6 Speech Rhythm in the Perception–Action Cycle

LAURA VERGA, SONJA A. KOTZ, AND MICHAEL SCHWARTZE

6.1 Speech and the Temporal Coordination of Behavior

The main purpose of language is to exchange information in communication with others. In the case of spoken language, all acoustic information is conveyed through the spectrally and temporally fluctuating patterns of the sound signal (Martin, 1972; Rosen, 1992). These evolve over time and into dialogue in an alternation of speaking and listening. In this rhythmic aspect, speech is ultimately a form of adaptive sensorimotor behavior that is shaped by our perpetual interaction with the environment. This implies that we need to be able to flexibly adjust behavior due to rapidly changing circumstances, goals, and demands. The successful use of speech therefore not only requires situational balancing of goals and demands but also adequate use of limited neural and cognitive resources. This necessitates precise temporal coordination of dynamics as diverse as turn-taking, signal encoding and decoding, interfacing with short-term and long-term memory, allocating attention, and combining temporally distinct speech elements to optimize communication. Yet, how temporal coordination is achieved within and between speakers and listeners remains a challenging question, as perhaps best expressed in Lashley's original reflection on the problem of serial order in behavior (Lashley, 1951).

Some acoustic signal dynamics have been shown to directly map onto neurophysiological signatures in speech processing. Compelling evidence showed that neural oscillatory activity mirrors rhythmic features of the speech signal (Giraud et al., 2007; Ding et al., 2016; see Chapters 3 and 5). The links between the acoustic signal and brain rhythms imply that speech is at least in parts both a reflection and a driver of neural signaling and cognitive mechanisms such as attention allocation that themselves are not restricted to speech. However, so far, it is still mostly unclear how acoustic rhythms interact with neural rhythms and the dynamic allocation of memory and attention or across domains, as indicated by visual enhancement effects (Schroeder and Lakatos, 2008; Schroeder et al., 2008). A better understanding of how these interactions shape and are themselves shaped by speech rhythms would therefore not only improve any explanation of human speech capacities but could also inform new perspectives on

https://doi.org/10.1017/9781009295888.008 Published online by Cambridge University Press

biological and cognitive aspects of ontogenetic and phylogenetic speech development. This extended approach might provide explanations on how successful temporal coordination emerges from the "composite of many interactive systems" that make up cerebral functions, thus delivering a crucial building block to answering the question of how the human brain factors time into the "problem of serial order in behavior" (Lashley, 1951, p. 135).

6.2 Linking Production and Perception Cycles to Temporal Processing

Considering the complex rhythmic structure of speech, one basic domain-general mechanism that might interleave with speech processing is temporal processing, that is, a mechanism underlying the encoding, decoding, and use of temporal information in motor and nonmotor behavior (Buhusi and Meck, 2005; Ivry and Schlerf, 2008). Other possible interactions between "systems" emerge from the need to rapidly activate and flexibly allocate attention and memory resources to particular speech events as they unfold in time (Large and Jones, 1999; Schroeder and Lakatos, 2008). Here, speaking and listening to speech share a common denominator in the encoding and decoding of temporal structure, the dynamic allocation of attention and memory, and the temporal coordination of these processes with the overarching goal to ensure stable speech processing and consequently successful communication.

Temporal processing has been attributed to a distributed brain network (Ivry and Schlerf, 2008; Merchant et al., 2013; Wiener, 2024). The main nodes of this network are the basal ganglia, the supplementary motor area (SMA) and prefrontal cortical areas, and the cerebellum, each of which seems to support specific aspects of temporal processing. Current views on these classically denoted motor areas extend to perception as well (e.g., Petacchi et al., 2005; Chen et al., 2008; Baumann et al., 2015), and thus questions emerged as to if, how, and why this potentially domain-general temporal processing network might interface with speech production and perception (Schirmer, 2004; Buhusi and Meck, 2005; Kotz and Schwartze, 2010).

Answering these questions relates back to the notion of speech communication as a form of adaptive behavior. The intricate interplay of production and perception in conversation can be subsumed but also further differentiated considering the organizational structure of the perception–action (PA) cycle framework (Fuster, 1995, 2004; Fuster and Bressler, 2015). The PA cycle is a neurofunctional implementation of the fundamental principle of biological adaptation, expressed in the notion that environmental events change the activity of neural receptors to generate adaptive motor behavior, which, in turn, generates environmental events that induce further changes and thus give rise to a self-contained periodic cycle or function circle (Uexküll, 1926, p. 126; Fuster and Bressler, 2015). The PA cycle framework suggests that the development of the human prefrontal cortex (PFC) during the course of evolution added unparalleled capacities for temporal coordination and the timely (predictive) activation of memory to the PA cycle (Fuster and Bressler, 2015). As the PFC coordinates motor and perceptual brain networks, it essentially exerts a "temporal syntactic function," which Fuster operationally defined as the "bridging of temporally separate elements of a behavioral gestalt," such as words forming an utterance (Fuster, 1995, p. 175). In interaction with other brain areas,

https://doi.org/10.1017/9781009295888.008 Published online by Cambridge University Press

this bridging function denotes the PFC as the highest node within executive and perceptual hierarchies that govern the temporal coordination of all goal-directed behavior (Fuster and Bressler, 2015).

