Speech Perception from a Neurophysiological Perspective

Anne-Lise Giraud, David Poeppel

Chapter 9
Speech Perception from a Neurophysiological Perspective

Anne-Lise Giraud and David Poeppel

9.1 Introduction: Terminology and Concepts

Of all the signals human auditory cortex has to process, the one with the most compelling relevance to the listener is arguably speech. Parsing and decoding speech—the conspecific signal affording the most rapid and most precise transmission of information—must be considered one of the principal challenges of the auditory system. This chapter concentrates on what speech perception entails and what the constituent operations might be, emphasizing a neurophysiological perspective.

Research on speech perception is profoundly interdisciplinary. The questions range from (1) characterizing the relevant properties of the acoustic signal (*acoustic phonetics*, *engineering*) to (2) identifying the various (neurophysiological, neurocomputational, psychological) subroutines that underlie the perceptual analysis of the signal (*neuroscience*, *computation*, *perceptual psychology*) to (3) understanding the nature of the representation that forms the basis for creating meaning (*linguistics*, cognitive *psychology*). The entire process comprises—at least—a mapping from mechanical vibrations in the ear to abstract representations in the brain.

One terminological note merits emphasis. *Speech perception* refers to the mapping from sounds to internal linguistic representations (roughly, words). This is not coextensive with *language comprehension*. Language comprehension can be mediated by ear (speech perception), but also by eye (reading, sign language, lip reading), or by touch (Braille). Thus, *speech perception proper comprises a set of auditory processing operations prior to language comprehension*. The failure to distinguish

A.-L. Giraud (✉)
Inserm U960, Département d'Etudes Cognitives, Ecole Normale Supérieure,
29 rue d'Ulm, 75005 Paris, France
e-mail: anne-lise.giraud@ens.fr

D. Poeppel
Department of Psychology, New York University, 6 Washington Place, New York, NY 10003, USA
e-mail: david.poeppel@nyu.edu

D. Poeppel et al. (eds.), *The Human Auditory Cortex*, Springer Handbook of Auditory Research 43, DOI 10.1007/978-1-4614-2314-0_9,

between speech and language has led to much unfortunate confusion; because the goal is to identify the critical component operations that underlie speech (and ultimately language) comprehension, a meticulous subdivision of the relevant cognitive science and linguistics terminology is essential. How does this translate into research practice? Insofar as we are interested in studying properties of words that are central to comprehension, but abstract and independent of the input modality, we would aim to find features that are stable across auditory, visual, or tactile presentation. In contrast, when we study speech perception, we are interested in the attributes that underlie the transformation from an acoustic signal to the possible internal representations. Because speech perception can thus be viewed as a subroutine of language comprehension in which the computation of meaning is not required, it can be approached, at least in part, by investigating the perception of isolated speech sounds (e.g., vowels or consonant-vowel syllables) or single words.

Current models of speech perception (and the associated neurobiological literature) tend to derive from studies of the perception of single speech sounds, syllables, or words. For example, the phenomenon of categorical perception (Liberman et al., 1967) as well as the work on vowel inventories (e.g., Näätänen et al., 1997) has stimulated an enormous literature on understanding sublexical perceptual processes. Aspects related to categorical perception have been examined and reviewed in detail (e.g., Harnad, 1987) and continue to motivate neurobiological studies on category formation and processing (Sharma & Dorman, 1999; Blumstein et al., 2005, Chang et al., 2010). Similarly, the experimental research on spoken word recognition (e.g., using tasks such as lexical decision, gating, priming, or shadowing) has laid the basis for prominent perception models, including the cohort model (Gaskell & Marslen-Wilson, 2002), the lexical access from spectra approach (Klatt, 1989), the TRACE model (McClelland & Elman, 1986), and others.

The literature has been ably reviewed and examined from different perspectives (Hawkins 1999; Cleary & Pisoni, 2001; Pardo & Remez, 2006), including from a slightly more linguistically motivated vantage point (Poeppel & Monahan, 2008; Poeppel et al., 2008). In addition, the related body of engineering research on automatic speech recognition has added important insights; this work, too, has been extensively reviewed (Rabiner & Juang, 1993). A recent book-length treatment of speech perception bridging acoustics, phonetics, neuroscience, and engineering is provided in Greenberg and Ainsworth (2006).

The goal of this chapter is to focus explicitly on the processing of naturalistic, connected speech, that is, *sentence level speech analysis*. The motivation for focusing on connected speech is threefold. First, there is a renewed interest in focusing on ecologically relevant, naturalistic stimulation. The majority of laboratory research places participants in artificial listening situations with peculiar task demands (e.g., categorical perception, lexical decision, etc.), typically unrelated to what the listener does in real life. That the execution of such task demands has a modulatory influence on the outcome of neurobiological experiments and leads to serious interpretive problems has been discussed at length (e.g., Hickok & Poeppel, 2000, 2004, 2007). Second, investigating speech perception using sentence level stimuli has a prominent history worth linking to; however, only in the last decade is it playing an

increasing role in cognitive neuroscience and neurophysiology (e.g., Scott et al., 2000; Luo & Poeppel, 2007; Friederici et al., 2010). Early and formative contributions to understanding speech research were made by focusing on signal-to-noise ratio and intelligibility of sentences. An influential monograph by Miller (1951) summarized some of this early work, which is also deeply influenced by engineering demands (for a recent discussion, see Allen, 2005). This early work highlighted the relevance of temporal parameters for speech. Third, some of the most provocative new insights into speech processing come from data on listeners exposed to sentence level input. As mentioned, the focus on single speech sounds placed a large emphasis on the relevance of detailed spectral cues (e.g., formant patterns) and short-term temporal cues (e.g., formant transitions) on recognition performance. In contrast, the recent work on sentence-level stimuli (i.e., materials with a duration exceeding 1–2 s), and using experimental task demands such as intelligibility, demonstrate the fundamental importance of long-term temporal parameters of the acoustic signal. A growing literature in human auditory neuroscience has identified attributes of the system that underlie processing of communicative signals at this level. Important new principles have been discovered.

The chapter proceeds as follows. First, some of the essential features of speech are outlined. Next, the properties of auditory cortex that reflect its sensitivity to these features are reviewed (Section 9.2) and current ideas about the processing of connected speech are discussed (Section 9.3). The chapter closes with a summary of speech processing models at a larger scale that attempt to capture many of these phenomena in an integrated manner (Section 9.4).

9.2 Processing Speech as an Acoustic Signal

9.2.1 Some Critical Cues

Naturalistic, connected speech is an aperiodic but quasi-rhythmic acoustic signal with complex spectrotemporal modulations, that is, complex variations of the frequency pattern over time. Figure 9.1 illustrates two useful ways to visualize the signal: as a waveform (A) and as a spectrogram (B). The waveform represents energy variation over time—the input that the ear actually receives. The outlined "envelope" (thick line) reflects that there is a temporal regularity in the signal at relatively low modulation frequencies. These modulations of signal energy (in reality, spread out across a filterbank) are below 20 Hz and peak roughly at a rate of 46 Hz (Steeneken & Houtgast, 1980; Elliott & Theunissen, 2009). From the perspective of what auditory cortex receives as input, namely the modulations at the output of each frequency channel of the filterbank that constitutes the auditory periphery (cf. Hall and Barker, Chapter 7), these energy fluctuations can be characterized by the modulation spectrum (Kanedera et al., 1999; Greenberg & Kingsbury, 1997). Importantly, these slow-energy modulations correspond roughly to the syllabic structure (or syllabic "chunking") of speech. The syllabic structure as reflected by the envelope, in turn,

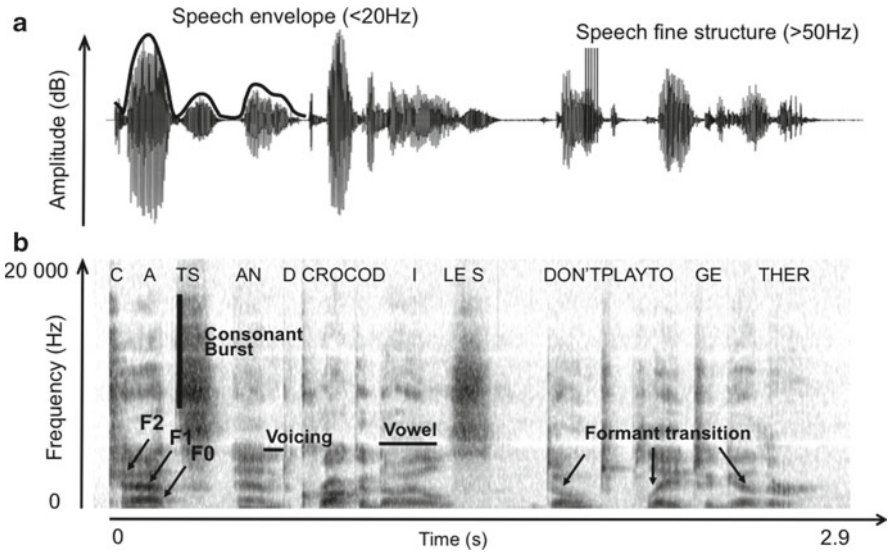

Fig. 9.1 Waveform (**a**) and spectrogram (**b**) of the same sentence uttered by a male speaker. Some of the key acoustic cues in speech comprehension are highlighted in black

is perceptually critical because it signals the speaking rate, it carries stress and tonal contrasts, and cross-linguistically the syllable can be viewed as the carrier of the linguistic (question, statement, etc.) or affective (happy, sad, etc.) prosody of an utterance (Rosen, 1992). As a consequence, a high sensitivity to envelope structure and envelope dynamics is critical for successful speech perception.

The second analytic representation, the spectrogram, decomposes the acoustic signal in the frequency, time, and amplitude domains (Fig. 9.1B). Textbook summaries often suggest that the human auditory system captures frequency information between 20 Hz and 20 kHz (and such a spectrogram is plotted here), but most of the information that is extracted for effective recognition lies below 8 kHz. It is worth remembering that speech transmitted over telephone landlines contains a much narrower bandwidth (200–3600 Hz) and is comfortably understood by normal listeners. A number of critical acoustic features can be identified in the spectrogram. The faintly visible vertical stripes represent the glottal pulse, which reflects the speaker's fundamental frequency, F0. This can range from approximately 100 Hz (male adult) to 300 Hz (child). The horizontal bands of energy show where in frequency space a particular speech sound is carried. The spectral structure thus reflects the articulator configuration. These bands of energy include the formants (F1, F2, etc.), definitional of vowel identity; high-frequency bursts associated, for example, with frication in certain consonants (e.g., /s/, /f/); and formant transitions that signal the change from a consonant to a vowel or vice versa.

The fundamental frequency (F0) conveys important cues about the speaker, for example, gender and size, and its modulation signals the prosodic contour of an

utterance (including, sometimes, lexical boundaries) and intonation (stress); F0 can also convey phonetic information (in tonal languages). The formants, mainly F1, F2, and F3, define the identity of vowels. The ratio between F1 and F2 is relatively characteristic of each vowel. Cues for vowel discrimination are thus mainly of spectral nature, if we assume that the auditory system computes F1/F2 ratios. It has also been suggested that the ratio of F3/F2 and F3/F1 can be computed online; this measure has high utility for speaker normalization (Monahan & Idsardi, 2010). It goes without saying that to compute such ratios, the auditory system must first extract the frequency structure of the sound.

Consonants are often associated with more transient acoustic properties, and with a broader spectral content. The energy bursts underlying consonants can range from partial obstructions of air flow (e.g., in fricatives such as /f/) to the release of energy after full occlusion (e.g., in stop consonants /p/, /t/, or /k/). Consonants can be discriminated either by the spectral content of their initial burst, that is, by the fast formant transitions that bridge consonant and vowels, or by the presence of voicing (Rosen, 1992), which corresponds to vocal chord vibrations occurring before and during the consonant burst. This means consonants are (or can be) discriminated on the basis of a mixture of spectral and temporal cues. All of these cues are present within the acoustic fine structure, that is, signal modulations at faster rates, say above 50 Hz. The capacity of the auditory brain to capture the speech fine structure is therefore important to recovering important details of the signal.

There exist excellent summaries of acoustic phonetics. Some emphasize the aspect of the productive apparatus (Stevens, 1998); others highlight a cross-linguistic perspective (Laver, 1994). There is a large body of data on the acoustic correlates of different attributes of speech, covered in dedicated textbooks (e.g., Pickett, 1999). Based on this brief and selective summary, two concepts merit emphasis: first, the extended speech signal contains critical information that is modulated at rates of less than 20 Hz, with the modulation peaking around 5 Hz. This low-frequency information correlates closely with the syllabic structure of connected speech. Second, the speech signal contains critical information at modulation rates higher than, say, 50 Hz. This rapidly changing information is associated with fine spectral changes that signal speech sound identity and other relevant speech attributes. Thus, there exist *two surprisingly different timescales concurrently at play in the speech signal*. This important issue is taken up in the text that follows.

Notwithstanding the importance of the spectral fine structure, there is a big caveat: speech can be understood, in the sense of being intelligible in psychophysical experiments, when the spectral content is replaced by noise and only the envelope is preserved. Importantly, this manipulation is done in separate bands across the spectrum, for example, as few as four separate bands (e.g., Shannon et al., 1995). Speech that contains only envelope but no fine structure information is called vocoded speech (Faulkner et al., 2000). Compelling demonstrations that exemplify this type of signal decomposition (Shannon et al., 1995; Ahissar et al., 1995; Smith et al., 2001) illustrate that the speech signal can undergo radical alterations and distortions and yet remain intelligible.

