ANNALS OF THE NEW YORK ACADEMY OF SCIENCES
Special Issue: *Speech Rhythm in Ontogenic, Phylogenetic, and Glossogenetic Development*
REVIEW

Rhythm in speech and animal vocalizations: a cross-species perspective

Andrea Ravignani,[1,2] Simone Dalla Bella,[3,4,5,a] Simone Falk,[3,6,a] Christopher T. Kello,[7,a] Florencia Noriega,[8,9,a] and Sonja A. Kotz[3,10,11]

[1]Artificial Intelligence Laboratory, Vrije Universiteit Brussel, Brussels, Belgium. [2]Institute for Advanced Study, University of Amsterdam, Amsterdam, the Netherlands. [3]International Laboratory for Brain, Music and Sound Research (BRAMS), Montréal, Quebec, Canada. [4]Department of Psychology, University of Montreal, Montréal, Quebec, Canada. [5]Department of Cognitive Psychology, Warsaw, Poland. [6]Laboratoire de Phonétique et Phonologie, UMR 7018, CNRS/Université Sorbonne Nouvelle Paris-3, Institut de Linguistique et Phonétique générales et appliquées, Paris, France. [7]Cognitive and Information Sciences, University of California, Merced, California. [8]Chair for Network Dynamics, Center for Advancing Electronics Dresden (CFAED), TU Dresden, Dresden, Germany. [9]CODE University of Applied Sciences, Berlin, Germany. [10]Basic and Applied NeuroDynamics Laboratory, Faculty of Psychology and Neuroscience, Department of Neuropsychology and Psychopharmacology, Maastricht University, Maastricht, the Netherlands. [11]Department of Neuropsychology, Max-Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

Addresses for correspondence: Andrea Ravignani, Artificial Intelligence Laboratory, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium. andrea.ravignani@gmail.com; Sonja A. Kotz, Basic and Applied NeuroDynamics Laboratory, Faculty of Psychology and Neuroscience, Department of Neuropsychology and Psychopharmacology, Maastricht University, Universiteitssingel 40, 6200 MD Maastricht, the Netherlands. sonja.kotz@maastrichtuniversity.nl

Why does human speech have rhythm? As we cannot travel back in time to witness how speech developed its rhythmic properties and why humans have the cognitive skills to process them, we rely on alternative methods to find out. One powerful tool is the comparative approach: studying the presence or absence of cognitive/behavioral traits in other species to determine which traits are shared between species and which are recent human inventions. Vocalizations of many species exhibit temporal structure, but little is known about how these rhythmic structures evolved, are perceived and produced, their biological and developmental bases, and communicative functions. We review the literature on rhythm in speech and animal vocalizations as a first step toward understanding similarities and differences across species. We extend this review to quantitative techniques that are useful for computing rhythmic structure in acoustic sequences and hence facilitate cross-species research. We report links between vocal perception and motor coordination and the differentiation of rhythm based on hierarchical temporal structure. While still far from a complete cross-species perspective of speech rhythm, our review puts some pieces of the puzzle together.

Keywords: speech rhythm; hierarchical; timing; time perception; rhythm cognition; bioacoustics

Introduction

The comparative, cross-species approach is a powerful method to understand the evolution of cognitive and communicative traits in our species.[1] Here, we use this approach to study vocal rhythm and investigate which similar traits can be found in other species, so to understand what is broadly shared across, for example, mammals, tetrapods, or vertebrates. We review several studies in the literature that are usually unconnected (see Table 1 for a glossary). In particular, we discuss the production and perception of rhythmic patterns in nonhuman species and in human development. We summarize several methods to measure rhythmic structure in vocalizations produced by humans and other species. We discuss the neural bases of speech rhythm, attempting to draw comparative links.

Rhythm processing requires (but is not limited to) the ability to produce and perceive individual

[a]These authors contributed equally.

doi: 10.1111/nyas.14166

Table 1. Alphabetical glossary of key terminology[a]

Term	Definition
Auditory grouping	While sounds are perceived as coming from a single source, similar sounds tend to group together
Babbling	Vocal experimentation and sound practice found in the early developmental stage of select species. In human infants, babbling precedes the emergence of first words (starting around 4–5 months of age, with variable duration, and lasting until the second and even third year of life)
Beat	Psychological process resulting in the perception and extraction of an isochronous pulse (not necessarily present in the physical signal) from a rhythmic sequence. The perception of a beat can result from metrical expectations, associated with different embedded periodicities
Durational categories	Classification or perception of different temporal intervals not as a continuum but as each belonging to a particular category. An indirect way of detecting durational categories in a dataset of durations is testing whether the distribution of durations is not uniform but multimodal, where each mode approximates the prototype of one category
Frontostriatal brain circuitry	Neural pathway(s) connecting cortical frontal lobe brain areas with the striatum, including the putamen and caudate nucleus
Grouping	Process of building basic patterns of sound events based on acoustic features, such as stress, loudness alternation, pitch variation, durational relationships, etc.
Hierarchical temporal structure	Clustering of signal events in time, such as peaks in the amplitude envelope, where smaller clusters are nested within larger clusters across timescales
Iamb	"Metrical form in speech alternating a weak (unstressed) syllable with a strong (stressed) syllable"
(Inter-event) interval	"Temporal duration encompassed by two events." Examples of events are the onset and offset times of a vocalization
Isochronous	"A series of events repeating at a constant rate"
Meter	Hierarchical organization of patterns of events based on spectral and structural properties
Rhythm	"Pattern of events in time," possibly with "a specific succession of durations" and accents as seen in speech
Stuttering	A developmental speech fluency disorder, which interrupts the rhythmic flow of speech and communication[9]
Timing	Perception and production of temporal relationships
Trochee	"Metrical form in speech alternating a strong syllable with a weak syllable"

[a]Definitions in quotation marks are reproduced verbatim from Refs. 7 and 8.

temporal intervals. Hence, we set off by briefly discussing literature on interval timing. Thorough treatments of interval timing are available elsewhere (see Refs. 2–6).

Human and nonhuman studies of vocal rhythm

Animal timing from the psychophysics literature

Timing and time perception has a long tradition in animal research. Rats, mice, pigeons, fish, and some primate species have all been studied in terms of their ability to estimate or reproduce temporal intervals in the millisecond-to-second range. A general finding of these studies is that predictions from the so-called scalar expectancy theory hold across species and domains (with some exceptions, see Ref. 4). Simply put, the theory predicts that timing sensitivity, corresponding to the accuracy in perceiving or reproducing time intervals, is inversely proportional to interval duration: animals, including humans, estimate longer intervals with less accuracy.

Research on timing and time perception is necessary but not sufficient to understand rhythm (for a parallel in music, see Ref. 10). In fact, timing research often focuses on the production or perception of individual time intervals. Rhythm instead focuses on patterns of temporal events, whose building blocks are individual time intervals. This is similar to perceiving individual frequencies that can be understood as a building block for perceiving the harmonies and timbres of sounds. However, the perception of individual frequencies is not sufficient to understand the perception of multiple overlapping frequencies. For instance, if humans

Ann. N.Y. Acad. Sci. xxxx (2019) 1–20 © 2019 The Authors. *Annals of the New York Academy of Sciences* published by Wiley Periodicals, Inc. on behalf of New York Academy of Sciences

listen to two tones at particular frequencies, they will hear a third one (the "missing fundamental"[11]) whose presence cannot be predicted by basic psychophysical data on individual frequency. Similarly, as reviewed here, simple timing patterns can be layered to create the perception of rhythmic structure that is not simply determined by its components (e.g., see Ref. 12).

Comparative experiments: training and testing animals on rhythm, meter, and prosody

Rhythm involves a series of time intervals, often at multiple timescales, that can combine to produce a hierarchical metrical structure.[13] The perception of rhythmic features, such as grouping, is usually studied in operant experiments.[14] Rats, budgerigars, and zebra finches have recently been tested in their capacity for metrical grouping. Rats, like humans, are capable of using pitch alternation in sound sequences to group them as trochees (high−low pairs); in contrast, unlike humans, rats cannot use durational alternation in sound sequences to group them as iambs (short−long pairs).[15] Zebra finches show similar discrimination capacities as rats.[16] Follow-up work has shown that, if thoroughly trained for a durational alternation, rats can indeed discriminate between iambs and trochees.[17] In a related experiment, although within a different setup, budgerigars could distinguish between iambic and trochaic meter, but required, to succeed, more than one cue among pitch, duration, loudness, and vowel quality.[18] Testing rats with stimuli identical to those used for budgerigars revealed a very different result: unlike parrots, rodents need all four cues to discriminate between prosodic patterns. Of these four cues, one was purely (duration) and two partly (loudness and pitch) rhythmic.[17] In summary, the ability to perceive and discriminate a simple metrical structure has been observed in several species, but more research is needed to fully determine the extent of these abilities across species.

Spontaneous individual vocal rhythms: what kind of temporal structure is contained in animals' call sequences and songs?

Several species have also been found to produce spontaneous vocal rhythms and are therefore particularly promising for comparative human–animal research,[19] including (1) laboratory rodents, such as mice, because biomedical research has thoroughly mapped their neurobiology;[20] (2) nonhuman primates, because of their phylogenetic relatedness to humans;[21] (3) songbirds, in particular, zebra finches, because they are an established model species for avian vocal flexibility and learning;[22] and (4) vocal learning mammals, such as seals, elephants, and bats, because they are the closest vocal learning animals to humans.[23] Below, we will briefly discuss examples of vocal rhythms in rodents, nonhuman primates, songbirds, and mammalian vocal learners.

Measures of vocal rhythm can be found in "transition probabilities"; these probabilities have become popular in birdsong and language research. Given a sequence of events, including event types A, B, C, etc., the transition probability between event types A and B is the probability that A is followed by B. A common application of this concept is to study sequences of discrete elements: in birdsong research, a low transition probability between notes A and B means that note A is rarely followed by B. With respect to measuring vocal rhythms, transition probabilities can be used in the temporal domain (see Ref. 8), with the caveat that time is continuous, so some discretization is necessary. For instance, one could calculate the transition probability from short to long call durations and vice versa (Fig. 1); if the former was high and the latter was low, short calls would often be followed by long calls, but long calls rarely would be followed by short calls.