Brain areas that connect to the PFC in support of this temporal syntactic function comprise neocortical areas such as primary sensory and association cortices but also the same subcortical areas that are implicated in temporal processing, namely the basal ganglia and the cerebellum that might exert a modulatory influence on the PA cycle (Fuster, 1995; Fuster and Bressler, 2015). Nevertheless, the parallel enlargement of the cerebellum and the PFC throughout evolution (MacLeod et al., 2003; Habas, 2021) and the well-documented cerebellar connectivity with the thalamus and basal ganglia as well as with central executive, default mode, salience, attentional, and language networks, stand exemplary for a wider network that might allow for more specialized interactions such as that of temporal processing and the PA cycle suggested here (Ramnani, 2006; Krienen and Buckner, 2009; Strick et al., 2009). It is of note, though, that within the PA cycle, temporal processing is not conceived as processing *of* time but processing *in* time (Fuster, 1995). Currently, it is also computationally largely unspecified how the PFC performs its temporal syntactic function. Nevertheless, the central notion of repeated and predictive activation that defines the PA cycle, potentially in a periodic fashion, points to the fundamental role of the corresponding neural rhythms in the temporal coordination of goal-directed behavior. In a similar vein, Lashley noted that temporal coordination of behavior may be achieved through "temporally spaced waves" that govern states of "facilitative excitation" (Lashley, 1951, p. 127). One may speculate that at least in those instances in which behavioral rhythms directly map onto neural rhythms, as shown for speech, neural oscillatory activity and temporal processing mechanisms instantiate a "master clock signal" that facilitates temporal coordination. This would essentially mean combining processing in time with processing of time, thus adding temporal specificity to the PA cycle with the aim to tune the PFC temporal syntactic function for optimal predictive adaptation.

6.3 The Complexity of Predicting "When"

Being able to predict *what* elements occur in an utterance is a powerful way to guide speech communication. However, to optimally adapt to an inherently dynamic environment, the human brain might likewise aim to predict specifically *when* something occurs (Schwartze and Kotz, 2013). Neural oscillations across different frequency bands offer a solution for the coding of temporal information and the generation of temporal predictions, directing neural and cognitive resources to specific points in time. Whereas associations or feedforward and feedback networks are good at predicting *what* occurs, *temporal* coordination and *temporal* prediction are achieved most directly and in an energy-efficient manner by oscillations and by oscillation-based synchrony (Buzsaki and Draguhn, 2004).

Due to the rhythmic complexity of speech, it is often difficult to maintain any strict differentiation of processing in time and of time or of *what* as opposed to *when* predictions. According to Rosen (1992), the acoustic signal carries rhythmic features on at least three distinct levels: envelope (fluctuations in overall amplitude at rates between 2 and 50 Hz),

https://doi.org/10.1017/9781009295888.008 Published online by Cambridge University Press

periodicity (50–500 Hz), and fine structure (0.5–10 kHz), each of which is defined by characteristic acoustic, auditory, and perceptual manifestations and their specific roles in linguistic contrasts. Correspondingly, Lashley discussed a series of hierarchies of organization for speech, ranging from the order of vocal movements to the discourse level (Lashley, 1951). This complexity leaves it open as to whether the interaction of temporal processing with the PA cycle would manifest in one or multiple rhythms and which feature or organizational level(s) would carry it. Human multitasking abilities seem to imply parallel processing in several PA cycles (Fuster and Bressler, 2015). However, if the notion of a "master clock signal" that governs temporal coordination holds true, it suggests a special status of one rhythm and that neural oscillatory activity in the respective frequency band is used for specific temporal prediction. This raises the issue of potential interference effects and confusion between production and perception as speakers almost simultaneously are listeners to their own speech and perceivers of their surroundings, including the environment and the interlocutor. Although different rhythms can in principle coexist in the same and different brain areas, but also interact with each other (Buzsaki and Draguhn, 2004), parallel processing of the same rhythm likely requires a more fine-grained differentiation of brain areas and connections implied in temporal processing and the PA cycle. For example, it has been shown that the dentate nucleus, the main output relay of the cerebellum, divides into motor and cognitive subcompartments with distinct connectivity patterns to the SMA (Dum and Strick, 2003; Akkal et al., 2007). Engagement of the SMA in temporal processing can be differentiated along an anterior–posterior axis (Schwartze et al., 2012). This organization is maintained for connections from the anterior SMA to the PFC and from the posterior SMA to premotor areas, and connections from both to the basal ganglia (Kotz et al., 2013). Although the picture is far from complete in this regard, such a configuration might allow parallel processing of the same rhythm in neighboring areas in production and perception.