Such findings have led to the idea that the temporal envelope, that is, temporal modulations of speech at relatively slow rates, is sufficient to yield speech comprehension (Scott et al., 2006; Loebach & Wickesberg, 2008; Souza & Rosen, 2009). When using stimuli in which the fine structure is compromised or not available at all, envelope modulations below 16 Hz appear to suffice for adequate intelligibility. The remarkable comprehension level reached by most patients with cochlear implants, in whom about 15–20 electrodes replace 3000 hair cells, remains the best empirical demonstration that the spectral content of speech can be degraded with tolerable alteration of speech perception (Roberts et al., 2011). A related demonstration showing the resilience of speech comprehension in the face of radical signal impoverishment is provided by sine-wave speech (Remez et al., 1981). In these stimuli both envelope and spectral content are degraded but enough information is preserved to permit intelligibility. Typically sine-wave speech preserves the modulations of the three first formants, which are themselves replaced by sine-waves centered on F0, F1, and F2. In sum, *dramatically impoverished stimuli remain intelligible insofar as enough information in the spectrum is available to convey temporal modulations at appropriate rates*.

9.2.2 Sensitivity of Auditory Cortex to Speech Features

9.2.2.1 Sensitivity to Frequency

This section reviews the equipment of auditory cortex to process spectral and temporal cues relevant to speech. Primary auditory cortex (A1) is organized as a series of adjacent territories (cf. Clarke and Morosan, Chapter 2), which retain cochlear tonotopy, much like visual cortex is organized as series of retinotopic regions (cf. Hall and Barker, Chapter 7). This means that the spectral content of speech signals that is physically decomposed by the basilar membrane in the cochlea and encoded in primary auditory neurons (cochlear filters) is still place-coded at the level of core auditory cortex, and possibly in some adjacent territories. A place code can be important to discriminate speech sounds that differ with respect to their spectral content. Tonotopic maps are organized in auditory cortex as multiple "mirrors," resulting in an alternation of regions coding high and low frequencies (Formisano et al., 2003; Petkov et al., 2006). One of these functionally early auditory territories seems to be specifically involved in the processing of periodicity pitch (Patterson et al., 2002; Bendor & Wang 2006; Nelken et al., 2008), which corresponds to a sensation of tonal height conveyed by the temporal regularity of a sound, rather than by its audiofrequency content (see Griffiths et al., 2010). This region is located in the most lateral part of Heschl's gyrus overlapping with a region that is sensitive to very low frequency sounds, that is, the frequencies that correspond to pitch percepts, usually referred to as "the pitch domain" (for some discussion, see Hall and Barker, Chapter 7). Experiments using magnetoencephalography (MEG) have implicated the same area when pitch is constructed binaurally (Chait et al., 2006),

extending the role of such an area to pitch analysis more broadly. The reason for the clustering of periodicity pitch and other pitch responses within this region is not well understood. A possible and parsimonious explanation could be that auditory neurons (not only cortical) with very low characteristic frequencies (CFs) respond equally well to an input from a cochlear filter with very low CF, and to the modulation at CF rate of other cochlear filters. In cortex, such an overlap can be envisaged as a transition between place and temporal coding principles (cf. Cariani and Micheyl, Chapter 13). Accordingly, the pitch domain corresponds to the lowest edge of the range of frequencies that can be decomposed by the basilar membrane's physical properties. With respect to speech processing, the pitch center should play an essential role in coding speaker identity and prosody/intonation contour. In line with this, functional magnetic resonance imaging (fMRI) studies in humans show, on the one hand, that the pitch center is more developed in right than left auditory cortex (Zatorre & Gandour, 2008), and on the other hand that identity of both vowels and speakers is better represented in right temporal cortex (Formisano et al., 2008), even though strong interactions across cortical hemisphere are necessary to complete complex speaker recognition tasks (von Kriegstein et al., 2010).

9.2.2.2 Sensitivity to Time

Most neurons in primary auditory cortex are sensitive to temporal properties of acoustic stimuli. Their discharge pattern easily phase-locks to pulsed stimuli of up to about 40–60 Hz (Bendor & Wang, 2007; Middlebrooks, 2008; Brugge et al., 2009). Yet, this ability is limited compared to subcortical neurons that can phase-lock to much higher rates. The ability to represent the temporal modulation of sounds by an "isomorphic" response pattern that precisely mimics the stimulus temporal structure with the discharges (Bendor & Wang, 2007) decreases from the periphery to auditory cortex. Whereas thalamocortical fibers can phase-lock up to around 100 Hz, neurons in the inferior colliculus, superior olive, and cochlear nucleus are able to follow even faster acoustic rates (Giraud et al., 2000; Joris et al., 2004). Thus, there is a dramatic temporal down-sampling from subcortical to cortical regions — and what follows from this architectural feature of cellular physiology is the need for different neural coding strategies. For acoustic modulations faster than 30–40 Hz, auditory cortical neurons respond only at the onset of stimulus, with remarkable precision (Abeles, 1982; Heil, 1997a, b; Phillips et al., 2002). In awake marmosets, Wang and colleagues identified two main categories of auditory cortical neurons. Whereas "synchronized" (phase-locking) neurons use a faithful temporal code (isomorphic) to represent stimulus temporal modulation, "unsynchronized" neurons use a rate code. In each of these categories, Bendor and Wang (2006) describe neurons that respond either by increasing (positive monotonic) or decreasing (negative monotonic) their discharge rate with stimulus modulation. Synchronized neurons that are able to phase-lock to the stimulus are essentially found in primary auditory cortex (A1). When moving away from A1, the proportion of unsynchronized "onset" neurons increases. Their response in several dimensions, that is, the

amount of spikes per time unit, the delay between stimulus and response onset, the duration of the spike train (Bendor & Wang, 2006), and the precise spike-timing (Kayser et al., 2010), may be used to form abstract temporal information and to perform more elaborate and integrated computations, such as speech segmentation, grouping, etc. (Wang, 2007).

While phase-locking seems to saturate around 40 Hz in A1, Elhilali et al. (2004) observed that primary auditory neurons can follow stimulus modulations at faster rates (up to 200 Hz) when fast modulations ride on top of a slow modulations. With respect to speech, this ability means that when carried by the speech envelope, aspects of the fine structure can be "isomorphically" encoded by auditory cortical neurons. This suggests that there may be two different mechanisms for encoding slow and fast temporal modulations (Ding & Simon, 2009). Slow-amplitude modulations gate fast phase-locking properties, because slow modulations permit a periodic reset of synaptic activity and a regeneration of the pool of neurotransmitters (synaptic depression hypothesis). Although periodic synaptic regeneration is plausible, one could question why individual auditory neurons would have fundamentally different properties and biophysical limitations than subcortical auditory neurons. It is conceivable that the specificity of auditory cortical neurons lies in the fact that they are more massively embedded in large corticocortical networks, which requires that they not only faithfully follow and code input but also temporally structure output transmission. In sum, *the role of the auditory cortex is not only to efficiently represent the auditory input efficiently*, *but also*, *and perhaps primarily*, *to convert input structure into a code that will possibly be matched with other types of representations*. As exposed in the text that follows, ensemble *neuronal oscillations may help by temporally structuring neuronal output and facilitating the "packaging" and transformations to more abstract neural codes* and representations, and pooling together neuronal ensembles according to endogenous principles.

9.2.2.3 Sensitivity to Spectrotemporal Modulations

Speech signals are characterized by modulations in both spectral and temporal domains. Two separate possible codes to represent complex stimuli such as speech have been implicated in the preceding text, a place code for spectral modulations and a temporal code for temporal modulations. Whether spectral and temporal modulations are encoded by a single or by distinct mechanisms remains an open question. The idea of a single code for spectrotemporal modulations is supported by the presence of neurons that respond to frequency modulations but not amplitude modulations (Gaese & Ostwald, 1995) and by complex responses to spectrotemporal modulations (Schönwiesner & Zatorre, 2009; Pienkowski & Eggermont, 2010). Luo et al. (2006, 2007b) and Ding and Simon (2009) tested, based on MEG recordings in human listeners, whether FM and AM used the same coding principles. Figure 9.2 schematizes the stimulus configuration and the hypothesized neural coding strategies (see legend). The authors argue that if coding equivalence (or similarity) is the case, cortical

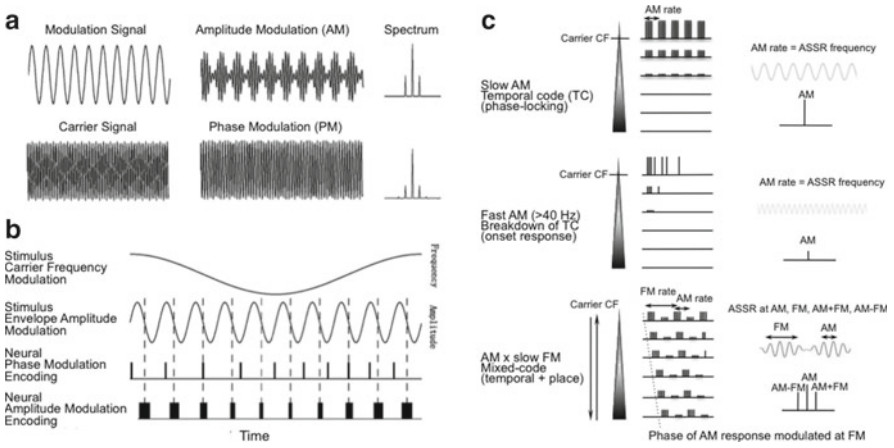

Fig. 9.2 Principles of amplitude and frequency modulations encoding in auditory cortex (Luo et al., Journal of Neurophysiology [2006], used with permission of APS). (**a**) In radio engineering modulation is used to encode acoustic stimuli, which can be either amplitude (AM; upper row) or phase modulated (PM; second row). (**b**) Proposals for neural AM and PM encoding. A stimulus is made of a frequency varying signal (upper row) and an amplitude modulation (second row). Using a PM encoding (third row), a neuron fires one spike per stimulus envelope cycle (dotted line) and the firing precise timing (phase) depends on the carrier frequency. Alternatively, using AM encoding (last row), a neuron changes its firing rate according to the instantaneous frequency of the carrier, while keeping constant the firing phase. (**c**) AM coding is illustrated in more detail in three different conditions, slow AM (upper row), fast AM (second row), and when AM and FM covary (last row). CF, characteristic frequency

responses as assessed by MEG should be the same when the carrier of slow AM is rapidly frequency modulated, or when a slowly changing carrier sound is amplitude modulated at fast rate (AM–FM comodulation experiments). Yet, they observed that only the phase of fast AM auditory responses (auditory steady state responses at 40 Hz) is modulated by slow FM, while both the phase and the amplitude of fast FM auditory responses (auditory steady state responses at 40 Hz) are modulated by slow AM. That AM and FM interact nonlinearly is beyond doubt. However, the mere fact that the spectral place-coding present in several auditory territories plays a more important role in FM processing than in AM processing could account for the asymmetry in the results. Whereas FM, by hypothesis, is encoded by a combination of place and temporal coding, AM is mostly encoded by temporal coding. Figure 9.2 depicts a model to characterize how AM and FM, critical features of speech signals, may plausibly be encoded, based on processing units that have a tonotopic axis and incorporate distinct thresholds for temporal stimulus modulations.

The asymmetric response pattern to fast and slow AM/FM might also depend on coding differences for fast and slow modulations. Whereas very slow frequency modulations are perceived as pitch variations, fast modulations are perceived as varying loudness. On the other hand, slow-amplitude modulations are perceived as variations of loudness, whereas fast modulations are perceived as roughness, or

flutter, or pitch. These sharp perceptual transitions could be underpinned by both the size and the place of the population recruited by each of these stimulus types. Whereas slow FM presumably allows for both a temporal and spatial segregation of cortical responses, entailing distinct percepts varying in pitch, fast FM presumably phase-locks together at FM the entire population stimulated by the varying carrier. A slight jitter in phase-locking could then account for the roughness of the sensation. In a similar way, fast AM is possibly no longer perceived as variations of loudness when the ability of neurons to phase-lock is overridden (beyond 40 Hz). Flutter (and then pitch sensations) for AM higher than 40 Hz superimposed on the primary spectral content of the modulated sound might reflect the additional excitation of neurons with very low CF (pitch neurons).

The spectral place-code, the transition from phase-locking to rate-coding for higher stimulus rates, and ensemble neuronal behavior, that is, the size of the population targeted by a stimulus, provide enough representational complexity to account for nonlinear neuronal responses to spectrotemporal acoustic modulations without invoking a specific AM/FM code.

9.2.2.4 Sparse Representations in the Auditory Cortex

The described response properties in auditory cortex need to be interpreted with caution. Electrophysiological recordings necessarily rely on a selection of neurons, a selection that is often biased toward units that *fire* in response to auditory stimuli (Atencio et al., 2009). Many neurons, however, are silent. The picture that arises from electrophysiological studies is one of "dense" coding because we extrapolate population behavior from a few recorded neurons. Hromádka and Zador (2009) argue that no more than 5% of auditory neurons fire above 20 spikes/s at any instant. These authors suggest that, rather than "dense," auditory responses are "sparse" and highly selective, which permits more accurate representations and a better discrimination of auditory stimuli. Sparse coding implies the existence of population codes relying strongly on topographical organization and spatial patterns. The behavioral relevance of such *mesoscopic* cortical organization has been recently demonstrated using functional neuroimaging (Formisano et al., 2008; Eger et al., 2009). Rather than looking at the mean of the response to repeated stimulation, these methods analyze the variance across trials and show that what differs from one trial to the next is meaningfully represented in the spatial pattern of the response. This spatial pattern can be distributed across functional areas, for example, across tonotopic auditory areas (Formisano et al., 2008; Chang et al., 2010). The bottom line of these studies is that percepts are individually encoded in mesoscopic neural response patterns on a millimeter–centimeter scale. The notion of hierarchical processing across several tonotopically organized functional regions is somewhat deemphasized by this new perspective, which is more compatible with an analysis-by-synthesis or Bayesian view in which higher and lower processing stages conspire to generate a percept (elaborated in the last section).