In mice, ultrasonic vocalizations exhibit quite stable transition probabilities in durations, especially for short-short and long-long transitions.[20] This temporal structure could be summarized by matrices as those in Figure 1C with values close to 1 in the diagonal, and values close to 0 elsewhere. Mice vocalizations also appear to be temporally organized in a hierarchical fashion.[20] However, more (operant) work is needed to test the existence of hierarchical organization at a neurocognitive level, that is, temporal events structured at different time scales, possibly embedding one level into the higher one, as opposed to structure appearing hierarchical as a byproduct of serial behavior or anatomical constraints (see Refs. 24 and 25 for a parallel discussion about recursion and cognition in behavior). Finally, taking a developmental perspective, the rhythm of mice vocalizations as pups is predictive of vocal rhythms in the same mice as adults.[20] These results suggest that there may be sensitive phases for rhythm development in infant mice, assuming that

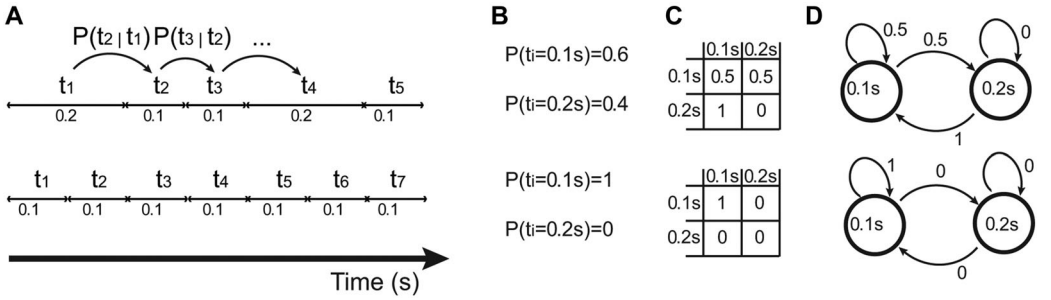

Figure 1. **Some of the possible ways of representing temporal patterns (bottom: isochronous, top: nonisochronous). (A) Time series of intervals, inducing transition probabilities, such as $P(t_2|t_1)$, which means the probability that the (inter-event) interval t_2 of length *x* ms follows an interval t_1 of length *y* milliseconds. (B) Individual probabilities of occurrence of a particular durational interval. (C) Transition matrices based on the transition probabilities described before. (D) A probabilistic finite state machine, which can also generate durational patterns as those seen in A and summarized in the transition matrices in C. Figure reproduced verbatim from Ref. 8, an Open Access article distributed under the terms of the Creative Commons Attribution License (CC BY).**

some learning is involved in mice vocal rhythms. Unlike common mice, Alston's singing mice do perform vocal duets: in this neotropical rodent, call timing is controlled by different neural circuitry depending on whether singing is performed in isolation or socially.[26–28]

Research on individual vocal rhythms in nonhuman primates is scarce: most work has investigated either group vocal rhythms[29,30] or individual nonvocal rhythms.[31–33] Focusing on individual vocal rhythms, early descriptive work remarked temporal regularities in gelada monkeys' vocalizations,[34] a claim which is intriguing but purely descriptive, unfortunately not supported by quantitative data or statistical inference. More recent work in macaques and orangutans noted a 5-Hz isochronous pattern during lip-smacking, facial movement, or vocalization.[35–37] Primate perspectives on speech rhythm can be found elsewhere (e.g., see Ref. 21), but it is clear that we do need to understand more about vocal rhythms in our closest living relatives, the primates.

Zebra finches have long been a model for vocal learning, though research in this species has historically focused on the spectral and combinatorial domains, rather than the temporal and rhythmic domains. The temporal dimension of zebra finches' songs has been explored recently, and the rhythms of their songs appear to be characterized by plasticity and interindividual variability, which are connected to learning and often in contrast to stereotypical calling.[38] Past methods used in birdsong research concluded strong stereotypy in zebra finches' rhythms, but this conclusion may have stemmed from analytical methods that focus on short time scales to the neglect of longer time scales.[38] We now find that zebra finches can flexibly time their unlearned calls.[39] In addition, zebra finches' songs exhibit a form of isochronous regularity: syllable onsets coincide, more often than not, with regular "beats" of an idealized isochronous grid (Fig. 2).[22] Evidence for the interplay between plasticity and regularity makes intuitive sense: an underlying isochronous grid can provide anchor points in time from which songs can be learned, structured, and flexibly varied.

The isochrony detection technique used in zebra finches has also been applied to a bat species

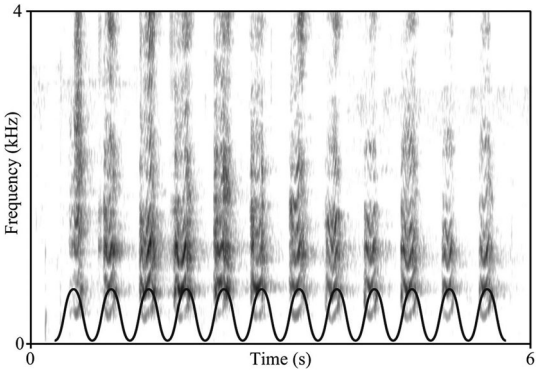

Figure 2. **Spectrogram showing the isochronous barking of a California sea lion. One possible way to visually detect isochronous regularity is to superimpose a metronomic grid, like the regular sinusoidal function shown here, to the spectrogram. Figure reproduced verbatim from Ref. 8, an Open Access article distributed under the terms of the Creative Commons Attribution License (CC BY).**

capable of vocal production learning. Surprisingly, the neotropical bat *Saccopteryx bilineata* exhibits isochronous rhythms not only in its echolocation calls, but also in male vocal displays (i.e., "songs") and pups' call sequences.[40] In addition, a (post-hoc) superimposed metronomic grid exhibited a tempo, which matched the wing-beat of the animals.[40] The finding of temporal similarities between vocal (calls) and nonvocal (wing-beat) rhythms in bats is one more comparative piece of evidence of the cross-modality of rhythm, which in humans entails audition, vision, movement, etc. (e.g., dance).

Collecting recordings of some species can be particularly challenging, as in adult male seals that "sing" underwater.[41] Underwater recordings consist of few microphones recording multiple, nonvisible sources, making it often difficult to attribute vocalizations to individuals.[41,42] In these cases, one can still indirectly test for rhythmicity by probing whether temporal structures of different song elements covary.[43] Alternatively, seal pups are often easier to record, as they mostly vocalize on land (as opposed to underwater). Analyses of temporal features of seal pups' vocalizations have shown some regularities and the emergence of durational categories over development, but larger sample sizes are needed to generalize (Fig. 3; see Ref. 44).

Babbling and stuttering

Vocal learners, such as harbor seals and *S. bilineata* bats,[45,46] are useful model species to study the potential interplay between rhythm and vocal learning ontogeny.[38,44] They may also give directions for research on early human vocal production. For instance, some species share similarities with humans in their earliest vocalizations called "babbling." Across different language contexts, human infants in their first year of life vocalize rhythmic chunks of repeated and then varied syllables (like da-da-da; e.g., see Refs. 47 and 48). Babbling features native language sound production and imitation of prosodic aspects of adult speech (e.g., see Refs. 48 and 49). Babbling as a kind of vocal play and imitation of adult calls, barks, trills, and songs is also observed in infant and juvenile pygmy marmosets,[50] sac-winged bat pups,[40,51] giant otters,[52] and zebra finches.[53] According to the Frame and Content theory,[54] babbling in human infants is a rhythmic motor training, which lays the grounds for basic syllable structure. Infants learn that vocalizing at different times during quasiperiodic cycles of mandibular opening and closure results in vowels, at maximal mandibular opening, and consonants, at maximal mandibular closure. However, it is still unclear how these early syllable rhythms in babbling contribute to later adult rhythms or later language capacities in general.[55] Results from nonhuman animals suggest that animal babbling is not linked in a simple way to later adult vocal production. For example, female sac-winged baby bats, during the babbling period, produce adult male songs and trills without producing them as adults.[51] In zebra finches, different brain circuits are active during juvenile babbling and later adult song production.[53] These results may inspire future research into human infant babbling by investigating the potential significance of early imitation capacity for later speech perception or the potential differences between neuronal circuits that are active during babbling and early/later speech production.

Finally, potential parallels between humans and nonhuman animals can be investigated in rhythm disorders in early vocal production. Stuttering, for example, is a speech fluency disorder typically emerging between the second and fourth year of life in humans.[56] Children show untypical disfluencies during speech production, such as silent blocks, syllable and sound repetitions, and prolongations. Stuttering-like behavior can also be observed in songbirds, such as zebra finches.[57,58] In humans, recent research links the disturbance of the rhythmic flow of speech in stuttering to faulty auditory−motor learning and erroneous temporal predictions, potentially originating from altered connectivity in subcortical−cortical timing circuits.[59–61] Interestingly, animal research points to a prominent role of basal ganglia dysfunction in stuttering zebra finches,[62] paralleling findings of impaired basal ganglia functioning in human children and adults who stutter.[63,64] More research is needed to unravel similarities in how rhythm contributes to the development of skilled speech motor control across species.

Vocal−motor entrainment in music and speech

Humans are generally highly skilled at processing complex temporal patterns, such as music and

speech. Most humans can perceive the regular beat of music, and detect stress in spoken utterances.[65] Notably, beat perception is often accompanied by a synchronized motor response. For example, the temporal features of musical patterns and their temporal regularity are particularly conducive to movement.[66] Our proclivity to move to music manifests when we move to its beat, which can happen spontaneously or deliberately, by foot or hand tapping, and in dance or synchronized walking. These skills are widespread in the general population.[67,68] A compelling body of evidence from experimental psychology and cognitive neuroscience indicates that rhythm and movement are tightly linked.[69–72] Matching movements to a beat is possible because the temporal dynamics of rhythmic sound lead to the perception of the beat,[73] a process linked to internal neurocognitive self-sustained oscillations.[74,75] The underlying process, called *entrainment*, generates temporal expectancies, which drive motor control, by allowing the alignment of movements to anticipated event times.

The human ability to spontaneously synchronize to music[76] and to simpler rhythmic stimuli[77] contrasts with the lack of evidence on spontaneous motor synchronization in spoken utterances (though see Refs. 78 and 79). Yet, there is some evidence that the accent structure of rhythmical speech, as found in, for example, children's poetry, can entrain movement even when participants are not explicitly instructed to move to speech.[66] Prominences in speech (stress patterns), akin to musical beats, may indeed represent a target of synchronized movement. Speech rhythm is particularly salient in poems, songs, and children's games ("metrical speech"), characterized by words and phrases that are molded into regularly recurring metrical patterns.[80,81] For example, in English or German, rhythm is conveyed by accentual patterns whereby strong and weak positions are filled by prominent (i.e., stressed, see Ref. 82) and non-prominent (i.e., unstressed) syllables. Like in music, speech patterns can evoke a subjective impression of isochrony.[83] This observation is striking, though, given that interstress intervals are typically quite variable in speech (coefficients of variations $>30\%$ of the average interstress interval[84,85]), as compared with expressive music (around 10–30% for inter-beat intervals in performed expressive music[86]). Moreover, speech meter in conversational speech is clearly less strict and regular than musical meter.[87] Higher regularity is found in metrical speech, however, such as poetry,[88–91] and group speech production, such as prayers and chanting (i.e., choral speaking[92]).

