Chapter 6
Speech Perception Under Adverse Listening Conditions

Stephen C. Van Hedger and Ingrid S. Johnsrude

Abstract Perceiving and understanding spoken language is something that most listeners take for granted, at least in favorable listening conditions. Yet, decades of research have demonstrated that speech is variable and ambiguous, meaning listeners must constantly engage in active hypothesis testing of what was said. Within this framework, even relatively minor challenges imposed on speech recognition must be understood as requiring the interaction of perceptual, cognitive, and linguistic factors. This chapter provides a systematic review of the various ways in which listening environments may be considered adverse, with a dual focus on the cognitive and neural systems that are thought to improve speech recognition in these challenging situations. Although a singular mechanism or construct cannot entirely explain how listeners cope with adversity in speech recognition, overcoming listening adversity is an attentionally guided process. Neurally, many adverse listening conditions appear to depend on higher-order (rather than primary) representations of speech in cortex, suggesting that more abstract linguistic knowledge and context become particularly important for comprehension when acoustic input is compromised. Additionally, the involvement of the cingulooperecular (CO) network, particularly the anterior insula, in a myriad of adverse listening situations may indicate that this network reflects a general indication of cognitive effort. In discussing the various challenges faced in the perception and understanding of speech, it is critically important to consider the interaction of the listener's cognitive resources (knowledge and abilities) with the specific challenges imposed by the listening environment.

S. C. Van Hedger (✉)
Department of Psychology, Huron University College, London, ON, Canada

Department of Psychology & Brain and Mind Institute, University of Western Ontario, London, ON, Canada

I. S. Johnsrude
Department of Psychology & Brain and Mind Institute, University of Western Ontario, London, ON, Canada
e-mail: svanhedg@uwo.ca

National Centre for Audiology & School of Communication Sciences and Disorders, University of Western Ontario, London, ON, Canada

L. L. Holt et al. (eds.), *Speech Perception*, Springer Handbook of Auditory Research 74, https://doi.org/10.1007/978-3-030-81542-4_6

Keywords Speech perception · Attention · Working memory · Cognitive control · Intelligibility · Context · Aging · Listening effort · Cinguloopercular network · Dual-stream auditory processing · Hierarchical representations

6.1 Introduction

Efficient and accurate speech recognition is essential for communication, although people often take this skill for granted. It is difficult to fully appreciate the degree to which individuals rely on speech communication to enrich and provide the essentials of life, and it is similarly difficult to appreciate the processes that support speech recognition across variable listening environments. Diverse listening challenges such as novel voices, speech accents, and a wide range of background noise pose unique perceptual, cognitive, and linguistic demands that often must be solved. This chapter provides an overview of how listeners perceive speech under a variety of adverse listening conditions, with an emphasis on the cognitive and neurobiological foundations that support perception under these conditions. In reviewing the ways in which listeners must overcome listening challenges, this chapter emphasizes that different adverse conditions place different demands on cognitive resources, and so one must consider the specific challenges of a given listening environment to understand how listeners may achieve successful comprehension.

The phrase "adverse listening conditions" might evoke an image of trying to carry on a conversation while sitting on an active airplane runway. What is meant by "adverse" is more varied, more mundane, and more plausible. Imagine, for example, trying to converse with a cashier as they are ringing up items in a crowded grocery store. To successfully perceive the cashier's speech, one must engage in several processes. First, the complex sound wave hitting the ears, which is a mixture of all the audible sounds in the store, must be perceptually organized into discrete sound sources in different locations, based on a variety of cues that enable perceptual grouping and segregation (Darwin and Carlyon 1995). One important cue is harmonicity: the frequency components of the cashier's voice occur at regular harmonic intervals in the spectrum, and one can use this cue to work out which frequency components belong together. This is made more difficult if other sounds, like the "beep" that accompanies the scanning of each grocery item, contain similar frequency components – these can effectively obliterate (energetically mask) the original components from the voice, which would then need to be perceptually restored using knowledge and contextual information. Somehow the noisy and variable speech sounds produced by the cashier are mapped in one's speech/language system onto linguistic representations that are organized (grouped and segregated) into words and phrases, evoking meaning. Several cashiers and customers in the environment may be talking at the same time, making it difficult to determine which words were produced by one's conversational partner and which came from elsewhere, since all of it is processed to some degree by the speech/language system. In other words, masking intelligible speech produces perceptual and cognitive

interference (informational masking) (Kidd and Colbourn 2017). If one is deeply familiar with the topic of conversation, and if the linguistic material is very simple and predictable, that will help. If the topic is hard to identify, or if an esoteric word is used, or if the interlocutor has an unfamiliar accent, that adds to the perceptual and cognitive challenge. If one has a hearing impairment, or is an older person, that adds to the challenge as well.

This chapter explores such challenges and what they may involve in more detail. Specifically, the chapter will first explore the cognitive processes underlying successful speech comprehension (Sects. 6.2.1 and 6.2.2), and how this depends on the neurobiology of the human brain (Sect. 6.2.3). From there, the chapter will detail different types of adverse conditions, and the cognitive resources that may be required to overcome them (Sect. 6.3). The role of attention in speech comprehension will then be specifically highlighted, with an emphasis on how the role of attention may differ dramatically depending on listening conditions (Sect. 6.4). The chapter will then introduce the idea of listening effort and explain it as an interaction between the demands imposed by the listening situation, and the unique constellation of cognitive abilities an individual listener brings to bear (Sect. 6.5). Finally, potentially fruitful directions of future research will be identified (Sect. 6.6).

6.2 Important Speech Features for Effective Comprehension

It may be difficult to appreciate the variety of processes required to successfully understand a signal as rich and complex as fluent speech. Just as one is not consciously aware of the complexities of other systems, such as the mechanisms supporting breathing or balance, speech understanding often feels like it occurs effortlessly and automatically. To put the complexity of speech understanding in perspective, briefly consider some of the steps involved in conversing with another individual. One presumably starts with a linguistic thought, which then must be transformed into a physiological code (moving one's lips, tongue, and vocal cords to produce the intended speech), using the distinctive articulations characteristic of a particular individual's accent, idiolect (speech habits peculiar to an individual), and voice. This signal, which now exists as compressions and rarefactions in the air, mixes with other acoustic energy in the environment, creating a complex waveform which impinges on the eardrum of the listener and is transduced into electrical impulses in the auditory nerve in the cochlea. The listener must analyze this complex sound to perceptually organize the auditory scene into discrete sources, segregating the target signal from any background, and mapping sounds onto linguistic representations, eventually resulting in understanding. This *speech chain* (Denes and Pinson 1993) unfolds extremely quickly in naturalistic settings and is aided by listeners' remarkable abilities to segment speech into meaningful units (Sect. 6.2.1), listeners' abilities to hold parts of speech in working memory and use context to improve comprehension (Sect. 6.2.2), and the neurobiology of listeners' auditory systems (Sect. 6.2.3).

6.2.1 Segmentation

One of the most fundamental components of comprehending speech is the parsing of a continuous speech signal into discrete words and phrases. This process is so well rehearsed that many individuals (who do not study speech for a living) are surprised to discover that the boundaries of words are not actually represented by silence or other reliable acoustic markers in the waveform. For example, consider a relatively long single word in English – *unimaginatively* – which contains seven syllables. Even if this word does not appear in everyday conversation, native speakers of English will generally have little trouble grouping these syllables together, easily parsing the word from other words that make up a phrase or sentence, such as *He spoke unimaginatively*. One can experience the issue of speech segmentation firsthand by listening to naturalistic speech from an unfamiliar language. In this exercise, one may get sense of where word boundaries exist, but this will be largely driven by how one segment speech in one's native language. Indeed, this is precisely what Cutler and Norris (1988) demonstrated in a seminal paper. English speakers tend to demarcate lexical items using the rhythmic patterns of their native language, with strong syllables being more likely to correspond to the beginning of a word in English. English listeners in their study were slower to detect a target word embedded in nonsense disyllables when there were two strong syllables (e.g., detecting *mint* in the nonsense disyllable *mintayve*) compared to a strong and a weak syllable (e.g., detecting *mint* in the nonsense disyllable *mintesh*). Thus, listeners must learn the appropriate cues to parse a continuous speech stream into discrete lexical items, but these cues are not universal across languages and are not necessarily reflected in the acoustics of the speech signal.

Further research indicates that a host of other statistical characteristics of a known language, in addition to stress patterns, are used to segment speech into words. Phoneme sequence constraints, or phonotactics, describe the permissible combinations of phonemes in a language at various points in a word, such as onsets and offsets. Listeners have implicit knowledge of the phonotactics of their native language, and word boundaries are inferred when phoneme transitional probabilities are low. For example, the sequence "I'd love lunch" (phonetic notation: /ajdləvlənʧ/) would be heard by English speakers as having a boundary between the /vl/ sequence. This is because the phoneme sequence /vl/ cannot occur at the beginnings of words in English. Rather, /vl/ can only occur in the middle of words (e.g., "unraveling"), or at word offsets (e.g., "unravel") – arguably even then with a schwa (a weak, unstressed vowel, such as the "a" in "about") between the /v/ and /l/. As such, in the example sequence /ajdləvlənʧ/, the only possible perceptual organization that would not leave nonword fragments (e.g., /ənʧ/) would place a word boundary between /ləv/ and /lənʧ/ (Norris et al. 1997). In addition to acoustic and lexical information, semantic information and context can also drive segmentation – in fact, according to Mattys et al. (2005), knowledge-based lexical and semantic cues are the most important for driving perception, followed by segmental cues such as phonotactics, with stress being perhaps the weakest cue to segmentation. The

problem of determining word boundaries is thus complex, requires the balancing of multiple, sometimes conflicting, constraints, and draws on both the acoustics of the signal and prior language-specific linguistic knowledge.

6.2.2 Working Memory and Use of Context

The effective comprehension of speech requires a kind of active hypothesis testing of what was said. The acoustics of speech do not cleanly map onto linguistic categories – a single acoustic event can have multiple phonetic interpretations depending on the speaker and the context of the listening environment, and a single phonetic category can have multiple acoustic realizations. This lack of invariance in speech (Liberman et al. 1967) means that there is a many-to-many mapping between any acoustic event and its linguistic meaning, which poses a computational problem to the listener. As such, working memory – the ability to temporarily store, maintain, and manipulate information in service of complex cognitive tasks (Baddeley and Hitch 1974) – may be particularly important for effectively weighing possible interpretations of incoming speech until the most appropriate interpretation can be selected.

For example, consider the vowels /I/ and /ɛ/, as heard in the words "bit" and "bet." These vowels in American English are highly similar with respect to their formant frequencies, making them particularly confusable. This means that a listener may rely on working memory to understand a spoken sentence in which the intended utterance is not immediately apparent. In the sentence, "The [bill/bell] was so large that it took me by surprise, even though I had previously been to that church," both interpretations of the bracketed words are plausible until the final word, which ultimately provides strong evidence for "bell." Even in this simple example, it should be apparent how working memory is an important component of effective speech comprehension, especially as ambiguity is increased or the strength of the meaningful context in which the ambiguous utterance is decreased, as is often the case in adverse listening conditions.