9.3 Cortical Processing of Speech as a Continuous Stream

Experimental research on the neural basis of speech has tended to focus on processing individually presented speech sounds, such as vowels, syllables, or single words. This approach has led to good progress, and the findings underpin most current models of speech. That being said, a large part of naturalistic speech comes at the listener as a continuous stream, in phrases and sentences, and not well "prepackaged" in perceptual units of analysis. Indeed, this *segmentation* problem remains a major challenge to contemporary models of adult and child speech perception as well as automatic speech recognition. Interestingly, in psychophysical research on speech, especially through the 1950s, a large body of work studied speech perception and intelligibility using phrasal or sentential stimuli (see, e.g., Miller [1951] for a summary of many experiments and Allen [2005] for a review of the influential work of Fletcher and others). There exist fascinating findings based on that work, for example, on the role of signal-to-noise ratio, but the feature that arises is that speech as a continuous signal has principled and useful temporal properties that merit attention and that may play a key role in the problem of speech parsing and decoding.

9.3.1 The Discretization Problem

In natural connected speech, speech cues are embedded in a continuous acoustic flow. Their on-line analysis and representation by spatial, temporal and rate neural encodings (see Cariani and Micheyl, Chapter 13) needs to be read out (decoded) by mechanisms that are unlikely to be continuous. The first step of these neural parsing and read-out mechanisms should be the discretization of the continuous input signal and its initial neural encoding. The generalization that perception is "discrete" has been motivated and discussed in numerous contexts (e.g., Pöppel, 1988; Van Rullen & Koch, 2003). There is an important distinction between *temporal integration* versus *discretization*, which for expository purposes is glossed over in this chapter.

One particular hypothesis about a potential mechanism for chunking speech and other sounds is discussed here, namely that cortical oscillations could be efficient instruments of auditory cortex output discretization, or discrete sampling. Neural oscillations reflect synchronous activity of neuronal assemblies that are either intrinsically coupled or coupled by a common input. They are typically measured in animal electrophysiology by local field potential recordings (review: Wang, 2010). The requirements for measuring oscillations and spiking activity are different. When spiking is looked for, the experimenter typically tracks a response to a stimulus characterized by a fast and abrupt increase in firing rate. Oscillations, on the other hand, can be observed in the absence of stimulation, and are modulated by stimulation in a less conspicuously causal way. The selection bias is therefore much stronger when measuring spikes than oscillations, because spiking reflects activity of either a single neuron or a small cluster of neurons selective to certain types of

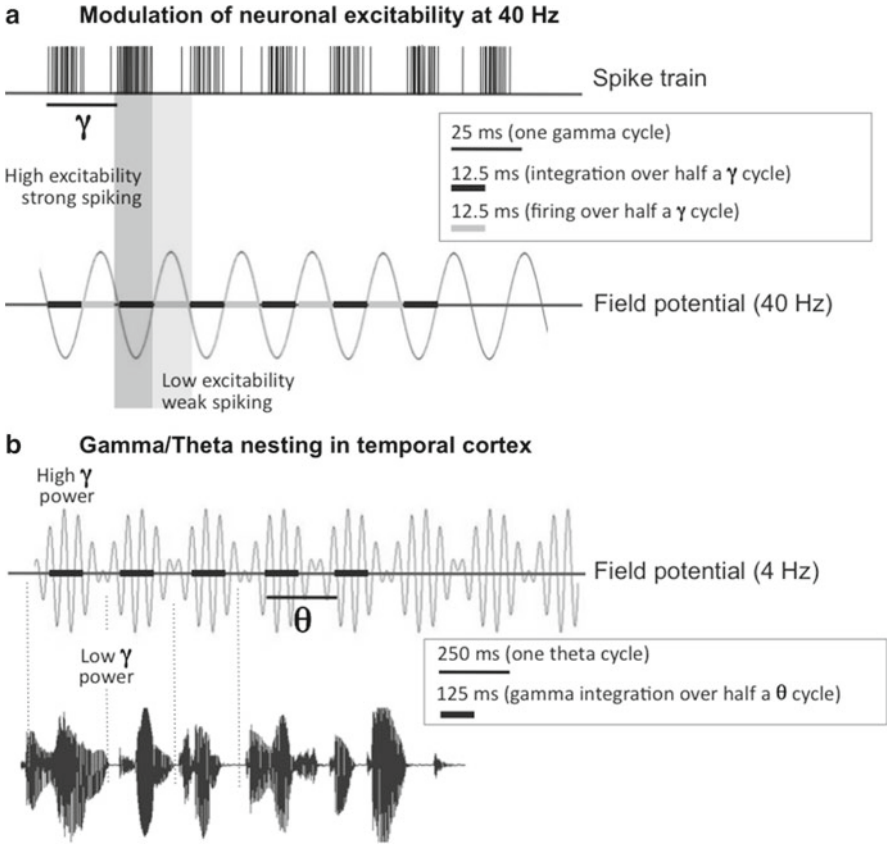

Fig. 9.3 The temporal relationship between speech and brain oscillations. (**a**) Gamma oscillations periodically modulate neuronal excitability and spiking. The hypothesized mechanism is that neurons fire for about 12.5 ms and integrate for the rest of the 25-ms time window. Note that these values are approximate, as we consider the relevant gamma range for speech to lie between 28 and 40 Hz. (**b**) Gamma power is modulated by the phase of theta rhythm (about 4 Hz). Theta rhythm is reset by speech resulting in keeping the alignment between brain rhythms and speech bursts

stimuli and ready to fire at the right moment. Cortical oscillations are proposed to shape spike-timing dynamics and to impose phases of high and low neuronal excitability (Britvina & Eggermont, 2007; Schroeder and Lakatos, 2009a, b; Kayser et al., 2010). The assumption that it is oscillations that cause spiking to be temporally clustered derives from the observation that spiking tends to occur in the troughs of oscillatory activity (Womelsdorf et al., 2007). The principle is illustrated in Figure 9.3A. It is also assumed that spiking and oscillations do not reflect the same aspect of information processing. Whereas spiking reflects axonal activity, oscillations are said to reflect mostly dendritic synaptic activity (Wang, 2010).

Neuronal oscillations are ubiquitous in cerebral cortex and other brain regions, for example, hippocampus, but they vary in strength and frequency depending on their location and the exact nature of their neuronal generators (Mantini et al., 2007).

In human auditory cortex, at rest, approximately 40 Hz activity (low gamma band) is strong and can be measured using stereotactic electroencephalography (EEG) in epileptic patients, MEG, or concurrent EEG and fMRI (Morillon et al., 2010). Neural oscillations in this range are endogenous in the sense that one can observe a spontaneous grouping of spikes at approximately 40 Hz even in the absence of acoustic stimulation. This gamma activity is thought to be generated by a ping-pong interaction between pyramidal cells and inhibitory interneurons (Borgers et al., 2005; Borgers & Kopell, 2008), or even just among interneurons that are located in superficial cortical layers (Tiesinga & Sejnowski, 2009). In the presence of a stimulus, this patterning at gamma frequencies becomes more pronounced, and clustered spiking activity is propagated to higher hierarchical processing stages (Arnal et al., 2011). Input to auditory cortex is conveyed by thalamocortical fibers contacting cells in layer IV. Unlike visual cortex, auditory cortical layer IV does not contain spiny stellate cells, which are the primary target of thalamocortical input, but rather pyramidal cells (Binzegger et al., 2007; da Costa & Martin, 2010). Whereas spiny stellates are small neurons with a modest dendritic tree, forming a horizontal coat of interdigitated ramifications, pyramidal cells are essentially vertical elements, reaching far below and above the layer where their cell bodies are found. Although it is unclear why cortical canonical microcircuits might be differently organized in the auditory and visual cortices (see Atencio et al., 2009), it is possible that this more vertical architecture emphasizes sequential/hierarchical processing over spatial integrative processing, meeting more closely critical requirements of speech processing, where analysis of the temporal structure is as important as spectral analysis.

By analogy with the proposal of Elhilali et al.(2004) that fast responses are gated by slower ones, it is interesting to envisage this periodic modulation of spiking by ensemble oscillatory activity as an endogenous mechanism to ensure sustained excitability of the system. This endogenous periodicity, however, could also reflect the alternation of dendritic integration and axonal transmission, which needs to be slowed down in the cortex due to the large amount of data to integrate, and the relatively long time lags between inputs signaling a common single event, possibly even through different sensory channels. In ecological situations, speech perception relies on the integration of visual and auditory inputs that are naturally shifted by about 100 ms (see van Wassenhove and Schroeder, Chapter 11). Integration of audiovisual speech requires data accumulation over a larger time window than the one allowed for by gamma oscillations. Such integration could occur under the patterning of oscillations in the theta range. In the next section, a potential role of theta activity in speech processing is thus outlined.

9.3.2 Speech Analysis at Multiple Timescales

Based on linguistic, psychophysical, and physiological data as well as conceptual considerations, it has been proposed that speech is analyzed in parallel at multiple

timescales (Poeppel, 2001, 2003; Boemio et al., 2005; Poeppel et al., 2008). The central idea is that both local-to-global and global-to-local types of analyses are carried out concurrently (multitime-resolution processing). The concept is related to reverse hierarchy theories of perception (Hochstein & Ahissar, 2002; Nahum et al., 2008). The principal motivations for such a hypothesis are twofold. First, a single, short temporal integration window that forms the basis for hierarchical processing, that is, increasingly larger temporal analysis units as one ascends the processing system, fails to account for the spectral and temporal sensitivity of the speech processing system and is hard to reconcile with behavioral performance. Second, the computational strategy of analyzing information on multiple scales is widely used in engineering and biological systems, and the neuronal infrastructure exists to support multiscale computation (Canolty & Knight, 2010). According to the view summarized here, speech is chunked into segments of roughly featural or phonemic length, and then integrated into larger units, as segments, diphones, syllables, words. In parallel, there is a fast global analysis that yields coarse inferences about speech (akin to Stevens' 2002 "landmarks" hypothesis), and that subsequently refines segmental analysis. Segmental and suprasegmental analyses could be carried out concurrently and "packaged" for parsing and decoding due to neuronal oscillations at different rates. Considering a mean phoneme length of about 25–80 ms and a mean syllabic length of about 150–300 ms, dual-scale segmentation is assumed to involve two sampling mechanisms, one at about 40 Hz (or, more broadly, in the low gamma range) and one at about 4 Hz (or in the theta range). Electrophysiological evidences in favor of this hypothesis are discussed later.

Schroeder and Lakatos (2009a, b) argue that oscillations determine phases of high and low excitability on pyramidal cells. This means that with a period of approximately 25 ms, gamma oscillations provide a 10- to 15-ms window for integrating spectrotemporal information (low spiking rate) followed by a 10- to 15-ms window for propagating the output (high spiking rate) (see, for illustration Fig. 9.3A.). However, a 10- to15-ms window of integration might be too short to characterize an approximately 50 ms phoneme. This raises the question of how many gamma cycles are required to encode phonemes correctly. This question has so far only been addressed using computational modeling (Shamir et al., 2009). Using a pyramidal interneuron gamma (PING) model of gamma oscillations (Borgers et al., 2005) that modulate activity in a coding neuronal population, Shamir et al. (2009) show that the shape of a sawtooth input signal designed to have the typical duration and amplitude modulation of a diphone (~50 ms; typically a consonant–vowel or vowel–consonant transition) can correctly be represented by three gamma cycles, which act as a three-bit code. This code has the required capacity to distinguish different shapes of the stimulus and is therefore a plausible means to distinguish between phonemes. That 50-ms diphones could be correctly discriminated with three gamma cycles suggests that phonemes could be sampled with one/two gamma cycles. This issue is critical, as *the frequency of neural oscillations in the auditory cortex might constitute a strong biophysical determinant with respect to the size of the minimal acoustic unit that can be manipulated for linguistic purposes*.

The notion of speech analysis at multiple timescales is useful because it allows the move from strictly hierarchical models of speech perception (e.g., Giraud & Price, 2001) to more complex models in which simultaneous extraction of different acoustic cues permits simultaneous high-order processing of different information from the same input signal. That speech *should* be analyzed in parallel at different timescales derives, among other reasons, from the observation that articulatory–phonetic phenomena occur at different timescales. It was noted previously (Fig. 9.1) that the speech signal contains events of different durations: short energy bursts and formant transitions occur within a 20- to 80-ms timescale, whereas syllabically carried information occurs over 150–300 ms. The processing of both types of events could be accounted for either by a hierarchical model in which smaller acoustic units (segments) are concatenated into larger units (syllables) or by a parallel model in which both temporal units are extracted independently, and then combined. A degree of independence in the processing of long (slow modulation) and short (fast modulation) units is observed at the behavioral level. For instance, speech can be understood well when it is first segmented into units up to 60 ms and when these local units are temporally reversed (Saberi & Perrott, 1999; Greenberg & Arai, 2001). This observation rules out the idea that speech processing relies solely on hierarchical processing of short and then larger units, as the correct extraction of short units is not a prerequisite for comprehension. Overall, there appears to be a grouping of psychophysical phenomena such that some cluster at thresholds of approximately 50 ms and below and others cluster at approximately 200 ms and above (a similar clustering is observed for temporal properties in vision; Holcombe 2009). Importantly, nonspeech signals are subject to similar thresholds. For example, 15–20 ms is the minimal stimulus duration required for correctly identifying upward versus downward FM sweeps (Luo et al., 2007a). By comparison, 200-ms stimulus duration underlies loudness judgments. In sum, physiological events at related scales form the basis for processing at that level. Gamma oscillations, for example, could act as an integrator such that all events occurring within about 15 ms are grouped, whereas events occurring within the next 15 ms are suppressed. Although it may sound inefficient to suppress half of the acoustic structure, an oscillatory mechanism could reflect a tradeoff between accurate signal extraction/representation and its on-line transmission to levels higher in the hierarchy, as well as ensuring the sustained excitability of the system.