6.4 Tracing a Fundamental Driving Rhythm

Lashley was not only positive that rudiments of every human behavioral mechanism can be traced down the evolutionary scale but also that they are represented in "primitive activities" (i.e., simple and evolutionarily ancient) of the nervous system (Lashley, 1951, p. 134). Several potential candidates have been put forward. However, following Lashley's rationale, a way to identify a likely driving rhythm is to reduce the rhythmic and organizational complexity of the speech signal to a simpler and evolutionary ancient level. The frame/content (f/c) theory of the origin of speech differentiates such a level, considering that speech is first and foremost a motor behavior that is expressed through serially organized movements while speaking (MacNeilage, 1998, 2008). Its central tenet is the mandibular cycle, that is, alternating opening and closing mouth movements, which constitute a general-purpose carrier or frame into which vowels and consonants are combined to form specific content (MacNeilage, 1998, 2008; see Chapter 2). Viewed at this level, continuous speech is carried by the biphasic cycles of a mandibular oscillation that can take simple contrastive forms such as *baba* but also all of the more complex forms found in speech (MacNeilage, 2008). Accordingly, opening the mouth to produce vowels

https://doi.org/10.1017/9781009295888.008 Published online by Cambridge University Press

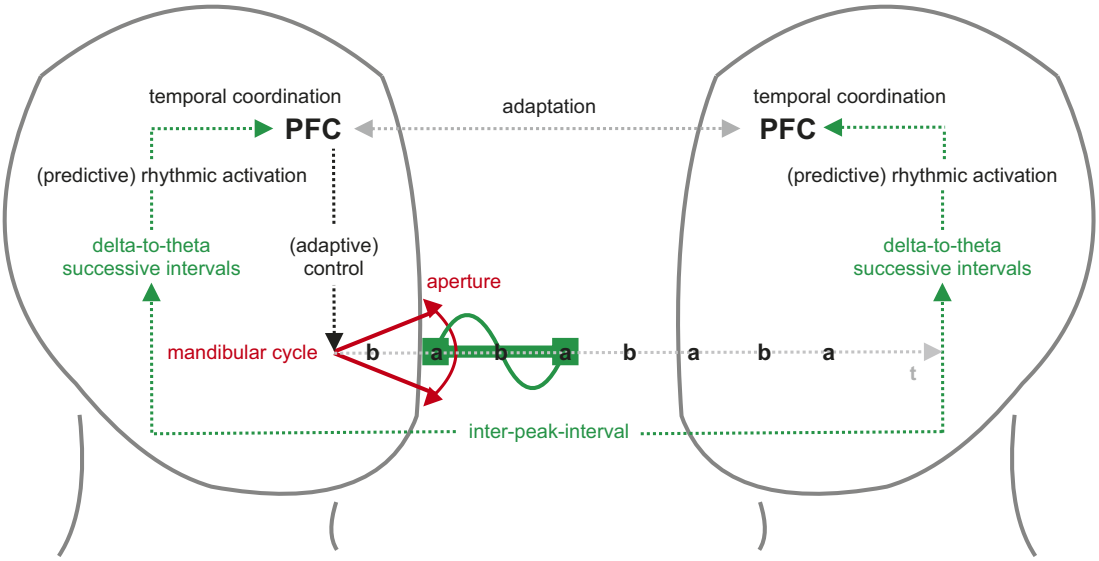


Figure 6.1 Syllabic rhythm in the speech perception–action cycle.
Speakers (left) produce an utterance with a syllabic rhythm through opening and closing movements of the mouth (mandibular cycle) and associated fluctuations in sound energy. Maximized oral aperture during vowel production generates energy peaks that define the rhythm's rate over successive inter-peak intervals. In speakers (through feedback, left) and listeners (right), this rhythm maps onto neural oscillations in the delta-to-theta frequency range (about 2–8 Hz). This direct mapping and/or the repeated use of interval-based temporal processing mechanisms allows for factoring "time" into behavior and into intrapersonal and interpersonal adaptation. Adaptation entails timely and predictive activation to optimally allocate neural and cognitive resources in production and perception. At the highest level of the underlying processing hierarchy, the PFC provides the temporal integration and coordination capacities that bridge the temporally separate elements of the utterance into one behavioral gestalt for monitoring, planning, and comprehension.

and closing it to produce consonants is regarded as the most fundamental organizational level of serial order in speech (MacNeilage, 2008).

The corresponding articulatory gestures and the resultant syllables are associated with a fluctuation in sound energy that reaches peaks with maxima of oral aperture, giving rise to a syllabic rhythm (Greenberg, 2006). During conversation, this syllabic rhythm is conveyed to the listener but also as feedback to the speaker, designating it as an anchor for temporal coordination and the driver of predictive adaptation (Figure 6.1). Syllabic rhythm is typically realized at rates that correspond to delta and theta frequencies (ca. 2–8 Hz), that is, a relatively slow neural rhythm with a period in the hundreds-of-milliseconds range. The correspondence between syllabic rhythm and neural activity is most directly expressed in the concept of the theta syllable, conceived as a unit of speech information that is defined by cortical function (Ghitza, 2013).

As an acoustically and neurally defined concept, the theta syllable rhythm does not draw on orthography, instead suggesting that information transmission in speech communication

https://doi.org/10.1017/9781009295888.008 Published online by Cambridge University Press

operates on syllabic "packets" (Ghitza and Greenberg, 2009). This has important implications for signal processing as it shifts the focus from abstract points in time such as word or sentence boundaries to the concrete physical markers of time that are instantiated by successive energy peaks, potentially defining one driving rhythm across speech production and perception on the one hand and speech-related temporal processing and temporal coordination mechanisms on the other (Schwartze and Kotz, 2013; Kotz and Schwartze, 2016).