Ambiguous speech material must be held online in some way until sufficient contextual information is received to disambiguate it. Context, broadly construed, is any information in, or related to, the environment that might constrain interpretation of an ambiguous utterance. Context can include other words in the utterance, or what was previously said, visual cues, or even shared history with the talker. Context influences the perception of speech across multiple levels of analysis, reflecting the inherent ambiguity of how acoustic patterns map onto linguistic categories, how words map onto meaning, and pragmatically how an utterance ought to be interpreted (often beyond its literal meaning).

Robust context effects have been observed at the level of phonemes, even for nonlinguistic context, such as sine waves presented in the frequency range of vowel formants (Holt 2005), which supports the idea that contextual influences on perception of sublexical elements may reflect a more general auditory process. Yet, at the

level of the talker, the influence of context depends on a listener's interpretation – that is, whether the listener attributes a particular sound to idiosyncratic variation in articulation (due to a talker's idiolect, or perhaps due to temporary articulatory constraints such as holding a pencil in their teeth) or due to principled changes that are linguistically informative (Kraljic et al. 2008). Such principled changes would include those due to the talker's dialect, coarticulatory effects, the articulators forming the next sound before the previous sound is completely produced, or other kinds of fine phonetic (or subphonemic) detail, signaling, for example, morphological complexity, utterance ending, register, and emotional state (Hawkins 2003). This suggests that the influence of context on phonetic perception is complex, depending both on the level at which the context is operating and the interpretation of the context in service of understanding meaning and talker-specific attributes.

Context is also critical for disambiguating words with multiple meanings and/or syntactic roles (Rodd et al. 2002, 2005). When interpreting an utterance, a listener must use the surrounding words to guide the selection of the appropriate syntactic role and semantic properties of each word. For example, in the phrase "the bank of the river," the initial word "the" indicates that "bank" is being used as a noun and not a verb, while the semantic properties of the word "river" indicate that "bank" is referring to the water's edge and not to, for example, an institution concerned with the borrowing and saving of money. These forms of ambiguity are ubiquitous in language. At least 80% of the common words in a typical English dictionary have more than one definition (Rodd et al. 2002), and many words, such as "run," have dozens of definitions. Each time one of these ambiguous words is encountered, the listener must hold the unfolding utterance in mind until they are able to select the appropriate meaning based on context.

Context is not limited to the auditory modality. In many everyday settings, the recognition of speech occurs in tandem with the processing of visual information, either in the environment or from the talker's face and gestures which establishes a specific context (incorporating a talker's sex, height, and facial attributes) for interpreting the speech signal. Listeners frequently make use of bimodal speech cues that are readily available in conversational settings and that tap existing knowledge. For example, if one is at a busy party and hears the sentence "I wanna eat the Grampa bunny's hearing aids!," knowing that what is shown in Fig. 6.1 is on the table in front of the 10-year-old talker (and that the "hearing aids" on the larger cake are made of marzipan, and that the child loves marzipan) would help enormously.

A wealth of research indicates that auditory and visual information complement each other in speech perception and that the facial gestures available in audiovisual speech make it more intelligible than auditory-alone speech. In one of the earliest publications on the topic, visual speech cues were noted to improve the signal-to-noise ratio (SNR) by up to 15 dB (Sumby and Pollack 1954), dramatically enhancing intelligibility. The use of visual speech information is especially advantageous when speech is semantically and syntactically complex (Reisberg et al. 1987) or when it is impoverished or degraded (Macleod and Summerfield 1990).

Linguistic information (phonotactic, lexical, semantic, syntactic, and facial/gestural), which is used to disambiguate speech, must be stored as long-term, stable

Fig. 6.1 The spoken sentence "I wanna eat the Grampa bunny's hearing aids!" makes a lot more sense when you know that the talker is a 10-year-old who loves marzipan, at a joint Easter birthday party for a 77-year-old man and his 8-year-old granddaughter and that the 10-year-old is looking at these cakes (particularly the one on the right; with "hearing aids" made of marzipan)

representations in the brain. Semantic knowledge and memory are required to comprehend spoken language, such as to interpret an utterance in the context of known facts and events. Information that has been stored about individual talkers can also facilitate intelligibility and comprehension. For example, voices of people that are personally known to a listener are substantially more intelligible than voices of strangers, when heard in a mixture with a competing talker (Johnsrude et al. 2013; Holmes et al. 2018), and better intelligibility also results when listeners are trained with voices in a lab (Nygaard and Pisoni 1998). Thus, long-term knowledge of a talker's articulatory patterns, developed through prior experience, can constrain the interpretation of speech.

6.2.3 Distributed Neurobiology for Effective Comprehension

At this point, it should be apparent that listeners use multiple cues to successfully comprehend fluent speech. This effective understanding of speech must be grounded in the neuroanatomy of the auditory system, as well as a more distributed language network, and so it is worth considering how speech comprehension is supported from a neurobiological perspective. Beginning with general auditory processing, anatomical and neurophysiological findings in nonhuman primates support the idea of multiple parallel streams of processing in the auditory system. Despite 25 million years of divergent evolution, the anatomical organization of cortical auditory system in rhesus macaque monkeys is often taken as a model for human cortical organization (Davis and Johnsrude 2007; Hackett 2011). Processing of auditory information is highly parallel (multiple computations at once) at various levels of the primate auditory system.

Even in the earliest cortical receiving areas (primary, or "core" auditory cortex), multiple representations of the input are available (Jones 2003). The organization of the cortical auditory system is cascaded, with hierarchical connections among auditory core, neighboring secondary or "belt" regions, and adjacent parabelt areas, suggesting at least three discrete levels of processing (Hackett 2011). A distributed, interconnected set of fields, in superior temporal gyrus and sulcus, in the inferior parietal lobule, and in prefrontal cortex, receive inputs from belt and parabelt regions, constituting a potential fourth stage of processing (Hackett 2011; see Fig. 6.2).

Accounts of speech processing in humans emphasize two main processing pathways that radiate out from primary auditory regions on the superior temporal plane (Hickok and Poeppel 2015; but also see Davis and Johnsrude 2007). The "dual-stream" account is based on the observation that temporal, parietal, and frontal connections of macaque auditory cortex are topographically organized. Anterior belt, parabelt, and associated anterior temporal-lobe regions interconnect with anterior and ventral frontal cortical sites (the ventral auditory stream). In contrast, more posterior belt, parabelt, and associated posterior temporal regions interconnect with more posterior and dorsal frontal cortical sites (the dorsal auditory stream) (Hackett 2011). These two routes have been given different putative functional roles. For the ventral stream, these include a role in lexico-semantic comprehension of speech, and in selective retrieval of contextual information associated with words (Hickok and Poeppel 2015). For the dorsal stream, these include motor-articulatory mapping of sound which might be particularly important for understanding when speech is acoustically degraded (Du et al. 2014).

Results of functional neuroimaging studies provide evidence that human speech perception may also be based on multiple hierarchical processing pathways consistent with a comparative neurobiological framework. Early functional magnetic resonance imaging (fMRI) investigations demonstrated that, for listeners hearing nonlinguistic stimuli, more complex sounds (amplitude and frequency-modulated tones, bandpass filtered noise) activated auditory regions beyond the core (belt and parabelt), whereas simpler sounds (pure tones) activated primarily the core (Giraud et al. 2000). Davis and Johnsrude (2003) investigated the hierarchical organization of the speech perception system used a converging-operations approach in which naturalistic sentence-length stimuli were processed three acoustically different ways, each applied parametrically to yield different levels of intelligibility. The

Fig. 6.2 (continued) four levels of processing, including core regions (darkest shading), belt regions (light shading), parabelt regions (hatching), and temporal and frontal regions that interconnect with belt and parabelt (dotted). (Adapted from Hackett et al. (2014).) Dotted lines indicate sulci that have been opened to show auditory regions. (**b**) Schematic of cortical areas in the macaque monkey that are metabolically active during processing of auditory, visual, and audiovisual stimuli. (From Poremba and Mishkin (2007)). (**c**) Model of hierarchical processing of speech summarizing neuroimaging data (see text; Davis and Johnsrude 2003; Okada et al. 2010; after Peelle et al. 2010). *CS* central sulcus, *IPL* inferior parietal lobule, *IPS* intraparietal sulcus, *ITG* inferior temporal gyrus, *MTG* middle temporal gyrus, *PFC* prefrontal cortex, *STG* superior temporal gyrus, *STS* superior temporal sulcus

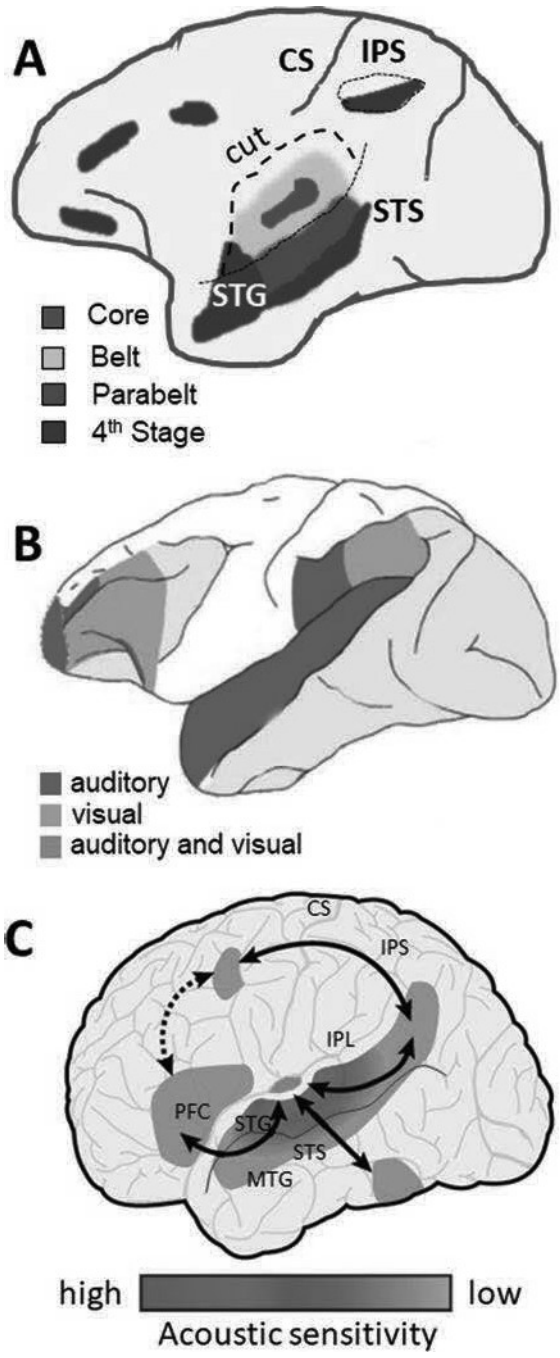

Fig. 6.2 Auditory-responsive cortex in the primate includes many anatomically differentiable regions. All brains show the brain from the side, with the front of the brain (frontal cortex) to the left of the page. (**a**) The anatomical organization of the auditory cortex is consistent with at least

(continued)

investigators were able to distinguish three levels of processing. Primary auditory regions were sensitive to any kind of sound, intelligible or not. Activity in more lateral, anterior, and posterior areas in the temporal lobe correlated with intelligibility, but also differed depending on acoustic characteristics (specifically, the type of distortion). More distant intelligibility sensitive regions in the middle and superior temporal lobes and in left inferior frontal gyrus (IFG) were not sensitive to the acoustic form of the stimuli, suggesting that more abstract, nonacoustic processing of speech is performed by these regions. These three levels of processing, reflecting progressive abstraction of the linguistic signal from the acoustic, appear to radiate out from primary auditory cortex in a fashion reminiscent of the anatomical organization of the auditory system in macaques (see Fig. 6.2).