9.3.3 Alignment of Neuronal Excitability with Meaningful Speech Events

An important requirement of the computational model mentioned previously (Shamir et al., 2009) is that ongoing gamma oscillations are phase-reset, for example, by a population of onset excitatory neurons. Without this onset signal the performance of the model drops. Ongoing intrinsic oscillations appear to be effective as a

segmenting tool only if they *align* with the stimulus. Schroeder and colleagues suggest that gamma and theta rhythms work together, and that the phase of theta oscillations determines the power and possibly also the phase of gamma oscillations (see Fig. 9.3B; Schroeder & Lakatos, 2008). This relationship is referred to as "nesting." Electrophysiology suggests that theta oscillations can be phase-reset by several means, in particular through multimodal corticocortical pathways (Arnal et al., 2009), but most probably by the stimulus onset itself. The largest cortical auditory evoked response measured with EEG and MEG, about 100ms after stimulus onset, could correspond to the phase reset of theta activity (Arnal et al., 2011). This phase reset would align the speech signal and the cortical theta rhythm, the proposed instrument of speech segmentation into syllable/word units. As speech is strongly amplitude modulated at the theta rate, this would result in aligning neuronal excitability with those parts of the speech signals that are most informative in terms of energy and spectrotemporal content (Fig. 9.3B; Giraud and Poeppel, submitted). There remain critical computational issues, such as the means to get strong gamma activity at the moment of theta reset. Recent psychophysical research emphasizes the importance of aligning the acoustic speech signal with the brain's oscillatory/quasi-rhythmic activity. Ghitza and Greenberg (2009) demonstrated that comprehension can be restored by inserting periods of silence in a speech signal that was made unintelligible by time-compressing it by a factor of 3. The mere fact of adding silent periods to speech to restore an optimal temporal rate, which is equivalent to restoring "syllabicity," improves performance even though the speech segments that remained available are not more intelligible. Optimal performance is obtained when 80-ms silent periods alternate with 40-ms time-compressed speech. These time constants allowed the authors to propose a phenomenological model involving three nested rhythms in the theta (5 Hz), beta, or low gamma (20–40 Hz) and gamma (80 Hz) domains (for extended discussion, see Ghitza, 2011).

9.3.4 Multitime-Resolution Processing: Asymmetric Sampling in Time

Poeppel (2003) attempted to integrate and reconcile several of the strands of evidence: first, speech signals contain information on at least two critical timescales, correlating with segmental and syllabic information; second, many nonspeech auditory psychophysical phenomena fall in two groups, with integration constants of approximately 25–50 ms and 200–300 ms; third, both patient and imaging data reveal cortical asymmetries such that both sides participate in auditory analysis but are optimized for different types of processing in left versus right; and fourth, crucially for the present chapter, neuronal oscillations might relate in a principled way to temporal integration constants of different sizes. Poeppel (2003) proposed that there exist hemispherically asymmetric distributions of neuronal ensembles with preferred shorter versus longer integration constants; these cell groups "sample" the input with different sampling integration constants (Fig. 9.4A). Specifically, left

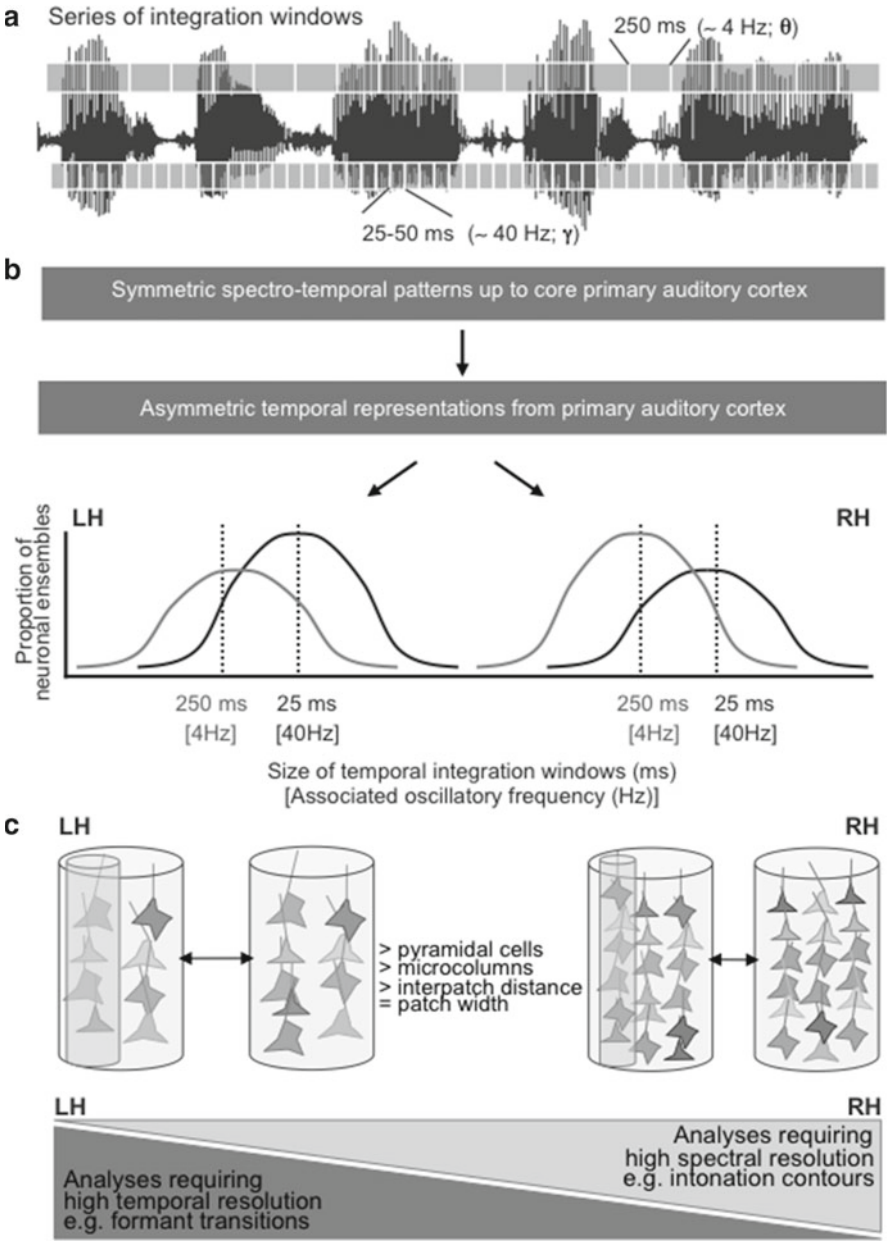

Fig. 9.4 (**a**) Temporal relationship between the speech waveform and the two proposed integration timescales (in ms) and associated brain rhythms (in Hz). (**b**) Proposed mechanisms for asymmetric speech parsing: left auditory cortex (LH) contains a larger proportion of neurons able to oscillate at gamma frequency than the right one (RH). (**c**) Differences in cytoarchitectonic organization between the right and left auditory cortices. Left auditory cortex contains larger pyramidal cells in superficial cortical layers and exhibits bigger microcolumns and a larger patch width and interpatch distance

auditory cortex has a relatively higher proportion of short term (gamma) integrating cell groups, whereas right auditory cortex has a larger proportion of long term (theta) integrating neurons (Fig. 9.4B). As a consequence, left hemisphere auditory cortex is better equipped for parsing speech at the segmental scale, and right auditory cortex for parsing speech at the syllabic timescale. This hypothesis, referred to as the asymmetric sampling in time (AST) theory, is illustrated in Figure 9.4 and accounts for a variety of psychophysical and functional neuroimaging results that show that left temporal cortex responds better to many aspects of rapidly modulated speech content while right temporal cortex responds better to slowly modulated signals including music, voices, and other sounds (Zatorre et al., 2002; Warrier et al., 2009). A difference in the size of the basic integration window between left and right auditory cortices would explain speech functional asymmetry by a better sensitivity of left auditory cortex to information carried in fast temporal modulations that convey, for example, phonetic cues. A specialization of right auditory cortex to slower modulations would grant it a better sensitivity to slower and stationary cues such as harmonicity and periodicity (Rosen, 1992) that are important to identify vowels, syllables, and thereby speaker identity. The AST theory is very close, in kind, to the spectrotemporal asymmetry hypothesis promoted by Zatorre (e.g., Zatorre et al., 2002; Zatorre & Gandour, 2008).

As mentioned above, the underlying physiological hypothesis is that left auditory cortex contains a higher proportion of neurons capable of producing gamma oscillations than right auditory cortex. Conversely, right auditory cortex contains more neurons producing theta oscillations. Consistent with this proposal, Hutsler and Galuske (2003) showed that the microcolumnar organization is different in the left and right auditory cortices (Fig. 9.4C). Left auditory cortex contains larger pyramidal cells in layer III and larger microcolumns. It could be the case that larger pyramidal cells produce oscillations at higher rates because the larger the cell the stronger the membrane conductance and the faster the depolarization/repolarization cycle. Pyramidal cell conductance may play a role in setting the rhythm at which excitatory/inhibitory circuits (PING) oscillate. This hypothesis, however, has to be verified using computational models.

To evaluate the plausibility of this model, four types of data are required. First, temporal integration over the short timescale (for both speech and nonspeech auditory signals) must be demonstrated. Second, evidence of temporal integration over the longer time scale is necessary. There exists a body of such evidence, some of which is reviewed by Poeppel (2003). Pitch judgments versus loudness judgments exemplify the two timescales, as do segmental versus syllabic processing timescales. Third, the information on these two timescales should interact, to yield perceptual objects that reflect the integrated properties of both modulation rates. This has not been widely tested, but there is compelling behavioral evidence in favor, discussed briefly later. Finally, there should be cerebral asymmetries in the cortical response properties, which are summarized.

Relevant psychophysical data testing interactions across timescales are sparse, but several studies have attempted to understand the relative contributions of different modulation rates. Elliott and Theunissen (2009) provide data showing that there

are interactions across bands with restricted temporal modulation frequencies, although they did not explicitly test the ranges of interest here. Chait et al. (submitted) show a striking interaction of two selected bands of speech signals in dichotic speech conditions: when both low (<8 Hz) and high (25–40 Hz) signals are presented concurrently, listeners' performance exceeds the predicted linear combination values, suggesting a clear interaction between the timescales of interest. Further, Saoud et al. (submitted) observed that speech comprehension is both faster and more accurate when the low-rate temporal envelope (0–4 Hz) of bisyllable words is presented through the left ear and the high temporal envelope (28–40 Hz) is presented to the right ear relative to the reverse dichotic situation. These results suggest (1) that the two timescales carry information that interacts synergistically to yield higher intelligibility representations of the input signal and (2) that comprehension is better when each auditory cortex receives speech information in a temporal format that matches its intrinsic oscillatory capacity. Recent fMRI evidence supports this conclusion (Saoud et al., 2012).

Despite a limited understanding of the psychophysics, a large number of imaging and neurophysiological studies have addressed the cerebral asymmetry predictions. For example, consistent with AST, Boemio et al. (2005), using temporally extended stimuli built from short segments of different durations, showed a striking rightwards asymmetry in superior temporal sulcus (STS) for these nonspeech stimuli when longer time segments were used (e.g., 300 ms), compared to the short-time-structure signals (e.g., 25 ms). Similarly, Overath et al. (2008) showed a significant rightward lateralization for auditory stimuli with increasing length of spectrotemporal time windows. Zaehle et al. (2004) tested speech and nonspeech signals and observed robust leftward lateralization for rapidly modulated auditory signals. Jamison et al. (2006) used nonspeech signals in an fMRI design and observed the predicted left/rapid–right/slow associations. The predictions have been tested for speech and nonspeech, and pitting spectral against temporal processing advantages (e.g., Obleser et al., 2008), including even in newborns (Telkemeyer et al., 2009). By and large, the predicted associations hold up well, and there is emerging consensus that temporal parameters of the sort discussed here play a central role in decoding auditory signals in the cortex.

Are the predicted asymmetric sampling properties truly architectural features of the system, or are the observed asymmetries driven into the system by properties of the stimuli employed? To verify that the sound analysis asymmetries are systemic properties, Giraud and colleagues (2007) measured the distribution of neuronal oscillations in subjects not exposed to input, that is, in a passive resting state. Using combined EEG/fMRI at rest, they discovered a stronger expression of gamma rhythm in left auditory cortex and a stronger expression in theta rhythm in right auditory cortex (Fig. 9.5A). Control analyses included analyses of other frequency bands in the alpha and low and high beta range. For these frequency bands there were no significant EEG/fMRI correlations in auditory cortex at rest and no detectable asymmetry.

