Trends in Neurosciences

CellPress
REVIEWS

Review

A Hierarchy of Autonomous Systems for Vocal Production

Yisi S. Zhang[1,*] and Asif A. Ghazanfar[1,2,3,*]

Vocal production is hierarchical in the time domain. These hierarchies build upon biomechanical and neural dynamics across various timescales. We review studies in marmoset monkeys, songbirds, and other vertebrates. To organize these data in an accessible and across-species framework, we interpret the different timescales of vocal production as belonging to different levels of an autonomous systems hierarchy. The first level accounts for vocal acoustics produced on short timescales; subsequent levels account for longer timescales of vocal output. The hierarchy of autonomous systems that we put forth accounts for vocal patterning, sequence generation, dyadic interactions, and context dependence by sequentially incorporating central pattern generators, intrinsic drives, and sensory signals from the environment. We then show the framework's utility by providing an integrative explanation of infant vocal production learning in which social feedback modulates infant vocal acoustics through the tuning of a drive signal.

Highlights

Vocal production can be understood in terms of a timescale-based hierarchy of autonomous dynamical systems representing biomechanics and central pattern generators, internal drives, and the environment.

Each level of the vocal system generates a dynamical aspect of vocal output at a specific timescale.

This proposed autonomous systems framework can be used to illuminate how social reinforcement shapes the drive to generate mature sounding vocalizations in developing animals.

A Simplified Framework for Vocal Production

Understanding the biology of vocal production – how all the mechanistic pieces are coordinated, and how this coordination is achieved over the course of development – is a formidable challenge. Here, we review what is known about the biomechanical and neural mechanisms of non-human animal vocal production. We focus on non-human species because we know so much more about their biology and development than we do for human speech production. Moreover, at the level of the central pattern generators (CPGs) and some forebrain pathways, many circuits appear homologous or at the very least analogous across species [1,2]. It will be apparent even in just the subset of studies that we describe that there is a problem of increasing complexity: As we learn increasingly more about the details of neural circuits and biomechanics in different species, finding a common ground is difficult. We therefore put forth what we think is a plausible framework for the integrative biology of vocal production across species: a hierarchical set of autonomous systems with feedback [3]. In a system like this, high-level states are slower and provide the scaffolding for the fast, lower-level states [3,4]. This approach has, for example, been used to model the fast acoustic features of speech as the result of the slower articulator movement [5,6].

Our hope is that this framework will provide an integrative perspective on adaptive vocal production. It emphasizes that, first, vocal production should not be viewed as a one-to-one mapping from single-neuron activity to sound. The neural activity at each level of this system is under the influence of its upper level and in turn determines (or at least strongly influences) how downstream neural activities unfold. The same set of neurons can be involved in various types of vocal output, and the aggregate activity of a large neural population across different structures is more relevant to the outcomes of the vocal production than any single neuron or small set of neurons. Second, the framework recognizes that vocal production is a consequence of the interplay between internal states and behavioral context [7]. The interactions between context, the internal state, and the vocal output have often been invoked to explain vocal acoustic structure or even vocalization types, but much of the research is based largely on speculation as to what the animal's internal states may be and

[1]Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08544, USA
[2]Department of Psychology, Princeton University, Princeton, NJ 08544, USA
[3]Department of Ecology & Evolutionary Biology, Princeton University, Princeton, NJ 08544, USA

*Correspondence:
yz9@princeton.edu (Y.S. Zhang) and
asifg@princeton.edu (A.A. Ghazanfar).

© 2019 Elsevier Ltd. All rights reserved.

how it could influence vocal output [8]. Any context comprises multiple types of external sensory cues interacting with fluctuating internal states (which generate visceral cues) [9,10].

The Scope of the Problem

Vocal production is intrinsically hierarchical in the time domain. Human conversation, for example, consists of a sequence of vocal exchanges between two individuals; the duration of the conversation can vary greatly but must be composed of at least one utterance from each participant. These utterances are themselves sequences but of words. Words are, in turn, composed of consonant–vowel sounds. A similar temporal hierarchical structure is found in the vocal output of other animals, and these utterances can also be exchanged systematically between individuals. In many species of birds, mated pairs sing duets [11]; for example, both sexes of the black-bellied wren (*Pheugopedius fasciatoventris*) initiate song production and both answer their mates' songs to form duets [12]. Among other vertebrates, there is a similar duetting among mated pairs of both Old and New World non-human primates [13,14]. In frogs [15] and rodents [16], there are structured vocal exchanges between males during territorial defense. Beyond mating and territorial defense, animals also exhibit vocal turn-taking as affiliative gestures, a form of 'grooming-at-a-distance' (marmoset monkeys: [17,18]; macaque monkeys: [19]; lemurs: [20]; and meerkats: [21]), much like humans seem to do [22]. Each individual's vocalizations often have multiple elements and these can be divided into smaller units based on their sequential structure and duration (e.g., motifs and syllables; Figure 1A–C).

The neural correlates of vocal production also have a seemingly hierarchical structure. Across vertebrates, vocal production requires a source of air power and a sound-producing organ (e.g., the larynx, syrinx, or swim bladder) [1,23], and the first-order innervation of these peripheral structures is thought to arise from a homologous set of brainstem structures [24]. In primates and other mammals, the neural circuitry related to respiratory power includes groups of neurons within the pons and medulla that generate rhythmic patterns [25–27]; the circuitry influencing laryngeal tension arises from the nucleus ambiguus located in the medulla [27,28]. Homologous circuitry is present in songbirds [29]. Figure 1D–F show only a portion of the neural pathways for vocal production in monkeys, birds, and fish. Some nodes in these vocal-motor networks are tightly correlated with the temporal patterns and frequency modulations of the vocalizations they produce; they are thus considered vocal CPGs [24,30,31]. The periaqueductal gray (PAG) in the midbrain provides descending control over these CPGs, and its activity influences the production of specific vocalizations [29,32–35]. The PAG receives inputs from higher-order motor structures in the forebrain that are critical for the initiation of vocalizations and the integration of sensory cues (such as the vocalizations of conspecifics) with internal states (e.g., arousal levels). In the human lineage, this forebrain circuitry and its relationship to brainstem structures have been elaborated upon considerably over the course of evolution [2,36,37].

Vocal Production from a Two-Level Hierarchical System

The minimal system that can autonomously generate motor behaviors consists of the appropriate effectors and self-generated neural activity that drives the movement of those biomechanics. Rhythmic motions, such as breathing, swallowing, walking, and swimming, are mainly driven by CPGs located within the spinal cord [38,39] and the brainstem [40,41]. The CPGs are intrinsically capable of producing rhythmic signals that directly influence the biomechanics of effectors by alternating the activation of flexor and extensor muscles. Vocal production is also driven by the coordinated activity of CPGs for multiple effectors (e.g., the diaphragm and larynx [24,30,31], and the result is the production of sound – or sound sequences – that often have a rhythmic structure [42,43]. These vocalizations exhibit some stable features in the temporal and/or spectral domains and can be organized into distinct call types.