Rodd et al. (2005, 2012) subsequently identified left dorsolateral frontal and posterior inferior temporal regions, even further from primary auditory cortex, which are recruited when listeners hear meaningful, intelligible sentences that contain words with more than one meaning, perhaps consistent with a fourth stage of processing; see Fig. 6.3. Binder et al. (2009) observed imaging results consistent with the idea that linguistic processes at higher processing stages are topographically further away from auditory cortex. In a meta-analytic study of 120 functional imaging reports, they observed that when people had to process the meaning of spoken or read words,

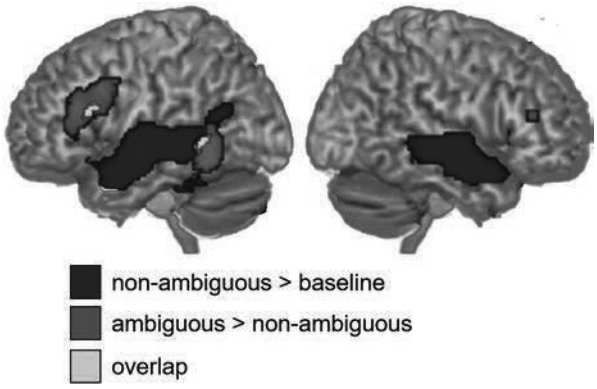

Fig. 6.3 Functional magnetic resonance imaging (fMRI) activation in response to spoken sentences with or without lexical ambiguity, shown superimposed on a brain structural image. The left hemisphere of the brain is shown on the left (front of the brain nearest left margin) and the right hemisphere on the right (front of the brain nearest right margin). Comparison between sentences without ambiguous words (e.g., "her secrets were written in her diary") and a baseline, energy-matched, noise condition revealed a large area of greater activation for the former condition (in blue) in left and right superior and middle temporal gyri, extending in the left hemisphere into posterior inferior temporal cortex and the left fusiform gyrus. Greater activation for this intelligible speech, compared to noise baseline, was also observed in both hemispheres in lingual gyrus, and in the dorsal part of the inferior frontal gyrus (pars triangularis). Greater activation for sentences with ambiguous words (e.g., "the *shell* was *fired* towards the *tank*") compared to matched sentences without (in red) was observed in left and right inferior frontal gyrus (IFG) (pars triangularis), and a region of the left posterior inferior temporal cortex. The yellow area indicates overlap between the two contrasts. (Adapted from Rodd et al. (2005))

activation clustered in seven distinct regions (Binder et al. 2009). Active regions included the inferior parietal lobule (the angular gyrus and some of the supramarginal gyrus); middle temporal gyrus; fusiform and adjacent parahippocampal regions; IFG, ventral and dorsal medial prefrontal cortex; and retrosplenial cortex.

Neuropsychological data are also consistent with this hierarchical framework. Damage from conditions such as stroke, in or near auditory cortex (particularly in the left hemisphere) in humans, can result in a condition called "word deafness," a type of agnosia in which spoken words are no longer recognized (Phillips and Farmer 1990). It is doubtful, however, that "word deafness" is entirely specific to speech, as it may also apply to some nonverbal sounds.

Farther from auditory cortex, damage results in language deficits at a higher linguistic or conceptual level. Lesion-symptom mapping (Bates et al. 2003) is a technique that allows researchers to combine behavioral and brain imaging maps of lesions from individuals with brain damage to identify the brain regions that, if damaged, are most likely to result in deficits on specific behavioral tasks. For example, using this technique in 64 individuals with left-hemisphere cortical damage, Dronkers et al. (2004; Turken and Dronkers 2011) established that a number of regions outside of primary auditory cortex, in the middle and superior temporal gyri and in the inferior frontal cortex, were commonly damaged in individuals who had difficulty understanding spoken sentences (Fig. 6.4). Again, areas that process

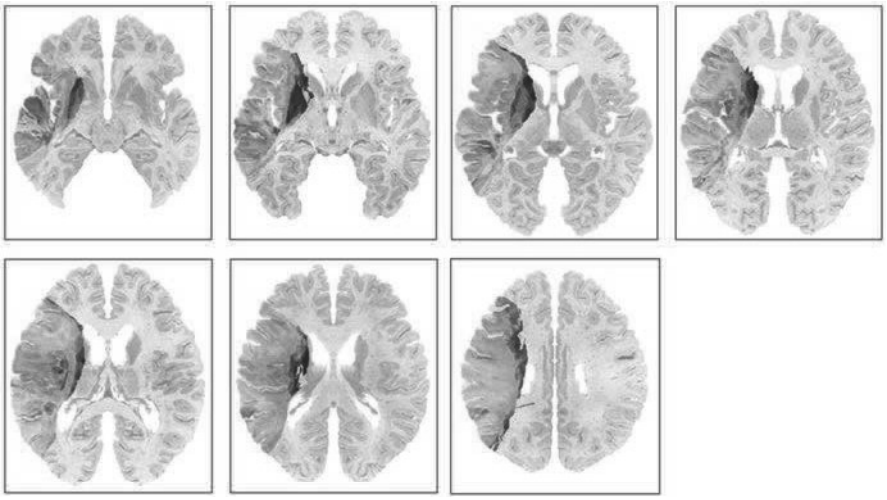

Fig. 6.4 Regions related to comprehension of spoken sentences (in red). The region of brain damage was mapped in each of 64 individuals with language disruption (aphasia) as a result of stroke in the left hemisphere of the brain. Different individuals showed difficulty with different aspects of language, depending on the location of the lesion. Areas in which damage related to impairment in the comprehension of spoken sentences are shown hot colors, with the strongest relationship shown in red. These data are superimposed on horizontal brain slices (in gray). In these images, left is on the left, and the front of the brain is at the top of each image. From left to right, and top to bottom, slices are progressively closer to the top of the brain. The red regions cover middle and superior temporal gyri and in the inferior frontal gyrus (IFG). (From Turken and Dronkers (2011); Fig. 1, panels 3–9)

meaning seem to be quite distant from auditory cortex. This is also demonstrated by a lesion-symptom mapping study conducted by Mesulam et al. (2012) on individuals with primary progressive aphasia, a neurodegenerative condition that presents as a loss of word meaning. They administered a comprehensive battery of language tests and examined the correlation between regional cortical atrophy and the magnitude of impairment on different tests. Impairment in auditory word comprehension correlated with atrophy in the anterior temporal region bilaterally, whereas impairment in sentence comprehension correlated with atrophy in orbitofrontal and lateral frontal regions, and in the inferior parietal lobule, all areas well away from auditory regions. To date, most work exploring the neurobiology of speech and language processing has examined responses to words or sentences presented in quiet conditions. How this network is altered when listening conditions are challenging will be discussed in Sect. 6.4.

6.3 The Cognitive Resources Recruited to Meet Challenges Resulting from Different Types of Adversity

The listening conditions of everyday life are highly variable. Sometimes speech is heard in quiet. More often, however, it is degraded or masked by other sounds. Such challenging situations increase processing demand (also referred to as processing load) when, for example, the stimulus is masked by interfering background noise or by speech from other talkers, or because the stimulus is degraded due to peripheral hearing loss. No specialized cognitive module fully accommodates the myriad of challenges one might encounter in everyday listening conditions – the mechanisms underlying the perception and understanding of speech are simply too distributed, and different challenges are met in different ways. Thus, while Sect. 6.3 discusses different types of adverse listening conditions separately for the sake of tractability, it should be remembered that in many real-world listening environments, more than one kind of listening challenge may be present at one time.

6.3.1 Masking

Masking can be defined as "the process by which the threshold of hearing for one sound is raised by the presence of another" (ANSI 2013). For example, the amplitude threshold for understanding a friend's speech will be increased if they are talking over a roaring waterfall or over a professor delivering a lecture, relative to a quiet environment. Yet, as this example highlights, the "masking sound" is always defined relative to the target speech and thus can be acoustically highly variable, ranging from broadband noise (as is the case with the waterfall) to a single talker (as is the case with the professor). As such, researchers have drawn a conceptual distinction between types of masking sounds to clarify whether the masking is

energetic or *informational* in nature (Brungart et al. 2001). These categories of masking, in addition to the mechanisms required to overcome them, are considered in Sects. 6.3.1.1, 6.3.1.2, and 6.3.1.3.

6.3.1.1 Energetic Masking

Energetic masking is thought to occur when the target sound and interfering sound overlap in time and frequency in the cochlea (e.g., Culling and Stone 2017), such as the detection of speech in broadband noise (like the waterfall example provided in the previous paragraph). Energetic masking poses a challenge for listeners because the masking noise interferes with the target speech at the level of the auditory nerve. Thus, energetic masking, as well as the mechanisms thought to provide a release from energetic masking, is typically discussed in terms of the auditory periphery, though in some cases the proposed explanations require some consideration of cognitive mechanism.

The effects of energetic masking also appear to be lessened when the masking noise is amplitude modulated, with optimal target speech intelligibility occurring around a 10 Hz modulation rate (Miller and Licklider 1950). This relative benefit of modulating the masker noise is presumably due to listeners being able to selectively process the target stimulus in the low-amplitude periods of the masker noise, which has been referred to as "dip listening" (Culling and Stone 2017). Importantly, with respect to a discussion of cognitive mechanism, dip listening appears to relate to masker familiarity, suggesting an influence of learning and memory on selective processing. Specifically, Collin and Lavandier (2013) demonstrated that masker modulations based on the same speech token are easier to cope with compared to masker modulations based on variable speech tokens. These findings suggest that the predictability of amplitude modulation in the masker stimulus is informative in modeling the relative benefit of dip listening, which points to a role of learning-driven familiarity on dip listening efficacy.