The left hemisphere dominance of gamma activity at rest was confirmed using a detailed anatomical approach in another concurrent EEG/fMRI data set (Morillon

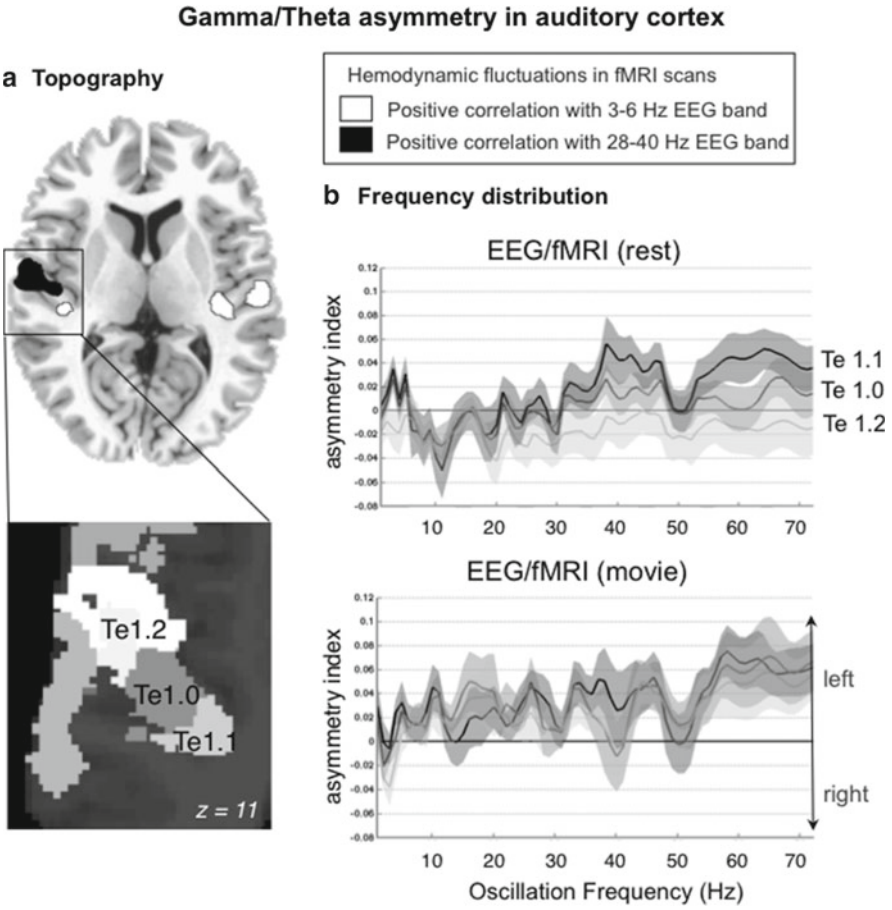

Fig. 9.5 (**a**) Experimental evidence for an asymmetry in cortical oscillations in the left and right auditory cortices at rest using combined EEG/fMRI (after Giraud et al., 2007). (**a**) Topographical distribution of EEG/fMRI coupling in the theta and low gamma bands. Note that both rhythms are expressed on both sides, but that a right/left dissociation can be seen at appropriate statistical threshold. (**b**) Correlations between EEG power and fMRI bold signal in three different cytoarchitectonic territories of Heschl's gyrus at rest and when subjects were watching a spoken movie (after Morillon et al., 2010). Asymmetry in the strength of EEG/fMRI correlation was maximal in Te 1.1 at rest (mostly within the gamma range) and increased in all three territories during audiovisual stimulation

et al., 2010). There, fMRI time series were extracted from various cytoarchitectonic territories along Heschl's gyrus and correlated with power variations of EEG over its entire spectrum (1–72 Hz). These data showed that the left dominance in spontaneous expression of gamma activity arises from the most posteromedial part of Heschl's gyrus (Te 1.1), and that it declines along its posteromedial to anterolateral axis (Fig. 9.5B). Because EEG/fMRI correlations are rather weak, these data were compared to MEG data at rest, from sensors that were pretested to be most responsive

to auditory input. The latter analyses confirmed the left-dominance of gamma rhythm at rest. However, unlike previous results, both the region of interest based on the EEG/fMRI approach and the MEG data did not give a consistent picture of spontaneous theta activity. The variance across experimental data underscores that more experiments are needed to validate, invalidate, or augment the AST proposal.

The EEG/fMRI experimental data show that oscillations in the delta band (1–3 Hz) become right-dominant during linguistic processing, while most other rhythms including beta activity become strongly left-dominant (Fig. 9.5B, lower panel). The delta/low theta rhythm has the temporal properties to underlie prosodic processing, as it corresponds to integration of speech signals in approximately 500-ms windows. This rate would be ideal to mediate prosodic operations such as extracting intonation contours indicative of speaker's emotional states, illocutionary intent, etc. It is thus possible that rather than theta, it is the delta rhythm that is predominantly right lateralized, while gamma and theta rhythms jointly underlie speech parsing in left auditory cortex. There is a lot of work in progress regarding this unresolved question.

9.4 Large-Scale Neurocognitive Models of Speech Processing

9.4.1 Emerging Consensus: Functional Neuroanatomic Models

Although the perceptual analysis of speech is rooted in the different anatomic subdivisions of auditory cortex in the temporal lobe, speech processing involves a large network that includes areas in parietal and frontal cortices, the relative activations of which strongly depend on the task performed. Several reviews have synthesized the state-of-the-art of functional neuroanatomy of speech perception (Scott & Johnsrude, 2003; Hickok & Poeppel, 2000, 2004, 2007; Rauschecker & Scott, 2009). We briefly summarize the main consensus findings (Fig. 9.6A) that are based on functional neuroimaging (fMRI, positron emission tomography [PET], MEG/EEG) and lesion data.

Departing from the classical model in which both a posterior (Wernicke's) and an anterior (Broca's) area form the anatomic network, it is now argued that speech is processed in parallel in at least two streams, a ventral stream for speech-to-meaning mapping (a "what" stream), and a dorsal stream for speech-to-articulation mapping (a "how" stream). Both streams converge on prefrontal cortex, with a tendency for the ventral pathway to contact ventral prefrontal cortex (BA 44/45, also referred to as Broca's area), and the dorsal pathway to contact dorsal premotor regions (Hickok & Poeppel, 2007; Rauschecker & Scott, 2009). The dual path network operates both in a feedforward (bottom-up) and feedback (top-down) manner—highlighting, in turn, the need for algorithmic theories that have appropriate primitives to permit such bidirectional processing in real time. An additional feature of the inclusion in current models of both ventral (temporal–frontal) and dorsal (temporal–parietal–frontal) streams has been a renewed appreciation for the subtlety

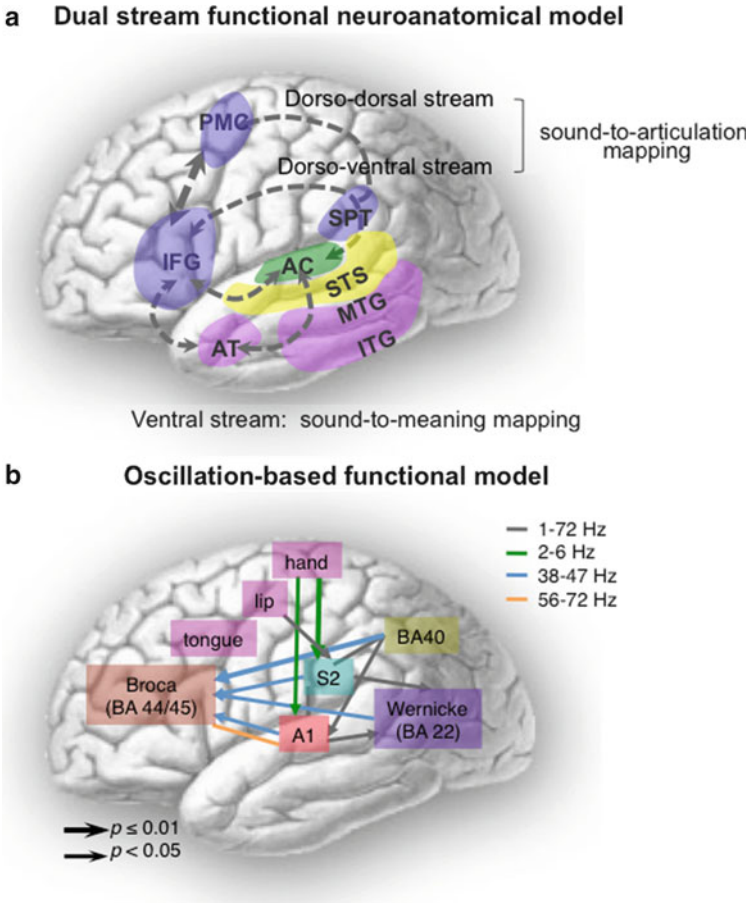

Fig. 9.6 Two functional neuroanatomical models of speech perception. (**a**) Model based on neuropsychology and functional neuroimaging data (PET and fMRI; after Hickok and Poeppel, 2007). (**b**) Model based on the propagation of resting oscillatory asymmetry during an audiovisual linguistic stimulation (a spoken movie). Modified from Giraud and Poeppel, 2012

of hemispheric specialization (Ueno et al., 2011). In particular, dorsal pathway structures (see Fig. 9.6A) appear much more strongly (left) lateralized, whereas the areas comprising the ventral processing stream(s), at least early on (e.g., superior temporal gyrus [STG], STS, medial temporal gyrus [MTG]), reveal robust bilateral contributions, whether assessed by hemodynamic or electrophysiological techniques. There is certainly no one-size-fits-all answer to hemispheric specialization for speech and language processing.

Historically, neuropsychological deficit-lesion research has been the main source of data regarding such anatomic models (Bates et al., 2003). In the context of dual stream proposals, the dorsal structures play a more central role in mediating output

related computations. Because output tasks (e.g., word repetition) are the most frequently used instruments in clinical work to assess poststroke performance, there is thus a natural tendency to overemphasize the degree of left hemisphere dominance for speech and language. While output operations are apparently strongly lateralized to the dominant left hemisphere, the operations underlying comprehension are much more bilateral (Giraud et al., 2004). Various aspects of comprehension, including the recognition of voice, of prosody, and of components of lexical semantics have been strongly implicated as right-hemisphere functions. In sum, statements about speech and language lateralization must be taken with caution, requiring reference to the specific subroutines under consideration. For a related electrophysiological perspective on language comprehension, see the "PARLO" model (Federmeier, 2007), in which top-down predictive processing and production are argued to be left lateralized and more bottom-up processes right lateralized.

The functional anatomy corresponds to stages of perceptual analysis that are required for recognition: analysis of the acoustic signal; transformation to a phonetic or phonological code in order to link to stored linguistic information; contact with the stored representations, e.g., words; contact with the conceptual information linked to lexical entries; and in addition, depending on the tasks, retrieval of the articulatory code underlying spoken output; and combination of items to yield phrases, that is, compositional operations.

In human auditory cortex, the acoustic analysis of speech is initiated bilaterally in Heschl's gyrus. Although there are presumably qualitative differences in the type of processing that is carried out on each side (as outlined previously), metabolic and hemodynamic responses reveal no compelling asymmetries in the acoustic processing of speech sounds at the level visible to these techniques. A new meta-analysis on sublexical speech perception confirms that bilateral regions are fully involved in initial analyses, with subsequent mapping to phonology more left lateralized (Turkeltaub & Coslett, 2010). Depending on the task, phonological processing involves regions that are either anteroventral or posterodorsal to Heschl's gyrus along the superior temporal gyrus (BA22; Davis et al., 2005; Davis & Johnsrude, 2007). Passive listening and intelligibility tasks tend to involve anteroventral regions where there might be relatively stable phonological representations, possibly organized in a topographic manner (e.g., syllable or vowel maps; Obleser et al., 2006; Chang et al., 2010). Activation may extend to more anterior and ventral regions of the left temporal lobe. Which subroutines are executed in the more anterior ventral territories is a subject of intense current investigation, and proposals range from the anterior temporal lobe (ATL) mediating conceptual storage (Patterson et al., 2007) to linguistic combinatorics (Brennan et al., 2010). Phonetic-to-lexical mapping typically activates the STS and the MTG. The posterior third of middle temporal gyrus appears to play a key intermediate role in both recognizing and activating words in their formal, linguistic guise (STS to MTG mapping), as has been reviewed in Hickok and Poeppel (2007) and Lau et al. (2008). Further, speech production tasks implicate MTG in lexical representations before articulation (Indefrey & Levelt, 2004). Finally, there are reasons to believe that the meaning of words is activated, preactivated, or selected in MTG (for recent review, see Lau et al., 2008)

The extent to which STS and MTG activation is bilateral in the context of processing word form and word meaning is unresolved. A growing body of data suggests that here, too, the bilateral contribution has been underestimated. For example, in an fMRI study, blood oxygenation level–dependent (BOLD) responses to vocoded speech before and after subjects had learned to understand its linguistic content were recorded and clearly bilateral activation of the MTG (BA21) was observed. Giraud et al. (2004) concluded that it was essentially the early, phonological, steps of analysis that were more lateralized, but not the semantic analysis.

