by the medium of the spoken word.

2. Encoding of speech in the inner ear and auditory nerve

Let us begin our journey by reminding ourselves about how speech sounds are generated, and what acoustic features are therefore elementary aspects of a speech sound that need to be encoded. When we speak, we produce both voiced and unvoiced speech sounds. Voiced speech sounds arise when the vocal folds in our larynx open and close periodically, producing a rapid and periodic glottal pulse train which may vary from around 80 Hz for a low bass voice to 900 Hz or above for a high soprano voice, although glottal pulse rates of somewhere between 125 to 300 Hz are most common for adult speech. Voiced speech sounds include vowels and voiced consonants. Unvoiced sounds are simply those which are not produced with any vibrating of the vocal folds. The manner in which they are created causes unvoiced speech sounds to have spectra typical of noise, while the spectra of voiced speech sounds exhibit a “harmonic structure”, with regular sharp peaks at frequencies corresponding to the overtones of the glottal pulse train. Related to these differences in the waveforms and spectra is the fact that, perceptually, unvoiced speech sounds do not have an identifiable pitch, while voiced speech sounds do have a clear pitch of a height that corresponds to their fundamental frequency, which corresponds to the glottal pulse rate. Thus, we can sing melodies with voiced speech sounds, but we cannot whisper a melody.

When we speak, the different types of sound sources, whether unvoiced noises or voiced harmonic series, are shaped by resonances in the vocal tract, which we must deftly manipulate by dynamically changing the volume and the size of the openings of a number of cavities in our throat, mouth and nose, which we do by articulatory movements of the jaw, soft palate, tongue and lips. The resonances in our vocal tracts impose broad spectral peaks on the spectra of the speech sounds, and these broad spectral peaks are known as formants. The dynamic pattern of changing formant frequencies encodes the lion's share of the semantic information in speech. Consequently, to interpret a speech stream that arrives at our ears, one might think that our ears and brains will chiefly need to examine the incoming sounds for broad peaks in the spectrum to identify formants. But in order to detect voicing and to determine voice pitch the brain must also look either for sharp peaks at regular intervals in the spectrum to identify harmonics, or, alternatively, look for periodicities in the temporal waveform. Pitch information provided by harmonicity, or, equivalently periodicity, is a vital cue to help identify speakers, gain prosodic information, or to determine the "tone" of a vowel in tonal languages like Chinese or Thai which use pitch contours to distinguish between otherwise identical, homophonic syllables. Encoding information about these fundamental features, formants and harmonicity or periodicity, is thus an essential job of the inner ear and auditory nerve. They do this as they translate incoming sound waveforms into a "tonotopically organized" pattern of neural activity which represents differences in acoustic energy across frequency bands by means of a so-called "rate-place code". Nerve fibers which are tuned to systematically different preferred, or "characteristic" frequencies are arranged in an orderly array. Differences in firing rates across the array encode peaks and valleys in the frequency spectrum, which conveys information about formants, and, to a lesser extent, about harmonics.

This concept of tonotopy is quite central to the way all sounds, not just speech sounds, are usually thought to be represented along the lemniscal auditory pathway. All the stations of the lemniscal auditory pathway shown in Figure 1, from the cochlea to primary auditory cortex, contain at least one, and sometimes several "tonotopic maps", i.e. arrays of frequency tuned neurons arranged in a systematic array from low to high preferred frequency. It is therefore worth examining this notion of tonotopy in some detail to understand its origin, and to ask what tonotopy can and cannot do to represent fundamental features of speech.

In the mammalian brain, tonotopy arises quite naturally from the way sounds are transduced into neural responses by the basilar membrane and organ of corti in the cochlea. When sounds are transmitted from the ear drum to the inner ear via the ossicles, the mechanical vibrations are transmitted to the basilar membrane via fluid filled chambers of the inner ear. The basilar membrane itself has a stiffness gradient, being stiff at the “basal” end, near the ossicles, and floppy at the far end, the “apex”. Sounds transmitted through the far end have little mechanical resistance from the stiffness of the basilar membrane, but have to displace more inert fluid column in the inner ear. Sounds traveling through the near end face less inertia, but more stiffness. The upshot of this is that one can think of the basilar membrane as a bank of mechanical “spring-mass” filters, with filters tuned to high frequencies at the base, and to increasingly lower frequencies toward the apex. Tiny, highly sensitive hair cells which sit on the basilar membrane then pick up these frequency filtered vibrations and translate them into electrical signals, which are then encoded as trains of nerve impulses (also called action potentials or spikes) in the bundles of auditory nerve fibers that connect the inner ear to the brain. Thus, each nerve fiber in the auditory nerve is frequency tuned, and the sound frequency it is most sensitive to is known as its “characteristic frequency” (CF).

The cochlea, and the basilar membrane inside it, is curled up in a spiral, and the organization of the auditory nerve mirrors that of the basilar membrane: inside it we have something that could be described as a “rate-place” code for sounds, where the amount of sound energy at the lowest audible frequencies (ca 50 Hz) is represented by the firing rates of nerve fibers right at the center, and increasingly higher frequencies are encoded by nerve fibers which are arranged in a spiral around that center. Once the auditory nerve reaches the cochlear nuclei, this orderly spiral arrangement “unwraps” to project systematically across the extent of the nuclei, creating tonotopic maps which are then passed on up the auditory pathway by orderly anatomical connections from one station to the next. What this means for encoding of speech in the early auditory system is that formant peaks of speech sounds, and maybe also the peaks of harmonics, should be represented by systematic differences in firing rates across the tonotopic array. The human auditory nerve contains about 30,000 such nerve fibers, each capable of firing anywhere between zero and several hundred spikes a second. So there are many hundreds of thousands of nerve impulses per second available to represent the shape of the sound spectrum across the tonotopic array. And indeed, there is quite a lot of experimental evidence that systematic firing rate differences across this array of nerve fibers are is not a bad first order approximation of what goes on in the auditory system (Delgutte 1997), but as ever so often in neurobiology, the full story is a lot more complicated.

Thanks to decades of physiological and anatomical studies on experimental animals by dozens of teams, the mechanisms of sound encoding in the auditory nerve are now known in sufficient detail that it has become possible to develop computer models that can predict the activity of auditory nerve fibers to arbitrary sound inputs (Zhang et al. 2001; Heinz et al. 2002; Sumner et al. 2002; Meddis and O'Mard 2005; Zhang and Carney 2005; Ferry and Meddis 2007), and here we shall use the model of Zilany et al. (2014) to look at the encoding of speech sounds in the auditory nerve in a little more detail.

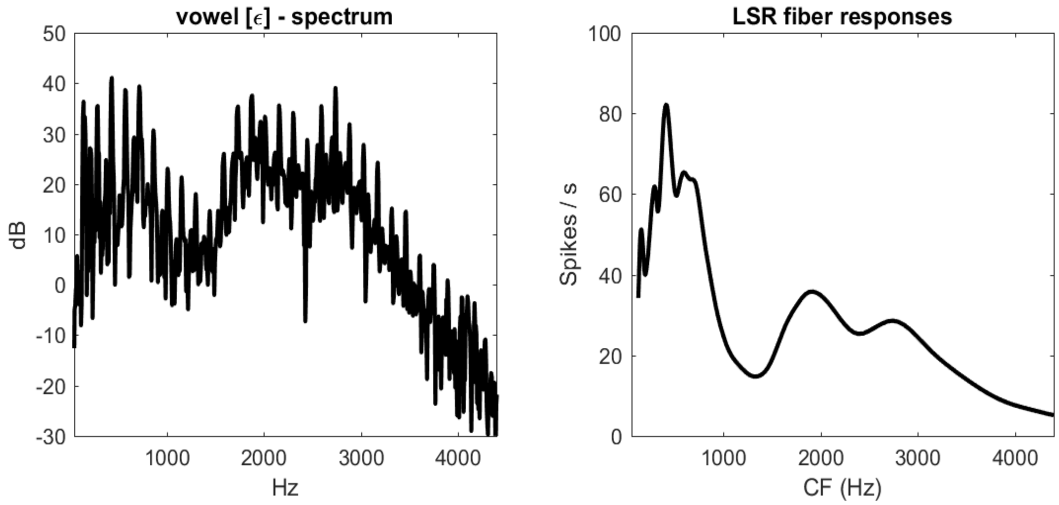

Figure 2: A power spectrum representing an instantaneous spectrogram (left) and a simulated distribution of firing rates for an auditory nerve fiber (right) both for the vowel [ɛ] in [hɛːd] "head".

The left panel of Figure 2 shows the power spectrum of a recording of the spoken vowel [ɛ], as in "head" (IPA [hɛːd]). The spectrum shows many sharp peaks at multiples of about 145 Hz – the harmonics of the vowel. These sharp peaks ride on top of broad peaks centered around ca. 500, 1850 and 2700 Hz – the formants of the vowel. The right panel of the figure shows the distribution of firing rates of "low spontaneous rate" (LSR) auditory nerve fibers in response to the same vowel, according to the auditory nerve fiber model by Zilany et al. (2014). Along the X axis we plot the CF of each nerve fiber, and along the Y axis we expect the average number of spikes the fiber would be expected to fire per second when presented with the vowel [ɛ] at a sound level of 65 dB SPL, the sort of sound level that might be typical during a calm conversation in a quiet background.

Comparing the spectrogram on the left with the distribution of firing rates on the right, it is apparent that the broad peaks of the formants are well reflected in the firing rate distribution, if anything perhaps more visibly so than in the spectrum, but most of the harmonics are not. Indeed, only the lowest three harmonics are visible, the others have been "ironed out" by the fact that the frequency tuning of cochlear filters is often broad compared to the frequency interval between individual harmonics, and becomes broader for higher frequencies. Only the very lowest harmonics are therefore "resolved" by the rate-place code of the tonotopic nerve fiber array, and we should think of tonotopy as well adapted to representing formants, but poorly adapted to representing pitch or voicing information. If you bear in mind that many telephones will high-pass speech at 300 Hz, thereby effectively cutting off the lowest harmonic peak, then there really is not much information about the harmonicity of the sound left reflected in the tonotopic firing rate distribution. But there are important additional cues to voicing and pitch as we shall see shortly.

The firing rates of auditory nerve fibers increase monotonically with increasing sound level, but these fibers do need a minimum threshold sound level, and they cannot increase their firing rates indefinitely when sounds keep getting louder. This gives auditory nerve fibers a limited “dynamic range”, which usually covers 50 dB or less. At the edges of the dynamic range, the formants of speech sounds cannot be effectively represented across the tonotopic array because the neurons in the array either fire not at all (or not above their “spontaneous” firing rates), or because they all fire as fast as they can. However, people can usually understand speech well over a very broad range of sound levels. To be able to code sounds effectively over a wide range of sound levels, the ear appears to have evolved different types of auditory nerve fibers, some of which specialize for hearing quiet sounds, with low thresholds but also relatively low saturation sound levels, and others which specialize for hearing louder sounds, with higher thresholds and higher saturation levels. Auditory physiologists call the more sensitive of these fiber types “high spontaneous rate” (HSR) fibers, given that these auditory nerve fibers may fire nerve impulses at fairly elevated rates (some 30 spikes/s or so) even in the absence of any external sound, and the less sensitive fibers are the LSR fibers which we have already encountered, and which fire only a handful of spikes/s in the absence of sound. There are also medium spontaneous rate fibers, which, as you might expect, lie in the middle between HSR and LSR fibers in sensitivity and spontaneous activity. You may of course wonder why these auditory nerve fibers would fire any impulses if there is no sound to encode, but it is worth bearing in mind that the amount of physical energy in relatively quiet sounds is minuscule, and that the sensory cells that need to pick up those sounds cannot necessarily distinguish a very quiet external noise from internal physiological noise that comes simply from blood flow or random thermal motion inside the ear at body temperature. Auditory nerve fibers operate right at the edge of this physiological “noise floor”, and the most sensitive cells are also most sensitive to the physiological background noise, which gives rise to their high spontaneous firing rate.

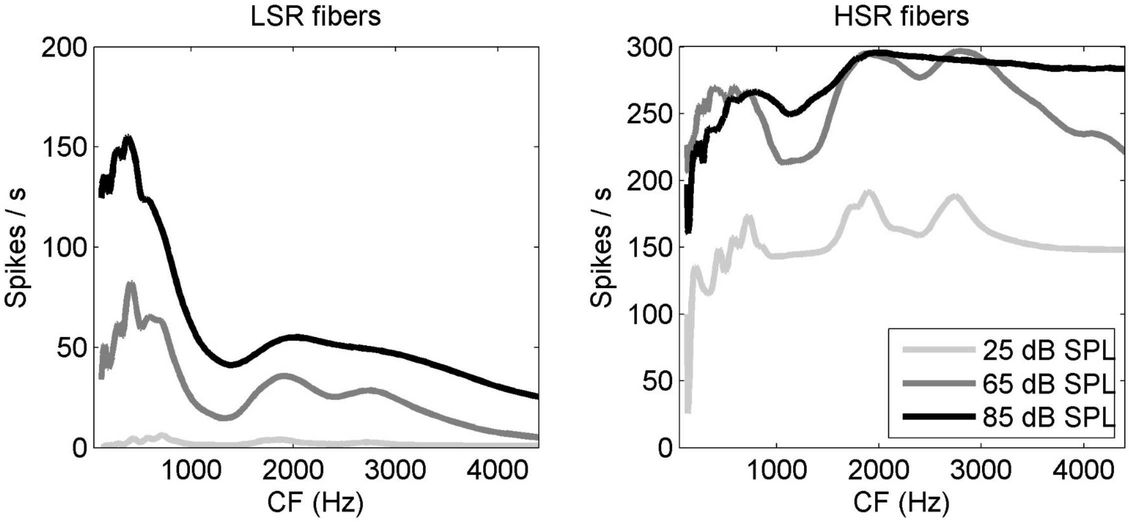

Figure 3: Firing rate distributions in response to the vowel [ɛ] in "head"[hɛːd] for low spontaneous rate fibers (left) and high spontaneous rate fibers (right) at 3 different sound intensities.