In spite of the higher variability of temporal patterns in speech compared with music, the temporal dynamics of metrical speech can still induce expectations about upcoming events.[93,94] The substrate of this mechanism lies in the quasi-rhythmic properties of the speech signal that engage oscillatory behavior in the brain.[95] Like music, speech patterns are thus capable of driving dynamic attending,[73] underpinned by neurocognitive self-sustained oscillations,[93,96] which phase-lock to the temporal dynamics of syllabic nuclei in speech.[5,94,97,98] Accurate prediction of the next verbal event (a stressed syllable) affords a certain degree of motor synchronization to the prominent stress pattern in speech, as observed in recent finger tapping studies.[85,99,100] Interestingly, concurrent synchronized movement can enhance verbal expectations, as found in prosodically diverse languages, such as German (a lexical stress language) and French (a non-stress language).[99,101] For example, finger tapping aligned to accented syllables of spoken utterances benefits the encoding and detection of subtle word changes.[99,101] Thus, coupling movement to the temporal dynamics of metrical speech can enhance verbal processing and memorization. This effect is reminiscent of more ecological situations in which hand clapping or stamping to metrical speech (e.g., children's lore)[102] is part of games that may enhance children's social and verbal skills.[103] Moreover, the aforementioned effects of synchronized movement may pave the way to innovative rhythm-based interventions currently under investigation for fostering language acquisition and learning in developmental populations with speech and language disorders, such as dyslexic[104] or autistic children.[105]

The link between rhythm and movement and the ability to couple movement to auditory prominences is ubiquitous in humans. This ability requires little learning, is associated with high flexibility, as humans can adapt to a wide range of tempos even quite far from their preferred movement rates, occurs within a variety of rhythmic stimuli, simple and complex rhythms, and also crossmodally.[106,107] The question as to whether

Ann. N.Y. Acad. Sci. xxxx (2019) 1–20 © 2019 The Authors. *Annals of the New York Academy of Sciences* published by Wiley Periodicals, Inc. on behalf of New York Academy of Sciences

other species are capable of synchronization to the beat, and if so, to what extent as compared with humans, has fueled research in the last decade. One intriguing hypothesis (the vocal learning—beat perception and synchronization hypothesis[87,108]) postulates that synchronization to a beat is a byproduct of the vocal learning mechanisms that are shared by several bird and mammal species, including humans. In keeping with this hypothesis, a strong link between motor and auditory brain areas is expected to underpin both vocal production and synchronization. There is evidence that these abilities are linked in humans.[109] This hypothesis received support by the finding that nonhuman animal species, namely sulfur-crested cockatoos[110,111] and other bird species that are vocal learners, can also synchronize.[112,113] Motor synchronization in vocal learners is quite flexible (i.e., adapting to a wider range of tempos), occurs with complex auditory signals, and is crossmodal,[110,111] thus displaying some of the properties of human synchronization. Recent evidence shows, however, that synchronization to a beat may extend to nonvocal learning species. There is evidence that a chimpanzee can tap above chance, though quite inflexibly, with a 600-ms metronome,[114,115] and a California sea lion can bob her head to the beat of a variety of auditory stimuli.[116,117] Thus, whether synchronization to beat is selectively associated with vocal learning across species is still an open question.[118,119]

Interactive rhythms during human speech development

As there are some parallels in the development of human and animal rhythmic vocalizations, the question arises to what extent vocal rhythms in interaction are also comparable across species. Only certain animals, though, utter specific pup-directed vocalizations by making them shorter, more repetitive, or more specialized than adult-directed vocalizations (see, Ref. 120 for male zebra finches; Ref. 121, for free-ranging female rhesus macaque; and Ref. 122, for North Atlantic right whale mother–calf pairs). In humans, vocal style changes dramatically in infant–adult interaction. There are at least two functions of rhythmic structure of human infant–adult interaction that may play a pivotal role for infants and young children to acquire speech and language skills: (1) rhythmic vocalizations and imitation subserving *communicative alignment* in early parent–infant interaction, and (2) *temporal predictions* about linguistic structure derived from rhythmic cues in infant-directed communication. These aspects could be further explored in the animal domain.

Older interlocutors across cultures display a distinct infant-directed speech register, no matter if they are female, male, parent, sibling, or a stranger. Their utterances are shorter and higher pitched, and they contain distinct melodic contours and more repetition.[123–125] These salient alterations in speech, as well as songs, chants, and rhythmic vocal play[126,127] contribute to an overall highly rhythmic character of infant-directed communication. According to evolutionary hypotheses, rhythmic traits of adult–infant interaction are an ancestral part of human child-rearing practice, whose primary goal was to foster infants' and mothers' capacity to affiliate and align with each other and to develop mutual understanding and experience sharing beyond symbolic communication.[128] In line with this idea, Jaffe and colleagues[129] found that infant's attachment (at 12 months) is predicted by temporal coordination patterns in turn-taking with familiar and especially unfamiliar adults at 4 months of age. Overall, from the age of 2 months on, turn-taking structure between mother and infant vocalizations is already observable with only a 30–40% overlap between reciprocal vocalizations. The most frequent exchange structure features two to three turns, and pauses (gaps) under 1 second.[130,131]

Mutual alignment is considered a key aspect of adult verbal interaction.[132] Early rhythmic and temporal alignment between mothers and preverbal infants could hence be a precursor of the sophisticated verbal alignment skills needed in later life. In a 2-year longitudinal study on mother−infant coordination, Abney and colleagues[133] identified a hierarchical temporal structure as a key aspect of alignment patterns in mother−infant interaction. Hierarchical temporal structure (see below) was extracted from the waxing and waning of amplitude in the acoustic signal thereby identifying hierarchically nested bouts of temporal clusters across timescales. Generally, mothers emphasize the hierarchical temporal structure of infant-directed speech and singing compared with adult-directed communication.[134] Abney and colleagues[133] found that the hierarchical temporal structure of mothers and infants' vocalizations was well aligned during

mother–infant interactions. In addition, preverbal vocalizations of infants (e.g., vocalic and syllabic sequences) were overall temporally better coordinated with their mother's vocalizations than any nonverbal vocalization (e.g., laughter and cries).

Adult listeners use temporal predictions to better attend to and process phonological, lexical, semantic, and syntactic structure in their interlocutor's speech.[135–138] Higher repetitiveness, greater metrical stability, shorter utterances, and enhanced utterance-final lengthening in infant-directed speech are all temporal cues, which could help infants to generate *temporal predictions about upcoming linguistic structure*. In infant-directed speech, temporal cues particularly emphasize phrase boundary information through enhanced preboundary lengthening and longer postboundary pauses.[124,139] These cues provided by adults help infants to direct their attention to phrase edges. Indeed, infants at 8 months of age more easily segment words at phrase-final versus medial positions in speech.[140] Infants are also able to generate temporal predictions from a regular beat structure, such as found in music.[141–143] As a musical stimulus, infant-directed singing may particularly support beat-related predictions in caregiver–infant communication. It features clearer metrical structure than speech[126] and therefore may better direct infants' attention toward words associated with a beat. First results showed a trend for infants at 11 months of age to process word-related information in song better than in speech.[144] Rhythmic structure may also facilitate infants' discrimination of infant-directed singing from speech,[145] and foster the development of more musical and more speech-related sound processing. Yet, unique contributions of the rhythm of singing to infants' language and musical skills still await further investigation.

Techniques for quantifying rhythmic structure

Rhythm as temporal hierarchy in human and nonhuman vocalization

Rhythm and timing in speech, as in complex animal vocalizations, have hierarchical temporal structure. We know where this structure comes from in speech: units of perception and production are built up hierarchically.[146] Phonemes are grouped together to form syllables, which are grouped together to form words, which are grouped together to form phrases, and so on. We have many ways of knowing about units of speech perception and production, including behavioral and neural experiments, linguistic inquiry, and our own intuitions. We know much less about the hierarchical structure of animal vocalizations because we do not have the luxury of linguistic inquiry and intuition and experimental methods are limited relative to speech. As a result, we do not have *a priori* units of perception and production that we can map onto recordings of animal vocalizations, as we can with speech recordings, although various methods for segmenting animal vocalizations have been studied.[147–149]

Regardless of whether we know the units or not, we can measure and quantify hierarchical temporal structure directly in the acoustic signal that results from vocalization. This structure is different from symbolic hierarchical expressions, as in linguistic research symbolic expressions do not specify timing or temporal durations. Linguistic representations must be elaborated to include temporal structure, which is influenced by the durations of linguistic units, and also prosodic factors such as stress and intonation. Most generally, smaller linguistic units correspond with shorter units of perception or production, which are sequenced together to form larger units, with the possibility of longer durations between larger units. This elaboration only indicates probabilistic, relative relations in temporal structure (Fig. 3), but it leads us to quantitative metrics that we can measure in the acoustic signal.

In particular, we can quantify the *degree* of hierarchical temporal structure. By doing so, we can show an indirect relationship with the putative linguistic units expressed as nested speech units, without needing to map individual units onto specific segments of the speech signal. With this indirect relationship established, we can quantify the degree of hierarchical temporal structure in recordings of animal vocalizations using the same method, and thereby compare the rhythmic structures of speech and animal vocalizations to learn more about their similarities and differences.

Hierarchical temporal structure in the acoustic signals of speech and animal vocalizations can be measured through the amplitude envelope,[150] which quantifies the bursts and lulls in acoustic energy. The timing and duration of the bursts are captured by clustering in peak events in the

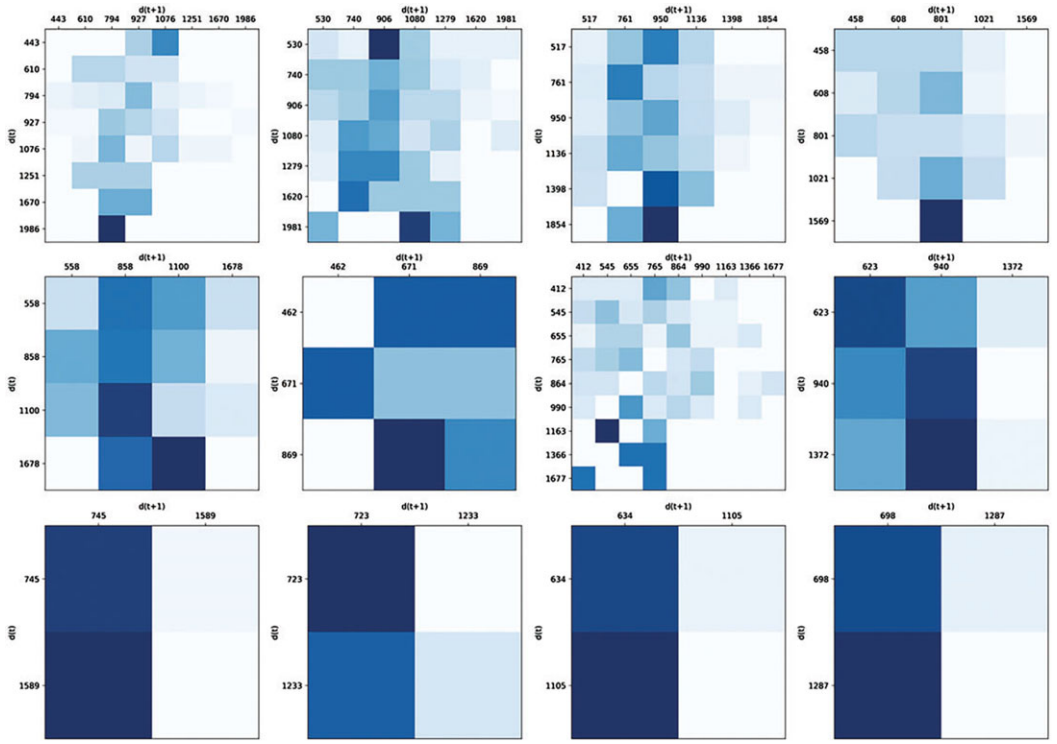

Figure 3. **Transition matrices for one individual seal pup across days, each showing the probability that a call of a given duration is followed by another call of the same (diagonal) or another duration. Each matrix represents 1 day, with calendar days progressing from left to right and top to bottom. Each row and column of one matrix represents the centroid of a durational category in milliseconds: leftmost columns and upmost rows are shorter (400—700 ms) categories; the further down and right, the longer the category. Shades of blue represent transition probabilities; that is, the probability, within a sequence of seal pup calls, that a specific category on the vertical axis is followed by a specific category on the horizontal axis. Darker blue corresponds to a higher transition probability; for instance, in the bottom-right matrix, the dark square means that an element of a durational category centered at 1287 ms is very likely to be followed by an element of a durational category centered at 698 ms, but very unlikely to be followed by an element of the same category centered at 1287 milliseconds. Notice how, over days, the number of categories shrinks (though see Ref. 44) and the transitions from one to the other become more predictable. Figure cropped from Ref. 44, an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License; additional details in the original paper.**

amplitude envelope across a wide range of timescales. Smaller clusters are nested within larger clusters across timescales, and nesting can be quantified using Allan factor (AF) variance.[151] AF variance measures temporal clustering of events at a given timescale by measuring variance in event counts for adjacent time windows. AF functions are computed by measuring AF variance across a range of timescales, to gauge the degree to which clustering increases across timescales.