In contrast to identification or "what"-type tasks mediated by ventral stream temporal lobe regions, the dorsal stream structures of auditory cortex (as well as parietal and frontal lobes) play a more critical role in sensorimotor aspects of speech processing. However, there are conflicting hypotheses about the dorsal stream's contributions, ranging from (1) processing spectral changes over time ("how" pathway) to (2) extracting relevant sound features and matching them with stored templates of motor responses ("do" pathway) to (3) transforming auditory representations of speech into motor programs for speech gestures. The data motivating the differing research questions derive mostly from imaging studies and neuropsychological patient data. Electrophysiological experiments have, to date, contributed less to the discussions of concurrent processing streams and the differential role of dorsal structures. Two brain regions have lately received special attention, Spt (Sylvian parietotemporal) and intraparietal sulcus (IPS). Imaging experiments in which subjects are required to generate overt or covert articulated outputs typically activate regions that are posterior to Heschl's gyrus, the posterior planum temporale, and the supramarginal gyrus located just above Heschl's gyrus in the parietal operculum (Kell et al., 2010). The two latter regions merge in area Spt, which is argued to carry out the sensorimotor transformations underlying speech and other vocal tract activities (Hickok & Poeppel, 2007). A different line of research, explicitly testing feedforward and feedback auditory processing in audiovocal integration in musicians, has implicated the IPS (and its connectivity to STS) in the computations linking perception and production (Zarate & Zatorre, 2008; Zarate et al., 2010). This aspect of dorsal pathway function is briefly revisited in Section 9.4.3.

9.4.2 Broadening the Empirical Scope: an Oscillation-Based Functional Model

The chapter has emphasized a neurophysiological perspective, and especially the potential role of neuronal oscillations as "administrative mechanisms" to parse and decode speech signals. Does such a focus converge with the functional anatomic models mentioned above? Recent experimental research has begun addressing this issue directly and developed a functional anatomic model solely derived from recordings of neuronal oscillations. Based on analyses of the sources of oscillatory activity, that is, brain regions showing asymmetric theta/gamma activity at rest and under linguistic stimulation, Morillon et al. (2010) propose a new functional model

of speech and language processing (Fig. 9.6B) that links elegantly to the textbook anatomy (illustrated in Fig. 9.6A). This model is grounded in a "core network" showing left oscillatory dominance at rest (no linguistic stimulation, no task), encompassing auditory, somatosensory, and motor cortices, and BA40 in inferior parietal cortex. The strongest asymmetries are observed in motor cortex and in BA40, which hence presumably play an important causal role in left hemispheric dominance during language processing. Critically, the proposed core network does not include Wernicke's (BA22) and Broca's (BA44/45) areas, despite the fact that both are classically related to speech and language processing. Interestingly, whereas these areas show no sign of asymmetry at rest, they "inherit" left dominant oscillatory activity during linguistic processing from the putative core regions. The model argues that posterior superior temporal cortex (Wernicke's area) inherits its profile from auditory and somatosensory cortices, while Broca's area inherits its profile from all posterior regions including auditory, somatosensory, Wernicke, and BA40. This model specifies that posterior regions share their oscillatory activity over the whole range of frequencies examined (1–72 Hz), while Broca's area inherits only the gamma range of the posterior oscillatory activity. This might reflect that oscillatory activity in Broca's area does not exclusively pertain to language. Finally, an important feature of the model is the influence of the motor lip and hand areas on auditory cortex oscillatory activity on the delta/theta scale, which underlines the importance of syllable and co-speech gesture production rates, on the receptive auditory sampling, and its asymmetric implementation. This model is compatible with a hardwired alignment of speech perception and production capacities at a syllable but not at a phonemic scale, suggesting that sensory/motor alignment at the phonemic scale is presumably acquired. Using an approach entirely driven by oscillations, this model is largely consistent with the previous one, but places a new emphasis on hardwired auditory–motor interactions, and on a determinant role of BA40 in language lateralization, which remains to be clarified.

9.4.3 The Role of the Auditory Cortex in Speech Production

Arguing that auditory cortex lies at the basis of speech perception is hardly surprising or insightful, yet it is worth remembering that the literature on speech recognition has been most deeply influenced by the *motor theory of speech perception* (Liberman et al., 1967; Corballis, 2009). This theory holds that listeners recover the intended articulatory gestures of the speaker, that is, properties of a motoric representation, and a substantial literature argues that motor cortical areas show activation in the relevant situations (Wilson et al., 2004; Pulvermueller et al., 2006). A different perspective can be characterized as the *sensory theory of speech production*. The idea is developed in some detail in Hickok et al. (2011). In the latter view, somatosensory (vocal tract configuration) and auditory (spectrotemporal) goals lie at the basis of speech production, from which claim it follows that *auditory cortex is centrally involved in production as well as perception*. In contrast, the causal role of motor

cortical structures for perception is thereby challenged. Both models depicted in Figure 9.6 underscore that brain systems for speech perception and speech production are intimately linked at both functional and anatomical levels.

What is the hypothesized role of auditory cortical regions in production tasks? Both imaging (fMRI) and electrophysiological (MEG) studies suggest that speech decoding structures in human auditory cortex are preactivated during speech planning (e.g. Kell et al., 2010; Tian & Poeppel, 2010), presumably through input from premotor cortex as well as parietal areas. These areas provide feedback to the motor system for the control of speech production. The discussion surrounding the contribution of sensory areas such as auditory cortex to production is largely embedded in the framework of internal forward models. Such models have been elaborated in detail by Guenther and colleagues for production (Guenther, 2006; Guenther et al., 2006) and receive support in electrophysiological studies demonstrating the predictive aspect of production via efference copies (Eliades & Wang, 2008; Tian & Poeppel, 2010), and align well with large-scale psycholinguistic models of perception and production (Hickok et al., 2011).

9.4.4 A Predictive (Bayesian) View on Speech Processing

When processing continuous speech, as outlined in Section 9.3, the brain needs to simultaneously carry out acoustic and linguistic operations: at every instant there is both acoustic input to be processed and meaning to be calculated from the preceding input. Discretization using phases during which cortical neurons are either highly or weakly receptive to input is one computational principle that could ensure constant alternation between sampling the input and matching this input onto higher-level, more abstract representations. The Bayesian perspective on this issue assumes that the brain decodes sounds by constantly generating inferences about what is and will be said, on the basis of the quickest and crudest neural representation it can make with an acoustic input (Poeppel et al., 2008). Discretization at multiple timescales and Bayesian speech decoding principles are gathered in the conceptual model proposed in Figure 9.7 (adapted from Poeppel et al., 2008). In this model, neural representations of speech sounds are activated via both (1) a bottom-up process and (2) a higher-order prior based on previous input, knowledge of language, etc. These assumptions may correspond to coarse "preactivation" of representations, which subsequently accelerate the match between representation and input. Such priors can theoretically be formed at every representational level, acoustic, phonological, lexical, etc. Figure 9.7 illustrates, in three horizontal levels, the mapping from an acoustic input on the left to an output lexical item (or string of words) on the right. The boxes at the bottom exemplify putative types of analyses that are required for successful recognition. Something like these proposed analyses must be correct on logical grounds—and this chapter argues that the multitime-resolution analysis plays one helpful role in the overall process. The three boxes in the middle level make reference to which cortical areas are implicated for some of the operations.

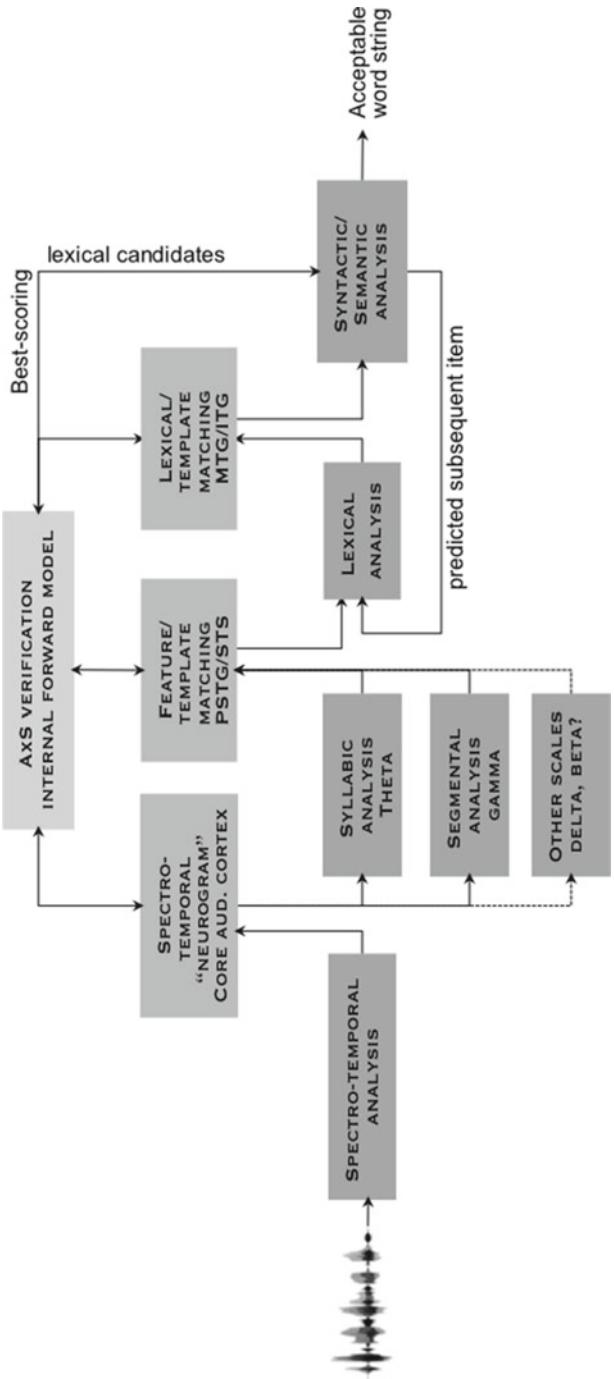

Fig. 9.7 Block-diagram of hypothesized operations taking place during speech perception within an analysis-by-synthesis framework (after Poeppel et al., 2008). The lower boxes represent different levels of sound to word mapping. The top box corresponds to a hypothetical level where an internal forward model in formed. The intermediate boxes show possible computations resulting from the interaction with bottom and top levels operations. The internal forward model is updated periodically with each new neuro-sample that is available (possibly every 30 ms and 200 ms). A detailed spectrotemporal analysis of the acoustic signal is available in bilateral primary auditory areas. Segmental (gamma) and syllabic-size (theta) samples are then formed close to auditory cortex, in the superior temporal gyrus (STG), and in the superior temporal sulcus (STS), respectively. The mapping between featural information and lexical entries occurs in the STS, and lexical processes in the medial temporal gyrus (MTG). Further compositional semantic and syntactic steps are carried out in prefrontal cortex. Top-down forward model signals reach many of the aforementioned steps/regions

Note that in this visualization, it is ventral stream areas that are principally implicated. The box on top identifies two of the putative types of "heuristics" or algorithms that are under consideration: the internal forward models mentioned previously, and analysis-by-synthesis (cf. Poeppel et al., 2008), an algorithm for perception suggested in the 1950s that takes small bits of input and generates, sequentially, the hypothesized output compatible with an input string, iteratively yielding better matches. On both of these concepts of processing, much of perception is actually achieved by a form of internal prediction and/or production, yet these models are rather different from motor theories. A proposal in very similar spirit to the one exemplified in Figure 9.7 is the "reverse hierarchy theory," a conceptualization developed to meet certain challenges in visual object recognition (Hochstein & Ahissar, 2002) and recently extended to speech processing (Nahum et al., 2008).

In their experiment, which effectively illustrates the tension between bottom-up and top-down components of speech decoding, Giraud et al. (2004) contrasted functional brain images in which identical vocoded stimuli could be either understood or not depending on previous experience. Before exposure to the corresponding natural speech stimuli, participants perceived vocoded speech as noise, whereas after exposure they perceived it as speech and could reconstruct the meaning from degraded sounds. At a behavioral level this exemplifies that perceiving linguistic content in speech is not merely the result of acoustic processing. At the functional neuroimaging level, very little neural activation corresponds to speech comprehension per se; the essential part of the process corresponds to auditory search, which reflects iterative matching between hypothesis and incoming input.

It is difficult to characterize the neurophysiological processes underlying top-down control on speech processing using the auditory modality alone, precisely because top-down and bottom-up influences concurrently operate on the same neuronal target. Van Wassenhove et al. (2005) designed a study using natural audiovisual speech where it is the visual modality that primes the auditory modality (see van Wassenhove and Schroeder, Chapter 11). Because when we speak the onset of visual movements leads the auditory onset by about 150 ms, the brain can infer/predict auditory input from visual movements. Using this ecological audiovisual setting, it is possible to record with EEG or MEG in humans both the response to the visual input (the predictor) and the impact of visual prediction on auditory response to speech. Van Wassenhove et al. (2005) showed that the early auditory response is accelerated by visual input, with the degree of temporal facilitation related to how informative the facial configuration was. For example, seeing a speaker with the mouth in a bilabial configuration (i.e., poised to say /ba/ or /pa/ or /ma/) leads to up to 25 ms of facilitation because such a small set of auditory targets is possible. Using an identical setting, that is, videos of a speaker pronouncing syllables, Arnal et al. (2009, 2011) have refined this approach, showing that facilitation also involves a reduction in the amplitude of the response. Critically, both latency and amplitude reductions are proportional to the informational value contained in the visual input. Syllables starting with a bilabial consonant, for example, /pa/ /ma/, are more informative, hence more predictive, than when the consonant is formed at

the back, for example, /ga/, /ka/. This shows that predictions made by the brain on the basis of rather crude sensory information strongly influence speech processing. Bayesian models of cortical responses stipulate that at each level of the hierarchy the neural response that is propagated forward reflects the difference between a prediction and the actual input (Friston, 2010). If correctly predicted, a stimulus therefore gives rise to a smaller cortical response than if unexpected. This phenomenon could be accounted for by the size to neuronal population that responds to a stimulus. When a speech stimulus is not predicted, the brain could respond with a large response reflecting the involvement of a broader neuronal population. This neural strategy, although ensuring that the brain does not miss a stimulus, is both cognitively costly and imprecise. As soon as a stimulus is either recognized or correctly anticipated the size of the recruited neuronal population drops, reflecting a more precise, focal activation in auditory cortex. Other accounts have recently been advanced for such phenomena (Wacongne et al., 2011).