6.3.1.2 Informational Masking

Informational masking is the term for all other forms of masking that are not energetic. As the signal is physically not interfered with at the periphery, informational masking is thought to operate at a more central (rather than peripheral) level. Consequently, it is more frequently discussed in terms of underlying cognitive mechanisms. Research has indicated the conditions under which informational masking is thought to occur (e.g., see Kidd and Colbourn 2017). Broadly defined, informational masking can be thought of as an increased challenge in understanding due to the perceived similarities between a target and masker stimulus, even when the target and masker stimuli do not overlap in frequency or time. As such, cognitive processes such as selective attention, divided attention, and working memory are important factors in understanding both informational masking and how to mitigate it.

One of the clearest demonstrations of informational masking comes from studies in which listeners misattribute entire words or phrases spoken by a masker talker to the target talker (Brungart et al. 2001), as this kind of pattern cannot be explained in terms of poor audibility resulting from energetic masking. For example, in a popular paradigm known as the coordinate response measure (CRM; Bolia et al. 2000), listeners hear two or more talkers simultaneously say a sentence with the structure "Ready [call sign] go to [color] [number] now." Participants must listen for their designated call sign on each trial and then navigate to the appropriate coordinate (in a color-number grid). In order to succeed at the task, participants must not confuse the coordinates of the target talker with those spoken by the masker(s).

In these kinds of listening situations, research has established a relationship between accurate speech recognition and cognitive functioning, at least for older individuals with hearing loss. For example, Humes et al. (2006) investigated younger (non-hearing-impaired) and older (hearing-impaired) listeners' performance on the CRM, finding that situations in which listeners were required to divide attention resulted in consistently worse performance for the older, compared to younger, listeners. Moreover, individual differences in short-term memory and working memory (operationalized as an average score of forward and backward digit span) were related to accurate speech recognition among the older listeners. These results suggest that the attentional demands of informational masking may lead to population differences among older and younger listeners who differ in hearing impairment, although individual differences in short-term and working memory may provide a particular benefit for hearing-impaired older listeners, presumably due to a better control of attention and an ability to actively maintain a greater number of hypotheses about what was said by each talker, which may help to resolve ambiguity.

More broadly, working memory appears to be important for release from informational masking when the linguistic content of the target and masker are semantically confusable. In an experiment by Zekveld et al. (2013), listeners had to detect a target sentence that was played simultaneously with stationary noise, amplitude-modulated noise modeled on a speech envelope, or a single talker. The target sentence, moreover, could be preceded by a word that was semantically related to the sentence or an unrelated nonword. Results demonstrated that working memory positively related to sentence comprehension under specific conditions – namely, when the target sentence was preceded by a semantically related word and when the masker stimulus was a single talker. These findings highlight how higher working memory may help listeners to more effectively use a meaningful cue to attend to a target talker, at least in situations where the masker is easily confusable with the target.

Given the demands of informationally masked speech on working memory and attention, it is possible that interventions aimed at improving the functioning of these cognitive constructs may result in better speech comprehension. In support of this framework, Ingvalson et al. (2015) found improvements in both reading span (a memory span task thought to index working memory) and speech-in-noise perception after 10 days of training on backward digit span, although the speech-in-noise

tasks used nonspeech environmental sounds (e.g., dog barks) as an informational masker, which may have different properties than speech used as an informational masker. In contrast, other work did not reveal a benefit of working memory training on speech-in-noise performance (Wayne et al. 2016). In this experiment, the masker stimulus was another talker. More research in this area is clearly needed.

Another means of improving listeners' abilities to understand informationally masked speech, which has received considerable empirical support, is to increase the perceptual familiarity with one of the talkers. Listeners can learn talker characteristics that lead to advantages in understanding informationally masked speech (Nygaard and Pisoni 1998). Importantly, this talker-familiarity advantage is not simply driven by heightened attention to, or salience of, the familiar talker; listeners also show enhanced performance when a novel talker is the target stimulus and the familiar talker is the masking stimulus. This suggests that familiarity may more broadly allow for the segregation of similar talkers into distinct auditory streams (Johnsrude et al. 2013).

To conclude, informational masking appears to pose a problem for listeners because the target and masker signals are often confusable in terms of linguistic content, which places particular demands on listeners' working memory and attention abilities to successfully parse these signals. Training programs that specifically target working memory have shown some transfer to perceiving informationally masked speech, but the evidence for this transfer is mixed. Long-term familiarity with a talker may improve speech intelligibility and reduce the demands of working memory and attention in part because listeners are more effectively able to orient their attention toward (or away from) the familiar talker, allowing greater segregation of auditory streams.

6.3.1.3 Spatial Release from Masking

One well-studied means of dealing with both energetic and informational masking is to use spatial cues to segregate the target speech from the masker stimulus, assuming such cues are present. Revisiting the scenario of conversing with a cashier in a crowded grocery store, spatial release from masking would help one differentiate the speech of one's cashier from, say, a cashier at another register, simply because these two sound sources are physically separated in space. Spatial release from masking is a particularly effective means of improving speech intelligibility across a wide range of masker stimuli. This is because the spatial separation between a talker and masker signal will result in both sounds reaching one's ears at slightly different times, with different loudness levels, and even different distributions of frequency components, due to the fact that sounds may be altered by the "acoustic shadow" of one's head, as well as by the shape of one's ears. These differences provide several cues that listeners may use to effectively segregate sound sources and improve comprehension.

For example, if a target and a masker sound are spatially separated, one ear may receive a more favorable SNR than the other. Listeners appear to be able to select

the ear with a higher SNR – an ability also referred to as "better-ear listening" (Edmonds and Culling 2006). The precise way in which listeners are able to ultimately select the more favorable ear is not completely understood, though it appears to be a "sluggish" process, meaning listeners cannot rapidly shift to take advantage of the relatively more favorable SNR (Culling and Mansell 2013). Further, selective attention to a given ear may alter the physiological response of the outer hair cells in the unattended ear (Srinivasan et al. 2014), altering the effective SNR in that ear.

A second way listeners can separate sounds based on spatial location is through binaural unmasking. Binaural unmasking occurs because, if the target and masker are at different locations, the phase or level difference between the two ears will be different for the two different sounds.

A study by Kidd et al. (2010) examined the acoustic factors that influence spatial release from both informational and energetic masking. In their paradigm, which assessed speech intelligibility using the CRM (see Sect. 6.3.1.2), the target speech stimuli were filtered into several frequency bands. Importantly, the authors found the greatest spatial release from masking when the stimulus was presented at full bandwidth (not filtered), suggesting an integration of binaural cues (phase and level differences) across different frequency regions help to improve performance. The next best spatial release from masking, however, was found for low-frequency components, suggesting that phase differences may be more important than level differences. In a second study, in which energetic and informational masking were varied and listeners could only rely on timing differences between the ears, the authors (Kidd et al. 2010) found large spatial release from masking only when there was significant informational masking. Taken together, these results highlight the importance of considering the extent to which a masker is energetic or informational, as well as the relative contribution of different cues to spatial localization in characterizing the speech intelligibility benefits that may arise from spatial release of masking.

6.3.2 Unfamiliar Talker

Even in favorable listening environments, with little to no masking noise, speech perception can pose a challenge if the talker is unfamiliar. The extent to which understanding an unfamiliar talker poses a challenge, in many cases, depends on the relative difference in accent between the speaker and the listener (Adank et al. 2009). This is because listeners who encounter a nonnative accent or an unfamiliar native accent must rapidly adapt to this variation in speaking, which often permeates multiple levels of the hierarchy of speech. For example, perceiving nonnative accented speech can be challenging when speakers produce contrasts that are not present in their native language, such as the /r/–/l/ contrast in English for native Japanese speakers (Bradlow et al. 1997). At a more suprasegmental level, nonnative speakers sometimes cannot produce the native stress and intonation patterns that help listeners parse the speech signal into meaningful words and phrases (Guion

et al. 2004). However, it should be noted that improvements in the production of native-like speech can be observed in adults after training (Lim and Holt 2011), highlighting the importance of learning and plasticity in the sensorimotor representations of nonnative speech categories.

Despite these challenges, listeners are often able to adapt to accented speech rather quickly, at least in specific listening environments. In a speeded word comprehension task, Clarke and Garrett (2004) found that listeners were initially slower to respond to nonnative accented speech, suggesting that there is an additional processing cost for comprehending unfamiliar speech. This relative slowdown, however, was rapidly attenuated (but not eliminated) over the course of just a few trials. This rapid accommodation, however, has been found to interact with the background noise of the listening environment. Under "quiet" listening conditions, the relative processing cost between accented and non-accented speech is small or sometimes not observed at all (Floccia et al. 2006) and can be mitigated through relatively little experience with the unfamiliar talker. Yet, in more adverse listening situations (such as the introduction of energetic or informational maskers), the relative difference between familiar and unfamiliar accented speech becomes significantly more pronounced.

This interaction between the noisiness of the listening environment and the understanding of unfamiliar speech has also been found with computer-synthesized speech, suggesting that it reflects a broader principle of unfamiliarity with the particular phonetic variation of a given talker rather than specific idiosyncrasies with a particular type of accent (Pisoni et al. 1985). In this experiment, the researchers compared several text-to-speech synthesizers to natural speech, finding that the relative difference in comprehension between synthetic and natural speech was magnified under more adverse (noisier) listening conditions. Moreover, the authors found that semantic and syntactic contexts were important components of intelligibility, which means that the relative challenges to comprehension posed by speaker unfamiliarity can be reduced by constraining the possibilities of a given speech token.

What cognitive mechanisms allow listeners to adapt to unfamiliar talkers in these listening situations? The observation that listeners show rapid improvements in understanding an unfamiliar talker suggests a kind of internal calibration, dependent on the degree to which stored phonological and lexical representations overlap with the incoming speech signal (Van Engen and Peelle 2014). This internal calibration may depend in part on working memory (Janse and Adank 2012), but the generalizability of this claim is unclear given that it was supported by a study using older listeners as participants, who may face unique challenges in speech perception and thus may recruit cognitive resources differently (see Sect. 6.3.3). Indeed, among younger listeners, the role of working memory has been less strongly supported in unfamiliar speech recognition and may be mediated by general vocabulary knowledge (Banks et al. 2015).

Successful adaptation to an unfamiliar talker may require inhibitory mechanisms. This is because an unfamiliar talker may pronounce a given word in a manner that more strongly aligns with a different representation for a listener. For example, using the /r/-/l/ contrast from above, a native Japanese speaker may

pronounce “rake” closer to “lake,” and thus a listener must inhibit “lake” to facilitate understanding, especially in situations where the context of the accompanying speech does not clearly constrain the interpretation of the word (e.g., the sentence “The rake/lake is big”). In support of this hypothesis, Banks et al. (2015) demonstrated that inhibitory control – measured through a Stroop Task – predicted the speed and efficacy of adapting to an unfamiliar talker, though in their paradigm the unfamiliar speech was simultaneously presented with speech-shaped background noise. Although this choice in experimental design is certainly justified, especially given the augmented effects of unfamiliar talker adaptation when listening in a noisy environment, an important consideration in any discussion of mechanism is whether adaptation to an unfamiliar talker in noise reflects the same cognitive processes as those required in less adverse listening conditions. As such, it will be important to clarify in future research whether the cognitive mechanisms that allow an individual to adapt to an unfamiliar talker are identical (and just more heavily recruited) in noisy environments, or whether the presence of an unfamiliar talker in conjunction with noise results in an emergent set of required cognitive processes.