Falk and Kello[134] analyzed recordings of German mothers either singing a song or telling a story to their infants, compared with the same mothers singing or storytelling to adults. AF functions showed a greater degree of nested clustering in infant-directed versus adult-directed speech and song, particularly in timescales ranging from hundreds of milliseconds to more than 10 seconds. Follow-up analyses showed that AF functions reflected the greater degree of prosodic exaggeration in infant-directed speech. Prosodic exaggeration is known to increase the variability in the acoustic durations of units of speech production, and AF variance captures this variability across a range of timescales. The authors analyzed hand-coded durations of linguistic units ranging from syllables to words and phrases to overall variability in speaking rate. The slopes of AF functions

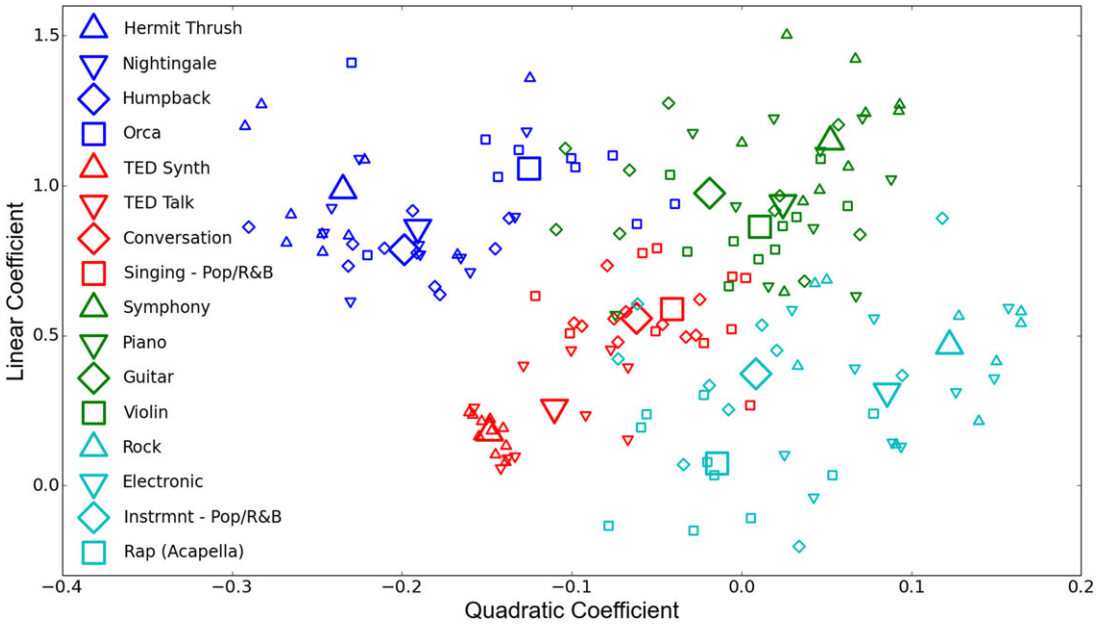

Figure 4. **Each AF function for each sound recording from Ref. 152 was quantified in terms of a linear coefficient that corresponded to the degree of hierarchical temporal structure, and a quadratic coefficient that corresponded to the amount and direction of change in hierarchical temporal structure as a function of timescale. Four different categories of sound recordings were analyzed—animal vocalization, speech, classical music, and popular music—each with four subcategories (see the legend). The scatter plot shows a point representing the curvature (*x*-axis) and slope (*y*-axis) of the AF function for each of 10-example recordings per subcategory. One can see that the four main categories have mostly distinct hierarchical temporal structures as measured by AF functions, and in some cases, the subcategories are further distinguished within the main categories. Large symbols indicate the centroid of each subcategory. Figure reproduced verbatim from Ref. 152.**

were shown to account for significant variability in all these linguistic units. This result provides supporting evidence that hierarchical temporal structure maps onto linguistic units as they are expressed in speech production.

Kello and colleagues[152] also applied AF analysis to a wide range of speech, music, and animal vocalization recordings. Results were consistent with those of Falk and Kello[134] and also extended them by showing that nested clustering is enhanced by musical composition. Moreover, AF functions were classified using support vector machines, and the results revealed a natural taxonomy of complex acoustic signals, where recordings within a given category yielded AF functions that could be separated from other categories (Fig. 4).

The AF function category most relevant to the current discussion corresponds to *conversational interactions*. In Ref. 152 and two subsequent studies (Ref. 153 and Schneider, Ramirez-Aristizabal, Gavilan, and Kello, unpublished data) dozens of recordings of various types of conversational interactions, in both English and Spanish, have all yielded AF functions with a common slope and bend. The same slope and bend was found for jazz improvisations, which have been likened to conversations.[154] Most notably, recordings of animal vocalizations from killer whales communicating with each other in pods yielded AF functions with the same basic shape as those for recordings of conversational interactions. Animal vocalizations from humpback whales, nightingales, and hermit thrushes were different—these animals do not use their songs in the service of vocal interactions, and AF functions did not follow the pattern common to conversational interactions. Instead, these other animal vocalizations fell into their own distinct pattern, closer to a monologue or solo song in terms of hierarchical temporal structure. Ravignani and colleagues[44] applied the same AF analysis to recordings of harbor seal pups, a species that employs vocal interactions similar to killer whales, and these recordings also yielded the same communicative AF function shape.

The observed commonality in so many different recordings of communicative interactions suggests an intriguing hypothesis: both human and nonhuman communicative interactions of all kinds may manifest the same, unique kind of hierarchical temporal structure that depends on the particular communicative function, and less so on the species or means of sound production. Such a result, if corroborated, would indicate that speech, music, and animal vocalizations all follow a common pattern of hierarchical temporal structure. If true, this could have implications for both segmentation and learning of incoming communicative stimuli.

Rhythms as distributions of inter-event intervals

Wildlife recordings often have contributions from diverse sounds thereby obscuring the signal of interest. Having a low signal to noise ratio limits the applicability of techniques acting directly over the waveform. In these situations, an alternative is to annotate the recordings with the onset or offset times and investigate the temporal structure of these events.[44,155,156] In this section, we present another method (in addition to the AF analyses above and other techniques described in detail elsewhere[7]) for characterizing and comparing temporal patterns in a series of events that was recently proposed in Ref. 157. This technique consists of characterizing timing as distributions of the logarithm of the inter-event intervals (IEIs) and comparing the distributions with the symmetric Kullback–Leibler divergence (sKL-divergence). We explore the scope of this method based on its strength to describe temporal structures in four datasets: random, isochronous, hierarchical, and speech. We start by describing the datasets and then discuss this technique.

The datasets consist of time series of events represented by pulses. The isochronous series has a pulse every 0.2 seconds. The random series is a Poisson process with a rate $\lambda = 12$ pulses per second. The hierarchical series is composed of triplets of double pulses. All these artificial sets—random, isochronous, and hierarchical—are 10-s long with a sampling rate of 1 kHz (i.e., a temporal resolution of 1 millisecond). Additionally, the isochronous and the hierarchical series are jittered with Gaussian noise with a standard deviation of 0.005 seconds. The speech dataset comes from "The north wind and sun dataset" corpus, consisting of recordings of the fable in 18 different languages. For our analysis, we use the position of the syllable centers annotated by Jadoul and colleagues.[158] This annotated speech dataset contains the syllable centers of all 18 languages.

The distribution of the logarithm of the IEI (log-IEI) highlights the typical IEI in the time series (Fig. 5). The isochronous signal has an event every 0.2 s, so its distribution of log-IEI shows a single peak at 0.2 s (Fig. 5). The distribution of the random series ranges over various time scales around $1/\lambda$. The hierarchical series presents three peaks—the shortest corresponds to the IEI within the double pulses, the middle one corresponds to the IEI between the double pulses within the triplet, and the third one corresponds to the intervals between the triplets. The IEIs of the speech dataset are spread with 50% of the IEIs between 0.19 and 0.25 seconds.

The downside of the distributions of IEI is that they are not sensitive to high-order temporal

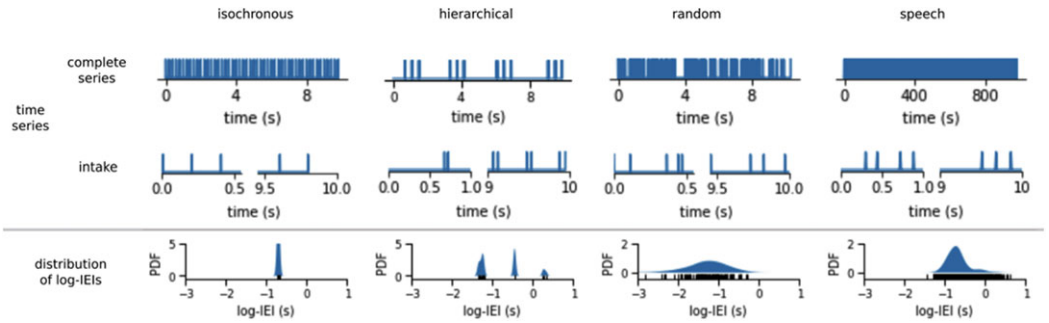

Figure 5. **Characterization of the temporal structure of four time series (columns)—isochronous, hierarchical, random, and speech—with distributions of inter-event intervals (IEIs). The top two rows show the full time series and a "zoom" on (i.e., intake of) the beginning and end of the time series. The bottom row shows the distribution of the logarithm base 10 of the IEI (log-IEI).**

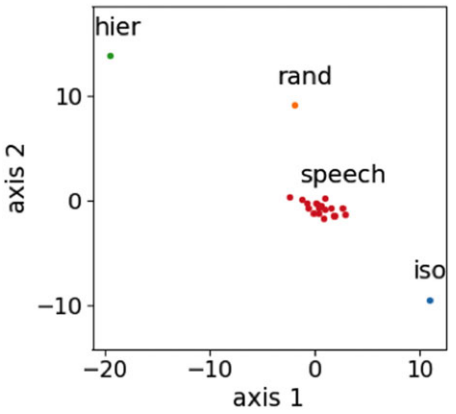

Figure 6. Two-dimensional scaling of the distances between the distributions of inter-event intervals (IEI). Pairwise distances computed using the symmetric Kullback–Leibler divergence (sKL-divergence) and scaled to a two-dimensional space using Scikit-learn MDS method.[159] This method takes pairwise distances between elements and locates them in an n-dimensional space. Here, we have chosen $n = 2$. There is one point for each distribution coming from the datasets isochronous (iso), hierarchical (hier), random (rand), and the 18 speech datasets. Neither the location of the points nor the axes are relevant per se; what counts here are the distances between the points—these are such that they try to preserve the sKL-divergences.

structures, as randomizing the IEIs would yield the same distributions. The advantage of these distributions is that they are easy to interpret and can be compared using the sKL-divergence (Fig. 6).[157] The Kullback–Leibler divergence measures the similarity between two probability distributions. The divergence is smaller the more similar two probability distributions are, being zero only for identical distributions.

Comparing the distributions of our datasets using the sKL-divergence, we obtain 210 distances. By projecting these distances to a two-dimensional space, we observe that the distributions of the speech datasets are more similar to each other than each is to one of the other datasets (Fig. 6). This approach is described in detail in Ref. 157.

The IEI-related distributions can be visualized and used for computations by employing either the IEI itself, or its logarithm (as we do here). Using the logarithm is advantageous because it scales the IEI according to their magnitude, thereby dealing with different time scales simultaneously. This logarithmic scaling may also be quite plausible neurobiologically, at least for single intervals and musical rhythm.[3,160] However, sometimes, one may prefer to work with the IEI directly, for instance, for dealing with negative intervals arising from overlapping calls from different signalers.[161,162] The fact that this method can work with both IEIs and their logarithm makes it flexible to work with different types of datasets.