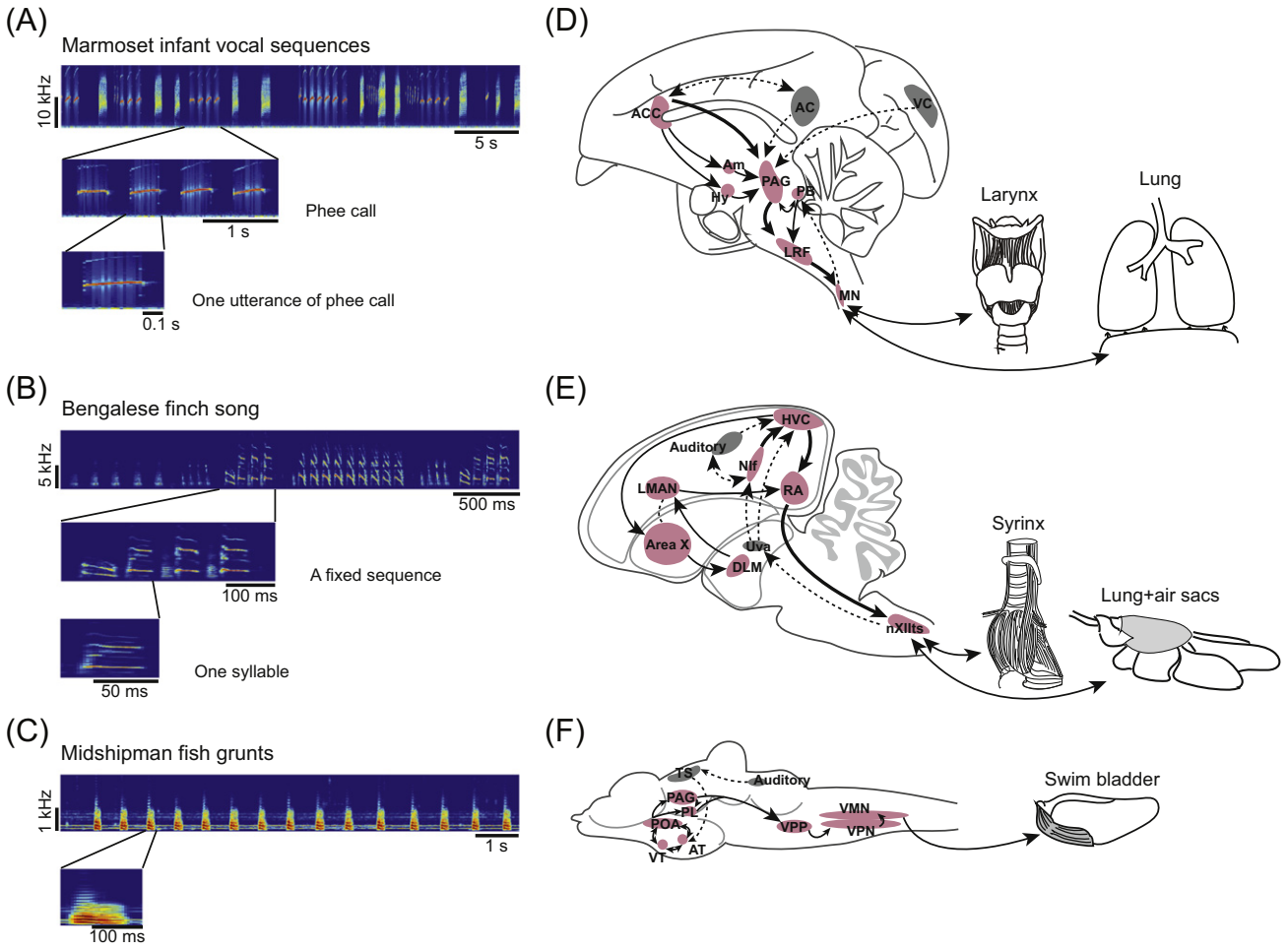

Figure 1. Temporal Hierarchies of Vocalizations, and the Corresponding Brain Structure and Vocal Apparatus. (A) A segment of infant marmoset monkey vocalization composed of utterances of different types of calls (contributed by Ghazanfar lab). (B) A segment of Bengalese finch song composed of fixed and variable syllable sequences (contributed by Yisi Zhang). (C) A segment of midshipman fish grunts (courtesy Dr Andrew Bass). (D) Vocal communication system and vocal apparatus of non-human primates. Abbreviations: AC, auditory cortex; ACC, anterior cingulate cortex; Am, amygdala; Hy, hypothalamus; LRF, lateral reticular formation; MN, motor nuclei; PAG, periaqueductal gray; PB, parabrachial nucleus; VC, visual cortex (adapted from [124,125]). (E) Song system and vocal apparatus of songbirds. Abbreviations: Area X, striatopallidal basal ganglia nucleus; DLM, dorsolateral thalamic nucleus; HVC, used as a proper name; LMAN, lateral magnocellular nucleus of the anterior nidopallium; NIf, nucleus interfacialis of the nidopallium; nXIIts, tracheosyringeal part of the hypoglossal motor nucleus; RA, robust nucleus of the arcopallium; Uva, nucleus uvaeformis (adapted from [126]). (F) Vocal network and vocal apparatus of fish. Abbreviations: AT, anterior tuberal nucleus; PL, paralemniscal midbrain tegmentum; POA, preoptic area; TS, torus semicircularis; VMN vocal motor nucleus; VPN, vocal pacemaker nucleus; VPP, vocal prepacemaker nucleus; VT, ventral tuberal nucleus (adapted from [127]). Unbroken lines indicate vocal pathways and broken lines indicate sensory pathways.

In primates and other mammals, the vocal apparatus consists of the lungs, the vocal folds of the larynx, and the vocal tract; the vocal tract consists of the oral and nasal cavities anterior to the larynx and whose shapes can be modified through articulations of the mouth and lips [23,44]. Sound production is through self-sustaining vocal fold oscillations induced by increases in subglottal air pressure and vocal fold tension [45–48]. A simple mechanism to vary fundamental frequencies, as demonstrated in marmoset monkeys [49,50] and zebra finches [51], is through manipulating two parameters, the air pressure and the vocal fold tension. As such, the spectral and temporal patterns emerge from the coordinated activity between respiratory and laryngeal (or syringeal for birds) muscles [49,51,52]. When coordinated by their respective CPGs, they

define the spectral properties of the sound and provide the temporal structure for vocal output [48–50]. Here's a specific example: In a vocal fish, the plainfin midshipman, pacemaker neurons in the hindbrain control the frequency of vocalizations, while 'pre-pacemaker' neurons that innervate them control the overall duration of the vocalization [53].

The emergent acoustic patterns, however, do not always slavishly reflect the output patterns of the CPGs; they are a joint consequence of the CPG patterns and the nonlinear dynamics of the tissue and biomechanical properties of the vocal apparatus. Typically, the production of different vocalizations is thought to occur through the differential assembly of neuronal populations that make up the vocal CPGs; it has been reasonably argued that there is a dedicated CPG assembly for each vocalization type [31]. Work in the domain of locomotion suggested, however, that this need not be the case. Distributed and coupled CPGs, instead of dedicated CPGs, yielded a successful model – with neurophysiological support – for switching between discrete modes of locomotion in the salamander: running and swimming [39], where the CPGs governing the limbs and body were coupled oscillators. The CPGs are usually modeled with low-dimensional nonlinear equations, a means to describe rich biologically relevant dynamics without the need to explicitly code the details of pattern formation [54,55]. Models like this are used to simulate marmoset monkey calls and birdsongs [52,56–58]. Together, the nonlinearities inherent in both the biomechanics of the vocal apparatus and the driving CPG activity are able to generate a set of acoustically discrete vocalizations with a repeated temporal pattern. They represent what we would refer to as a two-level hierarchical system (Figure 2A).