To get a sense of what these different auditory nerve fiber types may contribute to speech representations as different sound levels, we show in Figure 3 the firing rate distributions for the vowel [ɛ], much as in the right panel of Figure 2, but at 3 different sound levels (from a very quiet 25 dB SPL to a pretty loud 85 dB SPL, and for both LSR and HSR populations. As you can see, the LSR fibers (left panel) hardly respond at all at 25 dB, but the HSR fibers show clear peaks at the formant frequencies already at those very low sound levels. However, at the loud sound levels, most of the HSR fibers saturate, meaning that most of them are firing as fast as they can, so that the valleys between the formant peaks begin to disappear. One interesting consequence of this "division of labor" between HSR and LSR fibers for representing speech at low or high sound levels respectively is that it may provide an explanation why some people, particularly among the elderly, may complain of increasing inability to understand speech in situations with high background noise. Recent work from the work by Kujawa and Liberman (2015) has shown that, perhaps paradoxically, the less sound sensitive LSR fibers are actually more likely to be damaged during prolonged noise exposure. Patients with such selective fiber loss would still be able to hear quiet sounds quite well because their HSR fibers are intact, but they would find it very difficult to resolve sounds in high sound levels when HSR fibers are saturating and the LSR fibers that should encode spectral contrast at these high levels are missing. It has long been recognized that our ability to hear speech in noise tends to decline with age, even in those elderly who are lucky enough to retain normal auditory sensitivity (Stuart and Phillips 1996), and it has been suggested that cumulative, noise induced damage to LSR fibers such as that described by Kujawa and Liberman in their mouse model might pinpoint a possible culprit. Such "hidden hearing loss", which is not detectable with standard audiometric hearing tests that measure sensitivity to probe tones in quiet, can be a significant problem, for example by taking all the fun out of important social occasions, such as lively parties and get-togethers, which leads to significant social isolation. However, some recent studies have looked for, but failed to find, a clear link between greater noise exposure and poorer reception of speech in noise (Grinn et al. 2017; Grose et al. 2017), which would suggest that perhaps the decline in our ability to understand speech in noise as we age may be more to do with impaired representations of speech in higher, cortical centers than with impaired auditory nerve representations.

Of course, when you listen to speech, you don’t really want to have to ask yourself whether, given the current ambient sound levels, you should be listening to your HSR or your LSR auditory nerve fibers in order to get the best representation of speech formants, and one of the jobs of the auditory brainstem and midbrain circuitry is to combine information across these nerve fiber populations so that representations at midbrain and cortical stations will automatically adapt to changes both in mean sound level and in sound level “contrast” or variability, so that features like formants are efficiently encoded whatever the current acoustic environment happens to be (Dean et al. 2005; Rabinowitz et al. 2013; Willmore et al. 2016).

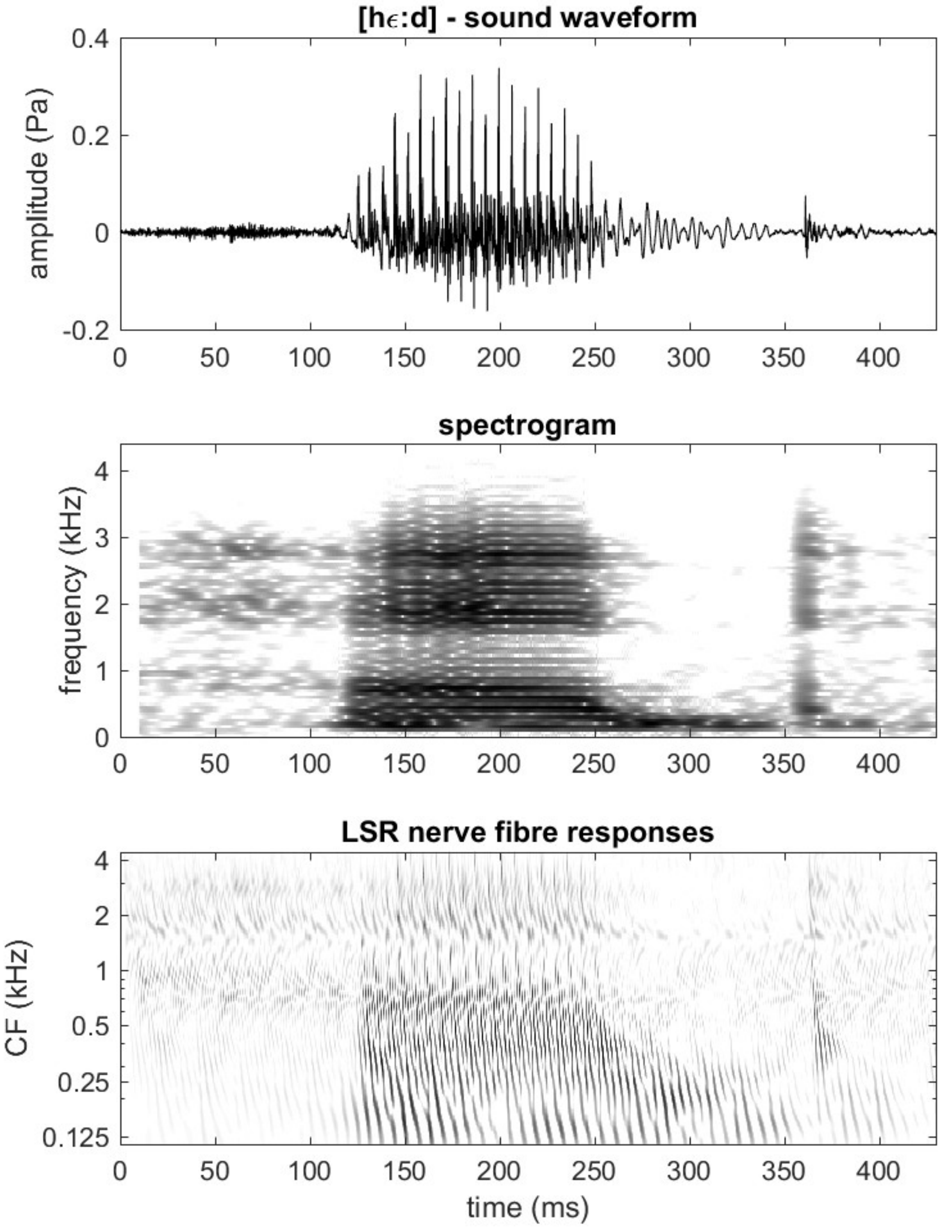

Figure 4: Waveform (top), spectrogram (middle) and simulated LSR auditory nerve fiber neurogram of the spoken word "head"[hɛːd].

As we have seen earlier, the tonotopic representation of speech sound spectra in the auditory nerve provides much information about speech formants, but not a great deal about harmonics which would reveal voicing or voice pitch. We probably owe much of our ability to nevertheless hear voicing and pitch easily and with high accuracy to the fact that, in addition to the small number of resolved harmonics, the auditory nerve delivers a great deal of so called “temporal fine structure information” to the brain. To appreciate what is meant by that, consider Figure 4, which shows the waveform (top), a spectrogram (middle) and an auditory nerve neurogram display (bottom) for a recording of the spoken word “head”. The neurogram was produced by computing firing rates of a bank of LSR auditory nerve fibers in response to the sound as a function of time using the model by Zilany et al. (2014). The waveform reveals the characteristic “energy arc” remarked upon by Greenberg (2006) for spoken syllables, with a relatively loud vowel flanked by relatively much quieter consonants. The voicing in the vowel is manifest in the large sound pressure amplitude peaks, which arise from the glottal pulse train, and which arise at regular interval of approximately 7 ms, that is a rate of approximately 140 Hz. This voice pitch is also reflected in the harmonic stack in the spectrogram, with harmonics at multiples of ~ 140 Hz, but this harmonic stack is not apparent in the neurogram. Instead we see that the nerve fiber responses rapidly modulate their firing rates to produce a temporal pattern of “bands” at time intervals which either directly reflect the 7 ms interval of the glottal pulse train (for nerve fibers with CFs below 0.2 kHz or above 1 kHz) or at intervals which are integer fractions (harmonics) of the glottal pulse interval. In this manner auditory nerve fibers convey important cues for acoustic features such as periodicity pitch by “phase locking” their discharges to salient features of the temporal fine structure of speech sounds with sub-millisecond accuracy.

As an aside, note that it is quite common for severe hearing impairment to be caused by an extensive loss of auditory hair cells in the cochlea, which can leave the auditory nerve fibers largely intact. In such patients it is now often possible to restore some hearing through cochlear implants which use electrode arrays implanted along the tonotopic array to deliver direct electrical stimulation to the auditory nerve fibers. The electrical stimulation patterns delivered by the twenty-odd electrode contacts provided by these devices are quite crude compared to the activity patterns created when the delicate dance of the basilar membrane is captured by some 3000 phenomenally sensitive auditory hair cells, but because coarsely resolving only a modest number of formant peaks is normally sufficient to allow speech sounds to be discriminated, the large majority of deaf cochlear implant patients do gain the ability to have pretty normal spoken conversations – as long as there is little background noise. Current cochlear implant processors are essentially incapable of delivering any of the temporal fine structure information which we have just described via the auditory nerve, and consequently cochlear implant users miss out on things like periodicity pitch cues which might help them separate out voices in a cluttered auditory scene. A lack of temporal fine structure can also affect the perception of dialect and affect in speech, as well as melody, harmony and timbre in music.

3. Subcortical pathways

As we have seen in Figure 1, the neural activity patterns we have just described are passed on first to the cochlear nuclei, and from then on through the superior olivary nuclei to the midbrain, thalamus and primary auditory cortex. We had also mentioned that each of these stations of the lemniscal auditory pathway has a tonotopic structure, so all we have learned in the previous section about tonotopic arrays of neurons representing speech formant patterns “neurogram style” thankfully still applies at each of these stations. But that is not to say that the neural representation of speech sounds does not undergo some transformations along these pathways. For example, the cochlear nuclei contain a variety of different neural cell types which receive different types of converging inputs from auditory nerve fibers, which may make them more or less sensitive to certain acoustic cues. So called “octopus” cells, for example, will collect inputs across a number of fibers across an extent of the tonotopic array, which will make them less sharply frequency tuned, but more sensitive to the temporal fine structure of sounds such glottal pulse trains (Golding and Oertel 2012). So called “bushy” cells in the cochlear nucleus are also very keen on maintaining temporal fine structure encoded in the timing of auditory nerve fiber inputs with very high precision, and passing this information on undiminished to the superior olivary nuclei (Joris et al. 1998). The nuclei of the superior olive receive converging (and, of course, tonotopically organized), inputs from both ears, which allows them to compute binaural cues to the direction that sounds may have come from (Schnupp and Carr 2009). Thus, firing rate distributions among neurons in the superior olive, and in subsequent processing stations, may provide information not just about formants or voicing of a speech sound, but also about whether the speech came from the left or right or straight ahead. This adds further “dimensions” to the neural representation of speech sounds in the brainstem, but much of what we have seen still applies: formants are represented by peaks of activity across the tonotopic array, and stimulus temporal fine structure is represented by the temporal fine structure of neural firing patterns. However, while the tonotopic representation of speech formants remains preserved throughout the subcortical pathways up to and including in the primary auditory cortex, temporal fine structure at fast rates of up to several hundred Hz is not preserved much beyond the superior olive. Maintaining sub-milisecond precision of firing patterns across a chain of chemical synapses and neural cell membranes which typically have temporal jitter and time constants in the milisecond range is not easy. To be up to the job, neurons in the cochlear nucleus and olivary nuclei have specialized synapses and ion channels which “more ordinary” neurons in the rest of the nervous system lack.

It is therefore generally thought that temporal fine structure cues to aspects such as the periodicity pitch of voiced speech sounds become “recoded” as one ascends the auditory pathway beyond the brainstem. Thus, from about the inferior colliculus onward, temporal fine structure at fast rates is increasingly less represented though fast and highly precise temporal firing patterns, but instead through neurons becoming “periodicity tuned” (Frisina 2001), meaning that their firing rates may vary as a function of the fundamental frequency of a voiced speech sound, in addition to depending on the amount of sound energy in a particular frequency band. Some early work on periodicity tuning in the inferior colliculus has led to the suggestion that this structure may even contain a “periodotopic map” (Schreiner and Langner 1988), with neurons tuned to different periodicities arranged along an orderly “periodotopic axis” running thought the whole length of the inferior colliculus, and with the periodotopic gradient more or less orthogonal to the tonotopic axis. Such an arrangement would be rather neat: periodicity being a major cue for sound features such as voice pitch, a periodotopic axis might, for example, physically separate out the representations of voices which differ substantially in pitch. But while some later neuroimaging studies seemed to support the idea of a periodotopic map in the inferior colliculus (Baumann et al. 2011), more recent, very detailed and comprehensive recordings with microelectrode arrays have shown conclusively that there are no consistent periotopic gradients running the width, breadth or depth of the inferior colliculus (Schnupp et al. 2015), nor are such periodotopic maps a consistent feature of primary auditory cortex (Nelken et al. 2008).