9.5 Summary

This chapter engages, at the outset, some potential terminological confusion. "Speech perception" is many things to many people, and the failure to distinguish carefully between terms that have overlapping, obtuse, or no definitions has led to some unfortunate misunderstandings in the literature. Section 9.2 summarizes some of the salient properties of the speech signal that lie at the basis of what human auditory cortex must process. Particular emphasis is placed on some temporal attributes, including the low modulation frequencies in speech that play a special role for intelligibility. The section covers the sensitivity to frequency and the sensitivity to time of cortical neurons. In addition, the high degree of tuning to spectrotemporal modulation is discussed. In Section 9.3, the chapter turns to the processing of speech as a continuous signal. One of the central challenges is here called the discretization problem: how does auditory cortex create chunks of the appropriate temporal granularity for further computation? The solution that is pursued in this chapter builds on the concept of neuronal oscillations. In particular, oscillations in multiple frequencies (theta, gamma) are argued to provide the right mechanisms to align with the speech signal and sample the speech signal at different rates. Multitime-resolution processing and the asymmetric sampling in time (AST) hypothesis are summarized. Section 9.4 outlines large-scale models. First, the consensus functional anatomic models are discussed. The dual stream model is highlighted, and new neuro-oscillatory data are reviewed that extend and strengthen such a multiple pathway approach to speech perception. Further, it is highlighted that auditory cortex plays a critical role in speech production, reversing the standard roles in the literature that emphasize the role of motor cortex and speech perception. In terms of functional analysis, the notion of an internal forward model is presented, building on the observation that much of perceptual analysis has a strong predictive component. Finally, audiovisual speech experiments are shown to test some of the predictions of these models.

A comprehensive and explanatory neurocognitive model of speech perception remains an ambitious goal. It is worth remembering that speech perception is a task that is executed with automaticity and great ease by even early learners, but that is handled surprisingly poorly by even the most sophisticated automatic devices. The brain appears to solve this very challenging problem by breaking it down into parts: it is broken down in space, by implementing the functional anatomy as multiple concurrent streams, and it is broken down in time, by implementing multitime-resolution mechanisms that analyze information on multiple scales concurrently. Like all models, surely the ones presented here are dramatically underspecified and will turn out to be naïve. That being said, one hopes that they are wrong in an interesting way, leading to new research questions and incremental progress on this foundational question about human perception.

Acknowledgments The preparation of this manuscript was supported by CNRS to A. L. G. and NIH 2R01DC05660 to D. P.

References

Abeles, M. (1982). Role of the cortical neuron: Integrator or coincidence detector? *Israel Journal of Medical Sciences*, 18(1), 83–92.

Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., & Merzenich, M. M. (2001). Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proceedings of the National Academy of Sciences of the USA*, 98(23), 13367–13372.

Allen, J. B. (2005). Articulation and intelligibility. *Synthesis Lectures on Speech and Audio Processing*, 1(1), 1–124.

Arnal, L. H., Morillon, B., Kell, C. A., & Giraud, A. L. (2009). Dual neural routing of visual facilitation in speech processing. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 29(43), 13445–13453.

Arnal, L. H., Wyart, V., & Giraud, A.L. (2011). Transitions in neural oscillations reflect prediction errors generated in audiovisual speech, *Nature Neuroscience*, 16(6), 794-801.

Atencio, C. A., Sharpee, T. O., & Schreiner, C. E. (2009). Hierarchical computation in the canonical auditory cortical circuit. *Proceedings of the National Academy of Sciences of the USA*, 106(51), 21894–21899.

Bates, E., Wilson, S. M., Saygin, A. P., Dick, F., Sereno, M. I., Knight, R. T., & Dronkers, N. F. (2003). Voxel-based lesion-symptom mapping. *Nature Neuroscience*, 6(5), 448–450.

Bendor, D., & Wang, X. (2006). Cortical representations of pitch in monkeys and humans. *Current Opinion in Neurobiology*, 16(4), 391–399.

Bendor, D., & Wang, X. (2007) Differential neural coding of acoustic flutter within primate auditory cortex. *Nature Neuroscience*, 10(6), 763–771.

Binzegger, T., Douglas, R. J., & Martin, K. A. (2007). Stereotypical bouton clustering of individual neurons in cat primary visual cortex. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 27(45), 12242–12254.

Blumstein, S. E., Myers, E. B., & Rissman, J. (2005). The perception of voice onset time: An fmri investigation of phonetic category structure. *Journal of Cognitive Neuroscience*, 17(9), 1353–1366.

Boemio, A., Fromm, S., Braun, A., & Poeppel, D. (2005). Hierarchical and asymmetric temporal sensitivity in human auditory cortices. *Nature Neuroscience*, 8(3), 389–395.

Borgers, C. & Kopell, N. J. (2008). Gamma oscillations and stimulus selection. *Neural Computations*, 20(2), 383–414.

Borgers, C., Epstein, S., & Kopell, N. J. (2005). Background gamma rhythmicity and attention in cortical local circuits: A computational study. *Proceedings of the National Academy of Sciences of the USA*, 102(19), 7002–7007.

Brennan, J., Nir, Y., Hasson, U., Malach, R., Heeger, D. J., & Pylkkänen, L. (2010). Syntactic structure building in the anterior temporal lobe during natural story listening. *Brain and Language*. doi:10.1016/j.bandl.2010.04.002

Britvina, T., & Eggermont, J. J. (2007). A Markov model for interspike interval distributions of auditory cortical neurons that do not show periodic firings. *Biological Cybernetics*, 96(2), 245–264.

Brugge, J. F., Nourski, K. V., Oya, H., Reale, R. A., Kawasaki, H., Steinschneider, M., & Howard, M. A., 3rd (2009). Coding of repetitive transients by auditory cortex on Heschl's gyrus. *Journal of Neurophysiology*, 102(4), 2358–2374.

Canolty, R. T., & Knight, R. T. (2010). The functional role of cross-frequency coupling. *Trends in Cognitive Sciences*, 14(11), 506–515.

Chait, M., Poeppel, D., & Simon, J. Z. (2006). Neural response correlates of detection of monaurally and binaurally created pitches in humans. *Cerebral Cortex*,, 16(6), 835–848.

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

Cleary, M., & Pisoni, D. B. (2001). Speech perception and spoken word recognition: Research and theory. In B. Goldstein (Ed.), *Handbook of perception* (pp. 499–534). Cambridge, MA: Blackwell.

Corballis, M. C. (2009). The evolution of language. *Annals of the New York Academy of Sciences*, 1156, 19–43.

da Costa, N. M., & Martin, K. A. C. (2010). Whose cortical column would that be? *Frontiers in Neuroanatomy*. doi: 10.3389/fnana.2010.00016.

Davis, M. H., & Johnsrude, I. S. (2007). Hearing speech sounds: Top-down influences on the interface between audition and speech perception. *Hearing Research*, 229(1–2), 132–147.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences. *Journal Experimental Psychology General*, 134(2), 222–241.

Ding, N., & Simon, J. Z. (2009). Neural representations of complex temporal modulations in the human auditory cortex. *Journal of Neurophysiology*, 102(5), 2731–2743. Eger, E., Michel, V., Thirion, B., Amadon, A., Dehaene, S., & Kleinschmidt, A. (2009). Deciphering cortical number coding from human brain activity patterns. *Current Biology*, 19(19), 1608–1615.

Elhilali, M., J. B. Fritz, Klein, D. J., Simon, J. Z., & Shamma, S. A. (2004). Dynamics of precise spike timing in primary auditory cortex. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 24(5), 1159–1172.

Eliades, S. J., & Wang, X. (2008). Neural substrates of vocalization feedback monitoring in primate auditory cortex. *Nature*, 453(7198), 1102–1106.

Elliott, T. M., & Theunissen, F. E. (2009). The modulation transfer function for speech intelligibility. *PLoS Computational Biology*, 5(3), e1000302. doi:10.1371/journal.pcbi.1000302

Faulkner, A., Rosen, S., & Smith, C. (2000). Effects of the salience of pitch and periodicity information on the intelligibility of four-channel vocoded speech: Implications for cochlear implants. *The Journal of the Acoustical Society of America*, 108(4), 1877–1887.

Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology*, 44(4), 491–505.

Formisano, E., Kim, D. S., Di Salle, F., van de Moortele, P. F., Ugurbil, K., & Goebel, R. (2003). Mirror-symmetric tonotopic maps in human primary auditory cortex. *Neuron*, 40(4), 859–869.

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

Friederici, A. D., Kotz, S. A., Scott, S. K., & Obleser, J. (2010). Disentangling syntax and intelligibility in auditory language comprehension. *Human Brain Mapping*, 31(3), 448–457.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127–138.

Gaese, B. H., & Ostwald, J. (1995). Temporal coding of amplitude and frequency modulation in the rat auditory cortex. *The European Journal of Neuroscience*, 7(3), 438–450.

Gaskell, M. G., & Marslen-Wilson, W. D. (2002). Representation and competition in the perception of spoken words. *Cognitive Psychology*, 45(2), 220–266.

Ghitza, O. (2011). Linking speech perception and neurophysiology: Speech decoding guided by cascaded oscillations locked to the input rhythm. *Frontiers in Psychology*, 2:130.

Ghitza, O., & Greenberg, S. (2009). On the possible role of brain rhythms in speech perception: Intelligibility of time-compressed speech with periodic and aperiodic insertions of silence. *Phonetica*, 66(1–2), 113–126.

Giraud, A.L., & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience*, in press.

Giraud, A. L., & Price, C. J. (2001). The constraints functional neuroimaging places on classical models of auditory word processing. *Journal of Cognitive Neuroscience*, 13(6), 754–765.

Giraud, A. Lorenzi, C., Ashburner, J., Wable, J., Johnsrude, I., Frackowiak, R., & Kleinschmidt, A. (2000). Representation of the temporal envelope of sounds in the human brain. *Journal of Neurophysiology*, 84(3), 1588–1598.

Giraud, A. L., Kell, C., Thierfelder, C., Sterzer, P., Russ, M. O., Preibisch, C., & Kleinschmidt, A. (2004). Contributions of sensory input, auditory search and verbal comprehension to cortical activity during speech processing. *Cerebral Cortex*, 14(3), 247–255.

Giraud, A. L., Kleinschmidt, A., Poeppel, D., Lund, T. E., Frackowiak, R. S., & Laufs, H. (2007). Endogenous cortical rhythms determine cerebral specialization for speech perception and production. *Neuron*, 56(6), 1127–1134.

Greenberg, S., & Ainsworth, W. A. (2006). *Listening to speech: An auditory perspective*. Mahwah, NJ: Lawrence Erlbaum.

Greenberg, S. & Arai, T. (2001). The relation between speech intelligibility and the complex modulation spectrum. *Proceedings of the 7th Eurospeech Conference on Speech Communication and Technology (Eurospeech-2001)*, 473-476.

Greenberg, S., & Kingsbury, B. E. D. (1997). The modulation spectrogram: In pursuit of an invariant representation of speech. *Proceedings of the 1997 IEEE International Conference on Acoustics*, *Speech*, *and Signal Processing (ICASSP '97)-Volume 3*

Griffiths, T. D., Kumar, S., Sedley, W., Nourski, K. V., Kawasaki, H., Oya, H., et al. (2010) Direct recordings of pitch responses from human auditory cortex. *Current Biology*, 20(12), 1128–1132.

Guenther, F. H. (2006). Cortical interactions underlying the production of speech sounds. *Journal of Communication Disorders*, 39(5), 350–365.

Guenther, F. H., Ghosh, S. S., & Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language*, 96(3), 280–301.

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

Hawkins, S. (1999). Reevaluating assumptions about speech perception: Interactive and integrative theories. In J. M. Pickett (Ed.), *The acoustics of speech communication* (pp. 232–288). Boston: Allyn and Bacon.

Heil, P. (1997a). Auditory cortical onset responses revisited. I. First-spike timing. *Journal of Neurophysiology*, 77(5), 2616–2641.

Heil, P. (1997b). Auditory cortical onset responses revisited. II. Response strength. *Journal of Neurophysiology*, 77(5), 2642–2660.

Hickok, G., & Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception. *Trends in Cognitive Sciences*, 4(4), 131–138.

Hickok, G., & Poeppel, D. (2004). Dorsal and ventral streams: A framework for understanding aspects of the functional anatomy of language. *Cognition*, 92(1–2), 67–99.

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

Hickok, G., Houde, J., & Rong, F. (2011). Sensorimotor integration in speech processing: Computational basis and neural organization. *Neuron*, 69(3), 407–422. Hochstein, S., &

Ahissar, M. (2002). View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron*, 36(5), 791–804.

Holcombe, A. O. (2009). Seeing slow and seeing fast: Two limits on perception. *Trends in Cognitive Sciences*, 13(5), 216–221.