6.5 Towards Cycles of Interaction

If the same low-frequency driving rhythm underlies the interaction of production and perception, temporal processing, and temporal coordination, it would have a broad range of implications for adaptive behavior. However, identification of this rhythm and the corresponding frequency range also imply specific limits and a potential differentiation of the role of oscillation- and interval-based temporal processing mechanisms.

Speaking and listening ultimately involve parallel acts of synchronizing information flow between the encoding and decoding capacities of speakers and listeners (Greenberg, 2012). In a similar vein, the f/c theory emphasizes the key role of the sociocultural dyad from an evolutionary perspective: "Language must have been a sociocultural invention. The first word only became a word when a receiver and a sender came to treat a particular sound complex as standing for a particular concept" (MacNeilage, 2008, p. 44). Temporal processing mechanisms may facilitate temporal coordination and synchronization also at this interpersonal level, which adds the rhythm of interaction as another temporal level to the threefold rhythmic differentiation (Rosen, 1992) of the acoustic signal. However, whereas oscillations and oscillation-based synchrony provide the most efficient means to synchronize encoding and decoding capacities at faster rates, interval-based temporal processing might be most efficient at rates in the hundreds-of-millisecond to seconds range and in guiding temporally predictive adaptation in the socially interactive PA cycles that are essential for the development of speech communication capacities.

Throughout the human lifespan, social interaction shapes and improves linguistic competence (Tomasello, 2000; Kuhl, 2007; Mundy and Jarrold, 2010; Dikker et al., 2017; Mundy, 2018) such as learning and proficiency in a nonnative language (Perani et al., 2003; Jeong et al., 2010, 2011; Consonni et al., 2013; Verga and Kotz, 2013, 2019; Kuhlen et al., 2017), while its absence can lead to linguistic deficits (Krashen, 1973; Fromkin et al., 1974; Curtiss, 1977). The reason is debated. However, several theories converge on defining a social partner as an attentional enhancer/modulator based on two interrelated characteristics of social contexts, namely that i) they are intrinsically multimodal (Gogate and Bahrick, 2001; Gogate et al., 2001), and ii) two partners mutually and dynamically join attention in a common ground (Hollich et al., 2000; Tomasello, 2000; Kuhl, 2007; Sage and Baldwin, 2010). In this shared conversational ground, a speaker and a listener can become temporally aligned at multiple levels, as indicated by converging speech rates (Schultz et al., 2016), phonetic realizations (Mukherjee et al., 2019), body postures (Shockley et al., 2003, 2007), gaze (Richardson and Dale, 2005; Richardson et al., 2007), but also neural activity (Stephens et al., 2010).

https://doi.org/10.1017/9781009295888.008 Published online by Cambridge University Press

By contrast, reduced listener/speaker coupling is associated with difficulties in speech understanding (e.g., Liu et al., 2021).

In conversations, this mutual exchange may assume a rhythmic quality through the smooth transition of turns between a speaker and a listener, which has been suggested to rely on avoiding both long silences and long overlaps (Stivers et al., 2009; Garrod and Pickering, 2015). Rudimental forms of turn-taking have been observed in interactive communication in many other species (Ghazanfar and Takahashi, 2014; Takahashi et al., 2016; Ravignani et al., 2019). In humans, it is assumed that interactive alignment provides the backbone for turn-taking: By allowing conversing partners to match their mental representations (Garrod and Pickering, 2004; Pickering and Garrod, 2004; Menenti et al., 2012), it facilitates the prediction not only of what is coming next (i.e., a response from the listener) but also – most importantly – of when the next conversational turn should start. This efficient cognitive strategy (Koban et al., 2019; Mukherjee et al., 2019) benefits from nonlinguistic convergence (e.g., shared gaze and attention; Richardson and Dale, 2005; Menenti et al., 2012). Yet, linguistic information deriving from the jaw movements subtending syllable production has been proposed to drive the mutual entrainment of endogenous oscillators in the speaker's and listener's brains (Wilson and Wilson, 2005; Kotz et al., 2018; see Figure 6.1). Thus, even in dyadic conversations, syllables emerge as basic fundamental units of speech in relation to mouth and jaw movements in an action–perception cycle (MacNeilage, 2008). Thus, dyadic mutual alignment is expected to elicit fronto-central brain activity, reflecting an action–perception feedback loop (Pulvermüller, 2018). Indeed, neural activity peaks in the mouth motor region (Glanz et al., 2018) and ventral premotor cortex (Wilson et al., 2004; Gordon et al., 2014; Glanz et al., 2018). Interestingly, similar orofacial movements such as lip-smacking in nonhuman animals have gained attention as possible precursors of cyclic speech components in the theta range (MacNeilage, 1998; Ghazanfar et al., 2013; Kotz et al., 2018). Lip-smacking is an affiliative signal observed in many primate species and is characterized by the rhythmic (vertical) opening and closing of the jaws. During the evolution of the human lineage, lip-smacking may have coupled with vocal signals to produce babbling (i.e., consonant-vocal speech-like sounds; MacNeilage, 1998, 2008; Ghazanfar and Takahashi, 2014).