Thus, tuning to periodicity (and, by implication, voicing and voice pitch), as well as to cues for sound source direction, is widespread among neurons in the lemniscal auditory pathway from at least the midbrain upwards, but neurons with different tuning properties appear to be arranged in clusters without much over-arching systematic order, and their precise arrangement can differ greatly from one individual to the next. Thus, neural populations in these structures are best thought of as a patchwork of neurons which are sensitive to multiple features of speech sounds, including pitch, sound source direction and formant structure (Bizley et al. 2009; Walker et al. 2011), without much discernible overall anatomical organization other than tonotopic order.

4. Primary auditory cortex

So far, in the first half of this chapter we have talked about how speech is represented in the inner ear, auditory nerve, and along the subcortical pathways. However, in order for speech to be perceived, the percolation of auditory information must reach the cortex. Etymologically, the word “cortex” is Latin for *rind*, which is fitting as the cerebral cortex covers the outer surface of the brain – much like a rind overs your favorite citrus fruit. Small mammals like mice and trees shrews are endowed with relatively smooth cortices, while the cerebral cortices of larger mammals including humans (*Homo sapiens*)—but even more impressively, African bush elephants (*Loxodonta africana*)—exhibit a high degree of cortical “folding” (Prothero and Sundsten 1984). The more folded, wrinkled, or crumpled your cortex, the more surface area can fit into your skull. This is important because a larger cortex (relative to body size) means more neurons, and more neurons generally means more computational power (Jerison 1973). For example, in difficult noisy listening conditions, the human brain appears to recruit additional cortical regions (Davis and Johnsrude 2003) which we shall come back to in the next few sections. In this section, we begin our journey through the auditory cortex by touching on the first cortical areas to receive auditory inputs: the "primary auditory cortex".

Anatomy and tonotopicity of the human primary auditory cortex

In humans the primary auditory cortex (A1) is located around a special wrinkle in the cortical sheet, known as "Heschl's gyrus" (HG). A gyrus, by the way, is the word used to describe a "ridge" where the cortical sheet is folded outward, while a sulcus describes an inward fold or valley. There are in fact multiple HG in each brain. First, all people have at least two HG, one in each cerebral hemisphere (the left and right halves of the visible brain). These are positioned along the superior aspect of each temporal lobe. In addition, some brains have a duplication in the HG, meaning that one or both hemispheres will have two ridges instead of one (Da Costa et al. 2011). This anatomical factoid can be useful for identifying A1 in real brains (as we shall see in Figure 5). However, the gyri are only used as landmarks: what matters is the sheet of neurons in and around HG, not whether that area is folded once or twice. This sheet of neurons area receives connections from the subcortical auditory pathways, most prominently via the medial geniculate body of the thalamus (see Figure 1 and the previous section). When the cortex is smoothed, *in silico,* using computational image processing, the primary auditory cortex can be shown to display the same kind of tonotopic maps that we observed in the cochlea and in subcortical regions. This has been known from invasive microelectrode recordings in laboratory animals for decades and can be confirmed to also be the case in humans using non-invasive MRI (Magnetic Resonance Imaging) by playing subjects stimuli at different tones and then modeling the optimal cortical responses to each tone. This use of functional MRI (fMRI) results in the kind of tonotopic maps shown in Figure 5.

Figure 5 depicts a flattened view of the left-hemisphere cortex colored in dark gray. Superimposed onto the flattened cortex is a tonotopic map (grayscale corresponding to the color bar on the bottom right). Over the surface of the tonotopic map, each point has a preferred stimulus frequency, in Hz, and if we follow the dotted arrow across HG, we find a gradient pattern of responses corresponding to low frequencies, high frequencies, and then low frequencies again. Given this tonotopic organization of the primary auditory cortex, which is in some respects not that different from the tonotopy seen in lower parts of the auditory system, we might therefore expect the nature of the representation of sounds (including speech sounds) in this structure to be still to a large extent "spectrogram like", that is, if we were to read out the firing rate distributions along the frequency axes of these areas while speech sounds are represented, then this "neurogram" of activity would exhibit dynamically shifting peaks and troughs that reflect the changing formant structure of the presented speech. That this is indeed the case has been shown in animal experiments by Engineer et al. (2008), who, in one set of experiments, trained rats to discriminate a large set of consonant-vowel syllables, and in another recorded neurograms for the same set of syllables from the primary cortices of anesthetized rats using microelectrodes. They found, firstly, that rats can learn to discriminate most American English syllables easily, but are more likely to confuse syllables which humans too might find more similar and easier to confuse (e.g. “sha” vs “da” is easy, but “sha” vs “cha” is harder). Second, Engineer et al. found that how easily rats can discriminate between two speech syllables can be predicted by how different the primary auditory cortex neurograms for these syllables are.

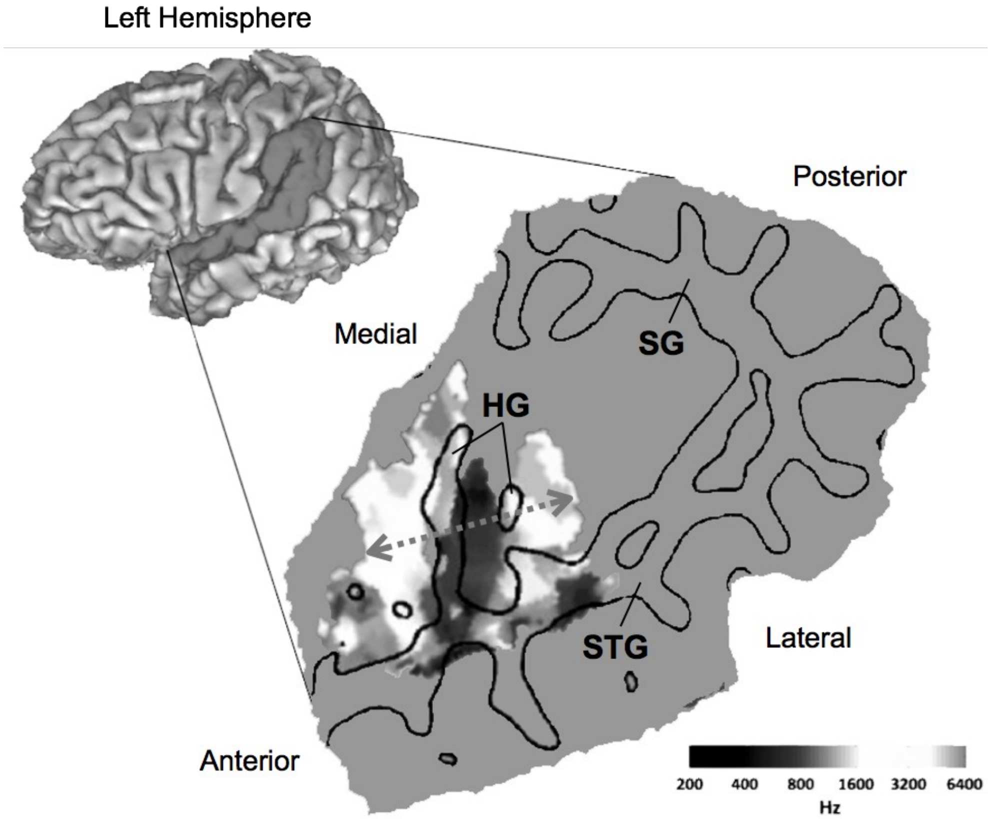

Figure 5: Tonotopic map. HG = Heschl’s Gyrus; STG = Superior Temporal Gyrus; SG = Supramarginal Gyrus; Hz = Hertz. Figure adapted with permission from Humphries et al. (2010).

These data would suggest that the representation of speech in primary auditory cortex is still a relatively “unsophisticated” time-frequency representation of sound features, with very little in the way of recognition or categorization or interpretation. Calling primary auditory cortex unsophisticated is, however, probably doing it an injustice, however, because other animal experiments indicate that neurons in primary auditory cortex can, for example, change their frequency tuning quickly and substantially if a particular task requires attention to be directed to a particular frequency band (Edeline et al. 1993; Fritz et al. 2003). Primary auditory cortex neurons can even become responsive to stimuli or events that aren’t auditory at all if these events are firmly associated with sound related tasks that an animal has learned to master (Brosch et al. 2005). Nevertheless, it is currently thought that the neural representations of sounds and events in primary auditory cortex are probably based on detecting relatively simple acoustic features, and are probably not specific to speech or vocalizations, given that primary cortex does not seem to have any obvious preference for speech over non-speech stimuli. In the human brain, to find the first indication of areas that appear to prefer speech stimuli to other, non-speech sounds, one must move beyond the tonotopic maps of the primary auditory cortex (Belin et al. 2000; Scott et al. 2000).

In the next few sections we will continue our journey through the auditory system into cortical regions that appear to make specialized contributions to speech processing, and which are situated in the temporal, parietal and frontal lobes. We will also talk about how these regions communicate with each other in noisy contexts and during self-generated speech, when information from the (pre-)motor cortex influences speech perception, and we will talk about representations of speech in time. Figure 6 introduces the regions and major connections to be discussed. You may want to refer back to it at times. In brief, we will consider the superior temporal gyrus (STG), the premotor cortex (PMC), and then loop back into the STG to discuss how brain regions in the auditory system work together as part of a dynamic network.

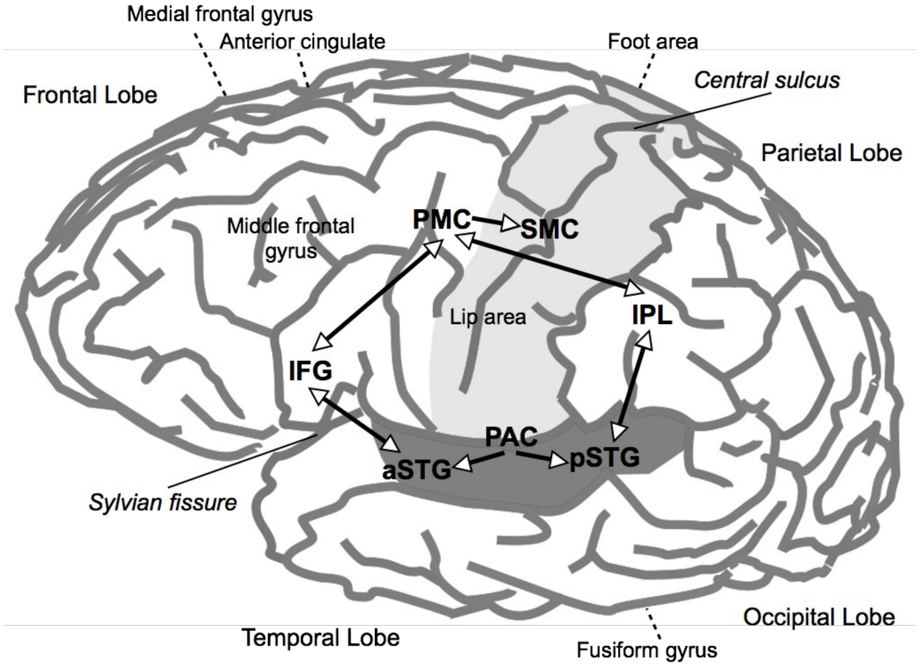

Figure 6: A map of cortical areas involved in the auditory representation of speech. PAC = Primary Auditory Cortex; STG = Superior Temporal Gyrus; aSTG = anterior STG; pSTG = posterior STG; IFG = Inferior Frontal Gyrus; PMC = Pre-Motor Cortex; SMC = Sensori-Motor Cortex; IPL = Inferior Parietal Lobule. Dashed lines indicate medial areas. Adapted with permission from Rauschecker and Scott (2009).

5. What does “higher order” cortex add?

All of the systems that we have reviewed so far on our journey along the auditory pathway have been general auditory processing systems. So, although important for speech processing, their function is not speech specific. For example the cochlea converts pressure waves into electrical impulses, whether the pressure waves encode a friendly “hello” or the sound of falling rain; the subcortical pathways process and propagate these neural signals to the primary auditory cortex, in ways that do not depend on whether they encode a phone conversation or barking dogs or noisy traffic; and the primary auditory cortex exhibits a tonotopic representation of an auditory stimulus, whether that stimulus was an extract of a Shakespearean soliloquy or of Ravel’s Boléro. In this section, we encounter a set of cortical areas that preferentially process speech over other kinds of auditory stimuli. We will also describe deeply revealing new work into the linguistic/phonetic representation of speech, obtained using surgical recordings in human brains.

