Temporal coding in the auditory cortex

Luc H Arnal, David Poeppel, Anne-Lise Giraud

Handbook of Clinical Neurology, Vol. 129 (3rd series)
The Human Auditory System
G.G. Celesia and G. Hickok, Editors

Chapter 5

Temporal coding in the auditory cortex

LUC H. ARNAL[1,2], DAVID POEPPEL[2], AND ANNE-LISE GIRAUD[1*]
[1]*Department of Neurosciences, University Medical Centre, Geneva, Switzerland*
[2]*Department of Psychology, New York University, New York, NY, USA*

OVERVIEW

Of all the signals the human auditory system has to process, the one with the most compelling relevance to the listener is arguably speech. Speech perception is learned and executed with automaticity and great ease, even by very young children, but is handled surprisingly poorly by even the most sophisticated automatic devices. Parsing and decoding speech can hence be considered one of the main challenges of the auditory system. This chapter focuses on how the human auditory cortex uses the temporal structure of the acoustic signal to extract phonemes and syllables, the two major types of events that need to be identified in connected speech.

Speech is a complex "multiplexed" acoustic signal exhibiting a quasiperiodic behavior at several timescales. The neural signals recorded from the auditory cortex using electroencephalography (EEG) or magnetoencephalography (MEG) also show a quasiperiodic structure, whether in response to speech or not. In this chapter we review recent neurophysiologic models of speech perception, grounded on the assumption that the quasiperiodic structure of collective neural activity in auditory cortex represents the ideal mechanical infrastructure to solve the speech demultiplexing problem, i.e., the fractioning of speech into linguistic constituents of variable size. The theoretic models presented here remain largely hypothetic. That being said, we believe that they constitute exciting new hypotheses, and should lead to new research questions and incremental progress on this foundational question about human perception.

The chapter proceeds as follows. First, some of the essential features of natural and speech auditory stimuli are outlined. Next, the properties of auditory cortex that reflect its sensitivity to these features are reviewed, and finally current ideas about the neurophysiologic mechanisms underpinning the processing of connected speech are discussed.

Timescales in auditory perception

Sounds are audible over a broad frequency range between 20 and 20 000 Hz. They enter the outer ear and travel through the middle ear to the inner ear, where they provoke the basilar membrane to vibrate at a specific location, depending on the sound frequency. Low and high frequencies induce vibrations of the apex and base of the basilar membrane, respectively. The deformation of the membrane upon acoustic stimulation provokes the deflection of inner hair cell ciliae, and the emission of a neural signal to cochlear neurons, subsequently transmitted to neurons of the cochlear nucleus in the brainstem. Each cochlear neuron is sensitive to a specific range of acoustic frequencies between 20 Hz and 20 kHz. Owing to their regular position along the basilar membrane, the cochlear neurons ensure the place coding of acoustic frequencies, also called "tonotopy," which is preserved up to the cortex (Moerel et al., 2013; Saenz and Langers, 2014).

Acoustic fluctuations below 20 Hz are not audible. They do not elicit place-specific response in the cochlea. Low frequencies <300 Hz are present in complex sounds as temporal fluctuations of audible frequencies, and are encoded through the discharge rate of cochlear neurons (Zeng, 2002). There is hence a range of frequencies from 20 to 300 Hz that are both place- and rate-coded at the auditory periphery. Temporal modulation of sounds in these frequencies typically elicits a sensation of pitch. Figure 5.1A and B summarizes the correspondence between categories of perceptual attributes and the sound modulation frequency (see also Nourski and Brugge, 2011 for a review). When sounds are modulated at very slow

*Correspondence to: Anne-Lise Giraud, Department of Neuroscience, University Medical Centre (CMU), 1, rue Michel-Servet, 1211 Geneva, Switzerland. Tel: +41-(0)223795547, Fax: +41-(0)223795452, E-mail: Anne-Lise.Giraud@unige.ch

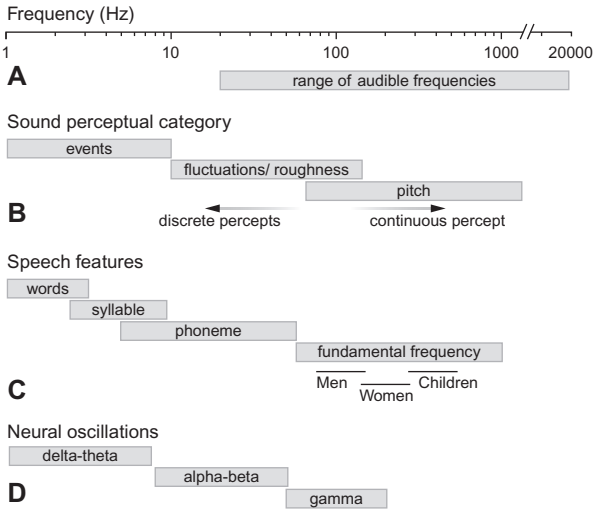

Fig. 5.1. (**A**) Scale of perceived temporal modulation. (**B**) Relevant psychophysical parameters (perceptual changes) of the spectrogram reflect the temporal constraints that superimpose on the structure of linguistic signals. (**C**) Temporal structure of linguistic features. (**D**) The length of linguistic features remarkably matches the frequency of oscillations that are observed at rest in the brain. Note that the frequency ranges at which auditory percepts switch from discrete (flutter) to continuous (pitch) roughly match the upper limit at which gamma rhythms can be entrained by the stimulus (≈200 Hz). (Modified from Joris et al., 2004, with permission from the American Physiological Society, and from Nourski and Brugge, 2011).

rates <10 Hz, a sequence of distinct events is perceived. When modulations accelerate from 10 to about 100 Hz distinct events merge into a single auditory stream, and the sensation evolves from fluctuating magnitude to a sensation of acoustic roughness (Fig. 5.1B).

Speech sounds are complex acoustic signals that involve only the lower part of audible frequencies (20–8000 Hz). They are "complex" in the sense that both their frequency distribution and their magnitude vary strongly and quickly over time. In natural speech, amplitude modulations (AMs) at slow (<20 Hz) and fast (>100 Hz) timescales are coupled (Fig. 5.2) and slower temporal fluctuations modulate the amplitude of spectral fluctuations. Current views suggest that slow modulations (<5 Hz) signal word and syllable boundaries (Hyafil et al., 2012), which are hence perceived as a sequence of distinct events, whereas phonemes (speech sounds) are signaled by fast spectrotemporal modulations (<30 Hz). They can be perceived as distinct events only when being discriminated from each other. Faster modulations, such as those imposed by the glottal pulse (100–300 Hz), indicate the voice pitch (Fig. 5.1C). Figure 5.1D shows how these perceptual events relate to the different frequency ranges of the EEG.