6.6 Conclusion

Speech communication generates acoustic and interpersonal rhythms that span multiple timescales. However, the mere fact that all speech communication is rhythmic does by no means imply that the role of the respective temporal information for speech processing is trivial. Speech serves to communicate meaning, and the acoustic signal encodes the respective information over time. Time and temporal coordination are thus not only crucial factors for the "logical and orderly arrangement of thought and action" that is at the heart of the problem of serial order of behavior (Lashley, 1951) but also for the reconstruction of orderly thought and action from the acoustic signal. The perception–action cycle framework (Fuster, 1995) defines the basis for the arrangement and the perceptual bridging of speech events over time. However, to continuously optimize predictive adaptation to the environment, this cycle may interface with other mechanisms and

https://doi.org/10.1017/9781009295888.008 Published online by Cambridge University Press

with other sensory domains. To this end, oscillation- and interval-based temporal processing mechanisms may guide a temporally specific variant of the perception–action cycle to integrate *what* and *when* aspects of speech for successful communication.

Box 6.1 Chapter Overview

Summary

Speech communication evokes rhythms that span multiple timescales. The acoustic signal evolves serially, necessitating adequate temporal coordination in production and perception. To optimize temporal coordination of neural and cognitive mechanisms, temporal processing mechanisms may exploit rhythmic speech features to guide a temporally specific variant of the perception–action cycle.

Implications

Rhythmic speech features indicate various neural and movement mechanisms. Rhythmic features in the delta-to-theta range might establish the common ground for interactions of evolutionarily ancient serial ordering, temporal processing, and higher-order temporal coordination mechanisms. This integrative perspective offers starting points for investigating the relation of speech and temporal processing capacities.

Gains

A concrete form of linguistic performance, speech communication directly reflects fundamental neural and cognitive mechanisms including rhythmic motor control, serial ordering, temporal coordination, or attention and memory allocation. A better understanding of how speech rhythm interfaces with temporal processing in perception and action can thus inform about key aspects of cognition.

REFERENCES

Akkal, D., Dum, R. P., and Strick, P. L. (2007). Supplementary motor area and presupplementary motor area: Targets of basal ganglia and cerebellar output. *Journal of Neuroscience*, 27, 10659–10673. https://doi.org/10.1523/JNEUROSCI.3134-07.2007

Baumann, O., Borra, R. J., Bower, et al. (2015). Consensus paper: The role of the cerebellum in perceptual processes. *Cerebellum*, 14, 197–220. https://doi.org/10.1007/s12311-014-0627-7

Buhusi, C. V., and Meck, W. H. (2005). What makes us tick? Functional and neural mechanisms of interval timing. *Nature Reviews Neuroscience*, 6, 755–765. https://doi.org/10.1038/nrn1764

Buzsaki, G., and Draguhn, A. (2004). Neural oscillations in cortical networks. *Science*, 304, 1926–1929. https://doi.org/10.1126/science.1099745

Chen, J. L., Penhune, V. B., and Zatorre, R. J., (2008). Listening to musical rhythms recruits motor regions of the brain. *Cerebral Cortex*, 18, 2844–2854. https://doi.org/10.1093/cercor/bhn042

https://doi.org/10.1017/9781009295888.008 Published online by Cambridge University Press

Consonni, M., Cafiero, R., Marin, D., et al. (2013). Neural convergence for language comprehension and grammatical class production in highly proficient bilinguals is independent of age of acquisition. *Cortex*, 49, 1252–1258. https://doi.org/10.1016/j.cortex.2012.04.009

Curtiss, S. (1977). *Genie: A psycholinguistic study of a modern-day "wild child."* New York: Academic Press.

Ding, N., Melloni, L., Zhang, H., Tian, X., and Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, 19, 158–164. https://doi.org/10.1038/nn.4186

Dikker, S., Wan, L., Davidesco, I., et al. (2017). Brain-to-brain synchrony tracks real-world dynamic group interactions in the classroom. *Current Biology*, 27, 1375–1380. https://doi.org/10.1016/j.cub.2017.04.002

Dum, R. P., and Strick, P. L. (2003). An unfolded map of the cerebellar dentate nucleus and its projections to the cerebral cortex. *Journal of Neurophysiology*, 89, 634–639. https://doi.org/10.1152/jn.00626.2002

Fromkin, V., Krashen, S. D., Curtiss, S., Rigler, D., and Rigler, M. (1974). The development of language in Genie: A case of language reacquisition beyond the critical period. *Brain and Language*, 1, 81–107. https://doi.org/10.1016/0093-934X(74)90027-3

Fuster, J. M. (1995). Temporal processing. In J. Grafman, K. J. Holyoak, and F. Boller (Eds.), *Structure and function of the human prefrontal cortex* (pp. 173–181). New York: New York Academy of Sciences.