Speech preferential areas

That areas of the brain exist which are necessary for the understanding of speech, but not for general sound perception, has been known since the 19th century, when the German neurologist, Carl Wernicke, associated the aphasia that bears his name with damage to the STG (Wernicke 1874). Wernicke's eponymous area was, incidentally, reinterpreted by later neurologists to refer only to the posterior third of the STG and adjacent parietal areas (Bogen and Bogen 1976), although some disagreement about its precise boundaries continues until this day (Tremblay and Dick 2016).

With the advent of fMRI at the end of the 20th century, the posterior STG was confirmed to respond more strongly to vocal sounds than to non-vocal sounds (e.g. speech, laughter, or crying as compared to the sounds of wind, galloping, or cars) (Belin et al. 2000). Neuroimaging also revealed a second, anterior area in the STG, which responds more to vocal than to non-vocal sounds (Belin et al. 2000). These voice-preferential areas can be found in both hemispheres of the brain. Additional studies have shown that it is not just the voice but also intelligible speech that excites these regions, although with speech processing being more specialized in the left hemisphere (Scott et al. 2000). Anatomically, anterior and posterior STG receive white-matter connections from the primary auditory cortex, and in turn feed two auditory processing streams, one antero-ventral, which extends into the inferior frontal cortex, and another postero-dorsal that curves into the inferior parietal lobule. The special function of these streams remains a matter of debate. For example, Rauschecker and Scott (2009) propose that the paths differ in processing “what” and “where” information in the auditory signal, where “what” refers to recognizing the cause of the sound (e.g. it's a thunderclap) and “where” to locating the sound's spatial location (e.g. to the west). Another, more linguistic suggestion is that the ventral stream is broadly semantic, whereas the dorsal stream might be described as more phonetic in nature (Hickok and Poeppel 2004). Whatever the functions, however, there do appear to be two streams diverging around the anterior and posterior STG.

Over the years, these early STG results have been replicated many times using neuroimaging (Price 2012). Each technique for observing activity of the human brain, whether non-invasive magnetoencephalography (MEG) or functional magnetic resonance imaging (fMRI), or invasive, surgical techniques like electrocorticography (ECoG), described in the next section, all have their respective limitations and shortcomings. It is therefore reassuring that the insights into the neuroanatomy of speech comprehension established by methods like MEG or fMRI, which can image the whole brain, are both confirmed and extended by studies using targeted surgical techniques like ECoG.

Auditory phonetic representations in the superior temporal gyrus

ECoG, which involves the placement of electrodes directly onto the surface of the brain, cannot easily record from the primary auditory cortex. This is because the primary auditory cortex is tucked away inside the Sylvian fissure, along the dorsal aspect of the temporal lobe. On the other hand, because ECoG measures the summed post-synaptic electrical current of neurons with millisecond resolution, it is sensitive to rapid neural responses at the timescale of individual syllables, or even individual phones. By contrast fMRI measures hemodynamic responses; these are changes in blood flow which are related to neural activity but occur on the order of seconds. In recent years, the use of ECoG has revolutionized the study of speech in auditory neuroscience. An exemplar of this can be found in a recent paper (Mesgarani et al. 2014).

Mesgarani et al. (2014) used ECoG to learn about the linguistic/phonetic representation of auditory speech processing in the STG of six epileptic patients. These patients listened passively to spoken sentences taken from the TIMIT corpus (Garofolo et al. 1993), while ECoG was recorded from their brains. These ECoG recordings were then analyzed to discover patterns in the neural responses to individual speech sounds (for a summary of the experimental setup, see Figure 7, panels A-C). The authors used a phonemic analysis of the TIMIT dataset to group the neural responses, at each electrode, according to the phoneme that caused it. For examples see panel D of Figure 7, which allows the comparison of responses to different speech sounds for a number of different sample electrodes labeled e1 to e5. The key observation here is that an electrode like e1 gives similar responses for /d/ and /b/ but not for /d/ and /s/, and the responses at each of the electrodes shown will respond strongly for some groups of speech sounds, but not others. Given these data we can ask the following questions. Do STG neurons group, or “classify”, speech segments through the similarity of their response patterns? And if so, which classification scheme do they use?

Linguists and phoneticians often analyze individual speech sounds into “feature” classes, based, for example, on either the “manner” or the “place of articulation” that is characteristic for that speech sound. Thus, /d/, /b/, and /t/ are all members of the “*plosive*” manner of articulation class because they are produced by an obstruction followed by a sudden release of air through the vocal tract, and /s/ and /f/ belong to the “*fricative*” class because both are generated by turbulent air hissing through a tight constriction in the vocal tract. At the same time, /d/ and /s/ also belong to the “*alveolar*” place of articulation class because for both phonemes the tip of the tongue is brought up toward the alveolar ridge just behind the top row of the teeth. In contrast, /b/ has a “*labial*” place of articulation because to articulate /b/ the airflow is constricted at the lips. Manner features are usually associated with particular acoustic characteristics. Plosives involve characteristically brief intervals of silence followed by a short noise burst, while fricatives exhibit sustained aperiodic noise spread over a wide part of the spectrum. Classifying speech sounds by place and manner of articulation is certainly very popular among speech scientists, and is also implied in the structure of the international phonetic alphabet (IPA), but it is by no means the only possible scheme. Speech sounds can also be described and classified according to alternative acoustic properties or perceptual features, such as loudness and pitch. An example feature that is harder to characterize in articulatory or acoustic terms is “sonority”. Sonority defines a scale of perceived loudness (Clements 1990) such that vowels are the most sonorous, glides are the next most sonorous, then liquids, nasals, and finally obstruents (i.e. fricatives and plosives). Despite the idea of sonority as a multi-tiered scale, phonemes are sometimes lumped into two groups of *sonorant* and *non-sonorant*, with everything but the obstruents counting as sonorants.

As these examples illustrate, there could in principle be many different ways in which speech sounds are grouped. To ask which grouping is “natural” or “native” for the STG, Mesgarani et al. (2014) used hierarchical clustering of neural responses to speech, examples of which can be seen in the ECoG recordings depicted in Figure 7, panel D. The results of the clustering analysis follows in Figure 7, panels E-G. Perhaps surprisingly, Mesgarani et al. (2014) discovered that the STG was organized primarily by manner of articulation features and secondarily by place of articulation features. The prominence of manner of articulation features can be seen by clustering the phonemes directly (Figure 7F). For example, on the right-side dendrogram we find neat clusters of plosives /d b g p k t/, fricatives /ʃ z s f θ/, and nasals /m n ŋ/. Manner of articulation features also stand out when the electrodes are clustered (Figure 7G). By following a column up, from the bottom dendrogram, one can find the “darkest” cells (those with the greatest selectivity for phonemes) and then follow these rows to the left to identify the phonemes to which the electrode signal was strongest. The electrode indexed by the leftmost column, for example, recorded neural activity that appeared selective for the plosives /d b g p k t/. In this way, one may also find electrodes that respond to both manner and place of articulation features. For example, the fifth column from the left responds to the bilabial plosives /b p/. Thus, the types of features that phoneticians have for a long time employed for classifying speech sounds turn out to be reflected in the criteria by which neural responses across the STG can be shown to group speech sounds. Mesgarani et al. (2014) argue that this pattern of organization, prioritizing manner over place of articulation features, is most consistent with auditory-perceptual theories of feature hierarchies (Stevens 2002; Clements 1985). Auditory-perceptual theories contrast, for instance, with articulatory or gestural theories, which Mesgarani et al. (2014) assert would have first prioritized the place of articulation features (Fowler 1986).

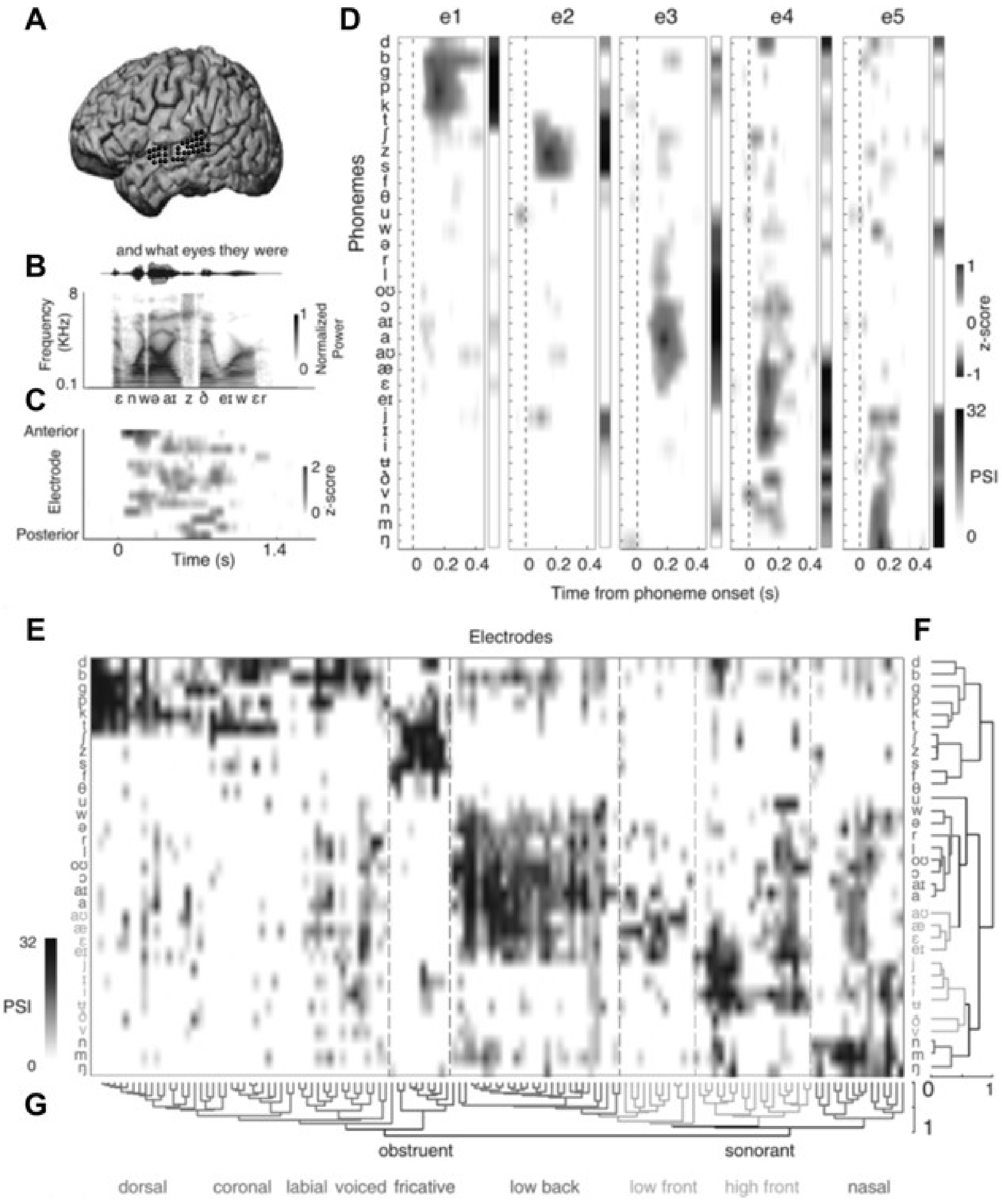

Figure 7: Feature-based representations in the human STG. Panel A shows left hemisphere cortex with black dots indicating ECoG electrodes. B shows an example acoustic stimulus ("and what eyes they were"), including orthography, waveform, spectrogram, and IPA transcription. C shows time-aligned neural responses to the acoustic stimulus. The electrodes (y-axis) were sorted spatially (anterior-to-posterior), with time (in seconds) along the x-axis. D shows sample phoneme responses by

electrode. For five electrodes (e1, …, e5), the plots show cortical selectivity for English phonemes (y-axis) as a function of time (x-axis) with phoneme onsets indicated by vertical dashed lines. The Phoneme Selectivity Index (PSI) is a summary over time of how selective the cortical response is for each phoneme. E shows phoneme responses (PSIs) for all electrodes, arranged for hierarchical clustering analyses. F and G show clustering analyses by phoneme and by electrode. These show how phonemes and electrodes are grouped, respectively, with reference to phonetic features. For example, F shows that electrodes can be grouped by selectivity to obstruents and sonorants. Adapted from Mesgarani et al. (2014) with permission of the publisher.

The clustering analyses in Figure 7 (panels F and G) are doubly rich: they at the same time support a broadly auditory-perceptual view of sound representations in the STG while also revealing limitations of that view. For instance, on the right-side cluster (panel F), we find that the phonemes /v/ and /ð/ do not cluster with the other fricatives /ʃ z s f θ/. Instead these fricatives, /v/ and /ð/, cluster with the sonorants. Moreover, /v/ and /ð/ are most closely clustered in a group of high front and central vowels and glides /j ɪ i ʉ/. This odd-grouping might reflect noise at some level of the experiment or analysis, however it raises the intriguing possibility that the STG actually groups /j ɪ i ʉ v ð/ together, and thus does not strictly follow established phonetic conventions. Therefore in addition to articulatory, acoustic, and auditory phonetics, studies like this on the cortical response to speech may pave the way to innovative *neural* feature analyses. We would however like to emphasize that these are early results in the field. The use of discrete segmental phonemes may, for example, be considered a useful first approximation to analyses using more complex, overlapping feature representations.