The temporal structure of speech sounds

Figure 5.2 illustrates two useful ways to visualize speech signals: 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 the "cochlear" filterbank) are below 20 Hz and peak roughly at a rate of 4–6 Hz (Steeneken and Houtgast, 1980; Elliott and Theunissen, 2009). From the perspective of what auditory cortex receives as input, namely the modulations at the output of each frequency

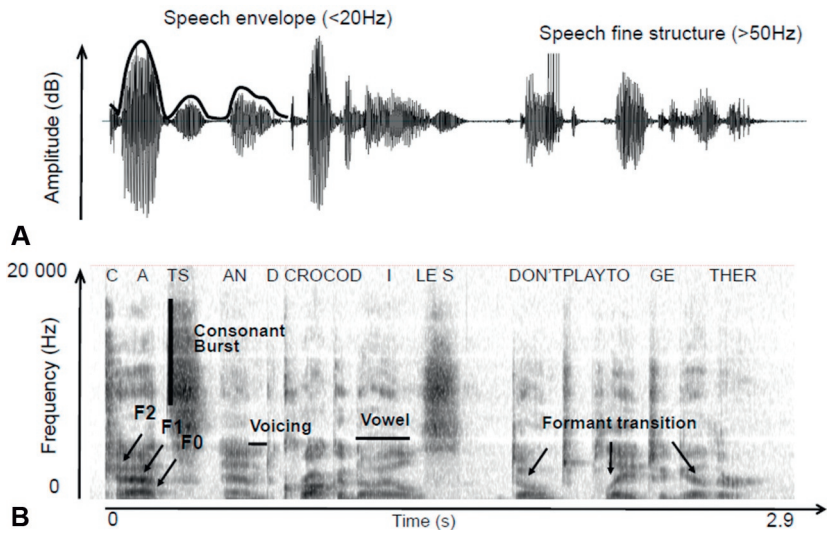

Fig. 5.2. (**A**) Waveform and (**B**) spectrogram of the same sentence uttered by a male speaker. Some of the key acoustic cues in speech comprehension are highlighted in black. (From Giraud and Poeppel, 2012b, with permission from Springer Science and Business Media.)

channel of the cochlear filterbank, these energy fluctuations can be characterized by the modulation spectrum (Kingsbury et al., 1998; Kanedera et al., 1999). At the shortest timescale (below 1 ms or equivalently above 1 kHz), the very fast temporal fluctuations are transformed into a spectral representation at the cochlea and the neural processing of these features is generally known as spectral processing. At an intermediate timescale (~70 Hz – 1 kHz), the temporal fluctuations are usually referred to as the temporal fine structure. The temporal fine structure is critical to the perception of pitch and interaural time differences that are important cues for sound source localization (Plack et al., 2005; Grothe et al., 2010). Temporal fluctuations on an even longer timescale (~1–10 Hz) are heard as a sequence of discrete events. Acoustic events occurring on this timescale include syllables and words in speech and notes and beats in music. Of course, there are no clear boundaries between these timescales; they are divided here based on human auditory perception.

The second analytic representation, the spectrogram, decomposes the acoustic signal in frequency, time, and amplitude domains (Fig. 5.2B). Although the human auditory system captures frequency information between 20 Hz and 20 kHz (as in Fig. 5.2), 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 an even narrower bandwidth (200–3600 Hz) and is comfortably understood by normal listeners.

A number of critical acoustic features can be identified in the speech 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; Fig. 5.1D). The horizontal bands of energy show where in the 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.

Notwithstanding the perceptual 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 structure is replaced by noise and only the envelope is preserved. For speech to remain intelligible, this manipulation should be done in separate bands across the spectrum. With training, it is still possible to grasp the speech content when the speech envelope, that is, temporal modulations of speech at relatively slow rates, is applied to only four separate frequency bands (e.g., Shannon et al., 1995). Such speech signals, containing only envelope but no fine structure information, are called vocoded speech (Faulkner et al., 2000). Compelling demonstrations, exemplified by this type of signal decomposition, illustrate that the speech signal can undergo radical alterations and distortions and yet remain intelligible (Shannon et al., 1995; Smith et al., 2002). Such findings have led to the idea that the temporal envelope is sufficient to yield speech comprehension (Rosen, 1992; Drullman et al., 1994a, b; Shannon et al., 1995; Giraud et al., 2004; Scott et al., 2006; Loebach and Wickesberg, 2008; Souza and 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 empiric 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 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.

Based on this brief and selective summary, two concepts merit emphasis: first, the extended speech signal contains critical information that is modulated at rates below 20 Hz, with the modulation peaking around 5 Hz (Edwards and Chang, 2013). This low-frequency information correlates closely with the syllabic structure of connected speech (Hyafil et al., 2012). 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 carry information about the speaker's gender or identity and other relevant speech attributes (Elliott and Theunissen, 2009). 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. In this chapter, we discuss the timescales longer than 5 ms (<200 Hz) with a focus on the timescale between 100 ms and 1 second (1–10 Hz). Temporal features that contribute to the spatial localization of sounds will not be discussed (see e.g. Grothe et al., 2010 for a review).

ENCODING OF SPECTROTEMPORAL FEATURES IN THE AUDITORY CORTEX

Sensitivity to temporal modulations in the primary auditory cortex

Speech temporal variations in signal frequency and magnitude are at the basis of articulated speech and require specific neural encoding and decoding properties. A critical property is the sensitivity of auditory neurons to temporal modulations. Auditory neurons in subcortical nuclei (cochlear nucleus, inferior colliculus, and medial geniculate nucleus of the thalamus) phase-lock very well with fast trains of stimuli presented at rates over 200 Hz (Fishman et al., 2000; Brugge et al., 2009; see also Joris et al., 2004 and Nourski and Brugge, 2011 for reviews on the topic). This property to adjust the response rate to the rate of acoustic modulations gradually diminishes along the auditory hierarchy (see Sharpee et al., 2011; for a review). In the primary auditory cortex, the firing rate of a large number of neurons is phase-locked to slow temporal modulations below 30 Hz (Liang et al., 2002; Malone et al., 2010; Yin et al., 2011). Progressive slowing of phase-locking properties across the auditory hierarchy has been observed across species, e.g., monkeys, cats, and ferrets, for both anesthetized and awake animals. The degree of neural phase locking to temporal AM is often characterized as a function of the modulation rate, using temporal modulation transfer functions.

In the auditory cortex, the encoding of temporal features is implemented using at least two complementary strategies. Some neurons are sensitive to slow modulation rates, and the degree of phase locking decreases with increasing rates. Other neurons are tuned to a preferred modulation rate and the degree of phase locking decreases when the modulation rate deviates from the preferred one. Some neurons even show multiple preferred modulation rates (Malone et al., 2010; Yin et al., 2011). Single-unit recordings from the primary auditory cortex of awake marmoset suggest that AM is encoded by different neural codes and/or different neurons (Wang, 2007; Wang et al., 2008). One population of neurons, called the synchronized population, encodes AM by spikes phase-locked to the stimulus envelope. A second neural population, called the non-synchronized population, encodes AM by the mean firing rate rather than the timing of spikes. The two neural populations in general do not spatially overlap. While the synchronized population encodes slow temporal modulations (mostly below 50 Hz), the non-synchronized population encodes faster temporal modulations (mostly above 50 Hz). Studies in awake macaque monkey, however, suggested that neurons might actually encode both slow temporal modulations by phase-locked activity and fast temporal modulations by the firing rate of non-phase-locked activity (Malone et al., 2010). In other words, the same neuron might encode slow and fast modulations, suggesting that primary auditory cortex neurons may carry multiplexed temporal and rate codes.