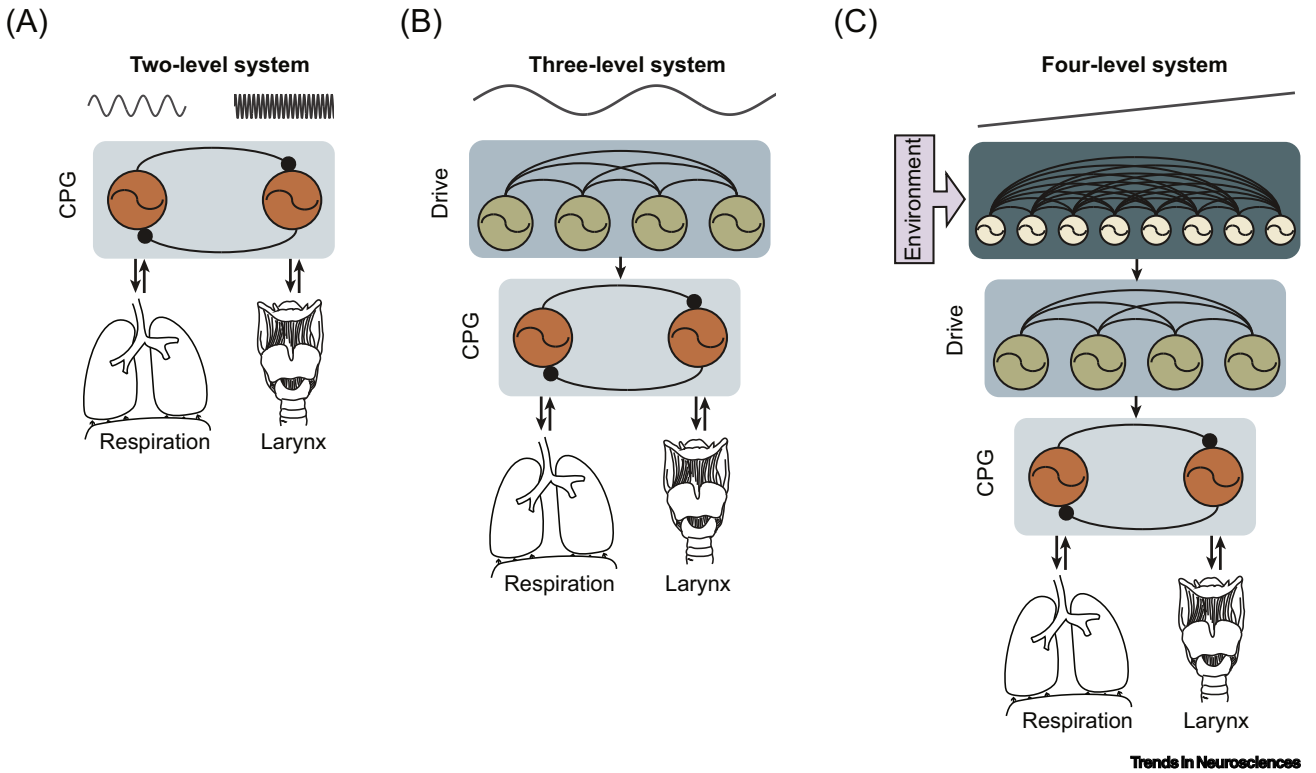

Figure 2. A Hierarchical Structure of the Vocal Production System Based on Timescales. (A) Two-level system: the vocal biomechanics (lungs and larynx) and two coupled central pattern generators (CPGs). (B) Three-level system: adding a drive signal on top of the CPGs enables the continuous production of various calls. (C) Four-level system: a fourth layer provides the modulation to the drive and allows animals to adjust the vocal output with respect to the environment.

On some temporal scales, the rhythm of vocalizations seems to be in a similar range across mammals. In humans, regardless of languages and contexts, the amplitude modulation of the speech signal consists of a rhythm that ranges between 3 and 8 Hz and is correlated with production of syllables [59,60]. The vocalizations and facial expressions of monkeys and apes also have this rhythmic structure [43,61–65]. For example, in marmoset monkeys, experimental interruptions of their contact calls (by playing back noise) reveal that the seemingly continuous longer elements of this vocalization are actually made up of smaller elements: marmoset vocalizations can only be interrupted at periodic time points [43]. These periodic intervals occur at a frequency of about 7 Hz. These data suggest that, just as human speech is built up from the elemental units of syllables, marmoset monkey vocalizations (and likely the vocalizations of other species) are built of multiple sequentially uttered units on the same timescale [43] (see [66] for review).

This suggests that humans and other mammals share a mechanistic substrate that produces 3–8-Hz vocal rhythms. We speculate that this substrate is at the two-level system we described (Figure 2A) and the set of CPGs involved is likely to be mostly overlapping with those involved in ingestive orofacial rhythms (like licking and chewing) for all mammals. These other orofacial movements are produced with a >2-Hz rhythm. The CPGs for all orofacial rhythms (including vocalizations) are located in the brainstem, specifically the pons and medulla [30,31,67]. These regions contain the cranial nuclei for sensory processing in the head and face and final motor outputs to orofacial muscles. Vocal and ingestive rhythms occur at frequencies faster than the respiratory rhythm and both the slower respiratory rhythm and the faster orofacial rhythms can be linked directly to separable patterns of CPG activity [67]. Yet, neurons within a CPG network can still participate in multiple orofacial behaviors. For example, in macaque monkeys, neural activity in the nucleus ambiguus of the medulla (which innervates laryngeal muscles) is modulated by vocalization, respiration, and/or swallowing [68,69]. Likewise, neurons in the medullary ventral respiratory group are modulated not only by respiration but also by swallowing and vocalizations [69]. Taken together, these data suggest that vocalizations and their rhythmicity are generated by a distributed network of CPGs that also participate in other rhythmic orofacial behaviors.

Complex Vocalization from a Three-Level Hierarchical System

Repetitively produced, rhythmic vocalizations driven by the two-level system do not necessarily represent ethologically relevant vocal behavior. A mechanism to produce different vocalizations at different times is needed. To enable that, one needs a third level. Let us consider the simplest situation regarding the timing of vocal output: when the animal is alone and spontaneously producing vocalizations (i.e., without an external triggering event such as another individual or a threat). Rats, for example, produce spontaneous ultrasonic vocalization bouts at a rate of about once every 7 or 8 s (~0.1 Hz) [70]. Adult marmoset monkeys produce contact calls around every 10 s (~0.1 Hz) when alone and with no conspecifics responding to them [18,71]. When in the hearing range of conspecifics, they exchange contact calls and their vocal production becomes antiphase locked to each other but with each marmoset maintaining a ~0.1-Hz rhythmic output [18]. Finally, the spontaneous speech of humans also has a similar slow temporal structure, whereby an individual will produce utterances of varying lengths roughly every 10 to 15 s (~0.1 Hz) [72] (though how this relates to turn-taking is not known). The temporal regularity of vocal production on a roughly similar timescale across species suggests that a common, slow oscillatory drive might exist within the mammalian vocal system.