6.3.3 The Effect of Aging

The discussion of adverse listening conditions thus far highlights that both external factors (such as the presence of energetic or informational maskers) and internal factors (such as the degree of overlap in accented speech with one’s mental representations) can influence the ease with which speech can be understood. This illustrates the importance of considering the interaction of the individual’s cognitive resources – knowledge and abilities – with the challenges imposed by the listening environment when discussing speech perception in adverse conditions.

Yet, in this framework one cannot simply assume that the individual’s abilities remain constant across the lifespan. For example, older listeners often have difficulties in understanding speech, especially when it is heard in a noisy environment (see Rogers and Peele, Chap. 9). A detailed discussion of how aging influences speech perception is beyond the scope of this chapter; however, research in this area has highlighted that the relationship between aging and speech perception is likely grounded in changes to both perceptual and cognitive processes. More specifically, age-related declines in sensory processing may increase the perceptual challenge of any given listening environment, which in turn may place greater demands on cognitive processes, such as selective attention and working memory, for successful comprehension (see Wayne and Johnsrude 2015 for a review). However, given that aging is also associated with declines in cognitive functioning, older listeners may have increased difficulties engaging these cognitive processes in service of speech understanding. Training programs designed to improve cognitive processes such as working memory among older listeners have generally produced null or minimal transfers to speech perception in adverse conditions (e.g., Wayne et al. 2016), and

consequently the best approach to reducing listening effort and increasing speech comprehension among elderly listeners is still actively debated.

6.4 Neuroimaging Evidence That Different Demands Recruit Different Systems

Just as the cognitive mechanisms that help listeners cope with adverse conditions depend on the particular elements of the listening environment, the neural mechanisms associated with speech recognition in adverse conditions depend on the specific factors that make a listening situation difficult. As such, it is inappropriate to think of any single brain area as responsible for accommodating "adverse conditions," broadly defined. Rather, neuroimaging research has identified consistent brain networks that are engaged as listeners cope with specific kinds of adverse conditions.

Before discussing these networks, it is important to highlight a methodological consideration in this research area. One of the most commonly used methods for investigating the neural underpinnings of speech perception in adverse conditions is fMRI, a noninvasive technique that provides relatively poor temporal resolution (given the lag of the hemodynamic response in response to neural activity) but good spatial resolution across the whole brain, making it particularly well-suited to studying networks serving complex behaviors such as speech perception. Yet, fMRI generates considerable acoustic noise that can energetically mask speech during image collection. To address this issue, researchers generally use a technique called "sparse scanning," in which the (noisy) process of image acquisition is confined to periods directly before and after, but not during, the presentation of speech (Hall et al. 1999).

Using fMRI sparse scanning, Davis and Johnsrude (2003) presented listeners with sentences that had three kinds of acoustic distortions (vocoded, interrupted, and energetically masked speech) applied to a varying degree, thus creating different levels of intelligibility. Whereas areas close to primary auditory cortex bilaterally were differentially activated for each of the acoustic distortion types – suggesting a kind of sound-form-based processing – the authors found several areas that were invariant to the acoustic distortions but sensitive to overall intelligibility, including the left IFG, hippocampus, and portions of the middle and superior temporal gyri. One explanation of these results, which supports a hierarchical view of speech processing, is that these acoustically invariant areas may modulate attention in service of understanding speech in adverse conditions. This hierarchy, however, does not necessarily imply "top-down" effects from frontal areas on auditory cortex; in fact, the timing of activation may be more parsimoniously understood as reflecting a feedforward process extending from auditory areas to a more distributed frontal and temporal network.

These findings highlight the importance of separating processing of the acoustic properties of distorted speech, from processing of intelligible speech. But speech

may vary in intelligibility and comprehensibility in very different ways. As discussed previously, speech may be difficult to understand because it is masked by noise that competes with the speech signal at the level of the auditory nerve (energetic masking); because the accent of the talker is different from the listener (accented speech); or because there is a competing talker whose speech may be confusable with the target talker (informational masking). If, at the same time, speech is challenging to understand at a linguistic level because it, for example, incorporates words with multiple meanings, or complex syntactic structures, these further add to the demand on cognitive resources. Given the differences in perceptual and cognitive processing required to successfully accommodate all these challenges, it is reasonable to expect differential neural involvement (Scott and McGettigan 2013).

Energetic masking has been associated with the broad recruitment of frontal and parietal regions, including the IFG, frontal operculum, and angular gyrus (Adank et al. 2012). Moreover, individual differences in cognition modulate the degree to which contextual cues benefit speech-in-noise perception, which is associated with differential activation in IFG and angular gyrus (Zekveld et al. 2012). Taken together, these findings suggest that perceiving speech in noise involves an interaction between auditory and frontoparietal areas, with factors such as context and individual differences in frontally mediated executive functions influencing the way in which these areas interact.

Informational masking, on the other hand, most prominently appears to recruit superior temporal areas (Mattys et al. 2012). This pattern of activity largely overlaps with the areas that are involved in processing clear speech without a masker, which makes sense given the similarity of the masker to the target. However, more extended activation has also been observed in situations where the masking speech is highly similar to the target speech (Nakai et al. 2005), including dorsolateral prefrontal cortex, anterior cingulate, and premotor areas. This in turn suggests that nonauditory regions, thought to underlie executive functions such as cognitive control, may be recruited depending on the perceived challenge of the adverse listening situation, above and beyond its acoustic and linguistic factors. This will be discussed in more detail in Sect. 6.4.3.

A distributed network of brain regions appears to be involved in accommodating accented speech. The regions involved may look more varied than they actually are because of the methodological difficulties in equating acoustic factors and comprehension difficulty across different participant samples. Put another way, there are many ways to operationalize accented speech, and these might pose different kinds of challenges to listeners depending on the particular accent of the participant sample. With this caveat in mind, the neural areas implicated in accommodating accented speech partially overlap with areas implicated in both energetic and informational masking (see Adank et al. 2015). Similar to informational masking, listening to accented speech results in greater activation of bilateral superior temporal areas (Adank et al. 2012), presumably due to greater auditory and phonological processing demands. Accented speech also engages regions around the supplementary motor area (SMA), left IFG, and frontal operculum, which has at least two

possible explanations, depending on the precise regions involved. One is that a network supporting cognitive control (the cinguloopercular network) has been activated due to the perceived difficulty of the task. This possibility will be discussed in Sect. 6.4.3. The other explanation is that, given the possible overlap between these areas and motor speech regions, listeners may recruit a speech motor network to simulate the production of the accented speech, direct attention to the most diagnostic features for successful recognition, and inhibit representations that may conflict with the auditory input (cf. Banks et al. 2015). This will be discussed further in Sect. 6.4.2.

6.4.1 Listening to Speech While Doing Something Else

When the sensory information at the ear is too ambiguous to support speech recognition by itself, knowledge-guided processes that help to interpret and repair the degraded signal are required. Many of these processes appear to be effortful and may not be recruited when attention is elsewhere. For example, imagine conversing with a friend at a hockey game. There are several potential energetic maskers (e.g., the synchronous roar of the crowd when a goal is scored) and informational maskers (e.g., the nearby conversations taking place), making speech comprehension more difficult and presumably effortful. Now, in this environment, imagine that, in the middle of the friend telling a story, one's attention is captured by the action of the hockey game. How well would the friend's speech be perceived? This is hard to study behaviorally, since it is difficult to measure perception of a stimulus to which a participant is not attending. Wild and colleagues (2012) used fMRI to compare processing of speech under full attention and under distraction. On every trial, young adult listeners with normal hearing attended to one of three simultaneously presented stimuli: an everyday, meaningful sentence (at one of four acoustic clarity levels), an auditory distracter, or a visual distracter. A post-scan recognition test showed that clear speech was processed even when not attended, but that attention greatly enhanced the processing of degraded speech. Furthermore, speech-sensitive cortex could be fractionated according to how speech-evoked responses were modulated by attention, and these divisions appeared to map onto the hierarchical organization of the auditory system, as discussed in Sect. 6.2.3. Only in middle temporal and frontal regions – regions corresponding to the highest stages of auditory processing – did activity appear to be enhanced by attention.

In a follow-up experiment, Ritz et al. (2016) pushed the paradigm, increasing the intelligibility of the degraded speech so that all words from sentences could be reported correctly when the sentences were attended (through the use of 12-band noise vocoding, referred to as NV12), and introducing a multiple object tracking (MOT) task with a parametrically varying number of moving dots to track (1, 3, 4, or 6). Both types of stimuli were presented on every trial, and the participant was cued at the beginning of each trial to attend to one or to the other. The results were striking and are shown in Fig. 6.5b. In anterior temporal cortex (yellow/orange