Although single auditory cortical neurons generally cannot phase lock to temporal modulations above 100 Hz (Wang, 2007), larger-scale recordings such as local field potentials and human MEG/EEG show phase locking to the stimulus modulations beyond 100 Hz. This suggests that macro-scale measurements that spatially integrate neuronal responses at the population level reflect the ability of auditory cortical regions to maintain a – spatially distributed – representation of higher temporal frequencies. In humans, intracranial recordings from the core auditory cortex and the lateral surface of posterolateral superior temporal gyrus show that phase-locked neural activity is most prominent below 50 Hz but remains measurable up to ~200 Hz (Brugge et al., 2009). Similarly, the human MEG response also shows phase locking up to 100 Hz (Lehongre et al., 2011; Miyazaki et al., 2013). That neural oscillations phase lock to exogenous entrainment up to this frequency possibly determines the transition between discrete and continuous auditory percepts (Fig. 5.1B).

Sensitivity to spectrotemporal modulations

Speech signals are characterized by modulations in both spectral and temporal domains (frequency modulations (FM) and AM). Two separate possible codes to represent complex stimuli such as speech have been implicated in the preceding text, a place code for FMs and a temporal code for AMs. Whether the encoding of frequency and AMs is implemented by a single or by distinct mechanisms is a long-lasting question, whether at the periphery or at the cortical level. The idea of a single code for spectrotemporal modulations is supported by the presence of neurons that respond to FMs but not AMs (Gaese and Ostwald, 1995) and by complex responses to spectrotemporal modulations (Pienkowski and Eggermont, 2009; Schonwiesner and Zatorre, 2009). Luo et al. (2006, 2007) and Ding and Simon (2009) tested, based on MEG recordings in human listeners, whether FM and AM used the same coding principles (Fig. 5.3). The authors argue that if coding equivalence (or similarity) is the case, cortical 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

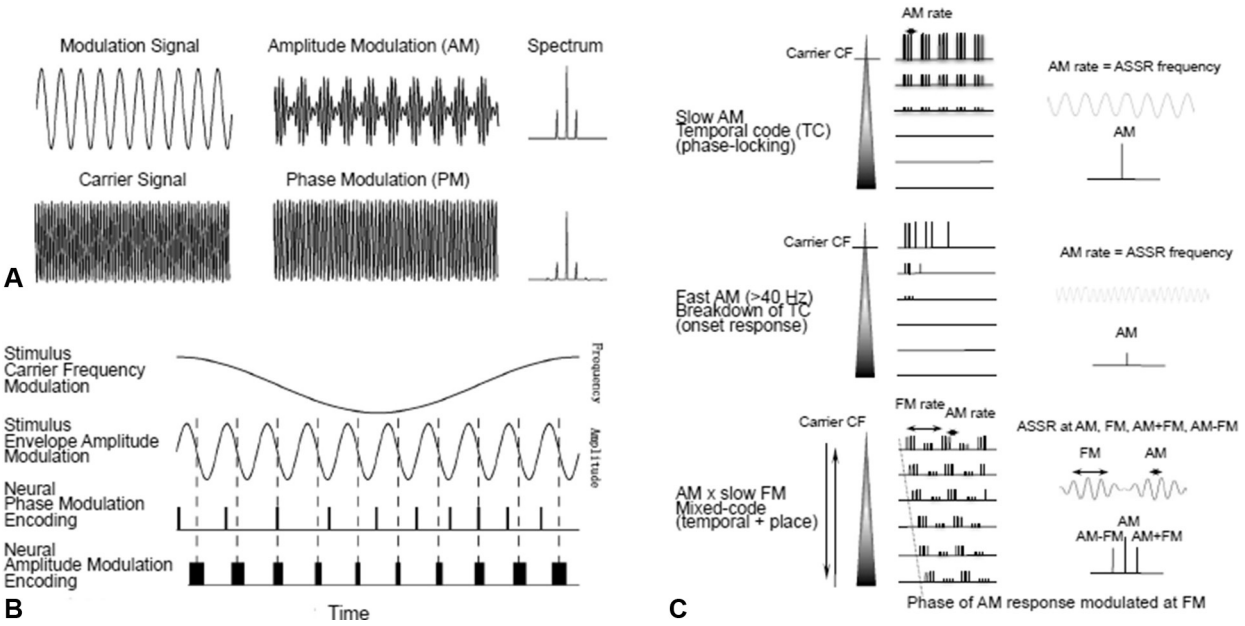

Fig. 5.3. Principles of amplitude and frequency modulation encoding in auditory cortex. (**A**) In radio engineering, modulation is used to encode acoustic stimuli, which can be either amplitude-modulated (AM; upper row) or phase-modulated (PM; second row). (**B**) Proposals for neural AM and PM encoding. 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 spikes' 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 details in three different conditions: slow AM (upper row), fast AM (second row), and when AM and PM covary (last row). CF, characteristic frequency. ASSR, auditory steady-state responses. (From Luo et al., 2006, with permission from the American Physiological Society.)

FM auditory responses (auditory steady-state responses at 40 Hz) are modulated by slow AM. That AM and FM interact non-linearly 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 (Barton et al., 2012). Whereas FM, by hypothesis, is encoded by a combination of place and temporal coding, AM is mostly encoded by temporal coding. As a consequence, 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 FMs 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 (Fig. 5.1). 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 the entire population stimulated by the varying carrier. 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 (pitch) neurons with very low characteristic frequency. 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, provides enough representational complexity to account for non-linear neuronal responses to spectrotemporal acoustic modulations without invoking a specific AM/FM code.

CORTICAL PROCESSING OF CONTINUOUS SOUND STREAMS

The discretization problem

Most of the aforementioned experimental settings involved artificial sinusoidally modulated stimuli. Yet, natural sounds, speech in particular, present temporal modulations that are not strictly periodic. On the other

hand, experimental research on speech has focused on the processing of individually presented speech sounds, such as vowels, syllables, or single words. The related findings underpin most current models of speech perception. However, in natural connected speech, speech information is embedded in a continuous acoustic flow, and sentences are not "pre-segmented" in perceptual units of analysis. Recent work on sentence-level stimuli (i.e., materials with a duration exceeding 1–2 seconds), using experimental tasks as intelligibility, demonstrate the fundamental importance of long-term temporal parameters of the acoustic signal. Online segmentation remains a major challenge to contemporary models of speech perception as well as automatic speech recognition.

Interestingly, a large body of psychophysical 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). Fascinating findings emerged from that work, emphasizing the role of signal-to-noise ratio in speech comprehension, but perhaps the most interesting feature is that connected speech has principled and useful temporal properties that may play a key role in the problem of speech parsing and decoding. Natural speech usually comes to the listener as a continuous stream and needs to be analyzed online and decoded by mechanisms that are unlikely to be continuous (Giraud and Poeppel, 2012a). The parsing mechanism corresponds to the discretization of the continuous input signal into subsegments of speech information that are read out, to a certain extent, independently from each other. The notion that perception is discrete has been extensively discussed and generalized in numerous sensory modalities and contexts (Pöppel, 1988; VanRullen and Koch, 2003; VanRullen et al., 2014). Here we discuss the hypothesis that neural oscillations constitute a possible mechanism for discretizing temporally complex sounds such as speech (Giraud and Poeppel, 2012a).

Analysis at multiple timescales

Speech is a multiplexed signal, that is, it interlinks several levels of complexity, and organizational principles and perceptual units of analysis exist at distinct timescales. Using data from linguistics, psychophysics, and physiology, Poeppel and colleagues 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). This assumption adds to the notion of reverse hierarchy (Hochstein and Ahissar, 2002; Nahum et al., 2008) and other hierarchic models in perception, which propose that the hierarchic complexification of sensory information (e.g., the temporal hierarchy) maps on to the anatomofunctional hierarchy of the brain (Giraud et al., 2000; Kiebel et al., 2008). The motivations for extending such a hypothesis are twofold. First, a single, short temporal window that forms the basis for hierarchic 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 biologic systems, and the neuronal infrastructure exists to support multiscale computation (Canolty and 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' "landmarks" hypothesis: Stevens, 2002), and that subsequently refines segmental analysis. Here, we propose that segmental and suprasegmental analyses could be carried out concurrently and "packaged" for parsing and decoding by neuronal oscillations at different rates.