However, what could be the biological basis for the much slower nearly 0.1-Hz rate of vocal production found in rats, marmoset monkeys, and humans? One link could be with the autonomic nervous system. In marmoset monkeys, spontaneous contact calls by adults are correlated with heart rate (a measure of arousal levels) [71]. This rhythmic arousal variation is the product

of the 'Mayer wave', an oscillation of the autonomic nervous system (specifically, sympathetic vasomotor tone); the Mayer wave appears to be common to all mammals [73]. The neural activity of the PAG, a brain structure essential for vocal production [74], is modulated by this 0.1-Hz oscillation [75]. This could causally account for the spontaneous production of vocalizations every 10 s or so by adult marmoset monkeys, and possibly other mammals [18,71]. Another possible source (and perhaps not unrelated) is the neocortex. The neocortex produces oscillations at various timescales [76,77] and this includes infraslow oscillations on the order of 0.1 Hz [78–80]. These very slow neural oscillations are distinct from other neural oscillations (e.g., 1–4-Hz delta rhythm and the much faster gamma rhythm); they travel across the neocortical sheet with a stereotypical spatiotemporal trajectory and are state dependent [81]. Moreover, there is at least one neurophysiological link between the Mayer wave (as measured by vasomotor tone) and neocortical activity: Infraslow fluctuations of neural activity occur as an envelope over fast gamma-band (30–80 Hz) activity, and this 0.1-Hz neural rhythm causes the 0.1-Hz oscillatory dilations and constrictions of arterioles [82]. This autonomic nervous system-related slow cortical rhythm represents the next level of our proposed hierarchical system (Figure 2B). It initiates the two-level system to produce vocal output at particular times, in this case, roughly every 10 s.

Importantly, the 0.1-Hz rhythm not only initiates vocal production but, in infants at least, it also influences the sequential structure of vocalizations. When separated from adults, newborn infant marmosets spontaneously produce long sequences of both immature- and mature-sounding vocalizations [49,83]. As is the case for human infants [84] and songbirds [42,85,86], the vocal output of marmoset infants is very rhythmic, tightly locked to the respiration rate around 1 Hz [83]. The time-varying spectral structure of infant marmoset vocal sequences also has a rhythm; however, it is at a rate that is an order of magnitude slower. Spectral entropy, a measure of the noisiness of the sound spectrum [49,87], fluctuates during infant vocal sequences at a 0.1-Hz frequency [83]. This spectral rhythm is in tight coherence with the infant's Mayer oscillation. The different and discrete call types produced in these sequences are likely the result of this arousal-linked 0.1-Hz oscillation driving the CPG oscillators and the subsequent vocal biomechanics through different regimes. Results from a computational model of rhythmic vocal output lend support to this idea [56]. Thus, our three-level hierarchical system can also modulate (in addition to initiating their coordinated activity) the CPGs controlling the vocal apparatus.

Context-Dependent Vocalization from a Four-Level Hierarchical System

Vocal output is context dependent. Spontaneous vocalizations described above are a special case of undirected contexts (no conspecifics are present) and thus reflect, in large part, the internal state of the animal. Most of the time, however, vocalizations involve another individual who responds in kind. In human vocal turn-taking, individuals become antiphase locked and entrained to each other through their speech [88]. Similar 'coupled oscillator' turn-taking behavior is observed in adult marmoset monkeys [18], macaque monkeys [89], meerkats [21], and potentially other species [66,89–93]. In both human and marmoset monkey turn-taking, vocal exchanges are separated by intervals within a small range, albeit at different timescales [18,94]. The coupling between individuals is modulated by the exchange of visual, auditory, and potentially other sensory signals. The processing of these sensory signals in the brain comprises the fourth level of the hierarchical system proposed here (Figure 2C).

Consider the following scenario: two marmoset monkeys can hear each other but cannot see each other. In this context, the two marmosets exchange contact 'phee' calls with antiphase locking, and mutually entrain to each other's call timing [18]. Manipulating the amplitude of contact call playback – a simulation of changing the physical distance of an out-of-sight caller – revealed that the marmoset vocal responses to nearby conspecifics were delayed and quieter; the opposite

was true for far-away conspecifics [17]. Of course, varying interindividual distances is not unique to marmoset monkeys. Zebra finches will adjust the intensity of their courtship songs according to their physical distance from a conspecific female [96]. Even human speakers are tacitly aware that as the distance between themselves and listeners is increased, vocal intensity must also be increased to maintain effective communication [97,98]. As a result, low-amplitude speech signals elicit high-amplitude responses from the listener. Adaptively adjusting speech amplitude is interpreted as a cooperative act requiring a high-level social skill [97,98]. However, a simple model incorporating auditory feedback can account for the behavior in humans as well as marmoset monkeys [17,95].

Changing the social distance between two individuals can also result in qualitative acoustic changes – the production of different call types [99]. In one study, an adult marmoset monkey was placed in four different contexts in which the marmoset could be (i) by itself, (ii) with a partner at the opposite corner of the room behind a curtain, (iii) with a visible partner at the opposite corner of the room, and (iv) with a partner close by. As the distance between two individuals decreased, marmoset monkeys switched from producing loud contact calls to producing softer, shorter, and noisier vocalizations [99]. Decreasing the distance between two individuals increased the robustness of auditory and visual signals from the conspecific. Together, the effect seems to be to increase the inhibition on the drive to produce vocalizations, which in turn reduces the overall time spent on calling and reduces the power to produce vocalizations. This is similar to what is observed for the spontaneous vocalizations of infants – the level of the drive changes what vocalizations are produced. The presence of increasing sensory cues coming from conspecifics seems to gradually diminish the dependence on the internal state fluctuation (i.e., the 0.1-Hz Mayer wave). This is supported by the decreasing coherence between vocal output and heart rate as a function of decreasing social distance [99]. In real life, the environment, including the proximity of conspecifics, slowly changes and continuously modulates the dynamics of the drive to vocalize. Relevant variables such as social distance represented by multiple sensory cues constitute the fourth (and final) level of the hierarchical set of autonomous systems (Figure 2C).

Using the Autonomous Systems Framework to Understand Vocal Development

How do infants, capable from postnatal day 1 of producing spontaneous vocalizations, develop adaptive vocal behavior? Sensory input from the social environment influences the structure of vocalizations very early in life. In humans, the acoustic structure of newborn crying reflects the acoustic structure of the ambient language environment [100]. In fairy wrens, the acoustic structure of parental nesting calls is learned prenatally by their offspring and partially reproduced postnatally in the offsprings' begging calls [101]. In marmoset monkeys, it is suggested that sensory experiences as early as in the womb contribute to shaping the temporal structure of the vocal output in early postnatal life [83]. The early vocal sequences of infant marmoset monkeys (within the first postnatal week), as well as their heart rate profiles, exhibited greater similarities in their sequential structure between dizygotic twins than with their non-twin siblings and non-relatives [83]. The similar autonomic modulation of the twins is likely through the interactions with their mothers both prenatally and postnatally. In humans, changes in maternal arousal levels can influence offspring both in the womb and through physical contact postnatally. For example, the cardiorespiratory dynamics of human infants will entrain to their mother's dynamics when they are laying on her body [102] and even during face-to-face communication [103]. These interindividual autonomic system influences can affect the vocal behavior of those infants [104]. Taken together, these studies suggest that the set of autonomous systems proposed here is working in concert very early in life and that vocal development does not proceed in discrete stages seemingly represented by the levels of the hierarchy (Figure 2).

Vocalizations continue to undergo changes over the course of postnatal development in many species. These changes include the improved ability to produce the correct sound in the right context, and this improvement is influenced by social factors. This is true for humans [105–107], birds [108–110], bats [111,112], and marmoset monkeys [49,113–115]. For example, newborn marmosets spontaneously produce many different call types when alone, some of which are not appropriate for the context or immature sounding [49]. Through development, marmosets gradually switch to producing only mature-sounding, energetically costly contact calls when they are out of sight or far away from conspecifics [99]. In order to produce such calls, laryngeal muscles providing the vocal fold tension and expiratory muscles providing the subglottal pressure have to be strongly contracted. How are the neural dynamics shaped to direct a strong input to the muscles? This could be achieved through tuning the drive signal (arousal state) via infant–parent interactions.