Fuster, J. M. (2004). Upper processing stages of the perception–action cycle. *Trends in Cognitive Sciences*, 8, 143–145. https://doi.org/10.1016/j.tics.2004.02.004

Fuster, J. M., and Bressler, S. L. (2015). Past makes future: Role of pFC in prediction. *Journal of Cognitive Neuroscience*, 27, 639–654. https://doi.org/10.1162/jocn_a_00746

Garrod, S., and Pickering, M. J. (2004). Why is conversation so easy? *Trends in Cognitive Science*, 8, 8–11. https://doi.org/10.1016/j.tics.2003.10.016

Garrod, S., and Pickering, M. J. (2015). The use of content and timing to predict turn transitions. *Frontiers in Psychology*, 6, 751. https://doi.org/10.3389/fpsyg.2015.00751

Ghazanfar, A. A., and Takahashi, D. Y. (2014). Facial expressions and the evolution of the speech rhythm. *Journal of Cognitive Neuroscience*, 26, 1196–1207. https://doi.org/10.1162/jocn_a_00575

Ghazanfar, A. A., Morrill, R. J., and Kayser, C. (2013). Monkeys are perceptually tuned to facial expressions that exhibit a theta-like speech rhythm. *Proceedings of the National Academy of Sciences*, 110, 1959–1963. https://doi.org/10.1073/pnas.1214956110

Ghitza, O. (2013). The theta-syllable: A unit of speech information defined by cortical function. *Frontiers in Psychology*, 4, 138. https://doi.org/10.3389/fpsyg.2013.00138

Ghitza, O., and 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. https://doi.org/10.1159/000208934

Giraud, A., Kleinschmidt, A., Poeppel, D., et al. (2007). Endogenous cortical rhythms determine cerebral specialization for speech perception and production. *Neuron*, 56, 1127–1134. https://doi.org/10.1016/j.neuron.2007.09.038

Glanz, I., Derix, J., Kaur, R., et al. (2018). Real-life speech production and perception have a shared premotor-cortical substrate. *Scientific Reports*, 8, 8898. https://doi.org/10.1038/s41598-018-26801-x

Gogate, L. J., and Bahrick, L. E. (2001). Intersensory redundancy and 7-month-old infants' memory for arbitrary syllable–object relations. *Infancy*, 2, 219–231. https://doi.org/10.1207/S15327078IN0202_7

Gogate, L. J., Walker-Andrews, A., and Bahrick, L. E. (2001). The intersensory origins of word-comprehension: An ecological-dynamic systems view. *Developmental Science*, 1, 1–37. https://doi.org/10.1111/1467-7687.00143

Gordon, R. G., Tranel, D., and Duff, M. C. (2014). The physiological basis of synchronizing conversational rhythms: The role of the ventromedial prefrontal cortex. *Neuropsychology*, 28, 624–630. https://doi.org/10.1037/neu0000073

https://doi.org/10.1017/9781009295888.008 Published online by Cambridge University Press

Greenberg, S. (2012). A multi-tier framework for understanding spoken language. In S. Greenberg and W. A. Ainsworth (Eds.), *Listening to speech: An auditory perspective* (pp. 411–434). Mahwah, NJ: Erlbaum.

Greenberg, S., Carvey, H., Hitchcock, L., and Chang, S. (2003). Temporal properties of spontaneous speech: A syllable-centric perspective. *Journal of Phonetics*, 31, 465–485. https://doi.org/10.1016/j.wocn.2003.09.005

Habas, C. (2021). Functional connectivity of the cognitive cerebellum. *Frontiers in Systems Neuroscience*, 15, 642225. https://doi.org/10.3389/fnsys.2021.642225

Hollich, G. J., Hirsh-Pasek, K., Golinkoff, R. M., et al. (2000). Breaking the language barrier: An emergentist coalition model for the origins of word learning. *Monographs of the Society for Research in Child Development*, 65, i–135.

Ivry, R. B., and Schlerf, J. E. (2008). Dedicated and intrinsic models of time perception. *Trends in Cognitive Sciences*, 12, 273–280. https://doi.org/10.1016/j.tics.2008.04.002

Jeong, H., Hashizume, H., Sugiura, M., et al. (2011). Testing second language oral proficiency in direct and semidirect settings: A social-cognitive neuroscience perspective. *Language Learning*, 61(3), 675–699.

Jeong, H., Sugiura, M., Sassa, Y., et al. (2010). Learning second language vocabulary: Neural dissociation of situation-based learning and text-based learning. *NeuroImage*, 50, 802–809. https://doi.org/10.1016/j.neuroimage.2009.12.038

Koban, L., Ramamoorthy, A., and Konvalinka, I. (2019). Why do we fall into sync with others? Interpersonal synchronization and the brain's optimization principle. *Social Neuroscience*, 14, 1–9. https://doi.org/10.1080/17470919.2017.1400463

Kotz, S. A., and Schwartze, M. (2010). Cortical speech processing unplugged: A timely subcortico-cortical framework. *Trends in Cognitive Sciences*, 14, 392–399. https://doi.org/10.1016/j.tics.2010.06.005

Kotz, S. A., and Schwartze, M. (2016). Motor-timing and sequencing in speech production. In G. Hickok and S. L. Luck (Eds.), *Neurobiology of language* (pp. 717–724). New York: Academic Press. https://doi.org/10.1016/B978-0-12-407794-2.00057-2

Kotz, S. A., Ravignani, A., and Fitch, W. T. (2018). The evolution of rhythm processing. *Trends in Cognitive Sciences*, 22, 896–910. https://doi.org/10.1016/j.tics.2018.08.002

Kotz, S. A., Anwander, A., Axer, H., and Knösche, T. R. (2013). Beyond cytoarchitectonics: The internal and external connectivity structure of the caudate nucleus. *PLoS ONE*, 8, e70141. https://doi.org/10.1371/journal.pone.0070141

Krashen, S. D. (1973). Lateralization, language learning, and the critical period: Some new evidence. *Language Learning*, 23, 63–74.