The notion that speech analysis occurs in parallel at multiple timescales justifies moving away from strictly hierarchic models of speech perception (e.g., Giraud and Price, 2001). Accordingly, the simultaneous extraction of different acoustic cues permits simultaneous high-order processing of different information from a unique 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. 5.2) that the speech signal contains events of different durations: short energy bursts and formant transitions occur within a 20–80-ms timescale, whereas syllabic information occurs over 150–300 ms. The processing of both types of events could be accounted for either by a hierarchic 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 certain 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 and Perrott, 1999; Greenberg and Arai, 2001). Because the correct extraction of short units is not a prerequisite

for comprehension, this rules out the notion that speech processing relies solely on hierarchic processing of short and then larger units. 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, non-speech 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., 2007). By comparison, 200 ms stimulus duration underlies loudness judgments. In sum, physiologic events at related scales form the basis for processing at that level. Therefore, the neuronal oscillatory machinery (together with motor constraints related to speech production; Morillon et al., 2010) presumably imposed strong temporal constraints that might have shaped the size of acoustic features selected to carry speech information. This is consistent with the notion that perception is discrete and that the exogenous recruitment of neuronal populations is followed by refractory periods that temporarily reduce the ability to optimally extract sensory information (Ghitza and Greenberg, 2009; Ghitza, 2011). According to this hypothesis, the temporally limited capacity of gamma oscillations to integrate information over time possibly imposes a lower limit to the phoneme length. This also suggests that oscillatory constraints in the delta-theta range possibly constrained the size of syllables to be roughly the size of a delta-theta cycle. Considering that the average length of phoneme and syllable is about 25–80 ms and 150–300 ms respectively (Figs. 5.1 and 5.2), the dual timescale segmentation requires two parallel 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).

Neural oscillations as endogenous temporal constraints

Neural oscillations correspond to synchronous activity of neuronal assemblies that are both intrinsically coupled and coupled by a common input. It was proposed that these oscillations reflect modulations of neuronal excitability that temporally constrain the sampling of sensory information (Schroeder and Lakatos, 2009a). The intriguing correspondence between the size of certain speech temporal units and the frequency of oscillations in certain frequency bands (Fig. 5.1) has elicited the intuition that they might play a functional role in sensory sampling (see below). Oscillations are evidenced by means of a spectrotemporal analysis of electrophysiologic recordings (see Wang, 2010, for a review). The requirements for measuring oscillations and spiking activity are different. The presentation of an exogenous stimulus typically results in an increase of spiking activity in those brain areas that are functionally sensitive to such inputs. Neural oscillations, on the other hand, can be observed in local field potential recordings in the absence of any external stimulation. Exogenous stimulation however typically modulates oscillatory activity, resulting either in a reset of their phase and/or a change (increase or decrease) in the magnitude of these oscillations (Howard and Poeppel, 2012).

Cortical oscillations are proposed to shape spike-timing dynamics and to impose phases of high and low neuronal excitability (Britvina and Eggermont, 2007; Schroeder and Lakatos, 2009a, b; Panzeri 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 specific phases (i.e., the trough) of oscillatory activity (Womelsdorf et al., 2007). 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). While both measures are relevant to address how sensory information is encoded in the brain, we believe that the ability of neural oscillations to temporally organize spiking activity supports the functional relevance of neural oscillations to solve the discretization problem and to permit the integration of complex sensory signals across time.

Neuronal oscillations are ubiquitous in the brain, but they vary in strength and frequency depending on their location and the exact nature of their neuronal generators (Mantini et al., 2007; Hyafil et al., 2012). The notion that neural oscillations shape the way the brain processes sensory information is supported by a wealth of electrophysiologic findings in humans and animals. On the one hand, stimuli that occur in the ideal excitability phase of slow oscillations (<12 Hz) are processed faster and with a higher accuracy (Lakatos et al., 2008; Busch et al., 2009; Henry and Obleser, 2012; Ng et al., 2012; Wyart et al., 2012). On the other hand, gamma-band 40-Hz activity (low gamma band) can be observed at rest in both monkey (Fukushima et al., 2012) and human auditory cortex. In humans, it can be measured using EEG, MEG, and with a more precise localization with concurrent EEG and functional magnetic resonance imaging (Morillon et al., 2010) and intracranial electroencephalographic recordings (stereotactic EEG (sEEG), Electrocorticography (EcoG)) in patients. Neural oscillations in this range are endogenous in the sense that one can observe a spontaneous spike clustering at approximately 40 Hz even in the absence of external stimulation. This gamma activity is thought to be generated by a

"ping-pong" interaction between pyramidal cells and inhibitory interneurons (Borgers et al., 2005, 2008), or even just among interneurons that are located in superficial cortical layers (Tiesinga and Sejnowski, 2009). Exogenous inputs usually increase gamma-band activity in sensory areas, presumably clustering spiking activity that is propagated to higher hierarchic processing stages (Arnal et al., 2011; Arnal and Giraud, 2012; Bastos et al., 2012). By analogy with the proposal of Elhilali et al. (2004) that slow responses gate faster ones, it is interesting to envisage this periodic modulation of spiking by oscillatory activity as an endogenous mechanism to optimize the extraction of relevant sensory input in time. Such integration could occur under the patterning of slower oscillations in the delta-theta range.

Alignment of neuronal excitability with speech timescales

Experimental exploration of how speech parsing and encoding is carried out by the brain is non-trivial. One approach has been to explore how neural responses can discriminate different sentences, assuming that the features of neural signals that are sensitive to such differences (e.g., frequency band, amplitude, phase) should reveal the features that are key to sentence decoding. Using this approach, it was shown that the phase of theta-band neural activity reliably discriminates different sentences (Luo and Poeppel, 2007). Specifically, when one sentence is repeatedly presented to listeners, the phase of ongoing theta-band activity follows a consistent phase sequence. When different sentences are played, however, different phase sequences are observed. Since theta-band (4–8 Hz) falls around the mean syllabic rate of speech (~5 Hz), the phase of theta-band activity likely tracks syllabic-level features of speech (Giraud and Poeppel, 2012a; Hyafil et al., 2012; Edwards and Chang, 2013). These findings support the notion that the syllabic timescale has adapted to a pre-existing cortical preference for temporal information in this frequency range. At this point, however, it is not clear whether the phase locking between the speech input and neural oscillations is necessary for speech intelligibility. On the one hand, sentences played backward (and therefore unintelligible) can similarly be discriminated on the basis of their phase course, which tempers the interpretation that these oscillations play a causal role in speech perception (Howard and Poeppel, 2011). On the other hand, two recent studies using distinct ways of acoustically degrading speech intelligibility demonstrate that the temporal alignment between the stimulus and delta-theta band responses is higher when the stimulus is intelligible (Peelle et al., 2013; Doelling et al., 2014). This, again, supports the notion that those neural oscillations that match the slow (syllabic) speech timescales are useful (if not necessary) for the extraction of relevant speech information.