The effect of parental feedback on changing vocal behavior is both immediate and cumulative. In one experiment, an infant marmoset and its parent were placed in the opposite corners of a room with an acoustically transparent curtain in between. Instead of turn-taking using contact calls as adult marmosets do in this scenario [18], infant marmosets produce various types of vocalizations. However, they do immediately respond with a more adultlike contact call following the parental call [116]. It was found that parental calls were contingent upon the dynamics of the acoustic change in infant calls: they are produced at the phase of the 0.1-Hz acoustic oscillation when the infant calls were at the transition point from less to more adultlike [116]. Thus, on a moment-to-moment basis, the contingent parental calls are affecting the dynamics of the drive to the CPGs in the infants; it is providing a reinforcement signal to the infants (Figure 3). Parent–infant interactions can lead to a long-term change in infant vocal behavior. In humans [105–107] and songbirds [108–110], contingent parental responses accelerate the development of infant vocalizations, making them sound more mature more quickly. Similarly, in marmoset monkeys, the timing of the transition from immature to mature-sounding contact calls is correlated with the amount of contingent parental vocal feedback [49].

A study in which the amount of contingent parental responses was experimentally manipulated revealed that such feedback has a causal influence on vocal learning in this primate species as it does in humans [115]. Other studies showed that infant marmosets with limited parental feedback continued producing a large number of immature calls [113,114]. Together, these results suggest that the contingent vocal feedback and parental interactions are not merely a perturbation to the infant vocal system but rather changing the stable state of the drive oscillator in order to keep producing higher-energy (mature sounding) calls (Figure 3). What could be the biological basis for the sustained change of drive oscillations? The dopamine (DA) neurons of mammalian PAG represent the social context and arousal in response to social cues [117,118]. Analogously, a recent study in songbirds has suggested that the DA neurons in the PAG of juveniles play a role in initiating song learning through the PAG–HVC pathway in the presence of a singing tutor [119]. In mammals, the PAG is also part of the pathway of vocal control. We thus hypothesize that DA signaling in the PAG modulated by cortical inputs is involved in vocal learning by marmoset monkeys (and perhaps other vocal learning animals) through contingent parental feedback.

Using the Hierarchy of Autonomous Systems as a Hypothesis-Testing Framework

The hierarchical autonomous systems framework for vocal production by its very nature suggests that components of vocal production can be, in some sense, isolated from each other. For example, the processes related to the production of a vocal sound (two-level system; Figure 2A) can be separated from the process that determines the initiation of a vocalization (three-level system;

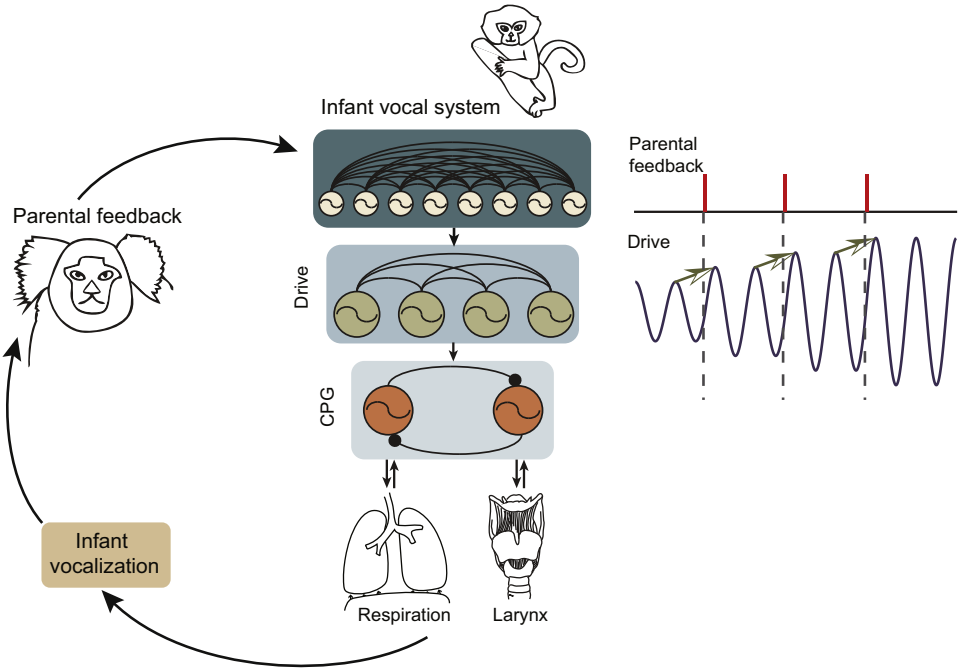

Figure 3. Vocal Development through Social Feedback. Here we illustrate vocal development through marmoset monkey parent–infant vocal interaction. The parent responds at the transitions where the infant starts producing more mature-sounding vocalizations [116]. As more mature-sounding vocalizations indicate a greater underlying drive signal [52], parental responses occur in the rising phase of the oscillatory drive. The contingent parental calls have a cumulative effect on the infant vocal production toward more energetically costly calls, accelerating vocal development on the timescale of days [115]. This process can be a consequence of shaping the drive signal of infant vocal production. The social feedback process described here compactly illustrates its cumulative influence on cumulative changes in the drive signal. Abbreviation: CPG, central pattern generator.

Figure 2B). Along the same lines, the type of vocalization produced (independent of its initiation) is determined by what is called for by the context (four-level system; Figure 2C). Neurophysiologic data from macaque monkeys trained to produce vocalizations on cue provide some support for this notion of separability, showing that sensory cues set up the decision to produce a vocalization in the ventrolateral prefrontal cortex while the anterior cingulate cortex reflects the degree to which the animal is motivated to produce a vocalization [120]. More specific predictions of the proposed multilevel framework could include testing the degree to which each level of the hierarchy functions as an autonomous system. For example, if the two-level system is autonomous, then experimental changes in laryngeal tension or the movement of other parts of the vocal apparatus should naturally affect the acoustic structure of vocalizations but it should not affect the rate or probability of producing a vocalization. Along the same lines, pharmacological manipulations of the autonomic nervous system (the three-level system) that affect arousal levels (e.g., beta blockers or beta agonists) should influence the timing of vocalizations independent of vocalization type. Finally, microstimulation, inactivation, or other manipulations of motor structures in the forebrain (the four-level system) should affect the probability of producing a particular type of vocalization without affecting its acoustic structure (i.e., novel vocal sounds will not be produced). Certainly, we are well aware that the biology of vocal production is much more complex than we presented it [7] (for example, we completely neglected the important role of hormones [121–123]), but despite these limitations, we would argue for the utility of the proposed framework, particularly for comparative studies.