Auditory phonetic representations in the sensory-motor cortex

From the STG, we turn now to a second cortical area. The vSMC, or ventral sensory-motor cortex, is better known for its role in speech production than speech comprehension (Bouchard et al. 2013). This part of cortex, near the ventral end of the SMC (see Figure 6), contains the primary motor and somatosensory areas, which send motor commands to and receive touch and proprioceptive information from the face, lips, jaw, tongue, velum, and pharynx. The vSMC plays a key role in controlling the muscles associated with these articulators, and is further involved in monitoring feedback from the sensory nerves in these areas when we speak. Less widely known is that vSMC also plays a role in speech perception. We know, for example, that a network including frontal areas becomes more active when the conditions for perceiving speech become more difficult (Davis and Johnsrude 2003), such as when there is background noise or the sound of multiple speakers overlaps (contrast easy listening conditions when distractions like these are absent). This context-specific recruitment of speech production areas might signal that they play an “auxiliary role” in speech perception, by providing additional computational resources when the STG is overburdened. As an auxiliary auditory system, which is primarily dedicated to coordinating the articulation of speech, we might ask how the vSMC represents heard speech. Does the vSMC represent the modalities of overt and heard speech similarly or differently? Is the representation of heard speech in the vSMC similar or different to that of the STG?

ECoG studies of speech production (Bouchard et al. 2013; Cheung et al. 2016) suggest that place of articulation features take primacy over the manner of articulation features in the vSMC, which is the reverse of what we described above for the STG (Mesgarani et al. 2014). Given that vSMC contains a map of body parts like the lips and tongue, it makes sense that this region be represented by place of articulation features, rather than by manner of articulation features. But does this representation in vSMC hold during both speech production and comprehension? Our starting hypothesis might be that, yes, the feature representations in vSMC will be the same regardless of task. There is even some theory to back this up. For example, there have been proposals, like the motor theory of speech perception (Liberman et al. 1967; Liberman and Mattingly 1985) or the analysis-by-synthesis theory (Stevens 1960) that view speech perception as a kind of active rather than passive process. Analysis-by-synthesis says that speech perception involves trying to match what you hear to what your own mouth, and other articulators, would have needed to do to produce what you heard. Speech comprehension would therefore involve the active process of covert speech production. Following this line of thought, we might suppose that what the vSMC does, when it is engaged in deciphering what your friend is asking you at a noisy cocktail party, is in some sense the same as what the vSMC does when it is used to articulate your reply. Because we know that place of articulation features take priority over manner of articulation features in the vSMC during a speech-production task (i.e. reading consonant-vowel syllables aloud), we might hypothesize that place of articulation features will similarly take primacy during passive listening. Interestingly, despite being predicted by theory, this prediction is wrong.

When Cheung et al. (2016) examined neural response patterns in the vSMC while subjects listened to recordings of speech, they found that, as in the STG, it was the *manner of articulation* features that took precedence. In other words, representations in vSMC were conditioned by task: during speech production the vSMC favored place of articulation features (Bouchard et al. 2013; Cheung et al. 2016); but during speech comprehension, the vSMC favoured manner of articulation features (Cheung et al. 2016). As we discussed above, the STG is also organised according to manner of articulation features when subjects listen to speech (Mesgarani et al. 2014). Therefore the representations in these two areas, STG and vSMC, appear to use a similar type of code when they represent heard speech.

To be more concrete, Cheung et al. (2016) recorded ECoG from the STG and vSMC of subjects performing two tasks. One task involved reading aloud from a list of consonant-vowel syllables (e.g. “ba”, “da”, “ga”), while the other task involved listening to recordings of people producing these syllables. Instead of using hierarchical clustering like Mesgarani et al. (2014) did in their study of the STG, Cheung et al. (2016) used a dimensionality-reduction technique called multidimensional scaling (MDS) but with the similar goal of describing the *structure* of phoneme representations in the brain during each task (Figure 8). For the speaking task, the dimensionality-reduced vSMC representations for eight sounds could be linearly separated into three place of articulation features: labial /p b/, alveolar /t d s ʃ/, and velar /k g/ (see Figure 8, panel D). The same phonemes could not be linearly separated into place of articulation features in the listening task (Figure 8, panel E), however they could be linearly separated into another set of features (Figure 8, panel G): voiced plosives /d g b/, voiceless plosives /k t p/, and fricatives /ʃ s/. These are the same manner of articulation and voicing features that characterize the neural responses in STG to heard speech (Figure 8, panel F). Again, the implication is that the vSMC has two codes for representing speech, suggesting that there are either two distinct but anatomically-intermingled neural populations in vSMC, or the same population of neurons is capable of operating in two very different representational modes. Unfortunately, the spatial resolution of ECoG electrodes is still too coarse to resolve this ambiguity, so other experimental techniques will be needed. For now, we can only say that during speech production, the vSMC uses a feature analysis that emphasizes place of articulation features, but during speech comprehension, the vSMC uses a feature analysis that instead emphasizes manner features and voicing. An intriguing possibility is that the existence of similar representations for heard speech in STG and vSMC may play an important role in the communication, or *connectivity*, between distinct cortical regions—a topic we touch on in the next section.

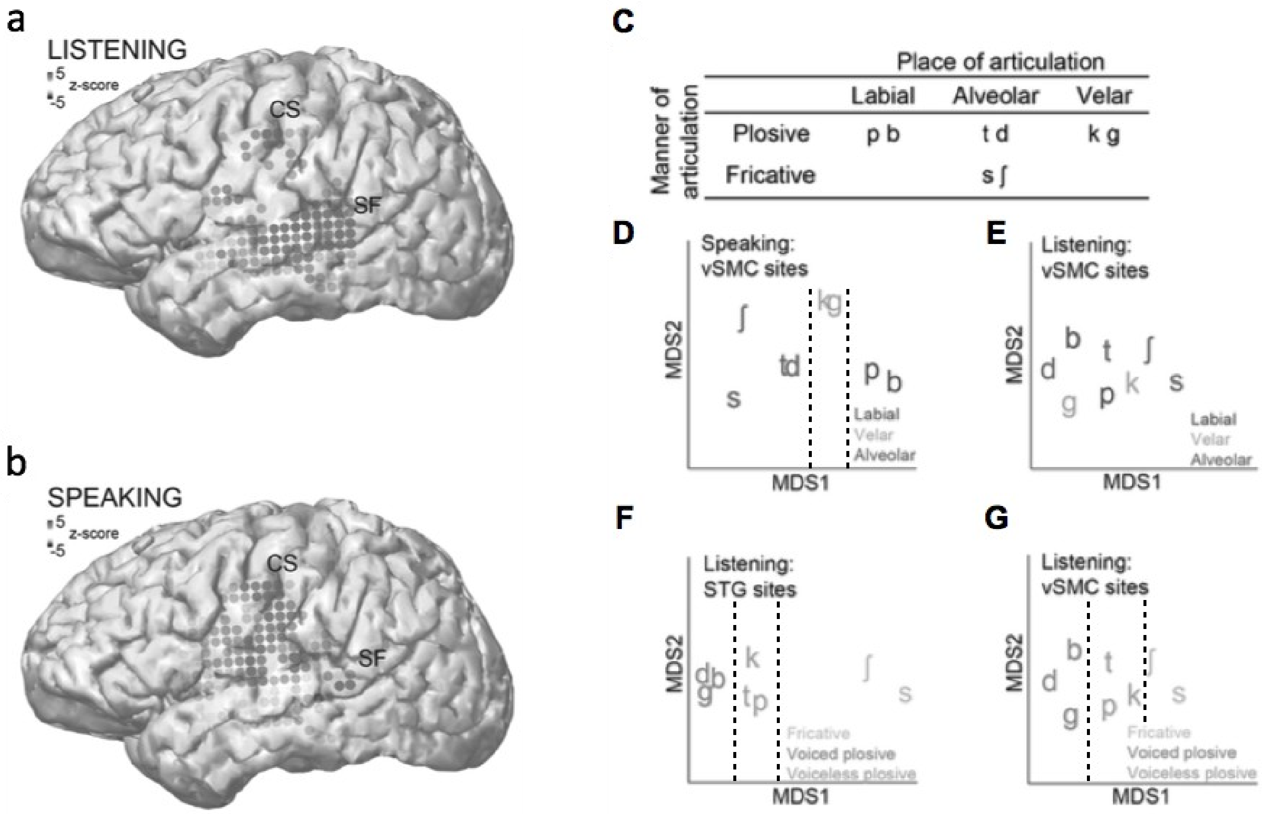

Figure 8: Feature-based representations in the human sensori-motor cortex. A and B show the most significant electrodes (gray dots) for the listening and speaking tasks. C presents a feature analysis of the consonant phonemes used in the experiments. The left phoneme in each pair is unvoiced and the right phoneme is voiced (e.g. /p/ is unvoiced and /b/ is voiced). D-G are discussed in the main text; each panel shows a low-dimensional projection of the neural data where distance between phoneme representations is meaningful (i.e. phonemes that are close to each other are represented similarly in the neural data). The dotted lines show how groups of phonemes can be linearly separated (or not) according to place of articulation, manner of articulation, and voicing features. Figure adapted with permission from Cheung et al. (2016).

6. Systems-level representations and temporal prediction

Our journey through the auditory system has focused on specific regions, and on the auditory representation of speech in these regions. However representations in the brain are not limited to isolated islands of cells, but also rely upon constellations of regions that relay information within a network. In this section, we touch briefly on the topic of systems-level representations of speech perception and on the related topic of temporal prediction, which is at the heart of why we have brains in the first place.

Auditory feedback networks

One way to appreciate the dynamic interconnectedness of the auditory brain is to consider the phenomenon of auditory suppression. Auditory suppression manifests, for example, in the comparison of STG responses when we listen to another person speak and when we speak ourselves, and thus hear the sounds we produce. Electrophysiological studies have shown that auditory neurons are suppressed in monkeys during self vocalization (Müller-Preuss and Ploog 1981; Eliades and Wang 2008; Flinker et al. 2010). This finding is consistent with fMRI and ECoG results in humans, showing that activity in the STG is suppressed during speech production compared to speech comprehension (Eliades and Wang 2008; Flinker et al. 2010). The reason for this auditory suppression is thought to be an internal signal (“efference copy”) received from another part of the brain, such as the motor or premotor cortex, which has “inside information” about external stimuli when those external stimuli are self-produced (von Holst and Mittelstaedt 1950). The brain’s use of this kind of inside information is not, incidentally, limited to the auditory system. Anyone who has failed to tickle themselves has experienced another kind of sensory suppression, again thought to be based on internally-generated expectations (Blakemore et al. 2000).

Auditory suppression in the STG is also a function of language proficiency. As an example, Parker Jones et al. (2013) explored the interactions between pre-motor cortex (PMC) and two temporal areas (sSTG and pSTG) when native and non-native English speakers performed speech-production tasks such as reading and picture naming in an MRI scanner. The fMRI data were then subjected to a kind of connectivity analysis, which can tell you which regions influenced which other regions of the brain. Technically, the observed signals were deconvolved to model the effect of the hemodynamic response, and the underlying neural dynamics were inferred by inverting a generative model based on a set of differential equations (Friston et al. 2003; Daunizeau et al. 2011). A positive connection between two regions, A and B, means that when the response in A is strong, the response in B will *increase* (i.e. B will have a positive derivative). Likewise, a negative connection means that when the response in A is strong, the response in B will *decrease* (B will have a negative derivative). Between the PMC and temporal auditory areas, Parker Jones et al. (2013) observed significant negative connections, implying that brain-activity in the PMC caused a decrease in auditory temporal activity consistent with “auditory suppression”. However auditory suppression was only observed in the native English speakers. In non-native speakers, there was no significant auditory suppression, but there was a positive effect between pSTG and PMC consistent with the idea of “error feedback”. The results suggest that PMC sends signal-canceling, top-down predictions to aSTG and pSTG. These top-down predictions are stronger if you are a native speaker and more confident about what speech sounds you produce. In non-native speakers, the top-down predictions cancelled less of the auditory input, and a bottom-up learning signal (“error”) was fed back from the pSTG to the PMC. Interestingly, as the non-native speakers became more proficient, the learning signals were observed to decrease, so that the most highly-proficient non-native speakers were indistinguishable from native speakers in terms of error feedback.

The example of auditory suppression argues for a systems-level view of speech comprehension that includes both auditory and premotor regions of the brain. Theoretically, we might think of these regions as being arranged in a functional hierarchy, with PMC located above both aSTG and pSTG. “Top-down” predictions may thus be said to descend from PMC to aSTG and pSTG, while “bottom-up” errors percolate in the opposite direction, from pSTG to PMC. We note that the framework used to interpret the auditory suppression results, “predictive coding”, subtly inverts the view that perceptual systems in the brain passively extract knowledge from the environment; instead, it proposes that these systems are actively trying to predict their sense experiences (Ballard et al. 1983; Mumford 1992; Kawato et al. 1993; Dayan et al. 1995; Rao and Ballard 1999; Friston and Kiebel 2009). In a foundational sense, predictive coding frames the brain as a forecasting machine, evolved to minimize surprises and to anticipate, and not merely react to, events in the world (Wolpert et al. 2001). This is not to say that people are necessarily prediction machines but rather to conjecture that perceptual systems in our brains, at least sometimes, predict sense experiences.