Neural oscillatory responses can also be entrained at much higher rates in the middle to high (40—200 Hz) gamma band (Fishman et al., 2000; Brugge et al., 2009). This could suggest that faster speech segments such as phonemic transitions could be extracted using the same encoding principle. High gamma responses in early auditory regions (Ahissar et al., 2001; Nourski et al., 2009; Mesgarani and Chang, 2012; Morillon et al., 2012) reflect the fast temporal fluctuations in the speech envelope. A recent EcoG study succeeded at reconstructing the original speech input using a combination of linear and non-linear methods to decode neural responses from high gamma activity in auditory cortical regions (Pasley et al., 2012). Therefore, the decoding of auditory activity on a large spatial scale (at the population level) demonstrates that the auditory cortex maintains a high-fidelity representation of temporal modulations up to 200 Hz. However, according to psychophysiologic findings described earlier, speech intelligibility mostly relies on the preservation of the low-frequency (<50 Hz) temporal fluctuations rather than on higher-frequency information. Therefore, whether it is necessary to maintain a representation of such acoustic features to correctly perceive speech remains unclear. The following section aims at clarifying the putative neural mechanisms underpinning the segmentation and the integration of auditory speech signals into an intelligible percept.

Parallel processing at multiple timescales

Schroeder and Lakatos (2009a, b) have argued that oscillations correspond to the alternation of phases of high and low neuronal excitability, which temporally constrain sensory processing. This means that gamma oscillations, which have a period of approximately 25 ms, provide a 10–15-ms window for integrating spectrotemporal information (low spiking rate) followed by a 10–15-ms window for propagating the output (high spiking rate; see, for illustration, Fig. 5.4A). However, because the average length of a phoneme is about 50 ms, a 10–15-ms window might be too short for integrating this information. Using a computational model of gamma oscillations generated by a pyramidal interneuron network (PING model: Borgers et al., 2005; Shamir et al., 2009) shows that the shape of a sawtooth input signal designed to have the typical duration and AM 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. Such a code has the capacity required to distinguish different shapes of the stimulus and is therefore a plausible means

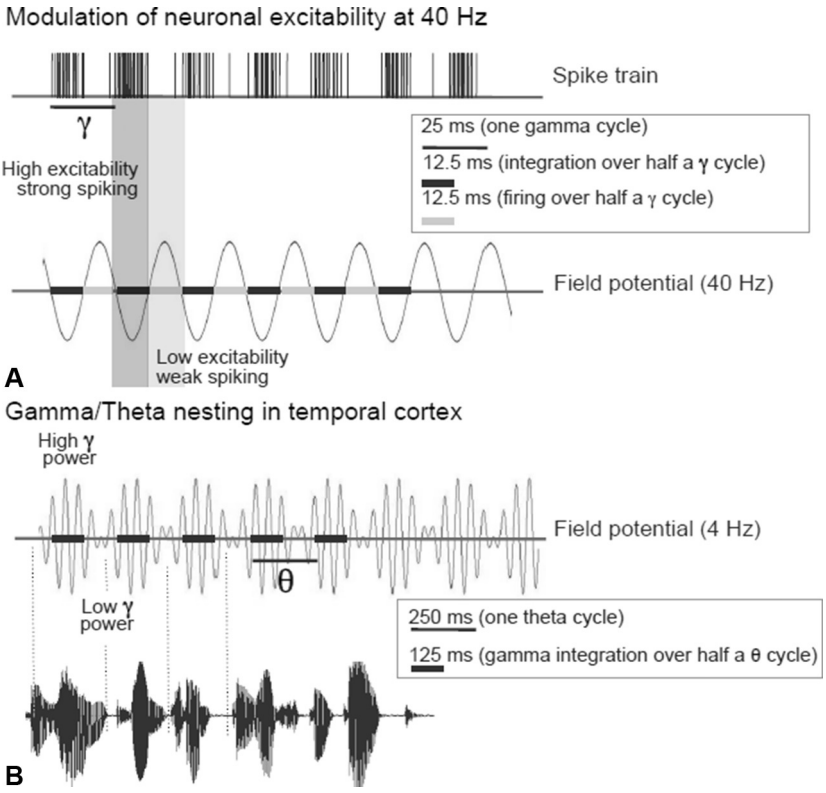

Fig. 5.4. The temporal relationship between speech and brain oscillations. (**A**) Gamma oscillations periodically modulate neuronal excitability and spiking. The hypothesized mechanism is that neurons fire for about 12.5 ms and integrate for the rest of the 25-ms time window. Note that these values 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 the theta rhythm (about 4 Hz). Theta rhythm is reset by speech, resulting in maintaining the alignment between brain rhythms and speech bursts.

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.

In a recent extension of this model, the parsing and encoding capacity of coupled theta and gamma oscillating modules was studied (Hyafil et al., 2012). In combination, these modules succeed in signaling syllable boundaries and to orchestrate spiking within syllabic windows, so that online speech decoding becomes possible with a similar accuracy as experimental findings using intracortical recordings in monkeys (Kayser et al., 2012).

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. In the absence of 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 (Fig. 5.4B; Schroeder et al., 2008). This cross-frequency relationship is referred to as "nesting." Electrophysiologic recordings suggest that theta oscillations can be phase-reset by several means, through multimodal corticocortical pathways (Lakatos et al., 2007; Arnal et al., 2009; Thorne et al., 2011) or through predictive top-down modulations, but most probably by the stimulus onset itself (Fig. 5.4B). 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. 5.4B). 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, while comprehension was drastically reduced by time-compressing speech signals by a factor of 3, comprehension was restored by artificially inserting periods of silence. The mere fact of restoring "syllabicity" by adding silent periods to speech improves performance, even though the speech segments that remained available are still compressed. Optimal performance is obtained when 80-ms silent periods alternate with 40-ms time-compressed speech signals. These time constants allowed the authors to propose a phenomenologic model involving three nested rhythms in the theta (5 Hz), beta, or low gamma (20–40 Hz) and gamma (80 Hz) domains (for an extended discussion, see Ghitza, 2011).

Parallel processing in bilateral auditory cortices

There is emerging consensus, based on neuropsychologic and imaging data, that speech perception is mediated bilaterally. 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 non-speech 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 are asymmetric distributions of neuronal ensembles between hemispheres with preferred shorter versus longer integration constants; these cell groups "sample" the input with different sampling integration constants (Fig. 5.5A). Specifically, left auditory cortex has a relatively higher proportion of short-term (gamma) integrating cell groups, whereas right auditory cortex has a larger long-term (theta) integrating proportion (Fig. 5.5B). As a consequence, left-hemisphere auditory cortex is likely better