Concluding Remarks

The hierarchical structure of the motor system has long been adopted to facilitate our understanding of motor planning and motor control. Vocal production is a concrete example of a motor behavior that can be well quantified due to the relatively low-dimensional nature of acoustic signals. In this review, we summarize findings in animal vocal production occurring at different timescales into a hierarchical autonomous systems framework to understand vocal production. Each hierarchy of this framework operates at a different temporal scale, starting with acoustic features within vocalizations, and proceeding to sequences of vocalizations, and finally to vocal interactions and context-dependent vocal production. Following this deduction, we gradually reveal the dynamics of higher-level brain regions that provide the manifolds for the dynamics of the lower level to unfold. Essentially, we propose that at least four levels are needed to fully understand the vocal production system, that is, from the lowest to the highest: the vocal biomechanics, the CPGs, the drive, and the sensory representation. With the help of the autonomous systems perspective and findings in recent experiments, nontrivial transitions in vocal behavior through development can be explained including those induced by social experience (see Outstanding Questions).

Outstanding Questions

Can the idea of distributed CPGs for vocal production be empirically validated, and if so, what is the neural basis for the versatile patterning of the vocal CPG circuit?

How do global fluctuations generated by the autonomic nervous system influence the dynamics of CPGs and other neural structures related to vocal production?

In what manner, and at what levels, does social reinforcement influence the neural dynamics and sensorimotor integration required for vocal production?

Acknowledgments

We thank Daniel Takahashi and Diana Liao for their influence in shaping the ideas presented in this manuscript. Midshipman grunts courtesy of Andy Bass. This work was supported by the National Institutes of Health-NINDS (R01NS054898 to A.A.G.).

Authors' Contributions

Y.S.Z. and A.A.G. wrote the manuscript.

Disclaimer Statement

The authors declare no competing interests.

References

1. Bass, A.H. (2014) Central pattern generator for vocalization: is there a vertebrate morphotype? *Curr. Opin. Neurobiol.* 28, 94–100
2. Hage, S.R. and Nieder, A. (2016) Dual neural network model for the evolution of speech and language. *Trends Neurosci.* 39, 813–829
3. Kiebel, S.J. *et al.* (2008) A hierarchy of time-scales and the brain. *PLoS Comput. Biol.* 4, e1000209
4. Flack, J.C. *et al.* (2013) Timescales, symmetry, and uncertainty reduction in the origins of hierarchy in biological systems. In *Cooperation and Its Evolution* (Sterelny, K. *et al.*, eds), pp. 45–74, MIT Press
5. Deng, L. *et al.* (2006) Structured speech modeling. *IEEE Trans. Audio Speech Lang. Process.* 14, 1492–1504
6. Dunbar, R.I. and Shultz, S. (2007) Evolution in the social brain. *Science* 317, 1344–1347
7. Gomez-Marin, A. and Ghazanfar, A.A. (2019) The life of behavior. *Neuron* 104, 25–36
8. Fischer, J. and Price, T. (2017) Meaning, intention, and inference in primate vocal communication. *Neurosci. Biobehav. Rev.* 82, 22–31
9. Barrett, L.F. and Simmons, W.K. (2015) Interoceptive predictions in the brain. *Nat. Rev. Neurosci.* 16, 419–429
10. Azzalini, D. *et al.* (2019) Visceral signals shape brain dynamics and cognition. *Trends Cogn. Sci.* 23, 488–509
11. Hall, M.L. (2009) A review of vocal duetting in birds. *Adv. Stud. Behav.* 40, 67–121
12. Logue, D.M. and Gammon, D.E. (2004) Duet song and sex roles during territory defence in a tropical bird, the black-bellied wren, *Thryothorus fasciatoventris*. *Anim. Behav.* 68, 721–731
13. Haimoff, E.H. (1986) Convergence in the duetting of monogamous Old World primates. *J. Hum. Evol.* 15, 51–59
14. Müller, A.E. and Anzenberger, G. (2002) Duetting in the Titi monkey *Callicebus cupreus*: structure, pair specificity and development of duets. *Folia Primatol.* 73, 104–115
15. Greenfield, M.D. and Rand, A.S. (2000) Frogs have rules: selective attention algorithms regulate chorusing in *Physalaemus pustulosus* (Leptodactylidae). *Ethology* 106, 331–347
16. Okobi, D.E. *et al.* (2019) Motor cortical control of vocal interaction in neotropical singing mice. *Science* 363, 983–988
17. Choi, J.Y. *et al.* (2015) Cooperative vocal control in marmoset monkeys via vocal feedback. *J. Neurophysiol.* 114, 274–283
18. Takahashi, D.Y. *et al.* (2013) Coupled oscillator dynamics of vocal turn-taking in monkeys. *Curr. Biol.* 23, 2162–2168
19. Arlet, M. *et al.* (2015) Grooming-at-a-distance by exchanging calls in non-human primates. *Biol. Lett.* 11, 20150711
20. Kulahci, I.G. *et al.* (2015) Lemurs groom-at-a-distance through vocal networks. *Anim. Behav.* 110, 179–186
21. Demartsev, V. *et al.* (2018) Vocal turn-taking in meerkat group calling sessions. *Curr. Biol.* 28, 3661–3666.e3663
22. Dunbar, R.I. (1993) Coevolution of neocortical size, group size and language in humans. *Behav. Brain Sci.* 16, 681–694
23. Ghazanfar, A.A. and Rendall, D. (2008) Evolution of human vocal production. *Curr. Biol.* 18, R457–R460
24. Bass, A.H. *et al.* (2008) Evolutionary origins for social vocalization in a vertebrate hindbrain–spinal compartment. *Science* 321, 417–421
25. Luthe, L. *et al.* (2000) Neuronal activity in the medulla oblongata during vocalization. A single-unit recording study in the squirrel monkey. *Behav. Brain Res.* 116, 197–210
26. Hage, S.R. and Jurgens, U. (2006) Localization of a vocal pattern generator in the pontine brainstem of the squirrel monkey. *Eur. J. Neurosci.* 23, 840–844
27. Jurgens, U. (2002) Neural pathways underlying vocal control. *Neurosci. Biobehav. Rev.* 26, 235–258
28. Hage, S.R. and Jurgens, U. (2006) On the role of the pontine brainstem in vocal pattern generation: a telemetric single-unit recording study in the squirrel monkey. *J. Neurosci.* 26, 7105–7115