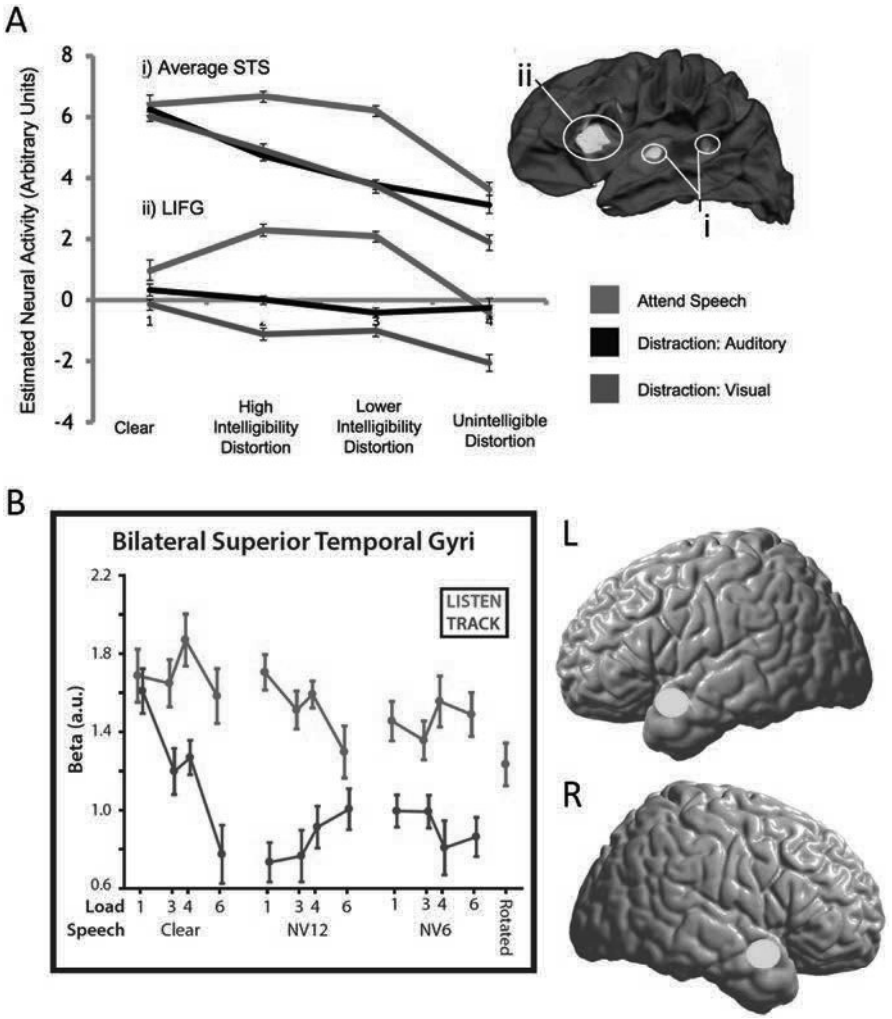

Fig. 6.5 (**a**) Activity in the left superior temporal sulcus (STS; i) and in left inferior frontal gyrus (IFG; ii) depends on attentional state and speech quality. In both regions, activity is enhanced when listeners attended to speech compared to when they attended to concurrently presented visual or auditory distractor, performing a target-detection task on these (Wild et al. 2012). Activity is particularly enhanced for degraded, but intelligible, distorted speech (the distortion was created through noise vocoding) (Shannon et al. 1995). When listening with no distractors, a pilot group could report 90% of the words from high-intelligibility distorted sentences, and 70% of the words from low-intelligibility distorted sentences. (Adapted from Wild et al. (2012)). (**b**) Activity in bilateral anterior temporal regions (shown schematically in yellow) depended on attentional state and speech quality in an unpublished study (Ritz et al., MSc thesis). As indicated by the red lines, activity was high when listeners were attending to speech, regardless of whether it was clear, very high-intelligibility 12-band noise-vocoded speech (NV12), or lower intelligibility 6-band NV speech (NV6). It was lower but still elevated when attending to "rotated speech" – this is completely unintelligible noise-vocoded speech. When listening with no distractors, a pilot group

(continued)

clusters in Fig. 6.5b), activity when attending to speech was uniformly quite high, whereas when the MOT task was attended, it was high only for clear speech, and only when one object was being tracked. The higher the MOT load, the lower the activity in this area. For degraded speech that was 100% intelligible (NV12 in Fig. 6.5b), a marked effect of attentional state was evident – even at the lowest MOT load (one object), activity when attending to speech was much higher than when attending to the MOT task. These results suggest that whereas these anterior temporal regions can process clear speech in the absence of attention, as long as the distractor task is not too demanding, processes involved in the comprehension of even lightly degraded speech critically require focused attention.

In a series of studies, Mattys and colleagues explored how additional concurrent processing load alters the processing of simultaneously presented spoken words. They found that processing speech under conditions of divided attention relies on different mechanisms compared to those involved in processing speech when attention is focused solely on speech. When listeners were required to listen to speech and perform a visual search task, they reweighted information in making perceptual decisions (Mattys et al. 2014). Moreover, they seemed to rely more on lexical semantic information for word segmentation, and on lexical knowledge for phoneme identification, than they would without a concurrent task. In contrast, they seemed to rely less on acoustic cues conveyed in fine phonetic detail. It may seem counterintuitive that, as load on central cognitive resources increases, listeners rely more, not less, on knowledge-guided factors (which presumably rely on the same central cognitive resources) for speech perception. This reweighting of cues may be due to poorer registration of the fine phonetic detail when distracted (Mattys and Palmer 2015). These studies are important because they indicate that attentional manipulations do not simply impair perception but instead qualitatively change perceptual decision criteria.

6.4.2 The Importance of Motor Representations

In a chapter focusing on speech perception in adverse listening conditions, it may seem initially inappropriate to devote a section to speech-motor representations. Yet, as briefly mentioned in the introduction of Sect. 6.4, certain kinds of adverse listening conditions (such as accented speech perception) have been associated with

---

could report 100% of the words from NV12 sentences, and 94% of the words from NV6 sentences. When sentences were heard while listeners focused on a distracting multiple object tracking (MOT) task (see text for details), activity was low even at the lowest level of MOT load (1 dot) when speech was even slightly degraded. When speech was clear and the MOT load was low (1 dot to track), there was no effect of attention in this area, and activity declined to the levels seen for degraded speech as tracking load increased. The y-axis is dimensionless beta weights (arbitrary units). On the x-axis, "load" (values 1–6) is the number of dots tracked during a concurrent MOT task. "Speech" is the speech type

the activation of motor and premotor cortex, suggesting that the mechanisms underlying the planning and production of speech may improve speech perception at least in some listening situations.

The broader discussion of how motor representations specifically relate to speech perception has a long history. In its strongest form, the motor theory of speech perception asserts that speech is not understood through the perception of its auditory components but rather through more abstract and invariant articulatory gestures (Liberman and Mattingly 1985). Although the motor theory of speech perception has been the subject of considerable debate (e.g., see Lotto et al. 2009), an increasing body of research has supported at least some motor involvement in the perception of speech, particularly in adverse listening conditions. Researchers have used transcranial magnetic stimulation (TMS) to alter the excitability of motor cortex and have demonstrated enhancement of motor-evoked potentials (MEPs) from lip and tongue muscles when listening to speech (Fadiga et al. 2002). These studies were conducted using clear speech, but subsequent work demonstrates that motor activation may contribute to categorical speech perception under adverse listening conditions. In an fMRI study, Du et al. (2014) asked participants to identify phoneme tokens presented at different SNRs. Activity correlated negatively with perceptual accuracy in left ventral premotor cortex and a region anterior to it (anatomically defined Broca's area). Furthermore, pattern-information analysis revealed that whereas phonemes could not be reliably discriminated in patterns of activity in bilateral auditory cortex except when the noise level was very low, representations of phonemes remained robust in ventral premotor and Broca's areas at much higher levels of noise. This suggests a role for motor regions in categorical perception of degraded speech sounds.

The involvement of motor representations in speech perception appears to depend on attention. Using TMS to temporarily disrupt motor areas associated with lip movements, Möttönen et al. (2014) demonstrated that auditory representations of lip- and tongue-articulated speech sounds (/ba/, /da/, and /ga/) were differentially modulated based on attention. When the sounds were attended to, the TMS-related modulation in auditory cortex was relatively early and strongly left lateralized; when the sounds were not attended to, the modulation in auditory cortex was later and not lateralized. These results thus support the hypothesis that motor cortex can influence the response properties of auditory cortex in the context of speech perception, but the precise interaction between these areas may critically depend on attention.

6.4.3 The Cinguloopercular Network

For nearly 20 years, it has been clear that several distinct tasks recruit a common network involving dorsolateral prefrontal cortex and anterior insula, dorsal anterior cingulate cortex, and the adjacent pre-supplementary motor area (Duncan 2010). This cinguloopercular (CO) network appears to become active whenever cognitive

demands are high, consistent with proposals that it is involved in cognitive control, specifically in performance monitoring (Dosenbach et al. 2006). This network appears to be recruited whenever a listener is attempting to understand speech that is challenging, either because the speech has been degraded or because a linguistic challenge, such as semantic ambiguity, has been imposed (Davis and Johnsrude 2003; Rodd et al. 2005). This elevated CO response, however, does not simply reflect challenge. Vaden et al. (2013) demonstrated that CO activity predicted word recognition on the next trial, which is similar to what has been noted in visuospatial tasks. The pattern of results suggests that the CO network is important for *adaptive* cognitive control. Furthermore, the results of Wild et al. (2012) indicate that this adaptive control may require focused attention on the difficult-to-understand speech signal.

6.5 Listening Effort

It has become increasingly evident to hearing-aid manufacturers and auditory researchers that "effortful listening" is an essential construct to consider. Two people might comprehend the same amount of speech in a given challenging listening situation, but one listener may feel that it was effortful and tiring, whereas another listener might have found it effortless. The first listener may alter their behavior to avoid such situations, or, if listening through a hearing prosthesis such as a hearing aid, may choose not to use it. Thus, the concept of "listening effort" may be a powerful predictor of behavior, independent of comprehension.

"Listening effort" is typically considered to be a unitary phenomenon and is studied as such. However, it has at least two different meanings. On the one hand, researchers write about listeners exerting effort. For example, Pichora-Fuller et al. (2016) define mental effort as the "deliberate allocation of mental resources to overcome obstacles in goal pursuit…" (p. 10S). In this sense, it is a process or brain activity. At the same time, listeners are aware of processing being fatiguing or effortful. In this sense, listening effort is a percept. Typically, listening effort is measuring using questionnaires – such subjective measures are focused on the explicit percept (e.g., Johnson et al. 2015). Physiological measures such as pupillometry and imaging (fMRI or EEG) have become more common tools (Peelle 2018). These may be sensitive either to mental exertion or the perception of difficulty or both; it is not presently clear.

As something that listeners perceive, listening effort may be most productively considered as an interaction between the perceptual, linguistic, or task challenges imposed by a listening situation and the cognitive resources that the listener brings to bear. Individual differences in cognitive resources (such as memory, perceptual learning, processing speed, fluid intelligence, and control processes) that permit one person to cope more efficiently or more successfully than another with the challenges imposed by a listening situation will have a strong influence on perceived effort (see Fig. 6.6). Although listening effort is usually measured in a