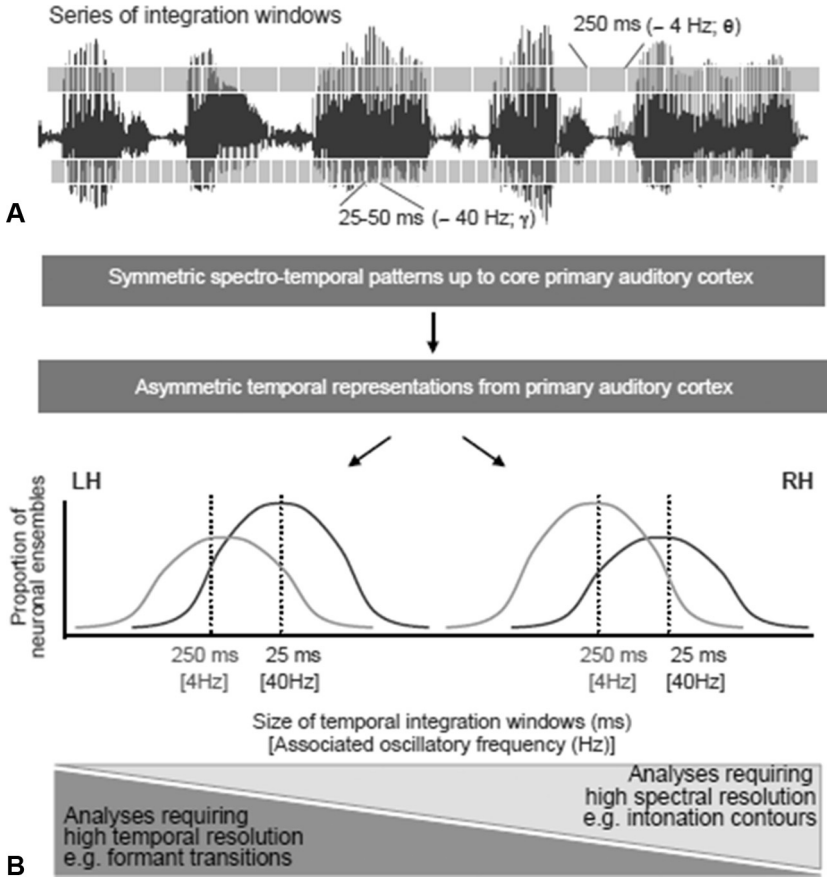

Fig. 5.5. The asymmetric sampling in time hypothesis. (**A**) Temporal relationship between the speech waveform and the two proposed integration timescales (in ms) and associated brain rhythms (in Hz). (**B**) Proposed mechanisms for asymmetric speech parsing: left auditory cortex (LH) contains a larger proportion of neurons able to oscillate at gamma frequency than the right one (RH). (From Giraud and Poeppel, 2012b, with permission from Springer Science and Business Media.)

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 summarized in Figure 5.5. It 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 and Gandour, 2008), which originally proposed that, whereas the left auditory cortex is better suited to process temporal information, the right auditory cortex is better at processing spectral information. While many psychophysics and neurophysiologic experiments seem to support this idea (see Poeppel, 2003; Poeppel et al., 2008 and Giraud and Poeppel, 2012a for reviews on the topic) there is a lot of work in progress regarding this unresolved question.

Dysfunctional oscillatory sampling

Additional evidence to support the notion that neural oscillations play an instrumental role in speech processing would be to show that dysfunctional oscillatory mechanisms result in speech-processing impairments. Dyslexia, which is a phonologic deficit, i.e., a deficit in processing speech sounds, presumably constitutes a good candidate to test this hypothesis. Temporal sampling mediated by cortical oscillations has recently been proposed to be a central mechanism in several aspects of dyslexia (Goswami, 2011). This proposal suggests that a deficit involving theta oscillations might impair the tracking of low temporal modulations in the syllabic range. In a complementary way, it was proposed recently that gamma oscillations might play a role in yielding an auditory phonemic deficit.

Interestingly, at around 30 Hz, the left-dominant phase-locking profile of auditory responses in MEG (auditory steady-state responses) was only present in subjects with normal reading ability (Lehongre et al., 2011). Because this response is absent in dyslexic participants, the authors suggested that the ability of their left auditory cortex to parse speech at the appropriate phonemic rate was altered. Those with dyslexia had a strong response at this frequency in right auditory cortex and therefore presented an abnormal asymmetry between left and right auditory cortices. Importantly, the magnitude of the anomalous asymmetry correlated with behavioral measures in phonology (such as non-word repetition and rapid automatic naming). Finally it was also shown that dyslexic readers had a stronger resonance than controls in both left and right auditory cortices at frequencies between 50 and 80 Hz. This supports the notion that these subjects had a tendency to oversample information in the phonemic range, this latter effect being positively correlated with a phonologic memory deficit. As a consequence, if dyslexia induces speech parsing at a wrong frequency, phonemic units would be sampled erratically, without necessarily inducing major perceptual deficits (Ramus and Szenkovits, 2008; Ziegler et al., 2009). As a consequence, the phonologic impairment could take different forms, with a stronger impact on the acoustic side for undersampling (insufficient acoustic detail per time unit) and on the memory side for oversampling (too many frames to be integrated per time unit).

Although important, the observation that oscillatory anomalies co-occur with atypical phonologic representations remains insufficient to establish a causal role of dysfunctional oscillatory sampling. Causal evidence that auditory sampling is determined by cortical columnar organization could be obtained from knockout animal models comparing neuronal activity to continuous auditory stimuli in sites with various degrees of columnar disorganization. However, such animal work can only indirectly address a specific relation to speech processing.

CONCLUSION

Time is an essential feature of speech perception. No speech sound can be identified without integrating the acoustic input over time, and the temporal scale at which such integration operates determines whether we are hearing phonemes, syllables, or words. The central idea of this chapter is that, unlike subcortical processing that faithfully encodes speech sounds in their precise spectrotemporal structure, processing in primary and association auditory cortices results in the discretization of spectrotemporal patterns, using variable temporal integration scales. By analogy with Heisenberg's uncertainty principle (Ozawa, 2003), speech representations cannot be precise in both time and space. The limited phase-locking capacity of the auditory cortex thus appears a likely counterpart to its spatial integration properties (across cortical layers

and functional regions). Speech processing through and across cortical columns containing complex recurrent circuits bears a cost on the temporal precision of speech representations, and integration at gamma scale could be a direct consequence of processing at the cortical column scale. In this chapter we argue that the auditory cortex uses gamma oscillations to integrate the speech auditory stream at the phonemic timescale, and theta oscillations to signal syllable boundaries and orchestrate gamma activity. Although the generation mechanisms are less well known for theta than for gamma oscillations, at present we see no alternative computational solution to the online speech segmentation and integration problem than invoking coupled theta and gamma activity. More research is needed to evaluate the detailed neural operations that are necessary to transform the acoustic input into linguistic representations, and it might turn out that non-oscillatory mechanisms succeed more efficiently in achieving these transformations.

References

Ahissar E, Nagarajan S, Ahissar M et al. (2001). Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. Proc Natl Acad Sci U S A 98: 13367–13372.

Allen JB (2005). Articulation and intelligibility. Synthesis Lectures on Speech and Audio Processing 1 (1): 1–124.

Arnal LH, Giraud AL (2012). Cortical oscillations and sensory predictions. Trends Cogn Sci 16: 390–398.

Arnal LH, Morillon B, Kell CA et al. (2009). Dual neural routing of visual facilitation in speech processing. J Neurosci 29: 13445–13453.

Arnal LH, Wyart V, Giraud AL (2011). Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. Nat Neurosci 14: 797–801.

Barton B, Venezia JH, Saberi K et al. (2012). Orthogonal acoustic dimensions define auditory field maps in human cortex. Proc Natl Acad Sci U S A 109: 20738–20743.

Bastos AM, Usrey WM, Adams RA et al. (2012). Canonical microcircuits for predictive coding. Neuron 76: 695–711.

Boemio A, Fromm S, Braun A et al. (2005). Hierarchical and asymmetric temporal sensitivity in human auditory cortices. Nat Neurosci 8: 389–395.