29. Schmidt, M.F. *et al.* (2012) Breathing and vocal control: the respiratory system as both a driver and a target of telencephalic vocal motor circuits in songbirds. *Exp. Physiol.* 97, 455–461
30. Barlow, S.M. *et al.* (2010) Central pattern generators for orofacial movements and speech. In *Handbook of Mammalian Vocalization* (Brudzynski, S.M., ed.), pp. 351–369, Academic Press
31. Hage, S.R. (2010) Localization of the central pattern generator for vocalization. In *Handbook of Behavioral Neuroscience* (19), pp. 329–337
32. Kittelberger, J.M. *et al.* (2006) Midbrain periaqueductal gray and vocal patterning in a teleost fish. *J. Neurophysiol.* 96, 71–85
33. Esposito, A. *et al.* (1999) Complete mutism after midbrain periaqueductal gray lesion. *Neuroreport* 10, 681–685
34. Jürgens, U. (1994) The role of the periaqueductal grey in vocal behaviour. *Behav. Brain Res.* 62, 107–117
35. Tschida, K. *et al.* (2019) A specialized neural circuit gates social vocalizations in the mouse. *Neuron* 103, 459–472.e4
36. Holstege, G. and Subramanian, H.H. (2016) Two different motor systems are needed to generate human speech. *J. Comp. Neurol.* 524, 1558–1577
37. Owren, M.J. *et al.* (2011) Two organizing principles of vocal production: implications for nonhuman and human primates. *Am. J. Primatol.* 73, 530–544
38. Grillner, S. and Wallen, P. (1985) Central pattern generators for locomotion, with special reference to vertebrates. *Annu. Rev. Neurosci.* 8, 233–261
39. Ijspeert, A.J. *et al.* (2007) From swimming to walking with a salamander robot driven by a spinal cord model. *Science* 315, 1416–1420
40. Bianchi, A.L. *et al.* (1995) Central control of breathing in mammals: neuronal circuitry, membrane properties, and neurotransmitters. *Physiol. Rev.* 75, 1–45
41. Jean, A. (2001) Brain stem control of swallowing: neuronal network and cellular mechanisms. *Physiol. Rev.* 81, 929–969
42. Sasahara, K. *et al.* (2015) A rhythm landscape approach to the developmental dynamics of birdsong. *J. R. Soc. Interface* 12, 20150802
43. Pomberger, T. *et al.* (2018) Precise motor control enables rapid flexibility in vocal behavior of marmoset monkeys. *Curr. Biol.* 28, 788–794.e783
44. Fitch, W.T. and Hauser, M.D. (1995) Vocal production in nonhuman primates – acoustics, physiology, and functional constraints on honest advertisement. *Am. J. Primatol.* 37, 191–219
45. van den Berg, J. (1958) Myoelastic-aerodynamic theory of voice production. *J. Speech Lang. Hear. Res.* 1, 227–244
46. Mindlin, G.B. and Laje, R. (2006) *The Physics of Birdsong*, Springer
47. Titze, I.R. and Martin, D.W. (1998) *Principles of Voice Production*, ASA
48. Elemans, C.P.H. *et al.* (2015) Universal mechanisms of sound production and control in birds and mammals. *Nat. Commun.* 6, 8978
49. Takahashi, D.Y. *et al.* (2015) The developmental dynamics of marmoset monkey vocal production. *Science* 349, 734–738
50. Teramoto, Y. *et al.* (2017) Vocal development in a Waddington landscape. *eLife* 6, e20782
51. Amador, A. *et al.* (2013) Elemental gesture dynamics are encoded by song premotor cortical neurons. *Nature* 495, 59–64
52. Zhang, Y.S. *et al.* (2019) Vocal state transition through laryngeal development. *Nat. Commun.* 10, 4592. https://doi.org/10.1038/s41467-019-12588-6
53. Chagnaud, B.P. *et al.* (2011) Vocalization frequency and duration are coded in separate hindbrain nuclei. *Nat. Commun.* 2, 346
54. Kelso, J.S. (1997) *Dynamic Patterns: The self-Organization of Brain and Behavior*, MIT Press
55. Thelen, E. and Smith, L.B. (1996) *A Dynamic Systems Approach to the Development of Cognition and Action*, MIT Press
56. Zhang, Y.S. and Ghazanfar, A.A. (2018) Vocal development through morphological computation. *PLoS Biol.* 16, e2003933
57. Laje, R. *et al.* (2002) Neuromuscular control of vocalizations in birdsong: a model. *Phys. Rev. E Stat. Nonlinear Soft Matter Phys.* 65, 051921
58. Gardner, T. *et al.* (2001) Simple motor gestures for birdsongs. *Phys. Rev. Lett.* 87, 208101
59. Chandrasekaran, C. *et al.* (2009) The natural statistics of audiovisual speech. *PLoS Comput. Biol.* 5, e1000436
60. Greenberg, S. *et al.* (2003) Temporal properties of spontaneous speech—a syllable-centric perspective. *J. Phon.* 31, 465–485
61. Lameira, A.R. *et al.* (2015) Speech-like rhythm in a voiced and voiceless orangutan call. *PLoS One* 10, e116136
62. Toyoda, A. *et al.* (2017) Speech-like orofacial oscillations in stump-tailed macaque (*Macaca arctoides*) facial and vocal signals. *Am. J. Phys. Anthropol.* 164, 435–439
63. Terleph, T.A. *et al.* (2018) An analysis of white-handed gibbon male song reveals speech-like phrases. *Am. J. Phys. Anthropol.* 166, 649–660
64. Bergman, T.J. (2013) Speech-like vocalized lip-smacking in geladas. *Curr. Biol.* 23, R268–R269
65. Ghazanfar, A.A. *et al.* (2012) Cineradiography of monkey lip-smacking reveals putative precursors of speech dynamics. *Curr. Biol.* 22, 1176–1182
66. Ghazanfar, A.A. and Takahashi, D.Y. (2014) The evolution of speech: vision, rhythm, cooperation. *Trends Cogn. Sci.* 18, 543–553
67. Moore, J.D. *et al.* (2014) How the brainstem controls orofacial behaviors comprised of rhythmic actions. *Trends Neurosci.* 37, 370–380
68. Chiao, G. *et al.* (1994) Neuronal activity in nucleus ambiguus during deglutition and vocalization in conscious monkeys. *Exp. Brain Res.* 100, 29–38
69. Larson, C.R. *et al.* (1994) Modification in activity of medullary respiratory-related neurons for vocalization and swallowing. *J. Neurophysiol.* 71, 2294–2304
70. Wright, J.M. *et al.* (2012) α-and β-Adrenergic receptors differentially modulate the emission of spontaneous and amphetamine-induced 50-kHz ultrasonic vocalizations in adult rats. *Neuropsychopharmacology* 37, 808
71. Borjon, J.I. *et al.* (2016) Arousal dynamics drive vocal production in marmoset monkeys. *J. Neurophysiol.* 116, 753–764
72. Henderson, A. *et al.* (1966) Sequential temporal patterns in spontaneous speech. *Lang. Speech* 9, 207–216
73. Julien, C. (2006) The enigma of Mayer waves: facts and models. *Cardiovasc. Res.* 70, 12–21
74. Lu, C.L. and Jürgens, U. (1993) Effects of chemical stimulation in the periaqueductal gray on vocalization in the squirrel monkey. *Brain Res. Bull.* 32, 143–151
75. Morris, K.F. *et al.* (2010) Respiratory and Mayer wave-related discharge patterns of raphé and pontine neurons change with vagotomy. *J. Appl. Physiol.* 109, 189–202
76. Buzsaki, G. (2004) Neuronal oscillations in cortical networks. *Science* 304, 1926–1929
77. Palva, S. and Palva, J.M. (2018) Roles of brain criticality and multiscale oscillations in temporal predictions for sensorimotor processing. *Trends Neurosci.* 41, 729–743
78. Fox, M.D. and Raichle, M.E. (2007) Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nat. Rev. Neurosci.* 8, 700
79. Chan, A.W. *et al.* (2015) Mesoscale infraslow spontaneous membrane potential fluctuations recapitulate high-frequency activity cortical motifs. *Nat. Commun.* 6, 7738
80. Leopold, D.A. *et al.* (2003) Very slow activity fluctuations in monkey visual cortex: implications for functional brain imaging. *Cereb. Cortex* 13, 422–433
81. Mitra, A. *et al.* (2018) Spontaneous infra-slow brain activity has unique spatiotemporal dynamics and laminar structure. *Neuron* 98, 297–305.e296
82. Mateo, C. *et al.* (2017) Entrainment of arteriole vasomotor fluctuations by neural activity is a basis of blood-oxygenation-level-dependent "resting-state" connectivity. *Neuron* 96, 936–948.e933
83. Zhang, Y.S. and Ghazanfar, A.A. (2016) Perinatally influenced autonomic nervous system fluctuations drive infant vocal sequences. *Curr. Biol.* 26, 1249–1260
84. MacNeilage, P.F. (2008) *The Origin of Speech*, Oxford University Press