Temporal prediction

The importance of “prediction” as a theme and as a hypothetical explanation for neural function also goes beyond explicit modeling in neural networks. We can invoke the idea of temporal prediction even when we do not know about the underlying connectivity patterns. Speech, for example, does not consist of a static set of phonemes. Rather speech is a continuous sequence of events, such that hearing part of the sequence gives you information about other parts that you have yet to hear. In phonology the sequential dependency of phonemes is called “phonotactics” and can be viewed as a kind of prediction. That is, if the sequence /st/ is more common than /sd/, because /st/ occurs in syllabic onsets, then it can be said that /s/ predicts /t/ (more than /s/ predicts /d/). This use of phonotactics for prediction is made explicit in machine learning, where predictive models (e.g. bigram and trigram models historically, or, more recently, recurrent neural networks) have played an important role in the development and commercial use of speech-recognition technologies (Jurafsky and Martin 2014; Graves & Jaitly 2014).

In neuroscience, the theme of prediction comes up in “masking” and “perceptual restoration” experiments. One remarkable ECoG study, by Leonard et al. (2016), played subjects recordings of words in which key phonemes were masked by noise. For example, a subject might have heard /fæ#tr/, where the /#/ symbol represents a brief noise burst masking the underlying phoneme. In this example, the intended word is ambiguous: it could have been /fæstr/ ‘faster’ or /fæktr/ ‘factor’. So, by controlling the context in which the stimulus was presented, Leonard et al. (2016) were able to manipulate subjects to hear one word or another. In the sentence “On the highway he drives his car much /fæ#tr/”, we expect the listener to perceive the word /fæstr/ ‘faster’. In another sentence, that expectation was modified so that subjects perceived the same noisy segment of speech as /fæktr/ ‘factor’. Leonard et al. (2016) then used a technique called “stimulus reconstruction”, by which it is possible to infer rather good speech spectrograms from intracranial recordings (Mesgarani et al. 2008; Pasley et al. 2012). Spectrograms reconstructed from masked stimuli showed that the STG had “filled in” the missing auditory representations (Figure 9). For example, when the context was modulated so that subjects perceived the ambiguous stimulus as /fæstr/ ‘faster’, the reconstructed spectrogram was shown to contain an imagined fricative [s] (Figure 9, panel E). When subjects perceived the word as /fæktr/ ‘factor’, the reconstructed spectrogram contained an imagined stop [k] (Figure 9, panel F). In this way, Leonard et al. (2016) demonstrated that auditory representations of speech are sensitive to their temporal context.

In addition to “filling in” missing phonemes, the idea of temporal prediction can be invoked as an explanation of how the auditory system accomplishes one of its most difficult feats: “selective attention”. Selective attention is often alluded to as the “cocktail party problem”, because many people have experienced the use of selective attention in a busy, noisy party to isolate one speaker’s voice from the cacophonous mixture of many. Mesgarani and Chang (2012) simulated this cocktail party experience (unfortunately without the cocktails) by simultaneously playing two speech recordings to their subjects, one in each ear. The subjects were asked to attend to the recording presented to a specific ear and ECoG was used to record neural responses from the STG. Using the same stimulus reconstruction technique that Leonard et al. (2016) used, Mesgarani and Chang (2012) took turns reconstructing the speech that was played to each ear. Despite the fact that acoustic energy entered both ears and presumably propagated up the subcortical pathway, Mesgarani and Chang (2012) found that, by the STG, only the attended speech stream could be reconstructed. To the STG, it was as if the unattended stream did not exist.

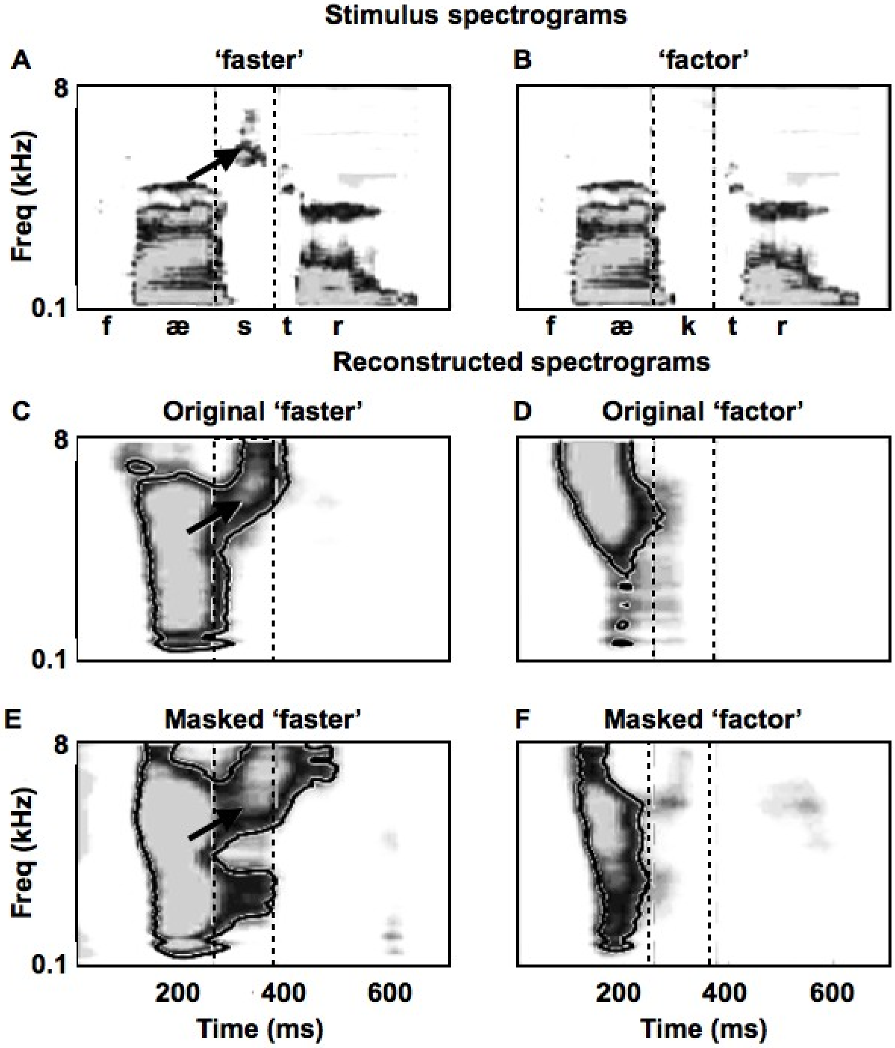

Figure 9: The human brain reinstates missing auditory representations. A and B show spectrograms for two words, /fæstr/ 'faster' and /fæktr/ 'factor'. The segments of the spectrograms for /s/ and /k/ are indicated by dashed lines. The arrow in A points to aperiodic energy in higher frequency bands associated with fricative sounds like [s], which is absent in B. C and D show neural reconstructions when subjects heard A and B. E and F show neural reconstructions when subjects heard the masked stimulus /fæ#tr/. In E, subjects heard "On the highway he drives his car much /fæ#tr/", which caused them to interpret the masked segment as /s/. In F, the context suggested that the masked segment should be /k/. Adapted with permission from Leonard et al. (2016).

We know from a second cocktail-party experiment (which again did not include any actual cocktails) that selective attention is sensitive to how familiar the hearer is to each speaker. In their behavioral study, Johnsrude et al. (2013) recruited a group of subjects that included multiple spouses. If you were a subject in the study, your partner's voice was sometimes the "target" (i.e. attended speech); your parter's voice was sometimes the "distractor" (i.e. unattended speech); and sometimes both target and distractor voices belonged to other subjects' spouses. Not only did Johnsrude et al. (2013) find that subjects were better at recalling semantic details of the attended speech when the target speaker was their partner, but subjects also performed better when their spouse played the role of distractor, compared to when both target and distractor roles were played by strangers. In effect, Johnsrude et al. (2013) amusingly showed that people are better at ignoring their own spouses than they are at ignoring strangers. Given that hearers can "fill in" missing information when it can be predicted from context (Leonard et al. 2016), it makes sense that subjects should comprehend the speech of someone familiar who they are better at predicting than a stranger. Given that native speakers are better than non-native speakers at suppressing the sound of their own voices (Parker Jones et al. 2013), it also makes sense that subjects should be better able to suppress the voice of their spouse—again assuming that their spouse's voice is more predictable to them than a stranger's. Taken together, this suggests that the mechanism behind selective attention is, again, prediction. So, while Mesgarani and Chang (2012) may be unable to reconstruct the speech of a distractor voice from ECoG recordings in the STG, it may be that "higher" brain regions will nonetheless contain a representation of the distractor voice for the purpose of suppressing it. An as-yet unproven hypothesis is that the increased neural activity in frontal areas, observed during noisy listening conditions (Davis and Johnsrude 2003), might be busy representing background noise or distractor voices, so that these sources may be filtered out of the mixed input signal. One way to test this might be to replicate Mesgarani and Chang (2012)'s cocktail party study, but with the focus on reconstructing speech from ECoG recordings taken from the auxiliary speech comprehension areas described by Davis and Johnsrude (2003) rather than from the STG.

In the next and final section, we turn from sounds to semantics and to the representation of meaning in the brain.

7. Semantic representations

Following a long tradition in linguistics that goes back to De Saussure (1989), speech may be thought of as a pairing of sound and meaning. In this chapter, our plan thus far has been to follow the so-called "chain of speech" (linking articulation, acoustics, and audition) deep into the brain systems involved in comprehending speech (cochlea, subcortical pathways, primary auditory cortex and beyond). We have asked how the brain represents speech at each stage and even how speech representations are dynamically linked in a network of brain regions. But we have not talked yet about meaning. This was largely dictated by necessity: much more is known about how the brain represents sound than meaning.

Indeed, it can even be difficult to pin down what meaning means. In this section, we will focus on a rather narrow kind of meaning that linguists refer to as semantics, and which should be kept distinct from another kind of meaning called pragmatics. Broadly speaking, semantics refers to literal meaning (e.g. “It is cold in here” as a comment on the temperature of the room) whereas pragmatics refers to meaning in context (“It is cold in here” as an indirect request that someone close the window). It may be true that much of what is interesting about human communication is contextual (we are social animals after all), but we shall have our hands full trying to come to grips with even a little bit of how the brain represents the literal meaning of words (lexical semantics). Moreover, the presentation that we give here views lexical semantics from a relatively new perspective grounded in the recent neuroscience and machine learning literatures, rather than from the linguistic (and philosophical) tradition of formal semantics (e.g. Aloni & Dekker 2016). This is important because many established results in formal semantics have yet to be explained neurobiologically. For future neurobiologists of meaning, there will be many important discoveries to be made.

Embodied meaning

Despite the difficulty of comprehending the totality of what some example of speech might mean to your brain, there are some relatively easy places to begin. One kind of meaning that a word might have, for instance, will relate to the ways in which you experience that word. Take the word “strawberry”. Part of the meaning of this word is the shape and vibrant color of strawberries that you have seen. Another is how it smells and feels in your mouth when you eat it. To a first approximation, we can think of the meaning of the word strawberry as the set of associated images, colors, smells, tastes, and other sensations that it can evoke. This is a very useful operational definition of “meaning” because it is to an extent possible to decode brain responses in sensory and motor areas and test whether these areas are indeed activated by words in the ways that we might expect, given the words’ meanings. To take a concrete example of how this approach can be used to distinguish the meaning of two words, consider the words “kick” and “lick”: they differ by only one phoneme, /k/ vs /l/. Semantically, however, the words differ substantially, including, for example, by the part of the body that they are associated with: the foot for “kick” and the tongue for “lick”. Since we know that the sensorimotor cortex contains a map of the body, the so-called “homunculus” (Penfield and Boldrey 1937), with the foot and tongue areas at opposite ends, the embodied-view of meaning would predict that hearing the word “kick” should activate the foot area, which is located near the very top of the head, along the central sulcus on the medial surface of the brain, whereas the word “lick” should active the tongue area, on the lateral surface almost all the way down the central sulcus to the Sylvian fissure. And indeed, these predictions have been verified now over a series of experiments (Pulvermüller 2005): when you hear a word like “kick” or “lick”, not only will your brain represent the sounds of these words through the progression of acoustic, phonetic and phonological representations in a hierarchy of auditory processing centers that we have been discussing in this chapter, but your brain will also represent the meaning of these words across a network of associations which certainly engage your sensory and motor cortices, and, as we shall see, many other cortical regions too.

The result of “kick” and “lick” is of fundamental importance because it gives us a leg up, so to speak, on the very difficult problem of trying to understand the representation of semantics in the brain. Of course, not all words are grounded in embodied semantics in the same way. For example, some words are abstract. Consider the word “society”. Questions like “What does a society taste like?” or even “What does a society look like?” are difficult to answer, because “societies” are not the kinds of things that we taste or see. “Societies” are not like “strawberries”. But even abstract words like “society” might contain embodied semantics that become apparent when we consider the ways in which metaphors link abstract concepts with concretely experienced objects (Lakoff and Johnson 1980). One feature of “societies”, we might assert, is that they have “insides” and “outsides”. In this respect, they are like a great many objects that we experience directly: “cups”, “bowls”, “rooms”. Therefore, it might be hypothesized that even such abstract words as “society” might have predictable effects on the sensory-motor system. Brain areas like the insula that respond to the physical “disgust” of fetid smells also respond to the social “disgust” of seeing an appalled look on someone else’s face (Wicker et al. 2003). There are limits, however, to the embodied view of meaning. Function words, like conjunctions and prepositions, are more difficult to associate with concrete experiences. As we have described it, the approach is also limited to finding meaning in the sensorimotor systems, which is unsatisfying as it ignores large swathes of the brain. In the next subsection, we turn to a more ambitious, if abstract, way of mapping the meaning of words that is not limited to finding meaning in the sensorimotor systems.