Time and rhythm: linking neural systems and behavior

From cross-species comparisons to evolutionary inference

Rhythms comprise features, such as intensity and duration, that fluctuate at somewhat equal time intervals in a complex and continuous auditory signal, such as human speech and music. Yet, an unresolved topic in time and rhythm research is why and how the ability to process temporal and rhythmic structure emerged in humans.[118,119,163] One idea ties rhythm processing to social synchronization across a number of species (for a review, see Ref. 13). Other research exploring the neurocognitive function of time and rhythm processing also points toward similarities of rhythmic and structural properties in speech and music[164,165] that are primarily denoted in vocal learners.[108] This coevolution of properties might reside in, and still rely on, frontostriatal brain circuitry[94,166] (see also, Ref. 167 on the evolution of structure), a system that engages in and monitors the acquisition of hierarchical pattern formation in multiple domains. This brain system also tags specific longer scale temporal attributes and synchronizes to temporal and structural cues found in speech and music (e.g., Refs. 5 and 168). However, it remains a mystery (1) how humans derived more complex structures in speech, language, and music from the temporal and sequencing properties of the frontostriatal system and (2) where the structural and functional boundaries lie within this system that separate human and nonhuman species. Consequently, a comparative approach to evaluate the computational proximity and extent of temporal and rhythmic sequences in species relying on an extended frontostriatal circuitry is called for.[13]

Human neurocognitive architecture of time and rhythm processing

The spatiotemporal properties of auditory signals reach the thalamus and cerebellum in the earliest stages of auditory processing. While precise and

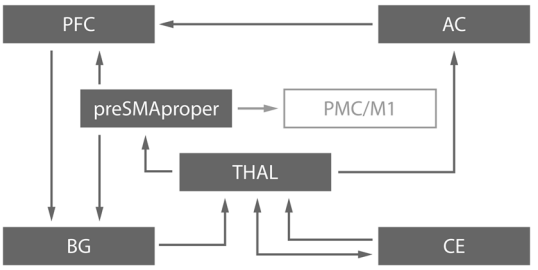

Figure 7. Cortico–subcortico–cortical neural circuitry underlying time and rhythm perception and production. CE, cerebellum; THAL, thalamus; AC, auditory cortex; PFC, prefrontal cortex; BG, basal ganglia; PMC, premotor cortex; M1, primary motor cortex; pre-SMA, presupplementary motor area. See also, Ref. 177.

continuous spatiotemporal information is sent via the thalamus to the auditory cortices where sensory and memory processes are initiated, the cerebellum projects salient events encoded in the auditory signal (onsets, offsets, and sharp energy changes) via the thalamus directly to frontal cortices (e.g., presupplementary motor area (SMA)). This latter trajectory is relevant for two reasons: (1) it attracts and maintains attention to salient changes in the auditory signal and (2) based on this dynamic attention modulation, prepares the frontostriatal system for the encoding of temporal interevent relations (intervals) that form the basic segmentation unit of sequences. The encoding and evaluation of the temporal cohesion of sequences require working memory and rely on the prefrontal cortex (PFC),[169] where temporal and memory information integrates.[5]

In production, the generation of a sequence engages the PFC. To start and continue this process, an interface of the pre-SMA and frontostriatal circuitry acts as a "pacemaker" and stabilizes a temporal grid for auditory sequence processing. Sequences adhere to a temporal architecture that integrates fast, short-range transitioning temporal events via the cerebellum and slower large-range intervals via the striatum (see also, Ref. 170 for different terminology). The actual initiation, timing, and triggering of auditory—motor sequences as for example found in speech engage the SMA-proper that controls these processes (e.g., see Refs. 171 and 172), followed by the premotor and primary motor cortices for the execution of sequences.

In sum, the described temporal architecture (Fig. 7) composed of fast, short-range and slower, long-range temporal information contributes both to perception and production of auditory—motor sequences, such as found in human speech and music.[5,173,174] Empirical evidence confirms that the ascribed temporal properties form the basis of temporal pattern formation found in simple and complex rhythm processing, which also relies on the same neural frontostriatal architecture as temporal processing per se.[175,176]

Shared neural circuitry, but where are cross-species boundaries?

While there is now ample evidence that several species of birds and mammals, including some nonhuman primates, rely on comparable frontostriatal circuitry (e.g., see Ref. 178) to acquire and produce simple and slightly more complex temporally structured sequences, vocal learning alone does not suffice to acquire hierarchical temporal structures found in human speech and music.[179] For example, zebra finches produce temporally structured syllable sequences[22] and can perceptually group auditory input.[16] Rhesus monkeys can produce single intervals and synchronize to a metronome,[180] while macaques display auditory grouping.[181,182] So far, though there is no evidence that any one of these species can form hierarchical temporal structure as found in human speech and music. One explanation, while still speculative, could be that the strict serial order of events in time does not yet define rule-based behavior beyond local dependencies.[94] Second, complex temporal and rule structure building may rely on an intricate relationship between frontostriatal and frontocerebellar circuitry, where the expansion of the neocerebellum reciprocally pushed the evolution of neocortex, such as the PFC.[183,184] This latter structural development is considered crucial for hierarchical structure building. Consequently, investigation of this frontostriatocerebellar interface in species producing and perceiving basic temporal structure is required to understand similarities and differences between simple and hierarchical temporal structure building in humans and other species.

General discussion and conclusions

Connecting fields, disciplines, and methods

This paper is a first attempt to summarize multiple approaches to understand the comparative and evolutionary nature of human speech rhythm. We

reviewed how animals from different taxonomic groups can produce and perceive temporal and rhythmic patterns with features relevant to human rhythm. We examined parallels between human and animal infant vocal production and interactive rhythms in order to better understand contributions of rhythm to human speech development. We found that social interaction in several species, including humans, produces a common pattern of temporal structure in vocalizations. We compared several techniques to measure temporal and rhythmic structure, both in human speech and animal vocalizations. We concluded by discussing the neural circuitry underlying speech rhythm and their relationship with nonvocal motor actions.

Admittedly, however, there is a big disconnect among, at least, five areas of scientific knowledge and research: (1) what we know of human speech rhythm, especially from a developmental perspective, (2) knowledge on how animals produce and perceive sounds which can be related to human speech rhythm, (3) techniques we can use to measure vocal rhythms behaviorally, within humans and across species, (4) comparative work on rhythmic, nonvocal movement, and (5) how our knowledge of the human nervous system relates to that of other species with respect to speech rhythm.

Future work

We suggest that future work should keep the current sparseness of these five approaches in mind and actively build bridges across them. Pragmatically, this would translate into designing experiments which span two, or more, of the five still loosely connected areas discussed above. For instance, researchers in animal bioacoustics (area 2 above) could perform analyses that are tightly matched to human vocal development (area 1). Likewise, behavioral metrics of vocal rhythmicity (area 4) would be even more valuable if usable as potential markers of neural processes or pathologies (area 5).

A recent, successful example of this kind of multidisciplinary work focused on rhythmic interactivity in a rodent.[26] Temporal features of songs of Alston's singing mice were investigated. The question was whether these features varied between isolated and interactive singing and were partly controlled by the cortex (as opposed to fully originating from subcortical structures of the brain). The authors of the study provided behavioral, pharmacological, and neural evidence that rhythmic vocal interactivity in Alston's mice stems from a cortico−subcortical circuitry. How this murine circuitry relates to the human circuitry in Figure 7 is still unknown. In addition to introducing this broad methodological approach,[26] the research tackled comparatively the question of how cortical control of vocalizations and turn-taking evolved in humans. Therefore, apart from its scientific contribution, this study shows that combining two or more of the areas above is indeed possible.

Two additional areas for future research, not discussed here, include the biology−culture interplay in, and the genetics of, speech rhythm. Studying the biology−culture interface can be used to reconcile old, unproductive nature versus nurture debates by potentially showing how cognitive biases and cultural transmission interact to deliver the rhythmic structure of speech. One possible method to study the biology−culture interplay in the laboratory is the iterated learning paradigm, where each participant learns a behavior which was produced (and modified) by a previous participant who learnt it the same way. Iterated learning experiments have been done to better understand linguistic morphology,[185] poetry,[186] and musical rhythm.[160,187−189] Iterated learning experiments where participants imitate and transmit nonsense syllable sequences[190] could be used to show whether, and if so how, cultural transmission amplifies domain-general biases resulting in rhythmic patterns of speech. This iterated learning approach can be integrated with neurophysiological measures.[191] Complementarily, tools and methodologies from genetics can be used to map the population genotypes to behavioral variability in rhythmic traits.[19,192] Initial work has been undertaken in special populations (e.g., those affected by Williams syndrome[193,194]), but could be extended to the whole population of one species, humans, or otherwise.

To conclude, the field of comparative rhythm research is rapidly growing and needs a multidisciplinary approach. This research field offers many low-hanging fruits, which are ready to be seized by colleagues interested in joining us.

Acknowledgments

A.R. has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant

agreement No. 665501 with the Research Foundation Flanders (FWO), Pegasus2 Marie Curie Fellowship 12N5517N. S.A.K (Co-PI) was supported by a Grant from the Portuguese Science Foundation (PTDC/MHC-PCN/0101/2014). C.K. received funding from National Science Foundation awards 1529127 and 1633722. S.D.B. received funding from Natural Sciences and Engineering Research Council of Canada (NSERC) award RGPIN-2019-05453.

Competing interests

The authors declare no competing interests.