85. Tchernichovski, O. *et al.* (2001) Dynamics of the vocal imitation process: how a zebra finch learns its song. *Science* 291, 2564–2569
86. Veit, L. *et al.* (2011) Learning to breathe and sing: development of respiratory-vocal coordination in young songbirds. *J. Neurophysiol.* 106, 1747–1765
87. Tchernichovski, O. *et al.* (2000) A procedure for an automated measurement of song similarity. *Anim. Behav.* 59, 1167–1176
88. Sacks, H. *et al.* (1978) A simplest systematics for the organization of turn-taking for conversation. *Language* 50, 696–735
89. Katsu, N. *et al.* (2019) Temporal adjustment of short calls according to a partner during vocal turn-taking in Japanese macaques. *Curr. Zool.* 65, 99–105
90. Pika, S. *et al.* (2018) Taking turns: bridging the gap between human and animal communication. *Proc. R. Soc. B Biol. Sci.* 285, 20180598
91. Fröhlich, M. (2017) Taking turns across channels: conversation-analytic tools in animal communication. *Neurosci. Biobehav. Rev.* 80, 201–209
92. Logue, D.M. and Krupp, D.B. (2016) Duetting as a collective behavior. *Front. Ecol. Evol.* 4, 7
93. Levinson, S.C. (2016) Turn-taking in human communication – origins and implications for language processing. *Trends Cogn. Sci.* 20, 6–14
94. Kuriki, S. *et al.* (1999) Motor planning center for speech articulation in the normal human brain. *Neuroreport* 10, 765–769
95. Takahashi, D.Y. *et al.* (2012) A computational model for vocal exchange dynamics and their development in marmoset monkeys. In *2012 IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL)*, pp. 1–2, IEEE
96. Brumm, H. and Slater, P.J. (2006) Animals can vary signal amplitude with receiver distance: evidence from zebra finch song. *Anim. Behav.* 72, 699–705
97. Johnson, C.J. *et al.* (1981) Effects of interpersonal distance on children's vocal intensity. *Child Dev.* 52, 721–723
98. Pelegrín-García, D. *et al.* (2011) Vocal effort with changing talker-to-listener distance in different acoustic environments. *J. Acoust. Soc. Am.* 129, 1981–1990
99. Liao, D.A. *et al.* (2018) Internal states and extrinsic factors both determine monkey vocal production. *Proc. Natl. Acad. Sci. U. S. A.* 115, 3978–3983
100. Mampe, B. *et al.* (2009) Newborns' cry melody is shaped by their native language. *Curr. Biol.* 19, 1994–1997
101. Colombelli-Négrel, D. *et al.* (2012) Embryonic learning of vocal passwords in superb fairy-wrens reveals intruder cuckoo nestlings. *Curr. Biol.* 22, 2155–2160
102. Van Puyvelde, M. *et al.* (2015) Whose clock makes yours tick? How maternal cardiorespiratory physiology influences newborns' heart rate variability. *Biol. Psychol.* 108, 132–141
103. Feldman, R. *et al.* (2011) Mother and infant coordinate heart rhythms through episodes of interaction synchrony. *Infant Behav. Dev.* 34, 569–577
104. Wass, S.V. *et al.* (2019) Parents mimic and influence their infant's autonomic state through dynamic affective state matching. *Curr. Biol.* 29, 2415–2422.e2414
105. Kuhl, P.K. *et al.* (2003) Foreign-language experience in infancy: effects of short-term exposure and social interaction on phonetic learning. *Proc. Natl. Acad. Sci. U. S. A.* 100, 9096–9101
106. Goldstein, M.H. and Schwade, J.A. (2008) Social feedback to infants' babbling facilitates rapid phonological learning. *Psychol. Sci.* 19, 515–523
107. Goldstein, M.H. *et al.* (2003) Social interaction shapes babbling: testing parallels between birdsong and speech. *Proc. Natl. Acad. Sci. U. S. A.* 100, 8030–8035
108. West, M.J. and King, A.P. (1988) Female visual displays affect the development of male song in the cowbird. *Nature* 334, 244–246
109. Carouso-Peck, S. and Goldstein, M.H. (2019) Female social feedback reveals non-imitative mechanisms of vocal learning in zebra finches. *Curr. Biol.* 29, 631–636.e3
110. Chen, Y. *et al.* (2016) Mechanisms underlying the social enhancement of vocal learning in songbirds. *Proc. Natl. Acad. Sci. U. S. A.* 113, 6641–6646
111. Prat, Y. *et al.* (2017) Crowd vocal learning induces vocal dialects in bats: playback of conspecifics shapes fundamental frequency usage by pups. *PLoS Biol.* 15, e2002556
112. Prat, Y. *et al.* (2015) Vocal learning in a social mammal: demonstrated by isolation and playback experiments in bats. *Sci. Adv.* 1, e1500019
113. Gultekin, Y.B. and Hage, S.R. (2017) Limiting parental feedback disrupts vocal development in marmoset monkeys. *Nat. Commun.* 8, 14046
114. Gultekin, Y.B. and Hage, S.R. (2018) Limiting parental interaction during vocal development affects acoustic call structure in marmoset monkeys. *Sci. Adv.* 4, eaar4012
115. Takahashi, D.Y. *et al.* (2017) Vocal learning via social reinforcement by infant marmoset monkeys. *Curr. Biol.* 27, 1844–1852. e1846
116. Takahashi, D.Y. *et al.* (2016) Early development of turn-taking with parents shapes vocal acoustics in infant marmoset monkeys. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 371 2015.0370
117. Cho, J.R. *et al.* (2017) Dorsal raphe dopamine neurons modulate arousal and promote wakefulness by salient stimuli. *Neuron* 94, 1205–1219.e1208
118. Matthews, Gillian A. *et al.* (2016) Dorsal raphe dopamine neurons represent the experience of social isolation. *Cell* 164, 617–631
119. Tanaka, M. *et al.* (2018) A mesocortical dopamine circuit enables the cultural transmission of vocal behaviour. *Nature* 563, 117–120
120. Gavrilov, N. *et al.* (2017) Functional specialization of the primate frontal lobe during cognitive control of vocalizations. *Cell Rep.* 21, 2393–2406
121. Kelley, D.B. (1986) Neuroeffectors for vocalization in *Xenopus laevis*: hormonal regulation of sexual dimorphism. *J. Neurobiol.* 17, 231–248
122. Marler, P. *et al.* (1988) The role of sex steroids in the acquisition and production of birdsong. *Nature* 336, 770
123. Bercovitch, F.B. *et al.* (1995) The endocrine stress response and alarm vocalizations in rhesus macaques. *Anim. Behav.* 49, 1703–1706
124. Hage, S.R. (2018) Dual neural network model of speech and language evolution: new insights on flexibility of vocal production systems and involvement of frontal cortex. *Curr. Opin. Behav. Sci.* 21, 80–87
125. Ackermann, H. *et al.* (2014) Brain mechanisms of acoustic communication in humans and nonhuman primates: an evolutionary perspective. *Behav. Brain Sci.* 37, 529–546
126. Zhao, W. *et al.* (2019) Inception of memories that guide vocal learning in the songbird. *Science* 366, 83–89
127. Feng, N.Y. and Bass, A.H. (2016) "Singing" fish rely on circadian rhythm and melatonin for the timing of nocturnal courtship vocalization. *Curr. Biol.* 26, 2681–2689