Borgers C, Epstein S, Kopell NJ (2005). Background gamma rhythmicity and attention in cortical local circuits: a computational study. Proc Natl Acad Sci U S A 102: 7002–7007.

Borgers C, Epstein S, Kopell NJ (2008). Gamma oscillations mediate stimulus competition and attentional selection in a cortical network model. Proc Natl Acad Sci U S A 105: 18023–18028.

Britvina T, Eggermont JJ (2007). A Markov model for interspike interval distributions of auditory cortical neurons that do not show periodic firings. Biol Cybern 96: 245–264.

Brugge JF, Nourski KV, Oya H et al. (2009). Coding of repetitive transients by auditory cortex on Heschl's gyrus. J Neurophysiol 102: 2358–2374.

Busch NA, Dubois J, VanRullen R (2009). The phase of ongoing EEG oscillations predicts visual perception. J Neurosci 29: 7869–7876.

Canolty RT, Knight RT (2010). The functional role of cross-frequency coupling. Trends Cogn Sci 14: 506–515.

Ding N, Simon JZ (2009). Neural representations of complex temporal modulations in the human auditory cortex. J Neurophysiol 102: 2731–2743.

Doelling KB, Arnal LH, Ghitza O et al. (2014). Acoustic landmarks drive delta-theta oscillations to enable speech comprehension by facilitating perceptual parsing. Neuroimage 85: 761–768.

Drullman R, Festen JM, Plomp R (1994a). Effect of reducing slow temporal modulations on speech reception. J Acoust Soc Am 95: 2670–2680.

Drullman R, Festen JM, Plomp R (1994b). Effect of temporal envelope smearing on speech reception. J Acoust Soc Am 95: 1053–1064.

Edwards E, Chang EF (2013). Syllabic (approximately 2–5 Hz) and fluctuation (approximately 1–10 Hz) ranges in speech and auditory processing. Hear Res 305: 113–134.

Elhilali M, Fritz JB, Klein DJ et al. (2004). Dynamics of precise spike timing in primary auditory cortex. J Neurosci 24: 1159–1172.

Elliott TM, Theunissen FE (2009). The modulation transfer function for speech intelligibility. PLoS Comput Biol 5: e1000302.

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. J Acoust Soc Am 108: 1877–1887.

Fishman YI, Reser DH, Arezzo JC et al. (2000). Complex tone processing in primary auditory cortex of the awake monkey. II. Pitch versus critical band representation. J Acoust Soc Am 108: 247–262.

Fukushima M, Saunders RC, Leopold DA et al. (2012). Spontaneous high-gamma band activity reflects functional organization of auditory cortex in the awake macaque. Neuron 74: 899–910.

Gaese BH, Ostwald J (1995). Temporal coding of amplitude and frequency modulation in the rat auditory cortex. Eur J Neurosci 7: 438–450.

Ghitza O (2011). Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input rhythm. Front Psychol 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: 113–126.

Giraud AL, Poeppel D (2012a). Cortical oscillations and speech processing: emerging computational principles. Nat Neurosci 15: 511–517.

Giraud AL, Poeppel D (2012b). Speech perception from a neurophysiological perspective. In: D Poeppel, T Overath, AN Popper et al. (Eds.), The Human Auditory Cortex, Springer, New York, pp. 225–260.

Giraud AL, Price CJ (2001). The constraints functional neuroimaging places on classical models of auditory word processing. J Cogn Neurosci 13: 754–765.

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

Giraud AL, Kell C, Thierfelder C et al. (2004). Contributions of sensory input, auditory search and verbal comprehension to cortical activity during speech processing. Cereb Cortex 14: 247–255.

Goswami U (2011). A temporal sampling framework for developmental dyslexia. Trends Cogn Sci 15: 3–10.

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

Grothe B, Pecka M, McAlpine D (2010). Mechanisms of sound localization in mammals. Physiol Rev 90: 983–1012.

Henry MJ, Obleser J (2012). Frequency modulation entrains slow neural oscillations and optimizes human listening behavior. Proc Natl Acad Sci U S A 109: 20095–20100.

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

Holcombe AO (2009). Seeing slow and seeing fast: two limits on perception. Trends Cogn Sci 13: 216–221.

Howard MF, Poeppel D (2011). Discrimination of speech stimuli based on neuronal response phase patterns depends on acoustics but not comprehension. J Neurophysiol 104: 2500–2511.

Howard MF, Poeppel D (2012). The neuromagnetic response to spoken sentences: Co-modulation of theta band amplitude and phase. Neuroimage 60: 2118–2127.

Hyafil A, Fontolan L, Gutkin B et al. (2012). Theoretical exploration of speech/neural oscillation alignment for speech parsing. FENS Abstract 6, S47.04.

Joris PX, Schreiner CE, Rees A (2004). Neural processing of amplitude-modulated sounds. Physiol Rev 84: 541–577.

Kanedera N, Arai T, Hermansky H et al. (1999). On the relative importance of various components of the modulation spectrum for automatic speech recognition. Speech Comm 28: 43–55.

Kayser C, Ince RA, Panzeri S (2012). Analysis of slow (theta) oscillations as a potential temporal reference frame for information coding in sensory cortices. PLoS Comput Biol 8: e1002717.

Kiebel SJ, Daunizeau J, Friston KJ (2008). A hierarchy of time-scales and the brain. PLoS Comput Biol 4: e1000209.

Kingsbury BED, Morgan N, Greenberg S (1998). Robust speech recognition using the modulation spectrogram. Speech Comm 25: 117–132.

Lakatos P, Chen CM, O'Connell MN et al. (2007). Neuronal oscillations and multisensory interaction in primary auditory cortex. Neuron 53: 279–292.

Lakatos P, Karmos G, Mehta AD et al. (2008). Entrainment of neuronal oscillations as a mechanism of attentional selection. Science 320: 110–113.

Lehongre K, Ramus F, Villiermet N et al. (2011). Altered low-gamma sampling in auditory cortex accounts for the three main facets of dyslexia. Neuron 72: 1080–1090.

Liang L, Lu T, Wang X (2002). Neural representations of sinusoidal amplitude and frequency modulations in the primary auditory cortex of awake primates. J Neurophysiol 87: 2237–2261.

Loebach JL, Wickesberg RE (2008). The psychoacoustics of noise vocoded speech: a physiological means to a perceptual end. Hear Res 241: 87–96.

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

Luo H, Wang Y, Poeppel D et al. (2006). Concurrent encoding of frequency and amplitude modulation in human auditory cortex: MEG evidence. J Neurophysiol 96: 2712–2723.

Luo H, Boemio A, Gordon M et al. (2007). The perception of FM sweeps by Chinese and English listeners. Hear Res 224: 75–83.

Malone BJ, Scott BH, Semple MN (2010). Temporal codes for amplitude contrast in auditory cortex. J Neurosci 30: 767–784.

Mantini D, Perrucci MG, Del Gratta C et al. (2007). Electrophysiological signatures of resting state networks in the human brain. Proc Natl Acad Sci U S A 104: 13170–13175.

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

Miller GA (1951). Language and communication, McGraw-Hill, New York.

Miyazaki T, Thompson J, Fujioka T et al. (2013). Sound envelope encoding in the auditory cortex revealed by neuromagnetic responses in the theta to gamma frequency bands. Brain Res 1506: 64–75.