References

1. Ravignani, A., H. Honing & S.A. Kotz. 2017. The evolution of rhythm cognition: timing in music and speech. *Front. Hum. Neurosci.* **11**. https://doi.org/10.3389/fnhum.2017.00303.
2. Buhusi, C.V. & W.H. Meck. 2005. What makes us tick? Functional and neural mechanisms of interval timing. *Nat. Rev. Neurosci.* **6:** 755.
3. Gibbon, J. 1977. Scalar expectancy theory and Weber's law in animal timing. *Psychol. Rev.* **84:** 279.
4. Lejeune, H. & J. Wearden. 2006. Scalar properties in animal timing: conformity and violations. *Q. J. Exp. Psychol.* **59:** 1875–1908.
5. Schwartze, M. & S.A. Kotz. 2013. A dual-pathway neural architecture for specific temporal prediction. *Neurosci. Biobehav. Rev.* **37:** 2587–2596.
6. Teki, S. 2016. A citation-based analysis and review of significant papers on timing and time perception. *Front. Neurosci.* **10**. https://doi.org/10.3389/fnins.2016.00330.
7. Ravignani, A. & P. Norton. 2017. Measuring rhythmic complexity: a primer to quantify and compare temporal structure in speech, movement, and animal vocalizations. *J. Lang. Evol.* **2**. https://doi.org/10.1093/jole/lzx002.
8. Ravignani, A. & G. Madison. 2017. The paradox of isochrony in the evolution of human rhythm. *Front. Psychol.* **8**. https://doi.org/10.3389/fpsyg.2017.01820.
9. American Psychiatric Publications. 2013. Diagnostic and statistical manual of mental disorders (DSM-5®). American Psychiatric Publications.
10. Bouwer, F.L., H. Honing & H.A. Slagter. 2019. Beat-based and memory-based temporal expectations in rhythm: similar perceptual effects, different underlying mechanisms. https://doi.org/10.1101/613398.
11. Smith, J.C., J.T. Marsh, S. Greenberg & W.S. Brown. 1978. Human auditory frequency-following responses to a missing fundamental. *Science* **201:** 639–641.
12. Trehub, S.E. & L.A. Thorpe. 1989. Infants' perception of rhythm: categorization of auditory sequences by temporal structure. *Can. J. Psychol.* **43:** 217.
13. Kotz, S., A. Ravignani & W.T. Fitch. 2018. The evolution of rhythm processing. *Trends Cogn. Sci.* **22:** 896–910.
14. ten Cate, C. & M. Spierings. 2019. Rules, rhythm and grouping: auditory pattern perception by birds. *Anim. Behav.* **151:** 249–257.
15. de la Mora, D.M., M. Nespor & J.M. Toro. 2013. Do humans and nonhuman animals share the grouping principles of the iambic-trochaic law? *Atten. Percept. Psychophys.* **75:** 92–100.
16. Spierings, M., J. Hubert & C. ten Cate. 2017. Selective auditory grouping by zebra finches: testing the iambic–trochaic law. *Anim. Cogn.* **20:** 665–675.
17. Toro, J.M. & M. Nespor. 2015. Experience-dependent emergence of a grouping bias. *Biol. Lett.* **11:** 20150374.
18. Hoeschele, M. & W.T. Fitch. 2016. Phonological perception by birds: budgerigars can perceive lexical stress. *Anim. Cogn.* **19:** 643–654.
19. Ravignani, A. 2019. Rhythm and synchrony in animal movement and communication. *Curr. Zool.* **65:** 77–81.
20. Castellucci, G.A., D. Calbick & D. McCormick. 2018. The temporal organization of mouse ultrasonic vocalizations. *PLoS One* **13:** e0199929.
21. Ghazanfar, A.A. 2013. Multisensory vocal communication in primates and the evolution of rhythmic speech. *Behav. Ecol. Sociobiol.* **67:** 1441–1448.
22. Norton, P. & C. Scharff. 2016. "Bird song metronomics": isochronous organization of zebra finch song rhythm. *Front. Neurosci.* **10**. https://doi.org/10.3389/fnins.2016.00309.
23. Ravignani, A. 2019. Humans and other musical animals. *Curr. Biol.* **29:** R271–R273.
24. Martins, M.D. 2012. Distinctive signatures of recursion. *Philos. Trans. R. Soc. B Biol. Sci.* **367:** 2055–2064.
25. Fitch, W. & M.D. Martins. 2014. Hierarchical processing in music, language, and action: Lashley revisited. *Ann. N.Y. Acad. Sci.* **1316:** 87–104.
26. Okobi, D.E., A. Banerjee, A.M. Matheson, *et al.* 2019. Motor cortical control of vocal interaction in neotropical singing mice. *Science* **363:** 983–988.
27. Banerjee, A., S.M. Phelps & M.A. Long. 2019. Singing mice. *Curr. Biol.* **29:** R190–R191.
28. Hage, S.R. 2019. Precise vocal timing needs cortical control. *Science* **363:** 926–927.
29. De Gregorio, C., A. Zanoli, D. Valente, *et al.* 2018. Female indris determine the rhythmic structure of the song and sustain a higher cost when the chorus size increases. *Curr. Zool.* **65:** 89–97.
30. Takahashi, D.Y., D.Z. Narayanan & A.A. Ghazanfar. 2013. Coupled oscillator dynamics of vocal turn-taking in monkeys. *Curr. Biol.* **23:** 2162–2168.
31. Dufour, V., N. Poulin, C. Curé & E.H. Sterck. 2015. Chimpanzee drumming: a spontaneous performance with characteristics of human musical drumming. *Sci. Rep.* **5:** 11320.
32. Dufour, V., C. Pasquaretta, P. Gayet & E.H. Sterck. 2017. The extraordinary nature of Barney's drumming: a complementary study of ordinary noise making in chimpanzees. *Front. Neurosci.* **11**. https://doi.org/10.3389/fnins.2017.00002.
33. Ravignani, A., V.M. Olivera, B. Gingras, *et al.* 2013. Primate drum kit: a system for studying acoustic pattern production by non-human primates using acceleration and strain sensors. *Sensors* **13:** 9790–9820.
34. Richman, B. 1987. Rhythm and melody in gelada vocal exchanges. *Primates* **28:** 199–223.

35. Toyoda, A., T. Maruhashi, S. Malaivijitnond & H. Koda. 2017. Speech-like orofacial oscillations in stump-tailed macaque (*Macaca arctoides*) facial and vocal signals. *Am. J. Phys. Anthropol.* **164:** 435–439.
36. Lameira, A.R., M.E. Hardus, A.M. Bartlett, *et al*. 2015. Speech-like rhythm in a voiced and voiceless orangutan call. *PLoS One* **10:** e116136.
37. Morrill, R.J., A. Paukner, P.F. Ferrari & A.A. Ghazanfar. 2012. Monkey lipsmacking develops like the human speech rhythm. *Dev. Sci.* **15:** 557–568.
38. Hyland Bruno, J. & O. Tchernichovski. 2017. Regularities in zebra finch song beyond the repeated motif. *Behav. Processes*. https://doi.org/10.1016/j.beproc.2017.11.001.
39. Benichov, J.I., E. Globerson & O. Tchernichovski. 2016. Finding the beat: from socially coordinated vocalizations in songbirds to rhythmic entrainment in humans. *Front. Hum. Neurosci.* **10**. https://doi.org/10.3389/fnhum.2016.00255.
40. Burchardt, L.S., P. Norton, O. Behr, *et al*. 2019. General isochronous rhythm in echolocation calls and social vocalizations of the bat *Saccopteryx bilineata*. *R. Soc. Open Sci.* **6:** 181076.
41. Sabinsky, P.F., O.N. Larsen, M. Wahlberg & J. Tougaard. 2017. Temporal and spatial variation in harbor seal (*Phoca vitulina* L.) roar calls from southern Scandinavia. *J. Acoust. Soc. Am.* **141:** 1824–1834.
42. Kershenbaum, A., J.L. Owens & S. Waller. 2019. Tracking cryptic animals using acoustic multilateration: a system for long-range wolf detection. *J. Acoust. Soc. Am.* **145:** 1619–1628.
43. Ravignani, A. 2018. Comment on "Temporal and spatial variation in harbor seal (*Phoca vitulina L.*) roar calls from southern Scandinavia" [J. Acoust. Soc. Am. 141, 1824-1834 (2017)]. *J. Acoust. Soc. Am.* **143:** 1–5.
44. Ravignani, A., C. Kello, K. de Reus, *et al*. 2019. Ontogeny of vocal rhythms in harbour seal pups: an exploratory study. *Curr. Zool*. **65:** 107–120.
45. Ravignani, A., W.T. Fitch, F.D. Hanke, *et al*. 2016. What pinnipeds have to say about human speech, music, and the evolution of rhythm. *Front. Neurosci*. **10**. https://doi.org/10.3389/fnins.2016.00274.
46. Vernes, S.C. 2017. What bats have to say about speech and language. *Psychon. Bull. Rev.* **24:** 111–117.
47. Oller, D.K. 2000. *The Emergence of the Speech Capacity*. Psychology Press.
48. Vihman, M.M. 2014. *Phonological Development: the First Two Years*. Boston, MA: Wiley-Blackwell.
49. Esteve-Gibert, N. & P. Prieto. 2013. Prosody signals the emergence of intentional communication in the first year of life: evidence from Catalan-babbling infants. *J. Child Lang.* **40:** 919–944.
50. Snowdon, C.T. & A.M. Elowson. 2001. 'Babbling' in pygmy marmosets: development after infancy. *Behaviour* **138:** 1235–1248.
51. Knörnschild, M., O. Behr & O. von Helversen. 2006. Babbling behavior in the sac-winged bat (*Saccopteryx bilineata*). *Naturwissenschaften* **93:** 451–454.
52. Mumm, C.A. & M. Knörnschild. 2014. The vocal repertoire of adult and neonate giant otters (*Pteronura brasiliensis*). *PLoS One* **9:** e112562.
53. Aronov, D., A.S. Andalman & M.S. Fee. 2008. A specialized forebrain circuit for vocal babbling in the juvenile songbird. *Science* **320:** 630–634.
54. MacNeilage, P.F. & B.L. Davis. 1990. Acquisition of speech production: the achievement of segmental independence. In *Speech Production and Speech Modelling*. W.J. Hardcastle & A. Marchal, Eds.: 55–68. Springer.
55. McGillion, M., J.S. Herbert, J. Pine, *et al*. 2017. What paves the way to conventional language? The predictive value of babble, pointing, and socioeconomic status. *Child Dev.* **88:** 156–166.
56. Yairi, E., N.G. Ambrose, E.P. Paden & R.V. Watkins. 2005. *Early Childhood Stuttering for Clinicians by Clinicians*. Austin, TX: Pro-ed.
57. Kobayashi, K., H. Uno & K. Okanoya. 2001. Partial lesions in the anterior forebrain pathway affect song production in adult Bengalese finches. *Neuroreport* **12:** 353–358.
58. Leonardo, A. & M. Konishi. 1999. Decrystallization of adult birdsong by perturbation of auditory feedback. *Nature* **399:** 466.
59. Civier, O., D. Bullock, L. Max & F.H. Guenther. 2013. Computational modeling of stuttering caused by impairments in a basal ganglia thalamo-cortical circuit involved in syllable selection and initiation. *Brain Lang.* **126:** 263–278.
60. Etchell, A.C., B.W. Johnson & P.F. Sowman. 2014. Behavioral and multimodal neuroimaging evidence for a deficit in brain timing networks in stuttering: a hypothesis and theory. *Front. Hum. Neurosci.* **8:** 467.
61. Falk, S., T. Müller & S. Dalla Bella. 2015. Non-verbal sensorimotor timing deficits in children and adolescents who stutter. *Front. Psychol.* **6:** 847.
62. Kubikova, L., E. Bosikova, M. Cvikova, *et al*. 2014. Basal ganglia function, stuttering, sequencing, and repair in adult songbirds. *Sci. Rep.* **4:** 6590.
63. Chang, S.-E. & D.C. Zhu. 2013. Neural network connectivity differences in children who stutter. *Brain* **136:** 3709–3726.
64. Giraud, A.-L., K. Neumann, A.-C. Bachoud-Levi, *et al*. 2008. Severity of dysfluency correlates with basal ganglia activity in persistent developmental stuttering. *Brain Lang.* **104:** 190–199.
65. Phillips-Silver, J., P. Toiviainen, N. Gosselin, *et al*. 2011. Born to dance but beat deaf: a new form of congenital amusia. *Neuropsychologia* **49:** 961–969.
66. Dalla Bella, S., A. Białuńska & J. Sowiński. 2013. Why movement is captured by music, but less by speech: role of temporal regularity. *PLoS One* **8:** e71945.
67. Sowiński, J. & S. Dalla Bella. 2013. Poor synchronization to the beat may result from deficient auditory-motor mapping. *Neuropsychologia* **51:** 1952–1963.
68. Tranchant, P., D.T. Vuvan & I. Peretz. 2016. Keeping the beat: a large sample study of bouncing and clapping to music. *PLoS One* **11:** e0160178.
69. Chen, J.L., V.B. Penhune & R.J. Zatorre. 2008. Moving on time: brain network for auditory-motor synchronization is modulated by rhythm complexity and musical training. *J. Cogn. Neurosci.* **20:** 226–239.
70. Grahn, J.A. & M. Brett. 2007. Rhythm and beat perception in motor areas of the brain. *J. Cogn. Neurosci.* **19:** 893–906.