Vector representations and encoding models

One difficulty of studying meaning is that it is not only difficult to represent inside the brain, it is difficult to represent at all. If you ask what the word “strawberry” means, we might point at a strawberry. If we know the activity in your visual system that is triggered by looking at a strawberry, then we can point to similar activity patterns in your visual system when you think of the word “strawberry” as another kind of meaning. You might imagine that it is harder to point to just any part of the brain and ask of its current state, ‘Is this a representation of “strawberry”?’. But it is not impossible. In this sub-section, we will, in as informal a way as possible, introduce the ideas of “vector representations” of words, and “encoding models” for identifying neural representations of vectors.

Generally speaking, an “encoding” model aims to predict how the brain will respond to a stimulus. Encoding models contrast with “decoding” models, which aim to do the opposite: guess which stimulus caused the brain response. The spectrogram reconstruction method that we mentioned in a previous section is an example of a decoding model (Mesgarani et al. 2008). An encoding model of sound would therefore try to predict the neural response to an audio recording. In a landmark study of semantic encoding, Mitchell et al. (2008) were able to predict fMRI responses to the meanings of concrete nouns, like “celery” and “airplane”. Unlike studies of embodied meaning, Mitchell et al. (2008) were able to predict neural responses that were not limited to the sensorimotor systems. For instance, they predicted accurate word-specific neural responses across bilateral occipital and parietal lobes, fusiform and middle frontal gyri, and sensory cortex; left inferior frontal gyrus; medial frontal gyrus and anterior cingulate (see Figure 6 for reference) (Mitchell et al. 2008). These encoding results highlight and expand upon something that was already implied by the idea that the meaning of a word might be distributed over multiple sensory systems. They expand the number of regions over which the meaning of a word might be distributed, to non-sensory systems like the anterior cingulate. An even greater expansion of these semantic regions can be found in more recent work (Huth et al. 2016).

So how does an encoding model work? If you are familiar with linear regression, then the model uses linear regression to map from a vector representation of a word to the intensity of a voxel. This model is repeated for each voxel representing the brain and trained on a subset of word embeddings before being tested on a compliment set of word embeddings, in order to evaluate the model’s ability to generalize beyond the words it was trained on. But what does “vector representation” and “word embedding” mean? This field is rather technical and jargon rich, but the key ideas are relatively easy to grasp. “Vector representations” or “word embeddings” represent each word by a “vector”, effectively a list of numbers. Similarly, “brain states” can be quantified by vectors or lists of numbers which represent the amount of activity seen in each voxel of the brain measured during a functional MRI scan. Once we have these vectors, using linear regression methods to try to identify relationships that might map one onto the other is mathematically quite straightforward. So the maths is not difficult and the brain activity vectors are measurable by experiment, but how do we obtain suitable “vector representations” for each word that we are interested in? Let us assume a vocabulary of exactly four words.

A list of four words:

1. “airplane”
2. “boat”
3. “celery”
4. “strawberry”

One way to encode each of these as a list of numbers is to simply assign each word with one number: “airplane” = [1], “boat” = [2], “celery” = [3], and “strawberry” = [4]. We have enclosed the numbers in square brackets to mean that these are lists. Note that it is possible to have only one item in a list. A good thing about this “encoding” of the words, as lists of numbers, is that the resulting lists are short and easy to decode: we only have to look them up in our memories or in a table. But this encoding does not do a very good job of capturing the differences in meanings between the words. For example, “airplane” and “boat” are both man-made vehicles that you could ride inside, whereas “celery” and “strawberry” are both edible parts of plants. A more involved semantic coding might make use of all of these descriptive features to produce the following representations.

Semantic-field encodings for four words

word	man-made	vehicle	ride-inside	edible	plant-part
airplane	1	1	1	0	0
boat	1	1	1	0	0
celery	0	0	0	1	1
strawberry	0	0	0	1	1

In this table, we have placed a “1” under the semantic description if the word along the row satisfies it. For example, an “airplane” is man-made, so the first number in its list is “1”, but “celery”, even if grown by humans, is not man-made, so the first number in its list is “0”. The full list for the word “boat” is [1, 1, 1, 0, 0], which is five numbers long. Is this a good encoding? It is certainly longer than the previous encoding (“boat” = [2]), and unlike the previous code it no longer distinguishes “airplane” from “boat” (both have the identical five-number codes). Finally, the codes are redundant in the sense that, as far as a linear-regression model is concerned, representing the word “boat” as [1, 1, 1, 0, 0] is no more expressive than representing it as [1, 0]. Still, we might like the more verbose listing, since we can interpret the meaning of each number, and we can solve the problem of “airplane” not differing from “boat” by adding another number to the list. That is, if we represented the words with six-number lists, then “airplane” and “boat” could be distinguished: “airplane” = [1, 1, 1, 0, 0, 0] and “boat” = [1, 1, 1, 0, 0, 1]. Now the last number of “airplane” is a “0” and the last number of “boat” is a “1”.

So far, our example may seem tedious and somewhat arbitrary: we had to come up with attributes such as “man-made” or “edible”, then consider their merit as “semantic feature dimensions” without any obvious objective criteria. However, there are many ways to *automatically* search for word embeddings without needing to dream up a large set of semantic fields. An incrementally more complex way is to rely on the context-words that each one of our target-words occurs with in a corpus of sentences. Consider a corpus that contains exactly four sentences.

1. “The boy rode on the airplane.”
2. “The boy also rode on the boat.”
3. “The celery tasted good.”
4. “The strawberry tasted better.”

Our target-words are, again, “airplane”, “boat”, “celery”, and “strawberry”. The context-words are “also”, “better”, “boy”, “good”, “on”, “rode”, “tasted”, and “the” (ignoring capitalization). If we create a table of target-words (rows) by context-words (columns), then we can count how many times each context-word occurred in a sentence with each target-word. This will produce a new set of word embeddings.

Context-word encodings of four words

word	also	better	boy	good	on	rode	tasted	the
airplane	0	0	1	0	1	1	0	2
boat	1	0	1	0	1	1	0	2
celery	0	0	0	1	0	0	1	1
strawberry	0	1	0	0	0	0	1	1

Unlike the previous semantic-field embeddings, which were constructed using our “expert opinions”, these context-word embeddings were learned from data (a corpus of four sentences). Learning a set of word embeddings from data can be very powerful. Indeed we can automate the procedure; and even a modest computer can process very large corpora of text to produce embeddings for hundreds of thousands of words in seconds. Another strength of creating word embeddings like these is that the procedure is not limited to concrete nouns, since context-words can be found for any target word—whether an abstract noun, verb, or even a function word. You might be wondering how context-words are able to represent *meaning*, but notice that words with similar meanings are bound to co-occur with similar context words. For example, an “airplane” and a “boat” are both vehicles that you ride in, so they will both occur quite frequently in sentences with the word “rode”; however, one will rarely find sentences that contain both “celery” and “rode”. Compared to “airplane” and “boat”, “celery” is more likely to occur in sentences containing the word “tasted”. As the English phonetician (Firth 1957; page 11) wrote: “You shall know a word by the company it keeps”.

With a reasonable “vector” representation for words like these, one can begin to see how you might be able to predict the brain activation for word meanings (Mitchell et al. 2008). Start with a fairly large set of words and their vector representations, and record the brain activity they evoke. Put aside some of the words (including perhaps the word “strawberry”) and use the remainder as a “training set” in order to find the best linear equation that maps from word vectors to patterns of brain activation. Finally, use that equation to predict “what the brain activation should have been” for the words you held back, and test how similar that predicted brain activation is to the one that is actually observed, and whether the activation patterns for “strawberry” is indeed more similar to that of “celery” than it is to that of “boat”. One similarity measure commonly used for this sort of problem is the cosine similarity, which can be defined for two *vectors* $\vec{p}$ and $\vec{q}$, according to the following formula:

$$similarity(\vec{p},\vec{q})=\frac{\vec{p}\cdot\vec{q}}{\sqrt{\sum_i p_i^2}\sqrt{\sum_i q_i^2}}$$

Now if we plug the context-word embeddings for each pair of words from our example four-word set into this equation, we end up with the following similarity scores. Note that numbers closer to “1” mean “more similar” and numbers closer to “0” mean “more dissimilar”. A perfect score of “1” actually means “identical”, which we see when we compare any word embedding with itself. You might also note that we have only populated the diagonal and upper triangle of this table, because the lower part is a reflection of the upper part, and therefore redundant.

Cosine similarities between four words

	airplane	boat	celery	strawberry
airplane	1	0.94	0.44	0.44
boat	--	1	0.41	0.41
celery	--	--	1	0.67
strawberry	--	--	--	1

As expected, the words “airplane” and “boat” received a very high similarity score (0.94), whereas “airplane” and “celery”, for example, received lower similarity scores (0.41). The score for “celery” and “strawberry”, however, were also more similar (0.67). Summary statistics like these, summarizing the similarity between two very long lists of numbers, are quick and easy to compute, even for very long lists of numbers. Exploring them also helps to build an intuition about how encoding models, like those of Mitchell et al. (2008), represent the meanings of words, and thus what the brain maps they discover represent. Specifically, Firth (1957)’s idea that the company a word keeps can be used to build up a semantic representation of the word has had a profound impact on the study of semantics recently, especially in the computational fields of natural language processing and machine learning (including deep learning). Mitchell et al. (2008)'s landmark study bridged natural language processing with neuroscience, in a way that finds common ground for both fields at the time of writing. Not only do we expect words that belong to similar semantic domains to co-occur with similar context-words, but if the brain is capable of statistical learning, as many believe, then this is exactly the kind of pattern that we should expect to find encoded in neural representations.

To summarize, we have only begun to scratch the surface of how linguistic meaning is represented in the brain. But figuring out what the brain is doing when it is interpreting speech is so important, and mysterious, that we have tried to illustrate a few recent innovations in enough detail that the reader might begin to imagine how to go further. Embodied meaning, vector representations, and encoding models are not the only ways to study semantics in the brain. They do, however, benefit from engaging with other areas of neuroscience, touching for example on the "homunculus" map in the somatosensory cortex (Penfield & Boldrey 1937). It is less clear, at the moment, how to extend these results from lexical to compositional semantics, or from literal meaning to metaphor. A more complete neural understanding of pragmatics will also be needed. Gladly, much work remains to be done. Because spoken language combines both sound and meaning, a full account of speech comprehension should explain how meaning is coded the brain. We hope that our readers will feel inspired to contribute the next exciting chapters in this endeavor.

Conclusion

Our journey through the auditory pathway has finally reached the end. It was a substantial trip, through the ear and auditory nerve, brainstem and midbrain, and many layers of cortical processing. We have seen how, along that path, speech information is initially encoded by some 30,000 auditory nerve fibers firing hundreds of thousands of impulses a second, and how their activity patterns across the tonotopic array encode formants, while their temporal firing patterns encode temporal fine structure cues to pitch and voicing. We have learned how, as these activity patterns then propagate and fan out over the millions of neurons of the auditory brainstem and midbrain, information from both ears will be combined to add cues to sound source direction. Furthermore, temporal fine structure information gets re-coded, so that temporal firing patterns at higher levels of the auditory brain no longer need to be read out with sub-milisecond precision, and information about the pitch and timbre of speech sounds is instead encoded by a distributed and multiplexed firing rate code. We have seen that the neural activity patterns at levels up to and including the primary auditory cortex are generally thought to represent predominantly physical acoustic or relatively low-level psycho-acoustic features of speech sounds, and that this is then transformed into increasingly phonetic representations at the level of the STG, and finally into semantic representations as we move beyond the STG into frontal and parietal brain areas.

Finally we have seen how notions of embodied meaning as well as of statistical learning are shaping our thinking about how the brain represents the meaning of speech.

By the time they reach these meaning-representing levels of the brain, the waves of neural activity racing up the auditory pathway will have have passed through at least a dozen anatomical processing stations, each comprising between a few hundreds of thousands to hundreds of millions of neurons, each richly and reciprocally interconnected both internally and with both the previous and the next levels in the processing hierarchy. We hope the reader will share our sense of awe when we consider that it takes a spoken word only a modest fraction of a second to travel through this entire, stunningly intricate network and to be transformed from sound wave to meaning.