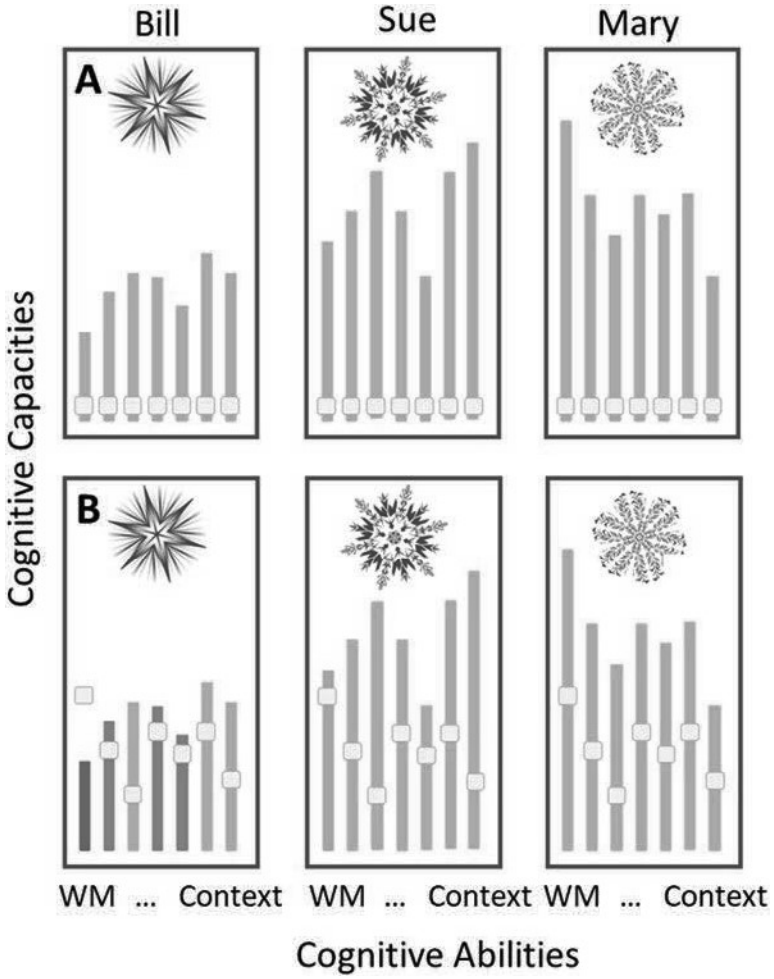

Fig. 6.6 (**a**) Three different individuals (represented by the unique purple snowflakes) and their distinct cognitive profile across seven putative abilities that are all relevant to speech perception in adverse conditions. Each bar is meant to represent an ability associated with speech perception, and the height of the bar indicates the strength of the ability. For example, the leftmost bar in each plot could be indexing working memory. (**b**) The seven white squares in each panel illustrate the cognitive demands imposed by a given listening situation. Note that the cognitive demands are the same across individuals. However, the degree to which each listener can respond to those demands depends on their individual cognitive profile. Demands fully occupy or outstrip several of the cognitive abilities for the listener on the left (highlighted in red). In contrast, the abilities of the listener on the right are more than adequate to cope with the demands – none of the squares are near the top of the ability bars (highlighted in green). The listener on the left will perceive effort (unless they give up), whereas the one on the left will find the listening situation effortless. This figure demonstrates how effort results from the interaction between the demands of a given listening situation (the position of the sliders) and an individual's cognitive abilities

unidimensional way (on a subjective questionnaire, or with pupillometry), it is probably not a unidimensional construct – different challenges are met in different ways. This framework enables researchers to cognitively and anatomically separate different processes, related to signal extraction, recovery, and repair that may contribute to the feeling of listening effort. At the same time, researchers can study factors that may alleviate listening effort, such as familiarity with someone's voice, or flexible and accurate use of meaningful context.

6.6 Chapter Summary

Speech is a complex and highly variable signal. Aspects of the speech signal itself (unfamiliar accents, semantic ambiguity, syntactic complexity), background signals (sound that either energetically or informationally masks a target speech signal), and listener-specific factors (selective attention and cognitive control abilities, familiarity with specific talkers or linguistic contexts) all contribute to how a given listening situation poses a challenge to recognition. Successful recognition of speech in adverse listening conditions therefore relies on interacting perceptual, cognitive, and linguistic factors. Some of these factors may be influenced considerably by learning, as seen with improved speech recognition for highly familiar talkers. Other factors, however, appear less susceptible to training, as seen with the mixed evidence of working memory training transferring to speech-in-noise perception. Although different adverse conditions place differential demands on cognitive resources, a consistent finding – supported behaviorally and neurally – is that adverse listening conditions place considerable demands on attention. Thus, compared to relatively clear listening condition, adverse listening conditions are served by the recruitment of additional brain networks – such as the CO network – even when both kinds of speech are equally intelligible. In this sense, the CO network may be viewed similarly to an "engine light" of a car, signaling an increase in mental effort but not specifically diagnosing the nature of the particular listening challenges in the moment. The emergence of "listening effort" as a construct, which represents the interaction between listening demands, and individual capacity across cognitive domains, may provide an important framework going forward for discussing speech perception in adverse listening conditions. Although the best operationalization of listening effort is still unclear and likely depends on the research question being addressed, it is clear that both listener-focused and signal-focused variables must be considered to fully understand speech perception in adverse listening conditions.

Compliance with Ethics Requirements Ingrid Johnsrude declares that she has no conflict of interest.

Stephen Van Hedger declares that he has no conflict of interest.

References

Adank P, Evans BG, Stuart-Smith J, Scott SK (2009) Comprehension of familiar and unfamiliar native accents under adverse listening conditions. J Exp Psychol Hum Percept Perform 35:520–529. https://doi.org/10.1037/a0013552

Adank P, Davis MH, Hagoort P (2012) Neural dissociation in processing noise and accent in spoken language comprehension. Neuropsychologia 50:77–84. https://doi.org/10.1016/j.neuropsychologia.2011.10.024

Adank P, Nuttall HE, Banks B, Kennedy-Higgins D (2015) Neural bases of accented speech perception. Front Hum Neurosci 9:1–7. https://doi.org/10.3389/fnhum.2015.00558

ANSI. (2013). *American National Standard Acoustical Terminology, ANSI S1.1-2013*. New York: American National Standards Institute.

Baddeley AD, Hitch G (1974) Working memory. Psychol Learn Motiv 8:47–89. https://doi.org/10.1016/S0079-7421(08)60452-1

Banks B, Gowen E, Munro KJ, Adank P (2015) Cognitive predictors of perceptual adaptation to accented speech. J Acoust Soc Am 137:2015–2024. https://doi.org/10.1121/1.4916265

Bates E, Wilson SM, Saygin AP et al (2003) Voxel-based lesion–symptom mapping. Nat Neurosci 6:448–450. https://doi.org/10.1038/nn1050

Binder JR, Desai RH, Graves WW, Conant LL (2009) Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. Cereb Cortex 19:2767–2796. https://doi.org/10.1093/cercor/bhp055

Bolia RS, Nelson WT, Ericson MA, Simpson BD (2000) A speech corpus for multitalker communications research. J Acoust Soc Am 107:1065–1066. https://doi.org/10.1121/1.428288

Bradlow AR, Pisoni DB, Akahane-Yamada R, Tohkura Y (1997) Training Japanese listeners to identify English / r / and / l /: IV. Some effects of perceptual learning on speech production. J Acoust Soc Am 101:2299–2310. https://doi.org/10.1121/1.418276

Brungart DS, Simpson BD, Ericson MA, Scott KR (2001) Informational and energetic masking effects in the perception of multiple simultaneous talkers. J Acoust Soc Am 110:2527–2538. https://doi.org/10.1121/1.1408946

Clarke CM, Garrett MF (2004) Rapid adaptation to foreign-accented English. J Acoust Soc Am 116:3647–3658. https://doi.org/10.1121/1.1815131

Collin B, Lavandier M (2013) Binaural speech intelligibility in rooms with variations in spatial location of sources and modulation depth of noise interferers. J Acoust Soc Am 134:1146–1159. https://doi.org/10.1121/1.4812248

Culling JF, Mansell ER (2013) Speech intelligibility among modulated and spatially distributed noise sources. J Acoust Soc Am 133:2254–2261. https://doi.org/10.1121/1.4794384

Culling JF, Stone MA (2017) Energetic masking and masking release. In: Middlebrooks J, Simon J, Popper A, Fay R (eds) The auditory system at the cocktail party. Springer handbook of auditory research, vol 60. Springer, Cham. https://doi.org/10.1007/978-3-319-51662-2_3

Cutler A, Norris D (1988) The role of strong syllables in segmentation for lexical access. J Exp Psychol Hum Percept Perform 14:113–121. https://doi.org/10.1037/0096-1523.14.1.113

Darwin CJ, Carlyon RP (1995) Auditory grouping. In: Moore BCJ (ed) The handbook of perception and cognition, vol 6, Hearing, 2nd edn. Academic Press, San Diego, pp 387–424

Davis MH, Johnsrude IS (2003) Hierarchical processing in spoken language comprehension. J Neurosci 23:3423–3431. https://doi.org/10.1523/JNEUROSCI.23-08-03423.2003

Davis MH, Johnsrude IS (2007) Hearing speech sounds: top-down influences on the interface between audition and speech perception. Hear Res 229:132–147. https://doi.org/10.1016/j.heares.2007.01.014

Denes PB, Pinson EN (1993) The speech chain: the physics and biology of spoken language. W.H. Freeman, New York

Dosenbach NUF, Visscher KM, Palmer ED et al (2006) A core system for the implementation of task sets. Neuron 50:799–812. https://doi.org/10.1016/j.neuron.2006.04.031

Dronkers NF, Wilkins DP, Van Valin RD et al (2004) Lesion analysis of the brain areas involved in language comprehension. Cognition 92:145–177. https://doi.org/10.1016/j.cognition.2003.11.002

Du Y, Buchsbaum BR, Grady CL, Alain C (2014) Noise differentially impacts phoneme representations in the auditory and speech motor systems. Proc Natl Acad Sci 111:7126–7131. https://doi.org/10.1073/pnas.1318738111

Duncan J (2010) The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. Trends Cogn Sci 14:172–179. https://doi.org/10.1016/j.tics.2010.01.004

Edmonds BA, Culling JF (2006) The spatial unmasking of speech: evidence for better-ear listening. J Acoust Soc Am 120:1539–1545. https://doi.org/10.1121/1.2228573

Fadiga L, Craighero L, Buccino G, Rizzolatti G (2002) Speech listening specifically modulates the excitability of tongue muscles: a TMS study. Eur J Neurosci 15:399–402. https://doi.org/10.1046/j.0953-816x.2001.01874.x

Floccia C, Goslin J, Girard F, Konopczynski G (2006) Does a regional accent perturb speech processing? J Exp Psychol Hum Percept Perform 32:1276–1293. https://doi.org/10.1037/0096-1523.32.5.1276

Giraud AL, Lorenzi C, Ashburner J et al (2000) Representation of the temporal envelope of sounds in the human brain. J Neurophysiol 84:1588–1598. https://doi.org/10.1152/jn.2000.84.3.1588

Guion SG, Harada T, Clark JJ (2004) Early and late Spanish–English bilinguals' acquisition of English word stress patterns. Biling (Camb Engl) 7:207–226. https://doi.org/10.1017/S1366728904001592

Hackett TA (2011) Information flow in the auditory cortical network. Hear Res 271:133–146. https://doi.org/10.1016/j.heares.2010.01.011

Hackett TA, de la Mothe LA, Camalier CR et al (2014) Feedforward and feedback projections of caudal belt and parabelt areas of auditory cortex: refining the hierarchical model. Front Neurosci. https://doi.org/10.3389/fnins.2014.00072

Hall DA, Haggard MP, Akeroyd MA et al (1999) "Sparse" temporal sampling in auditory fMRI. Hum Brain Mapp 7:213–223. https://doi.org/10.1002/(SICI)1097-0193(1999)7:3<213::AID-HBM5>3.0.CO;2-N