71. Janata, P., S.T. Tomic & J.M. Haberman. 2012. Sensorimotor coupling in music and the psychology of the groove. *J. Exp. Psychol. Gen.* **141:** 54.
72. Zatorre, R.J., J.L. Chen & V.B. Penhune. 2007. When the brain plays music: auditory–motor interactions in music perception and production. *Nat. Rev. Neurosci.* **8:** 547.
73. Large, E.W. & M.R. Jones. 1999. The dynamics of attending: how people track time-varying events. *Psychol. Rev.* **106:** 119.
74. Fujioka, T., L.J. Trainor, E.W. Large & B. Ross. 2012. Internalized timing of isochronous sounds is represented in neuromagnetic beta oscillations. *J. Neurosci.* **32:** 1791–1802.
75. Nozaradan, S., I. Peretz, M. Missal & A. Mouraux. 2011. Tagging the neuronal entrainment to beat and meter. *J. Neurosci.* **31:** 10234–10240.
76. Peckel, M., T. Pozzo & E. Bigand. 2014. The impact of the perception of rhythmic music on self-paced oscillatory movements. *Front. Psychol.* **5:** 1037.
77. Bouvet, C.J., M. Varlet, S. Dalla Bella, *et al.* 2019. Preferred frequency ratios for spontaneous auditory-motor synchronization: dynamical stability and hysteresis. *Acta Psychol. (Amst.)* **196:** 33–41.
78. Assaneo, M.F., P. Ripollés, J. Orpella, *et al.* 2019. Spontaneous synchronization to speech reveals neural mechanisms facilitating language learning. *Nat. Neurosci.* **22:** 627–632.
79. Bowling, D.L., C.T. Herbst & W.T. Fitch. 2013. Social origins of rhythm? Synchrony and temporal regularity in human vocalization. *PLoS One* **8:** e80402.
80. Kiparsky, P. & G. Youmans. 2014. *Rhythm and Meter: Phonetics and Phonology*. Vol. 1. Academic Press.
81. Ong, W.J. 1982, 2002. Orality and *L*iteracy: the *T*echnologizing of the *W*ord. Methuen & Co. Ltd.
82. Beckman, M.E. 1986. Stress and Non-Stress Accent. Netherlands Phonetic Archives. Dordrecht: Foris.
83. Lehiste, I. 1977. Isochrony reconsidered. *J. Phonet*. **5:** 253–263.
84. Dauer, R.M. 1983. Stress-timing and syllable-timing reanalyzed. *J. Phonet*. **11:** 51–62.
85. Lidji, P., C. Palmer, I. Peretz & M. Morningstar. 2011. Listeners feel the beat: entrainment to English and French speech rhythms. *Psychon. Bull. Rev.* **18:** 1035–1041.
86. Repp, B.H. 1998. A microcosm of musical expression. I. Quantitative analysis of pianists' timing in the initial measures of Chopin's Etude in E major. *J. Acoust. Soc. Am.* **104:** 1085–1100.
87. Patel, A.D. 2010. *Music, Language, and the Brain*. Oxford: Oxford University Press.
88. Lerdahl, F. 2001. The sounds of poetry viewed as music. *Ann. N.Y. Acad. Sci.* **930:** 337–354.
89. Obermeier, C., S.A. Kotz, S. Jessen, *et al.* 2016. Aesthetic appreciation of poetry correlates with ease of processing in event-related potentials. *Cogn. Affect. Behav. Neurosci.* **16:** 362–373.
90. Obermeier, C., W. Menninghaus, M. von Koppenfels, *et al.* 2013. Aesthetic and emotional effects of meter and rhyme in poetry. *Front. Psychol.* **4:** 10.
91. Tillmann, B. & W.J. Dowling. 2007. Memory decreases for prose, but not for poetry. *Mem. Cognit.* **35:** 628–639.
92. Cummins, F. 2009. Rhythm as entrainment: the case of synchronous speech. *J. Phonet.* **37:** 16–28.
93. Jones, M.R. 2009. Musical time. In *The Handbook of Music Psychology*. S. Hallam, I. Cross & M. Thaut, Eds.: 81–92. Oxford: Oxford University Press.
94. Kotz, S.A. & M. Schwartze. 2010. Cortical speech processing unplugged: a timely subcortico-cortical framework. *Trends Cogn. Sci.* **14:** 392–399.
95. Giraud, A.-L. & D. Poeppel. 2012. Cortical oscillations and speech processing: emerging computational principles and operations. *Nat. Neurosci.* **15:** 511.
96. Calderone, D.J., P. Lakatos, P.D. Butler & F.X. Castellanos. 2014. Entrainment of neural oscillations as a modifiable substrate of attention. *Trends Cogn. Sci.* **18:** 300–309.
97. Nozaradan, S., I. Peretz & A. Mouraux. 2012. Selective neuronal entrainment to the beat and meter embedded in a musical rhythm. *J. Neurosci.* **32:** 17572–17581.
98. Peelle, J.E. & M.H. Davis. 2012. Neural oscillations carry speech rhythm through to comprehension. *Front. Psychol.* **3:** 320.
99. Falk, S. & S. Dalla Bella. 2016. It is better when expected: aligning speech and motor rhythms enhances verbal processing. *Lang. Cogn. Neurosci.* **31:** 699–708.
100. Falk, S., T. Rathcke & S. Dalla Bella. 2014. When speech sounds like music. *J. Exp. Psychol. Hum. Percept. Perform.* **40:** 1491.
101. Falk, S., C. Volpi-Moncorger & S. Dalla Bella. 2017. Auditory-motor rhythms and speech processing in French and German listeners. *Front. Psychol.* **8:** 395.
102. Opie, P. & P. Opie. 1951. *The Oxford Dictionary of Nursery Rhymes*. Oxford: Clarendon Press.
103. Politimou, N., S. Dalla Bella, N. Farrugia & F. Franco. 2019. Born to speak and sing: musical predictors of language development in pre-schoolers. *Front. Psychol.* **10:** 948.
104. Schön, D. & B. Tillmann. 2015. Short- and long-term rhythmic interventions: perspectives for language rehabilitation. *Ann. N.Y. Acad. Sci.* **1337:** 32–39.
105. Wan, C.Y., L. Bazen, R. Baars, *et al.* 2011. Auditory-motor mapping training as an intervention to facilitate speech output in non-verbal children with autism: a proof of concept study. *PLoS One* **6:** e25505.
106. Repp, B.H. 2005. Sensorimotor synchronization: a review of the tapping literature. *Psychon. Bull. Rev.* **12:** 969–992.
107. Repp, B.H. & Y.-H. Su. 2013. Sensorimotor synchronization: a review of recent research (2006–2012). *Psychon. Bull. Rev.* **20:** 403–452.
108. Patel, A.D. 2006. Musical rhythm, linguistic rhythm, and human evolution. *Music Percept.* **24:** 99–104.
109. Dalla Bella, S., M. Berkowska & J. Sowiński. 2015. Moving to the beat and singing are linked in humans. *Front. Hum. Neurosci.* **9:** 663.
110. Patel, A.D., J.R. Iversen, M.R. Bregman & I. Schulz. 2009. Studying synchronization to a musical beat in nonhuman animals. *Ann. N.Y. Acad. Sci.* **1169:** 459–469.
111. Patel, A.D., J.R. Iversen, M.R. Bregman & I. Schulz. 2009. Experimental evidence for synchronization to a musical beat in a nonhuman animal. *Curr. Biol.* **19:** 827–830.
112. Hasegawa, A., K. Okanoya, T. Hasegawa & Y. Seki. 2011. Rhythmic synchronization tapping to an audio-visual

metronome in budgerigars. *Sci. Rep.* **1**. https://doi.org/10.1038/srep00120.
113. Schachner, A., T.F. Brady, I.M. Pepperberg & M.D. Hauser. 2009. Spontaneous motor entrainment to music in multiple vocal mimicking species. *Curr. Biol.* **19:** 831-836.
114. Hattori, Y., M. Tomonaga & T. Matsuzawa. 2013. Spontaneous synchronized tapping to an auditory rhythm in a chimpanzee. *Sci. Rep.* **3:** 1566.
115. Hattori, Y., M. Tomonaga & T. Matsuzawa. 2015. Distractor effect of auditory rhythms on self-paced tapping in chimpanzees and humans. *PLoS One* **10:** e0130682.
116. Cook, P., A. Rouse, M. Wilson & C.J. Reichmuth. 2013. A California sea lion (*Zalophus californianus*) can keep the beat: motor entrainment to rhythmic auditory stimuli in a non vocal mimic. *J. Comp. Psychol.* **127:** 1–16.
117. Rouse, A.A., P.F. Cook, E.W. Large & C. Reichmuth. 2016. Beat keeping in a sea lion as coupled oscillation: implications for comparative understanding of human rhythm. *Front. Neurosci.* **10**. https://doi.org/10.3389/fnins.2016.00257.
118. Ravignani, A. & P. Cook. 2016. The evolutionary biology of dance without frills. *Curr. Biol.* **26:** R878–R879.
119. Wilson, M. & P.F. Cook. 2016. Rhythmic entrainment: why humans want to, fireflies can't help it, pet birds try, and sea lions have to be bribed. *Psychon. Bull. Rev.* **23:** 1647–1659.
120. Chen, Y., L.E. Matheson & J.T. Sakata. 2016. Mechanisms underlying the social enhancement of vocal learning in songbirds. *Proc. Natl. Acad. Sci. USA* **113:** 6641–6646.
121. Whitham, J.C., M.S. Gerald & D. Maestripieri. 2007. Intended receivers and functional significance of grunt and girney vocalizations in free-ranging female rhesus macaques. *Ethology* **113:** 862–874.
122. Parks, S., L. Conger, D. Cusano & S. Van Parijs. 2014. Variation in the acoustic behavior of right whale mother–calf pairs. *J. Acoust. Soc. Am.* **135:** 2240–2240.
123. Fernald, A., T. Taeschner, J. Dunn, *et al.* 1989. A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *J. Child Lang.* **16:** 477–501.
124. Martin, A., Y. Igarashi, N. Jincho & R. Mazuka. 2016. Utterances in infant-directed speech are shorter, not slower. *Cognition* **156:** 52–59.
125. Papoušek, M., H. Papoušek & D. Symmes. 1991. The meanings of melodies in motherese in tone and stress languages. *Infant Behav. Dev.* **14:** 415–440.
126. Bergeson, T.R. & S.E. Trehub. 2002. Absolute pitch and tempo in mothers' songs to infants. *Psychol. Sci.* **13:** 72–75.
127. Trehub, S.E. & L. Trainor. 1998. Singing to infants: lullabies and play songs. *Adv. Infancy Res.* **12:** 43–78.
128. Dissanayake, E. 2000. *Art and Intimacy: How the Arts Began*. University of Washington Press.
129. Jaffe, J., B. Beebe, S. Feldstein, *et al.* 2001. Rhythms of dialogue in infancy: coordinated timing in development. *Monogr. Soc. Res. Child Dev.* **66:** i-viii, 1–132.
130. Gratier, M., E. Devouche, B. Guellai, *et al*. 2015. Early development of turn-taking in vocal interaction between mothers and infants. *Front. Psychol.* **6**. https://doi.org/10.3389/fpsyg.2015.01167.
131. Hilbrink, E.E., M. Gattis & S.C. Levinson. 2015. Early developmental changes in the timing of turn-taking: a longitudinal study of mother–infant interaction. *Front. Psychol.* **6**. https://doi.org/10.3389/fpsyg.2015.01492.
132. Pickering, M.J. & S. Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behav. Brain Sci.* **27:** 169–190.
133. Abney, D.H., A.S. Warlaumont, D.K. Oller, *et al*. 2017. Multiple coordination patterns in infant and adult vocalizations. *Infancy* **22:** 514–539.
134. Falk, S. & C.T. Kello. 2017. Hierarchical organization in the temporal structure of infant-direct speech and song. *Cognition* **163:** 80–86.
135. Cason, N. & D. Schön. 2012. Rhythmic priming enhances the phonological processing of speech. *Neuropsychologia* **50:** 2652–2658.
136. Quené, H. & R.F. Port. 2005. Effects of timing regularity and metrical expectancy on spoken-word perception. *Phonetica* **62:** 1–13.
137. Roncaglia-Denissen, M.P., M. Schmidt-Kassow & S.A. Kotz. 2013. Speech rhythm facilitates syntactic ambiguity resolution: ERP evidence. *PLoS One* **8:** e56000.
138. Rothermich, K., M. Schmidt-Kassow & S.A. Kotz. 2012. Rhythm's gonna get you: regular meter facilitates semantic sentence processing. *Neuropsychologia* **50:** 232–244.
139. Albin, D.D. & C.H. Echols. 1996. Stressed and word-final syllables in infant-directed speech. *Infant Behav. Dev.* **19:** 401–418.
140. Seidl, A. & E.K. Johnson. 2006. Infant word segmentation revisited: edge alignment facilitates target extraction. *Dev. Sci.* **9:** 565–573.
141. Cirelli, L.K., C. Spinelli, S. Nozaradan & L.J. Trainor. 2016. Measuring neural entrainment to beat and meter in infants: effects of music background. *Front. Neurosci.* **10:** 229.
142. Hannon, E.E. & S.E. Trehub. 2005. Tuning in to musical rhythms: infants learn more readily than adults. *Proc. Natl. Acad. Sci. USA* **102:** 12639–12643.
143. Winkler, I., G.P. Háden, O. Ladinig, *et al*. 2009. Newborn infants detect the beat in music. *Proc. Natl. Acad. Sci. USA* **106:** 2468–2471.
144. Lebedeva, G.C. & P.K. Kuhl. 2010. Sing that tune: infants' perception of melody and lyrics and the facilitation of phonetic recognition in songs. *Infant Behav. Dev.* **33:** 419–430.
145. Tsang, C.D., S. Falk & A. Hessel. 2017. Infants prefer infant-directed song over speech. *Child Dev.* **88:** 1207–1215.
146. Martin, J.G. 1972. Rhythmic (hierarchical) versus serial structure in speech and other behavior. *Psychol. Rev*. **79:** 487–509.
147. Kershenbaum, A., D.T. Blumstein, M.A. Roch, *et al*. 2016. Acoustic sequences in non-human animals: a tutorial review and prospectus. *Biol. Rev.* **91:** 13–52.
148. Kershenbaum, A., A.E. Bowles, T.M. Freeberg, *et al*. 2014. Animal vocal sequences: not the Markov chains we thought they were. *Proc. Biol. Sci.* **281:** 20141370.
149. Rohrmeier, M., W. Zuidema, G.A. Wiggins & C. Scharff. 2015. Principles of structure building in music, language and animal song. *Philos. Trans. R. Soc. B Biol. Sci.* **370:** 20140097.
150. Singh, N.C. & F.E. Theunissen. 2003. Modulation spectra of natural sounds and ethological theories of auditory processing. *J. Acoust. Soc. Am*. **114:** 3394–3411.
151. Allan, D.W. 1966. Statistics of atomic frequency standards. *Proc. IEEE* **54:** 221–230.