Krienen, F. M., and Buckner, R. L. (2009). Segregated fronto-cerebellar circuits revealed by intrinsic functional connectivity. *Cerebral Cortex*, 19, 2485–2497. https://doi.org/10.1093/cercor/bhp135

Kuhl, P. K. (2007), Is speech learning "gated" by the social brain? *Developmental Science*, 10: 110–120. https://doi.org/10.1111/j.1467-7687.2007.00572.x

Kuhlen, A. K., Bogler, C., Brennan, S. E., Haynes, J.-D. D. (2017). Brains in dialogue: Decoding neural preparation of speaking to a conversational partner. *Social Cognitive and Affective Neuroscience*, 12, 1–10. https://doi.org/10.1093/scan/nsx018

Large, E. W., and Jones, M. R. (1999). The dynamics of attending: How people track time-varying events. *Psychological Review*, 106, 119–159. https://doi.org/10.1037/0033-295X.106.1.119

Lashley, K. S. (1951). The problem of serial order in behavior. In L. A. Jeffress (Ed.), *Cerebral mechanisms in behavior* (pp. 112–136). New York: Wiley.

Liu, L., Ding, X., Li, H., et al. (2021). Reduced listener–speaker neural coupling underlies speech understanding difficulty in older adults. *Brain Structure and Function*, 226, 1571–1584. https://doi.org/10.1007/s00429-021-02271-2

https://doi.org/10.1017/9781009295888.008 Published online by Cambridge University Press

Martin, J. G. (1972). Rhythmic (hierarchical) versus serial structure in speech and other behavior. *Psychological Review*, 79, 487–509. https://doi.org/10.1037/h0033467

MacLeod, C. E., Zilles, K., Schleicher, A., Rilling, J. K., and Gibson, K. R. (2003). Expansion of the neocerebellum in Hominoidea. *Journal of Human Evolution*, 44, 401–429. https://doi.org/10.1016/s0047-2484(03)00028-9

MacNeilage, P. F. (1998). The frame/content theory of evolution of speech production. *Behavioral and Brain Sciences*, 21, 499–546. https://doi.org/10.1017/s0140525x98001265

MacNeilage, P. F. (2008). *The origin of speech*. New York: Oxford University Press.

Menenti, L., Pickering, M. J., and Garrod, S. C. (2012). Toward a neural basis of interactive alignment in conversation. *Frontiers in Human Neuroscience*, 6, 185. https://doi.org/10.3389/fnhum.2012.00185

Merchant, H., Harrington, D. L., and Meck, W. H. (2013). Neural basis of the perception and estimation of time. *Annual Review of Neuroscience*, 36, 313–336. https://doi.org/10.1146/annurev-neuro-062012-170349

Mukherjee, S., Badino, L., Hilt, P. M., et al. (2019). The neural oscillatory markers of phonetic convergence during verbal interaction. *Human Brain Mapping*, 40, 187–201. https://doi.org/10.1002/hbm.24364

Mundy, P. (2018). A review of joint attention and social-cognitive brain systems in typical development and autism spectrum disorder. *European Journal of Neuroscience*, 47, 497–514. https://doi.org/10.1111/ejn.13720

Mundy, P., and Jarrold, W. (2010). Infant joint attention, neural networks and social cognition. *Neural Networks*, 23, 985–997. https://doi.org/10.1016/j.neunet.2010.08.009

Perani, D., Abutalebi, J., Paulesu, E., et al. (2003). The role of age of acquisition and language usage in early, high-proficient bilinguals: An fMRI study during verbal fluency. *Human Brain Mapping*, 19, 170–182. https://doi.org/10.1002/hbm.10110

Petacchi, A., Laird, A. R., Fox, P. T., and Bower, J. M. (2005). Cerebellum and auditory function: An ALE meta-analysis of functional neuroimaging studies. *Human Brain Mapping*, 25, 118–128. https://doi.org/10.1002/hbm.20137

Pickering, M. J., and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27, 169–225.

Pulvermüller, F. (2018). Neural reuse of action perception circuits for language, concepts and communication. *Progress in Neurobiology*, 160, 1–44. https://doi.org/10.1016/j.pneurobio.2017.07.001

Ramnani, N. (2006). The primate cortico-cerebellar system: Anatomy and function. *Nature Reviews Neuroscience*, 7, 511–522. https://doi.org/10.1038/nrn1953

Ravignani, A., Verga, L., and Greenfield, M. D. (2019). Interactive rhythms across species: The evolutionary biology of animal chorusing and turn-taking. *Annals of the New York Academy of Sciences*, 1453(1), 14230. https://doi.org/10.1111/nyas.14230

Richardson, D. C., and Dale, R. (2005). Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive Science*, 29, 1045–1060. https://doi.org/10.1207/s15516709cog0000_29