Hromádka, T., & Zador, A. M. (2009). Representations in auditory cortex. *Current Opinion in Neurobiology*, 19(4), 430–433.

Hutsler, J., & Galuske, R. A. (2003). Hemispheric asymmetries in cerebral cortical networks. *Trends in Neurosciences*, 26(8), 429–435.

Indefrey, P., & Levelt, W. J. (2004) The spatial and temporal signatures of word production components. *Cognition*, 92(1–2), 101–144.

Jamison, H. L., Watkins, K. E., Bishop, D. V., & Matthews, P. M. (2006). Hemispheric specialization for processing auditory nonspeech stimuli. *Cerebral Cortex*,, 16(9), 1266–1275.

Joris, P. X., Schreiner, C. E., & Rees, A. (2004). Neural processing of amplitude-modulated sounds. *Physiological Reviews*, 84(2), 541–577.

Kanedera, N., Arai, T., Hermansky, H., & Pavel, M. (1999). On the relative importance of various components of the modulation spectrum for automatic speech recognition. *Speech Communication*, 28(1), 43–55.

Kayser, C., Logothetis, N. K., & Panzeri, S. (2010). Millisecond encoding precision of auditory cortex neurons. *Proceedings of the National Academy of Sciences of the USA*, 107(39), 16976–16981.

Kell, C. A., Morillon, B., Kouneiher, F., & Giraud, A. L. (2010). Lateralization of speech production starts in sensory cortices—a possible sensory origin of cerebral left dominance for speech. *Cerebral Cortex*,. doi:10.1093/cercor/bhq167

Klatt, D. H. (1989). Review of selected models of speech perception. In W. Marslen-Wilson (Ed.), *Lexical representation and process* (pp. 169–226). Cambridge, MA: MIT Press.

Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics: (De)constructing the N400. *Nature Reviews Neuroscience*, 9(12), 920–933.

Laver, J. (1994). *Principles of phonetics. Cambridge textbooks in linguistics*. New York: Cambridge University Press.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, 74(6), 431–461.

Loebach, J. L., & Wickesberg, R. E. (2008). The psychoacoustics of noise vocoded speech: A physiological means to a perceptual end. *Hearing Research*, 241(1–2), 87–96.

Luo, H., & Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron*, 54(6), 1001–1010.

Luo, H., Wang, Y., Poeppel, D., & Simon, J. Z. (2006). Concurrent encoding of frequency and amplitude modulation in human auditory cortex: MEG evidence. *Journal of Neurophysiology*, 96(5), 2712–2723.

Luo, H., Boemio, A., Gordon, M., & Poeppel, D. (2007a). The perception of FM sweeps by Chinese and English listeners. *Hearing Research*, 224(1–2), 75–83.

Luo, H., Wang, Y., Poeppel, D., & Simon, J. Z. (2007b). Concurrent encoding of frequency and amplitude modulation in human auditory cortex: Encoding transition. *Journal of Neurophysiology*, 98(6), 3473–3485.

Mantini, D., Perrucci, M. G., Del Gratta, C., Romani, G. L., &, Corbetta, M. (2007). Electrophysiological signatures of resting state networks in the human brain. *Proceedings of the National Academy of Sciences of the USA*, 104(32), 13170–13175.

McClelland, J. L., & Elman, J. L. (1986). The trace model of speech perception. *Cognitive Psychology*, 18(1), 1–86.

Middlebrooks, J. C. (2008). Auditory cortex phase locking to amplitude-modulated cochlear implant pulse trains. *Journal of Neurophysiology*, 100(1), 76–91.

Miller, G. A. (1951). *Language and communication*. New York: McGraw-Hill.

Monahan, P. J., & Idsardi, W. J. (2010). Auditory sensitivity to formant ratios: Toward an account of vowel normalization. *Language and Cognitive Processes*, 25(6), 808–839.

Morillon, B., Lehongre, K., Frackowiak, R. S., Ducorps, A., Kleinschmidt, A., Poeppel, D., & Giraud, A. L. (2010). Neurophysiological origin of human brain asymmetry for speech and language. *Proceedings of the National Academy of Sciences of the USA*, 107(43), 18688–18693.

Näätänen, R., Lehtokoski, A., Lennes, M., Cheour, M., Huotilainen, M., Iivonen, A., & Alho, K. (1997). Language-specific phoneme representations revealed by electric and magnetic brain responses. *Nature*, 385(6615), 432–434.

Nahum, M., Nelken, I., & Ahissar, M. (2008). Low-Level information and high-level perception: The case of speech in noise. *PLoS Biology*, 6(5), e126.

Nelken, I., Bizley, J. K., Nodal, F. R., Ahmed, B., King, A. J., & Schnupp, J. W. (2008). Responses of auditory cortex to complex stimuli: Functional organization revealed using intrinsic optical signals. *Journal of Neurophysiology*, 99(4), 1928–1941.

Obleser, J., Boecker, H., Drzezga, A., Haslinger, B., Hennenlotter, A., Roettinger, M., et al. (2006). Vowel sound extraction in anterior superior temporal cortex. *Human Brain Mapping*, 27(7), 562–571.

Obleser, J., Eisner, F., & Kotz, S. A. (2008). Bilateral speech comprehension reflects differential sensitivity to spectral and temporal features. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 28(32), 8116–8123.

Overath, T., Kumar, S., von Kriegstein, K., & Griffiths, T. D. (2008). Encoding of spectral correlation over time in auditory cortex. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 28(49), 13268–13273.

Pardo, J. S., & Remez, R. E. (2006). The perception of speech. In M. Traxler & M. A. Gernsbacher (Eds.), *The handbook of psycholinguistics*, 2nd ed. (pp. 201–248). New York: Academic Press.

Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience*, 8(12), 976–987.

Patterson, R. D., Uppenkamp, S., Johnsrude, I. S., & Griffiths, T. D. (2002). The processing of temporal pitch and melody information in auditory cortex. *Neuron*, 36(4), 767–776.

Petkov, C. I., Kayser, C., Augath, M., & Logothetis, N. K. (2006). Functional imaging reveals numerous fields in the monkey auditory cortex. *PLoS Biology*, 4(7), e215.

Phillips, D. P., Hall, S. E., & Boehnke, S. E. (2002). Central auditory onset responses, and temporal asymmetries in auditory perception. *Hearing Research*, 167(1–2), 192–205.

Pickett, J. M. (1999). *The acoustics of speech communication*. Boston: Allyn and Bacon.

Pienkowski, M., & Eggermont, J. J. (2010). Nonlinear cross-frequency interactions in primary auditory cortex spectrotemporal receptive fields: A Wiener-Volterra analysis. *Journal of Computational Neuroscience*, 28(2), 285–303.

Poeppel, D. (2001). Pure word deafness and the bilateral processing of the speech code. *Cognitive Science*, 25(5), 679–693.

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: Cerebral lateralization as asymmetric sampling in time. *Speech Communication*, 41(1), 245–255.

Poeppel, D., & Monahan, P. J. (2008). Speech perception: Cognitive foundations and cortical implementation. *Current Directions in Psychological Science*, 17(2), 80.

Poeppel, D., Idsardi, W. J., & van Wassenhove, V. (2008). Speech perception at the interface of neurobiology and linguistics. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 363(1493), 1071–1086.

Pöppel, E. (1988). *Mindworks: Time and conscious experience*. Boston: Harcourt Brace Jovanovich.

Pulvermüller, F., Huss, M., Kherif, F., Moscoso del Prado Martin, F., Hauk, O., & Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. *Proceedings of the National Academy of Sciences of the USA*, 103(20), 7865–7870.

Rabiner, L., & Juang, B. H. (1993). Fundamentals of speech recognition. Englewood Cliffs, NJ:Prentice-Hall

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

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science*, 212(4497), 947–949.

Roberts, B., Summers, R. J., & Bailey, P. J. (2011). The intelligibility of noise-vocoded speech: Spectral information available from across-channel comparison of amplitude envelopes. *Proc. R. Soc. B*, 278(1711), 1595–1600.

Rosen, S. (1992). Temporal information in speech: Acoustic, auditory and linguistic aspects. *Philosophical Transactions of the Royal Society of London B:*, *Biological Sciences*, 336(1278), 367–373.

Saoud, H., Josse, G., Bertasi, E., Truy, E., Chait, M., & Giraud, A-L. (2012). Brain-speech alignment enhances auditory cortical responses and pseech perception. *The Journal of Neuroscience*, in press.

Saberi, K., & Perrott, D. R. (1999). Cognitive restoration of reversed speech. *Nature*, 398(6730), 760.

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

Schroeder, C. E., & Lakatos, P. (2009a). Low-frequency neuronal oscillations as instruments of sensory selection. *Trends in Neurosciences*, 32(1), 9–18.

Schroeder, C. E., & Lakatos, P. (2009b). The gamma oscillation: Master or slave? *Brain Topography*, 22(1), 24–26.

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

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

Scott, S. K., Rosen, S., Lang, H., & Wise, R. J. (2006). Neural correlates of intelligibility in speech investigated with noise vocoded speech-a positron emission tomography study. *The Journal of the Acoustical Society of America*, 120(2), 1075–1083.

Shamir, M., Ghitza, O., Epstein, S., & Kopell, N. (2009). Representation of time-varying stimuli by a network exhibiting oscillations on a faster time scale. *PLoS Computational Biology*, 5(5), e1000370.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, 270(5234), 303.

Sharma, A., & Dorman, M. F. (1999).Cortical auditory evoked potential correlates of categorical perception of voice-onset time. *The Journal of the Acoustical Society of America*, 106, 1078–1083.

Smith, Z. M., Delgutte, B., & Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, 416(6876), 87–90.

Souza, P., & Rosen S. (2009). Effects of envelope bandwidth on the intelligibility of sine- and noise-vocoded speech. *The Journal of the Acoustical Society of America*, 126(2), 792–805.

Steeneken, H. J. M., & Houtgast, T. (1980). A physical method for measuring speech-transmission quality. *The Journal of the Acoustical Society of America*, 67(1), 318–326.

Stevens, K. N. (1998). *Acoustic phonetics*. Cambridge, MA: MIT Press.

Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *The Journal of the Acoustical Society of America*, 111(4), 1872–1891.

Telkemeyer, S., Rossi, S., Koch, S. P., Nierhaus, T., Steinbrink, J., Poeppel, D., & Wartenburger, I. (2009). Sensitivity of newborn auditory cortex to the temporal structure of sounds. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 29(47), 14726–14733.

Tian, X., & Poeppel, D. (2010). Mental imagery of speech and movement implicates the dynamics of internal forward models. *Frontiers in Psychology*. doi: 10.3389/fpsyg.2010.00166

Tiesinga, P., & Sejnowski, T. J. (2009). Cortical enlightenment: Are attentional gamma oscillations driven by ING or PING? *Neuron*, 63(6), 727–732.

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

Ueno, T., Saito, S., Rogers, T. T., & Lambon-Ralph, M. A. (2011) Lichtheim 2: synthesizing aphasia and the neural basis of language in a neurocomputational model of the dual dorsal-ventral language pathways. *Neuron* 72: 385–96.

Van Rullen, R., & Koch, C. (2003). Is perception discrete or continuous? *Trends in Cognitive Sciences*, 7(5), 207–213.

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the USA*, 102(4), 1181–1186.

von Kriegstein, K., Smith, D. R., Patterson, R. D., Kiebel, S. J., & Griffiths, T. D. (2010) How the human brain recognizes speech in the context of changing speakers. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 30(2), 629–638.

Wacongne, C., Labyt, E., van Wassenhove, V., Bekinchtein, T., Naccache, L., & Dehaene, S. (2011). Proceedings of the National Academy of Sciences, 108: 20754–9.

Wang, X. (2007). Neural coding strategies in auditory cortex. *Hearing Research*, 229(1–2), 81–93.

Wang, X. J. (2010). Neurophysiological and computational principles of cortical rhythms in cognition. *Physiological Reviews*, 90(3), 1195–1268.

Warrier, C., Wong, P., Penhune, V., Zatorre, R., Parrish, T., Abrams, D., & Kraus, N. (2009). Relating structure to function: Heschl's gyrus and acoustic processing. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 29(1), 61–69.

Wilson, S. M., Saygin, A. P., Sereno, M. I., & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nature Neuroscience*, 7(7), 701–702.

Womelsdorf, T., Schoffelen, J. M., Oostenveld, R., Singer, W., Desimone, R., Engel, A. K., & Fries, P. (2007). Modulation of neuronal interactions through neuronal synchronization. *Science* 316(5831), 1609–1612.

Zaehle, T., Wüstenberg, T., Meyer, M., & Jäncke, L. (2004). Evidence for rapid auditory perception as the foundation of speech processing: A sparse temporal sampling fmri study. *The European Journal of Neuroscience*, 20(9), 2447–2456.

Zarate, J. M., & Zatorre, R. J. (2008). Experience-dependent neural substrates involved in vocal pitch regulation during singing. *NeuroImage*, 40(4), 1871–1887.

Zarate, J. M., Wood, S., & Zatorre, R. J. (2010). Neural networks involved in voluntary and involuntary vocal pitch regulation in experienced singers. *Neuropsychologia*, 48(2), 607–618.

Zatorre, R. J., & Gandour, J. T. (2008). Neural specializations for speech and pitch: Moving beyond the dichotomies. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 363(1493), 1087–1104.

Zatorre, R. J., Belin, P., & Penhune, V. B. (2002). Structure and function of auditory cortex: Music and speech. *Trends in Cognitive Sciences*, 6(1), 37–46.