152. Kello, C.T., S. Dalla Bella, B. Médé & R. Balasubramaniam. 2017. Hierarchical temporal structure in music, speech and animal vocalizations: jazz is like a conversation, humpbacks sing like hermit thrushes. *J. R. Soc. Interface* **14:** 20170231.
153. Ramirez-Aristizabal, A.G., B. Médé & C.T. Kello. 2018. Complexity matching in speech: effects of speaking rate and naturalness. *Chaos Solitons Fract.* **111:** 175–179.
154. Sawyer, R.K. 2005. Music and conversation. *Music. Commun.* **45:** 60.
155. Jadoul, Y., B. Thompson & B. de Boer. 2018. Introducing Parselmouth: a Python Interface to Praat. *J. Phonet*. **71:** 1–15.
156. Ravignani, A. 2018. Spontaneous rhythms in a harbor seal pup calls. *BMC Res. Notes* **11:** 1–4.
157. Noriega, F.,, A.C. Montes-Medina & M. Timme. 2019. Quantitative analysis of timing in animal vocal sequences. Preprint. arXiv https://arxiv.org/abs/1902.07650.
158. Jadoul, Y., A. Ravignani, B. Thompson, *et al*. 2016. Seeking temporal predictability in speech: comparing statistical approaches on 18 world languages. *Front. Hum. Neurosci.* **10**. https://doi.org/10.3389/fnhum.2016.00586.
159. Pedregosa, F., G. Varoquaux, A. Gramfort, *et al*. 2011. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12:** 2825–2830.
160. Ravignani, A., B. Thompson, M. Lumaca & M. Grube. 2018. Why do durations in musical rhythms conform to small integer ratios? *Front. Comput. Neurosci.* **12**. https://doi.org/10.3389/fncom.2018.00086.
161. Demartsev, V., A. Strandburg-Peshkin, M. Ruffner & M. Manser. 2018. Vocal turn-taking in meerkat group calling sessions. *Curr. Biol.* **28:** 3661–3666.e3.
162. Stivers, T., N.J. Enfield, P. Brown, *et al*. 2009. Universals and cultural variation in turn-taking in conversation. *Proc. Natl. Acad. Sci. USA* **106:** 10587–10592.
163. Iversen, J.R. 2016. In the beginning was the beat: evolutionary origins of musical rhythm in humans. In *The Cambridge Companion to Percussion*. R. Hartenberger, Ed.: 281–295. Cambridge University Press.
164. Kotz, S.A. & M. Schmidt-Kassow. 2015. Basal ganglia contribution to rule expectancy and temporal predictability in speech. *Cortex* **68:** 48–60.
165. Harding, E.E., D. Sammler, M.J. Henry, *et al*. 2019. Cortical tracking of rhythm in music and speech. *Neuroimage* **185:** 96–101.
166. Kotz, S.A., M. Schwartze & M. Schmidt-Kassow. 2009. Non-motor basal ganglia functions: a review and proposal for a model of sensory predictability in auditory language perception. *Cortex* **45:** 982–990.
167. Lieberman, P. 2009. FOXP2 and human cognition. *Cell* **137:** 800–802.
168. Pastor, M.A., E. Macaluso, B. Day & R. Frackowiak. 2006. The neural basis of temporal auditory discrimination. *Neuroimage* **30:** 512–520.
169. Fuster, J.M. 2001. The prefrontal cortex—an update: time is of the essence. *Neuron* **30:** 319–333.
170. Teki, S., M. Grube, S. Kumar & T.D. Griffiths. 2011. Distinct neural substrates of duration-based and beat-based auditory timing. *J. Neurosci.* **31:** 3805–3812.
171. Hertrich, I., S. Dietrich & H. Ackermann. 2016. The role of the supplementary motor area for speech and language processing. *Neurosci. Biobehav. Rev.* **68:** 602–610.
172. Tourville, J.A. & F.H. Guenther. 2011. The DIVA model: a neural theory of speech acquisition and production. *Lang. Cogn. Process.* **26:** 952–981.
173. Kotz, S.A., R.M. Brown & M. Schwartze. 2016. Cortico-striatal circuits and the timing of action and perception. *Curr. Opin. Behav. Sci.* **8:** 42–45.
174. Kotz, S.A. & M. Schwartze. 2016. Motor-timing and sequencing in speech production: a general-purpose framework. In *Neurobiology of Language*. G. Hickok & S.L. Small, Eds.: 717–724. San Diego, CA: Academic Press.
175. Grahn, J.A. 2009. The role of the basal ganglia in beat perception: neuroimaging and neuropsychological investigations. *Ann. N.Y. Acad. Sci.* **1169:** 35–45.
176. Nozaradan, S., M. Schwartze, C. Obermeier & S.A. Kotz. 2017. Specific contributions of basal ganglia and cerebellum to the neural tracking of rhythm. *Cortex* **95:** 156–168.
177. Schwartze, M. 2012. *Adaptation to Temporal Structure*. Leipzig: Max Planck Institute for Human Cognitive and Brain Sciences.
178. Merchant, H., J. Grahn, L. Trainor, *et al*. 2015. Finding the beat: a neural perspective across humans and non-human primates. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370:** 20140093.
179. Fitch, W.T. 2009. The biology and evolution of rhythm: unraveling a paradox. In *Language and Music as Cognitive Systems*. P. Rebuschat, M. Rohrmeier, J.A. Hawkins & I. Cross, Eds.: 73–95. Oxford, UK: Oxford University Press.
180. Zarco, W., H. Merchant, L. Prado & J.C. Mendez. 2009. Subsecond timing in primates: comparison of interval production between human subjects and rhesus monkeys. *J. Neurophysiol.* **102:** 3191–3202.
181. Ayala, Y.A., A. Lehmann & H. Merchant. 2017. Monkeys share the neurophysiological basis for encoding sound periodicities captured by the frequency-following response with humans. *Sci. Rep.* **7:** 16687.
182. Honing, H. & H. Merchant. 2014. Differences in auditory timing between human and nonhuman primates. *Behav. Brain Sci.* **37:** 557–558.
183. MacLeod, C.E., K. Zilles, A. Schleicher, *et al*. 2003. Expansion of the neocerebellum in Hominoidea. *J. Hum. Evol.* **44:** 401–429.
184. Weaver, A.H. 2005. Reciprocal evolution of the cerebellum and neocortex in fossil humans. *Proc. Natl. Acad. Sci. USA* **102:** 3576–3580.
185. Kirby, S., H. Cornish & K. Smith. 2008. Cumulative cultural evolution in the laboratory: an experimental approach to the origins of structure in human language. *Proc. Natl. Acad. Sci. USA* **105:** 10681–10686.
186. deCastro-Arrazola, V. & S. Kirby. 2019. The emergence of verse templates through iterated learning. *J. Lang. Evol.* **4:** 28–43.
187. Jacoby, N. & J.H. McDermott. 2017. Integer ratio priors on musical rhythm revealed cross-culturally by iterated reproduction. *Curr. Biol.* **27:** 359–370.
188. Ravignani, A., T. Delgado & S. Kirby. 2016. Musical evolution in the lab exhibits rhythmic universals. *Nat. Hum. Behav.* **1:** 0007.

189. Ravignani, A., B. Thompson, T. Grossi, *et al*. 2018. Evolving building blocks of rhythm: how human cognition creates music via cultural transmission. *Ann. N.Y. Acad. Sci.* https://doi.org/10.1111/nyas.13610.
190. Edmiston, P., M. Perlman & G. Lupyan. 2018. Repeated imitation makes human vocalizations more word-like. *Proc. R. Soc. B Biol. Sci*. **285:** 20172709.
191. Lumaca, M., A. Ravignani & G. Baggio. 2018. Music evolution in the laboratory: cultural transmission meets neurophysiology. *Front. Neurosci.* **12**. https://doi.org/10.3389/fnins.2018.00246.
192. Gingras, B., H. Honing, I. Peretz, *et al*. 2015. Defining the biological bases of individual differences in musicality. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370:** 20140092.
193. Nazzi, T., S. Paterson & A. Karmiloff-Smith. 2003. Early word segmentation by infants and toddlers with Williams syndrome. *Infancy* **4:** 251–271.
194. Lense, M.D. & E.M. Dykens. 2016. Beat perception and sociability: evidence from Williams syndrome. *Front. Psychol.* **7**. https://doi.org/10.3389/fpsyg.2016.00886.