Richardson, D. C., Dale, R., and Kirkham, N. Z. (2007). The art of conversation is coordination: Common ground and the coupling of eye movements during dialogue. *Psychological Science*, 18, 407–413. https://doi.org/10.1111/j.1467-9280.2007.01914.x

Rosen, S. (1992). Temporal information in speech: Acoustic, auditory and linguistic aspects. *Philosophical Transactions of the Royal Society B*, 336, 367–373. https://doi.org/10.1098/rstb.1992.0070

Sage, K. D., and Baldwin, D. A. (2010). Social gating and pedagogy: Mechanisms for learning and implications for robotics. *Neural Networks*, 23, 1091–1098. https://doi.org/10.1016/j.neunet.2010.09.004

https://doi.org/10.1017/9781009295888.008 Published online by Cambridge University Press

Schirmer (2004). Timing speech: A review of lesion and neuroimaging findings. *Cognitive Brain Research*, 21, 269–287. https://doi.org/10.1016/j.cogbrainres.2004.04.003

Schroeder, C. E., and Lakatos, P. (2008). Low-frequency neuronal oscillations as instruments of sensory selection. *Trends in Neurosciences*, 32, 9–18. https://doi.org/10.1016/j.tins.2008.09.012

Schroeder, C. E., Lakatos, P., Kajikawa, Y., Partan, S., and Puce, A. (2008). Neural oscillations and visual amplification of speech. *Trends in Cognitive Sciences*, 12, 106–113. https://doi.org/10.1016/j.tics.2008.01.002

Schultz, B. G., O'Brien, I., Phillips, N., et al. (2016). Speech rates converge in scripted turn-taking conversations. *Applied Psycholinguistics*, 37, 1201–1220. https://doi.org/10.1017/S0142716415000545

Schwartze, M., and Kotz, S. A. (2013). A dual-pathway neural architecture for specific temporal prediction. *Neuroscience and Biobehavioral Reviews*, 37, 2587–2596. https://doi.org/10.1016/j.neubiorev.2013.08.005

Schwartze, M., Rothermich, K., and Kotz, S. A. (2012). Functional dissociation of pre-SMA and SMA-proper in temporal processing. *NeuroImage*, 60, 290–298. https://doi.org/10.1016/j.neuroimage.2011.11.089

Shockley, K., Santana, M.-V., and Fowler, C. A. (2003). Mutual interpersonal postural constraints are involved in cooperative conversation. *Journal of Experimental Psychology: Human Perception and Performance*, 29, 326–332. https://doi.org/10.1037/0096-1523.29.2.326

Shockley, K., Baker, A. A., Richardson, M. J., and Fowler, C. A. (2007). Articulatory constraints on interpersonal postural coordination. *Journal of Experimental Psychology: Human Perception and Performance*, 33, 201–208. https://doi.org/10.1037/0096-1523.33.1.201

Stephens, G. J., Silbert, L. J., and Hasson, U. (2010). Speaker–listener neural coupling underlies successful communication. *Proceedings of the National Academy of Sciences*, 107, 14425–14430. https://doi.org/10.1073/pnas.1008662107

Stivers, T., Enfield, N. J., Brown, P., et al. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26), 10587–10592. https://doi.org/10.1073/pnas.0903616106

Strick, P. L., Dum, R. P., and Fiez, J. A., 2009. Cerebellum and nonmotor function. *Annual Review of Neuroscience*, 32, 413–434. https://doi.org/10.1146/annurev.neuro.31.060407.125606

Takahashi D. Y., Fenley, A. R., and Ghazanfar, A. A. (2016). Early development of turn-taking with parents shapes vocal acoustics in infant marmoset monkeys. *Philosophical Transactions of the Royal Society B*, 371, 20150370. https://doi.org/10.1098/rstb.2015.0370

Tomasello, M. (2000). The social-pragmatic theory of word learning. *Pragmatics*, 10, 401–413.

Uexküll, J. V. (1926). *Theoretical biology*. New York: Harcourt, Brace & Co.

Verga, L., and Kotz, S. A. (2013). How relevant is social interaction in second language learning? *Frontiers in Human Neuroscience*, 7, 550. https://doi.org/10.3389/fnhum.2013.00550

Verga, L., and Kotz, S. A. (2019). Spatial attention underpins social word learning in the right fronto-parietal network. *NeuroImage*, 195, 165–173. https://doi.org/10.1016/j.neuroimage.2019.03.071

Wiener, M. (2024). Coordinate-based meta-analyses of the time perception network. In H. Merchant and V. de Lafuente (Eds.), *Neurobiology of interval timing: Advances in experimental medicine and biology* (Volume 1455). Cham: Springer. https://doi.org/10.1007/978-3-031-60183-5_12

Wilson, M., and Wilson, T. P. (2005). An oscillator model of the timing of turn-taking. *Psychonomic Bulletin & Review*, 12, 957–968. https://doi.org/10.3758/bf03206432

Wilson, S. M., Saygin, A. P., Sereno, M. I., and Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nature Neuroscience*, 7, 701–702. https://doi.org/10.1038/nn1263

https://doi.org/10.1017/9781009295888.008 Published online by Cambridge University Press