One last time we caution the reader to remember that the picture we have painted here of a feed forward hierarchical network which transforms acoustics to phonetics to semantics is a highly simplified one. It is well grounded in scientific evidence, but it is necessarily a rather selective telling of the story such as we understand it to date. Recent years have been a particularly productive time in auditory neuroscience, as insights from animal research, human brain imaging, human patient data and ECoG studies and artificial intelligence have begun to come together to provide the framework of understanding which we have attempted to outline here. But many important details remain unknown, and while we feel fairly confident that the insights and ideas we have presented here will stand the test of time, we must be aware that future work may not just complement and refine but even overturn some of the ideas which we currently put forward as our best approximations to the truth. One thing we are absolutely certain of though is that studying how human brains speak to each other will remain a profoundly rewarding intellectual pursuit for many years to come.

References

Aloni, M. and Dekker, P. (2016) *The Cambridge Handbook of Formal Semantics*. Cambridge: Cambridge University Press.

Ballard, D. H., Hinton, G. E. and Sejnowski, T. J. (1983) Parallel visual computation. *Nature* 306:21-26.

Baumann, S., Griffiths, T. D., Sun, L., Petkov, C. I., Thiele, A. and Rees, A. (2011) Orthogonal representation of sound dimensions in the primate midbrain. *Nature Neuroscience* 14:423-5.

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

Bizley, J. K., Walker, K. M., Silverman, B. W., King, A. J. and Schnupp, J. W. (2009) Interdependent encoding of pitch, timbre, and spatial location in auditory cortex. *Journal of Neuroscience* 29:2064-75.

Blakemore, S.-J., Wolpert, D. and Frith, C. (2000) Why can't you tickle yourself?. *Neuroreport* 11:R11-R16.

Bogen, J. E. and Bogen, G. (1976) Wernicke's region--where is it? *Annals of the New York Academy of Sciences* 280:834-843.

Bouchard, K. E., Mesgarani, N., Johnson, K. and Chang, E. F. (2013) Functional organization of human sensorimotor cortex for speech articulation. *Nature* 495:327-332.

Brosch, M., Selezneva, E. and Scheich, H. (2005) Nonauditory events of a behavioral procedure activate auditory cortex of highly trained monkeys. *Journal of Neuroscience* 25:6797-6806.

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

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

Clements, G. N. (1990) The role of the sonority cycle in core syllabification. *Papers in laboratory phonology* 1:283-333.

Da Costa, S., van der Zwaag, W., Marques, J. P., Frackowiak, R. S., Clarke, S. and Saenz, M. (2011) Human primary auditory cortex follows the shape of Heschl's gyrus. *Journal of Neuroscience* 31:14067-14075.

Daunizeau, J., David, O. and Stephan, K. E. (2011) Dynamic causal modelling: a critical review of the biophysical and statistical foundations. *NeuroImage* 58:312-322.

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

Dayan, P., Hinton, G. E., Neal, R. M. and Zemel, R. S. (1995) The Helmholtz machine. *Neural Computation* 7:889-904.

De Saussure, F., 1989 *Cours de linguistique générale: Édition critique*. Otto Harrassowitz Verlag.

Dean, I., Harper, N. and McAlpine, D. (2005) Neural population coding of sound level adapts to stimulus statistics. *Nature Neuroscience* 8:1684-1689.

Delgutte, B. (1997) Auditory neural processing of speech. In: Hardcastle, W. J. and Laver, J. (Eds.), *The Handbook of Phonetic Sciences*, Blackwell.

Edeline, J. M., Pham, P. and Weinberger, N. M. (1993) Rapid development of learning-induced receptive field plasticity in the auditory cortex. *Behavioral Neuroscience* 107:539-51.

Eliades, S. J. and Wang, X. (2008) Neural substrates of vocalization feedback monitoring in primate auditory cortex. *Nature* 453:1102-1106.

Engineer, C. T., Perez, C. A., Chen, Y. H., Carraway, R. S., Reed, A. C., Shetake, J. A., Jakkamsetti, V., Chang, K. Q. and Kilgard, M. P. (2008) Cortical activity patterns predict speech discrimination ability. *Nature Neuroscience* 11:603-8.

Ferry, R. T. and Meddis, R. (2007) A computer model of medial efferent suppression in the mammalian auditory system. *The Journal of the Acoustical Society of America* 122:3519-3526.

Firth, J. (1957) Papers in Linguistics, Oxford: Oxford University Press.

Flinker, A., Chang, E. F., Kirsch, H. E., Barbaro, N. M., Crone, N. E. and Knight, R. T. (2010) Single-trial speech suppression of auditory cortex activity in humans. *Journal of Neuroscience* 30:16643-16650.

Fowler, C. A. (1986) An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics* 14:3-28.

Frisina, R. D. (2001) Subcortical neural coding mechanisms for auditory temporal processing. *Hearing Research* 158(1-2):1-27.

Friston, K. and Kiebel, S. (2009) Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 364:1211-1221.

Friston, K. J., Harrison, L. and Penny, W. (2003) Dynamic causal modelling. *NeuroImage* 19:1273-1302.

Fritz, J., Shamma, S., Elhilali, M. and Klein, D. (2003) Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nature Neuroscience* 6:1216-23.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L. and Zue, V. (1993) TIMIT acoustic-phonetic continuous speech corpus. *Linguistic Data Consortium*.

Golding, N. L. and Oertel, D. (2012) Synaptic integration in dendrites: exceptional need for speed. *The Journal of physiology* 590:5563-5569.

Graves, A. and Jaitly, N. (2014) Towards end-to-end speech recognition with recurrent neural networks. *Proceedings of the 31st International Conference on Machine Learning, PMLR* 32(2):1764-1772.

Greenberg, S. (2006) A Multi-tier Framework for Understanding Spoken Language. In: Greenberg, S. & Ainsworth, W. A. (Ed.), *Listening to speech: an auditory perspective*, Lawrence Erlbaum Associates.

Grinn, S. K., Wiseman, K. B., Baker, J. A. and Le Prell, C. G. (2017) Hidden hearing loss? No effect of common recreational noise exposure on cochlear nerve response amplitude in humans. *Frontiers in Neuroscience* 11:465.

Grose, J. H., Buss, E. and Hall, J. W. (2017) Loud Music Exposure and Cochlear Synaptopathy in Young Adults: Isolated Auditory Brainstem Response Effects but No Perceptual Consequences. *Trends in Hearing* 21:1-18.

Heinz, M. G., Colburn, H. S. and Carney, L. H. (2002) Quantifying the implications of nonlinear cochlear tuning for auditory-filter estimates. *The Journal of the Acoustical Society of America* 111:996-1011.

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

von Holst, E. and Mittelstaedt, H. (1950) Das eafferenzprinzip. *Naturwissenschaften* 37:464-476.

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

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

Jerison, H. J., 1973 *Evolution of the brain and intelligence*. New York: Academic Press.

Johnsrude, I. S., Mackey, A., Hakyemez, H., Alexander, E., Trang, H. P. and Carlyon, R. P. (2013) Swinging at a cocktail party: Voice familiarity aids speech perception in the presence of a competing voice. *Psychological Science* 24:1995-2004.

Joris, P. X., Smith, P. H. and Yin, T. C. T. (1998) Coincidence Detection in the Auditory System: 50 Years after Jeffress. *Neuron* 21:1235-1238.

Jurafsky, D. and Martin, J. H., 2014 *Speech and language processing*. Pearson London.

Kawato, M., Hayakawa, H. and Inui, T. (1993) A forward-inverse optics model of reciprocal connections between visual cortical areas. *Network: Computation in Neural Systems* 4:415-422.

Kiebel, S. J., Von Kriegstein, K., Daunizeau, J. and Friston, K. J. (2009) Recognizing sequences of sequences. *PLoS computational biology* 5:e1000464.

Kujawa, S. G. and Liberman, M. C. (2015) Synaptopathy in the noise-exposed and aging cochlea: Primary neural degeneration in acquired sensorineural hearing loss. *Hearing research* 330:191-199.

Lakoff, G. and Johnson, M., 1980 *Metaphors we live by*. University of Chicago press.

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

Liberman, A. M., Cooper, F. S., Shankweiler, D. P. and Studdert-Kennedy, M. (1967) Perception of the speech code. *Psychological Review* 74:431.

Liberman, A. M. and Mattingly, I. G. (1985) The motor theory of speech perception revised. *Cognition* 21:1-36.

Meddis, R. and O'Mard, L. P. (2005) A computer model of the auditory-nerve response to forward-masking stimuli. *The Journal of the Acoustical Society of America* 117:3787-3798.

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

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

Mesgarani, N., David, S. V., Fritz, J. B. and Shamma, S. A. (2008) Phoneme representation and classification in primary auditory cortex. *The Journal of the Acoustical Society of America* 123:899-909.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A. and Just, M. A. (2008) Predicting human brain activity associated with the meanings of nouns. *Science* 320:1191-1195.

Müller-Preuss, P. and Ploog, D. (1981) Inhibition of auditory cortical neurons during phonation. *Brain Research* 215:61-76.

Mumford, D. (1992) On the computational architecture of the neocortex. *Biological cybernetics* 66:241-251.

Nelken, I., Bizley, J. K., Nodal, F. R., Ahmed, B., King, A. J. and Schnupp, J. W. (2008) Responses of auditory cortex to complex stimuli: functional organization revealed using intrinsic optical signals. *Journal of Neurophysiology* 99:1928-41.

Parker Jones, O., Seghier, M. L., Duncan, K. J. K., Leff, A. P., Green, D. W. and Price, C. J. (2013) Auditory--motor interactions for the production of native and non-native speech. *Journal of Neuroscience* 33:2376-2387.

Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., Shamma, S. A., Crone, N. E., Knight, R. T. and Chang, E. F. (2012) Reconstructing speech from human auditory cortex. *PLoS Biology* 10:e1001251.

Penfield, W. and Boldrey, E. (1937) Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation. *Brain* 60:389-443.

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

Prothero, J. W. and Sundsten, J. W. (1984) Folding of the cerebral cortex in mammals. *Brain, Behavior and Evolution* 24:152-167.

Pulvermüller, F. (2005) Brain mechanisms linking language and action. *Nature Reviews Neuroscience* 6:576-582.

Rabinowitz, N. C., Willmore, B. D. B., King, A. J. and Schnupp, J. W. H. (2013) Constructing noise-invariant representations of sound in the auditory pathway. *PLoS biology* 11:e1001710.

Rao, R. P. and Ballard, D. H. (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience* 2:79-87.

Rauschecker, J. P. and Scott, S. K. (2009) Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nature neuroscience* 12:718-724.

Schnupp, J. W. and Carr, C. E. (2009) On hearing with more than one ear: lessons from evolution. *Nature Neuroscience* 12:692-7.

Schnupp, J. W. H., Garcia-Lazaro, J. A. and Lesica, N. A. (2015) Periodotopy in the gerbil inferior colliculus: local clustering rather than a gradient map. *Front Neural Circuits* 9:37.

Schreiner, C. E. and Langner, G. (1988) Periodicity coding in the inferior colliculus of the cat. II. Topographical organization. *Journal of Neurophysiology* 60:1823-40.

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

Stevens, K. N. (1960) Toward a model for speech recognition. *The Journal of the Acoustical Society of America* 32:47-55.

Stevens, K. N. (2002) Toward a model for lexical access based on acoustic landmarks and distinctive features. *The Journal of the Acoustical Society of America* 111:1872-1891.

Stuart, A. and Phillips, D. P. (1996) Word recognition in continuous and interrupted broadband noise by young normal-hearing, older normal-hearing, and presbyacusic listeners. *Ear and Hearing* 17:478-489.

Sumner, C. J., Lopez-Poveda, E. A., O'Mard, L. P. and Meddis, R. (2002) A revised model of the inner-hair cell and auditory-nerve complex. *The Journal of the Acoustical Society of America* 111:2178-2188.

Tremblay, P. and Dick, A. S. (2016) Broca and Wernicke are dead, or moving past the classic model of language neurobiology. *Brain and Language* 162:60-71.

Walker, K. M., Bizley, J. K., King, A. J. and Schnupp, J. W. (2011) Multiplexed and robust representations of sound features in auditory cortex. *Journal of Neuroscience* 31:14565-76.

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

Wicker, B., Keysers, C., Plailly, J., Royet, J.-P., Gallese, V. and Rizzolatti, G. (2003) Both of us disgusted in My insula: the common neural basis of seeing and feeling disgust. *Neuron* 40:655-664.

Willmore, B. D. B., Schoppe, O., King, A. J., Schnupp, J. W. H. and Harper, N. S. (2016) Incorporating Midbrain Adaptation to Mean Sound Level Improves Models of Auditory Cortical Processing. *The Journal of neuroscience* 36:280-289.

Wolpert, D. M., Ghahramani, Z. and Flanagan, J. R. (2001) Perspectives and problems in motor learning. *Trends in Cognitive Sciences* 5:487-494.

Zhang, X. and Carney, L. H. (2005) Analysis of models for the synapse between the inner hair cell and the auditory nerve. *The Journal of the Acoustical Society of America* 118:1540-1553.

Zhang, X., Heinz, M. G., Bruce, I. C. and Carney, L. H. (2001) A phenomenological model for the responses of auditory-nerve fibers: I. Nonlinear tuning with compression and suppression.. *The Journal of the Acoustical Society of America* 109:648-670.

Zilany, M. S. A., Bruce, I. C. and Carney, L. H. (2014) Updated parameters and expanded simulation options for a model of the auditory periphery.. *The Journal of the Acoustical Society of America* 135:283-286. doi: 10.1121/1.4837815