Hawkins S (2003) Roles and representations of systematic fine phonetic detail in speech understanding. J Phon 31:373–405. https://doi.org/10.1016/j.wocn.2003.09.006

Hickok G, Poeppel D (2015) Neural basis of speech perception. In: Aminoff MJ, Boller F, Swaab DF (eds) Handbook of clinical neurology, 129th edn. Elsevier, pp 149–160

Holmes E, Domingo Y, Johnsrude IS (2018) Familiar voices are more intelligible, even if they are not recognized as familiar. Psychol Sci 29:1575–1583. https://doi.org/10.1177/0956797618779083

Holt L (2005) Temporally nonadjacent nonlinguistic sounds affect speech categorization. Psychol Sci 16:305–312. https://doi.org/10.1111/j.0956-7976.2005.01532.x

Humes LE, Lee JH, Coughlin MP (2006) Auditory measures of selective and divided attention in young and older adults using single-talker competition. J Acoust Soc Am 120:2926–2937. https://doi.org/10.1121/1.2354070

Ingvalson EM, Dhar S, Wong PCM, Liu H (2015) Working memory training to improve speech perception in noise across languages. J Acoust Soc Am 137:3477–3486. https://doi.org/10.1121/1.4921601

Janse E, Adank P (2012) Predicting foreign-accent adaptation in older adults. Q J Exp Psychol 65:1563–1585. https://doi.org/10.1080/17470218.2012.658822

Johnson J, Xu J, Cox R, Pendergraft P (2015) A comparison of two methods for measuring listening effort as part of an audiologic test battery. Am J Audiol 24:419–431. https://doi.org/10.1044/2015_AJA-14-0058

Johnsrude IS, Mackey A, Hakyemez H et al (2013) Swinging at a cocktail party: voice familiarity aids speech perception in the presence of a competing voice. Psychol Sci 24:1995–2004. https://doi.org/10.1177/0956797613482467

Jones EG (2003) Chemically defined parallel pathways in the monkey auditory system. Ann N Y Acad Sci 999:218–233. https://doi.org/10.1196/annals.1284.033

Kidd G, Colbourn HS (2017) Informational masking in speech recognition. In: Middlebrooks J, Simon J, Popper A, Fay R (eds) The auditory system at the cocktail party, Springer handbook of auditory research, 60th edn. Springer International Publishing, Cham, pp 75–109

Kidd G, Mason CR, Best V, Marrone N (2010) Stimulus factors influencing spatial release from speech-on-speech masking. J Acoust Soc Am 128:1965–1978. https://doi.org/10.1121/1.3478781

Kraljic T, Brennan SE, Samuel AG (2008) Accommodating variation: dialects, idiolects, and speech processing. Cognition 107:54–81. https://doi.org/10.1016/j.cognition.2007.07.013

Liberman AM, Mattingly IG (1985) The motor theory of speech perception revised. Cognition 21:1–36. https://doi.org/10.1016/0010-0277(85)90021-6

Liberman AM, Cooper FS, Shankweiler DP, Studdert-Kennedy M (1967) Perception of the speech code. Psychol Rev 74:431–461. https://doi.org/10.1037/h0020279

Lim SJ, Holt LL (2011) Learning foreign sounds in an alien world: videogame training improves non-native speech categorization. Cogn Sci 35:1390–1405. https://doi.org/10.1111/j.1551-6709.2011.01192.x

Lotto AJ, Hickok GS, Holt LL (2009) Reflections on mirror neurons and speech perception. Trends Cogn Sci 13:110–114. https://doi.org/10.1016/j.tics.2008.11.008

Macleod A, Summerfield Q (1990) A procedure for measuring auditory and audiovisual speech-reception thresholds for sentences in noise: rationale, evaluation, and recommendations for use. Br J Audiol 24:29–43. https://doi.org/10.3109/03005369009077840

Mattys SL, Palmer SD (2015) Divided attention disrupts perceptual encoding during speech recognition. J Acoust Soc Am 137:1464–1472. https://doi.org/10.1121/1.4913507

Mattys SL, White L, Melhorn JF (2005) Integration of multiple speech segmentation cues: a hierarchical framework. J Exp Psychol Gen 134:477–500. https://doi.org/10.1037/0096-3445.134.4.477

Mattys SL, Davis MH, Bradlow AR, Scott SK (2012) Speech recognition in adverse conditions: a review. Lang Cogn Process 27:953–978. https://doi.org/10.1080/01690965.2012.705006

Mattys SL, Barden K, Samuel AG (2014) Extrinsic cognitive load impairs low-level speech perception. Psychon Bull Rev 21:748–754. https://doi.org/10.3758/s13423-013-0544-7

Mesulam MM, Wieneke C, Thompson C et al (2012) Quantitative classification of primary progressive aphasia at early and mild impairment stages. Brain 135:1537–1553. https://doi.org/10.1093/brain/aws080

Miller GA, Licklider JCR (1950) The intelligibility of interrupted speech. J Acoust Soc Am 22:167–173. https://doi.org/10.1017/S0031182000023970

Möttönen R, van de Ven GM, Watkins KE (2014) Attention fine-tunes auditory-motor processing of speech sounds. J Neurosci 34:4064–4069. https://doi.org/10.1523/JNEUROSCI.2214-13.2014

Nakai T, Kato C, Matsuo K (2005) An fMRI study to investigate auditory attention: a model of the cocktail party phenomenon. Magn Reson Med Sci 4:75–82. https://doi.org/10.2463/mrms.4.75

Norris D, Mcqueen JM, Cutler A, Butterfield S (1997) The possible-word constraint in the segmentation of continuous speech. Cogn Psychol 34:191–243. https://doi.org/10.1006/cogp.1997.0671

Nygaard LC, Pisoni DB (1998) Talker-specific learning in speech perception. Percept Psychophys 60:355–376. https://doi.org/10.3758/BF03206860

Okada K, Rong F, Venezia J et al (2010) Hierarchical organization of human auditory cortex: evidence from acoustic invariance in the response to intelligible speech. Cereb Cortex 20:2486–2495. https://doi.org/10.1093/cercor/bhp318

Peelle JE (2018) Listening effort: how the cognitive consequences of acoustic challenge are reflected in brain and behavior. Ear Hear 39:204–214. https://doi.org/10.1097/AUD.0000000000000494

Peelle JE, Johnsrude IS, Davis MH (2010) Hierarchical organization for speech in human auditory cortex and beyond. Front Hum Neurosci 4:1–3. https://doi.org/10.3389/fnhum.2010.00051

Phillips DP, Farmer ME (1990) Acquired word deafness, and the temporal grain of sound representation in the primary auditory cortex. Behav Brain Res 40:85–94. https://doi.org/10.1016/0166-4328(90)90001-U

Pichora-Fuller MK, Kramer SE, Eckert MA et al (2016) Hearing impairment and cognitive energy. Ear Hear 37:5S–27S. https://doi.org/10.1097/AUD.0000000000000312

Pisoni DB, Nusbaum HC, Greene BG (1985) Perception of synthetic speech generated by rule. Proc IEEE 73:1665–1676. https://doi.org/10.1109/PROC.1985.13346

Poremba A, Mishkin M (2007) Exploring the extent and function of higher-order auditory cortex in rhesus monkeys. Hear Res 229:14–23. https://doi.org/10.1016/j.heares.2007.01.003

Reisberg D, McLean J, Goldfield A (1987) Easy to hear but hard to understand: a lip-reading advantage with intact auditory stimuli. In: Dodd B, Campbell R (eds) Hearing by eye: the psychology of lip-reading. Lawrence Erlbaum Associates, Inc., Hillsdale, pp 97–113

Ritz H, Wild C, Johnsrude IJ (2016) The effects of concurrent cognitive load on the processing of clear and degraded speech. In: 22nd annual meeting of the Organization for Human Brain Mapping

Rodd JM, Gaskell G, Marslen-Wilson W (2002) Making sense of semantic ambiguity: semantic competition in lexical access. J Mem Lang 46:245–266. https://doi.org/10.1006/jmla.2001.2810

Rodd JM, Davis MH, Johnsrude IS (2005) The neural mechanisms of speech comprehension: fMRI studies of semantic ambiguity. Cereb Cortex 15:1261–1269. https://doi.org/10.1093/cercor/bhi009

Rodd JM, Johnsrude IS, Davis MH (2012) Dissociating frontotemporal contributions to semantic ambiguity resolution in spoken sentences. Cereb Cortex 22:1761–1773. https://doi.org/10.1093/cercor/bhr252

Scott SK, McGettigan C (2013) The neural processing of masked speech. Hear Res 303:58–66. https://doi.org/10.1016/j.heares.2013.05.001

Shannon RV, Zeng FG, Kamath V et al (1995) Speech recognition with primarily temporal cues. Science 270:303–304. https://doi.org/10.1126/science.270.5234.303

Srinivasan S, Keil A, Stratis K et al (2014) Interaural attention modulates outer hair cell function. Eur J Neurosci 40:3785–3792. https://doi.org/10.1111/ejn.12746

Sumby WH, Pollack I (1954) Visual contribution to speech intelligibility in noise. J Acoust Soc Am 26:212–215. https://doi.org/10.1121/1.1907309

Turken AU, Dronkers NF (2011) The neural architecture of the language comprehension network: converging evidence from lesion and connectivity analyses. Front Syst Neurosci 5:1–20. https://doi.org/10.3389/fnsys.2011.00001

Vaden KI, Kuchinsky SE, Cute SL et al (2013) The cingulo-opercular network provides word-recognition benefit. J Neurosci 33:18979–18986. https://doi.org/10.1523/JNEUROSCI.1417-13.2013

Van Engen KJ, Peelle JE (2014) Listening effort and accented speech. Front Hum Neurosci 8:1–4. https://doi.org/10.3389/fnhum.2014.00577

Wayne RV, Johnsrude IS (2015) A review of causal mechanisms underlying the link between age-related hearing loss and cognitive decline. Ageing Res Rev 23:154–166. https://doi.org/10.1016/j.arr.2015.06.002

Wayne RV, Hamilton C, Huyck JJ, Johnsrude IS (2016) Working memory training and speech in noise comprehension in older adults. Front Aging Neurosci 8:1–15. https://doi.org/10.3389/fnagi.2016.00049

Wild CJ, Yusuf A, Wilson DE et al (2012) Effortful listening: the processing of degraded speech depends critically on attention. J Neurosci 32:14010–14021. https://doi.org/10.1523/JNEUROSCI.1528-12.2012

Zekveld AA, Rudner M, Johnsrude IS et al (2012) Behavioral and fMRI evidence that cognitive ability modulates the effect of semantic context on speech intelligibility. Brain Lang 122:103–113. https://doi.org/10.1016/j.bandl.2012.05.006

Zekveld AA, Rudner M, Johnsrude IS, Rönnberg J (2013) The effects of working memory capacity and semantic cues on the intelligibility of speech in noise. J Acoust Soc Am 134:2225–2234. https://doi.org/10.1121/1.4817926