Moerel M, De Martino F, Santoro R et al. (2013). Processing of natural sounds: characterization of multipeak spectral tuning in human auditory cortex. J Neurosci 33: 11888–11898.

Morillon B, Lehongre K, Frackowiak RS et al. (2010). Neurophysiological origin of human brain asymmetry for speech and language. Proc Natl Acad Sci U S A 107: 18688–18693.

Morillon B, Liegeois-Chauvel C, Arnal LH et al. (2012). Asymmetric function of theta and gamma activity in syllable processing: an intra-cortical study. Front Psychol 3: 248.

Nahum M, Nelken I, Ahissar M (2008). Low-level information and high-level perception: the case of speech in noise. PLoS Biol 6: e126.

Ng BS, Schroeder T, Kayser C (2012). A precluding but not ensuring role of entrained low-frequency oscillations for auditory perception. J Neurosci 32: 12268–12276.

Nourski KV, Brugge JF (2011). Representation of temporal sound features in the human auditory cortex. Rev Neurosci 22: 187–203.

Nourski KV, Reale RA, Oya H et al. (2009). Temporal envelope of time-compressed speech represented in the human auditory cortex. J Neurosci 29: 15564–15574.

Ozawa M (2003). Universally valid reformulation of the Heisenberg uncertainty principle on noise and disturbance in measurement. Phys Rev A 67: 042105.

Panzeri S, Brunel N, Logothetis NK et al. (2010). Sensory neural codes using multiplexed temporal scales. Trends Neurosci 33: 111–120.

Pasley BN, David SV, Mesgarani N et al. (2012). Reconstructing speech from human auditory cortex. PLoS Biol. http://dx.doi.org/10.1371/journal.pbio.1001251.

Peelle JE, Gross J, Davis MH (2013). Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. Cereb Cortex 23: 1378–1387.

Pienkowski M, Eggermont JJ (2009). Nonlinear cross-frequency interactions in primary auditory cortex spectro-temporal receptive fields: a Wiener-Volterra analysis. J Comput Neurosci 28: 285–303.

Plack CJ, Oxenham AJ, Fay RR et al. (2005). Pitch: neural coding and perception, Springer, New York.

Poeppel D (2001). New approaches to the neural basis of speech sound processing: introduction to special section on brain and speech. Cognit Sci 25: 659–661.

Poeppel D (2003). The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time'. Speech Comm 41: 245–255.

Poeppel D, Idsardi WJ, van Wassenhove V (2008). Speech perception at the interface of neurobiology and linguistics. Philos Trans R Soc Lond B Biol Sci 363: 1071–1086.

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

Ramus F, Szenkovits G (2008). What phonological deficit? Q J Exp Psychol 61: 129–141.

Remez RE, Rubin PE, Pisoni DB et al. (1981). Speech perception without traditional speech cues. Science 212: 947–949.

Roberts B, Summers RJ, Bailey PJ (2011). The intelligibility of noise-vocoded speech: spectral information available from across-channel comparison of amplitude envelopes. Proc Biol Sci 278: 1595–1600.

Rosen S (1992). Temporal information in speech: acoustic, auditory and linguistic aspects. Philos Trans R Soc Lond B Biol Sci 336: 367–373.

Saberi K, Perrott DR (1999). Cognitive restoration of reversed speech. Nature 398: 760.

Saenz M, Langers DR (2014). Tonotopic mapping of human auditory cortex. Hear Res 307: 42–52.

Schonwiesner M, Zatorre RJ (2009). Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. Proc Natl Acad Sci U S A 106: 14611–14616.

Schroeder CE, Lakatos P (2009a). Low-frequency neuronal oscillations as instruments of sensory selection. Trends Neurosci 32: 9–18.

Schroeder CE, Lakatos P (2009b). The gamma oscillation: master or slave? Brain Topogr 22: 24–26.

Schroeder CE, Lakatos P, Kajikawa Y et al. (2008). Neuronal oscillations and visual amplification of speech. Trends Cogn Sci 12: 106–113.

Scott SK, Rosen S, Lang H et al. (2006). Neural correlates of intelligibility in speech investigated with noise vocoded speech – a positron emission tomography study. J Acoust Soc Am 120: 1075–1083.

Shamir M, Ghitza O, Epstein S et al. (2009). Representation of time-varying stimuli by a network exhibiting oscillations on a faster time scale. PLoS Comput Biol 5: e1000370.

Shannon RV, Zeng FG, Kamath V et al. (1995). Speech recognition with primarily temporal cues. Science 270: 303–304.

Sharpee TO, Atencio CA, Schreiner CE (2011). Hierarchical representations in the auditory cortex. Curr Opin Neurobiol 21: 761–767.

Smith ZM, Delgutte B, Oxenham AJ (2002). Chimaeric sounds reveal dichotomies in auditory perception. Nature 416: 87–90.

Souza P, Rosen S (2009). Effects of envelope bandwidth on the intelligibility of sine- and noise-vocoded speech. J Acoust Soc Am 126: 792–805.

Steeneken HJ, Houtgast T (1980). A physical method for measuring speech-transmission quality. J Acoust Soc Am 67: 318–326.

Stevens KN (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. J Acoust Soc Am 111: 1872–1891.

Thorne JD, De Vos M, Viola FC et al. (2011). Cross-modal phase reset predicts auditory task performance in humans. J Neurosci 31: 3853–3861.

Tiesinga P, Sejnowski TJ (2009). Cortical enlightenment: are attentional gamma oscillations driven by ING or PING? Neuron 63: 727–732.

VanRullen R, Koch C (2003). Is perception discrete or continuous? Trends Cogn Sci 7: 207–213.

VanRullen R, Zoefel B, Ilhan B (2014). On the cyclic nature of perception in vision versus audition. Philos Trans R Soc Lond B Biol Sci 369: 20130214.

Wang X (2007). Neural coding strategies in auditory cortex. Hear Res 229: 81–93.

Wang XJ (2010). Neurophysiological and computational principles of cortical rhythms in cognition. Physiol Rev 90: 1195–1268.

Wang X, Lu T, Bendor D et al. (2008). Neural coding of temporal information in auditory thalamus and cortex. Neuroscience 157: 484–494.

Warrier C, Wong P, Penhune V et al. (2009). Relating structure to function: Heschl's gyrus and acoustic processing. J Neurosci 29: 61–69.

Womelsdorf T, Schoffelen JM, Oostenveld R et al. (2007). Modulation of neuronal interactions through neuronal synchronization. Science 316: 1609–1612.

Wyart V, de Gardelle V, Scholl J et al. (2012). Rhythmic fluctuations in evidence accumulation during decision making in the human brain. Neuron 76: 847–858.

Yin P, Johnson JS, O'Connor KN et al. (2011). Coding of amplitude modulation in primary auditory cortex. J Neurophysiol 105: 582–600.

Zatorre RJ, Gandour JT (2008). Neural specializations for speech and pitch: moving beyond the dichotomies. Philos Trans R Soc Lond B Biol Sci 363: 1087–1104.

Zatorre RJ, Belin P, Penhune VB (2002). Structure and function of auditory cortex: music and speech. Trends Cogn Sci 6: 37–46.

Zeng FG (2002). Temporal pitch in electric hearing. Hear Res 174: 101–106.

Ziegler JC, Pech-Georgel C, George F et al. (2009). Speech-perception-in-noise deficits in dyslexia. Dev Sci 12: 732–745.