Evolution of speech rhythm: a cross-species perspective

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

Author affiliations
[1]Artificial Intelligence Laboratory, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium
[2]Research Department, Sealcentre Pieterburen, Hoofdstraat 94a, 9968 AG Pieterburen, The Netherlands
[3] International Laboratory for Brain, Music and Sound Research (BRAMS), Montréal, QC, H3C 3J7, Canada
[4]Department of Psychology, University of Montreal, Pavillon Marie-Victorin, CP 6128 Succursale Centre-Ville, Montréal, QC, H3C 3J7, Canada
[5]Department of Cognitive Psychology, WSFiZ in Warsaw, Warsaw, 01-030, Poland
[6] Laboratoire de Phonétique et Phonologie, UMR 7018, CNRS / Université Sorbonne Nouvelle Paris-3, Institut de Linguistique et Phonétique générales et appliquées, 19 rue des Bernardins, 75005 Paris, France
[7]Cognitive and Information Sciences, University of California, Merced, 5200 N. Lake Road, Merced, CA, 95343, USA
[8]Chair for Network Dynamics, Center for Advancing Electronics Dresden (cfaed), TU Dresden, 01062 Dresden, Germany
[9]CODE University of Applied Sciences, Lohmühlenstraße 65, 12435 Berlin, Germany
[10]Basic and Applied NeuroDynamics Laboratory, Faculty of Psychology and Neuroscience, Department of Neuropsychology & Psychopharmacology, Maastricht University, Maastricht, The Netherlands
[11]Department of Neuropsychology, Max-Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

* These 4 authors contributed equally.

Corresponding authors contact information: Andrea Ravignani, Artificial Intelligence Lab, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium. andrea.ravignani@gmail.com; Sonja Kotz, Basi and Applied NeuroDynamics Laboratory, Maastricht University, Universiteitssingel 40, 6200 MD Maastricht, the Netherlands. sonja.kotz@maastrichtuniversity.nl

Short title: Speech rhythm across species

Keywords *(max 6):* Speech rhythm; hierarchical; timing; time perception; rhythm cognition; bioacoustics

Abstract *(190 words)*. Cognition and communication, at the core of human speech rhythm, do not leave a fossil record. However, if the purpose is to understand the origin and evolution of speech rhythm, alternative methods are available. A powerful tool is comparative approach: studying the presence or absence of cognitive/behavioral traits in other species, drawing conclusions on which traits are shared between species, and which are recent human inventions. Here we apply this approach to traits related to human speech rhythm. Many species exhibit temporal structure in their vocalizations but little is known about the range of rhythmic structures perceived and produced, their biological and developmental bases, and communicative functions. We review the literatures on human and non-human studies of rhythm in speech and animal vocalizations to survey similarities and differences. We report important links between vocal perception and motor coordination, and the differentiation of rhythm based on hierarchical temporal structure. We extend this review to quantitative techniques useful for computing rhythmic structure in acoustic sequences and hence facilitating cross-species research. While still far from a full comparative cross-species perspective of speech rhythm, we are closer to fitting missing pieces of the puzzle.

Main text (Word count: 7000 words)

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 apply this method to the study of speech rhythm, investigating which similar traits can be found in other species to understand what remains uniquely human. We review several literatures which are usually unconnected. In particular, we discuss the production and perception of rhythmic patterns in non-human species and in human development. We summarize several methods to measure rhythmic structure in vocalizations produced by humans and other animals. We discuss the neural bases of speech rhythm, attempting to draw comparative links, both between modalities and species. As 'rhythm' encompasses and transcend the ability to produce and perceive individual temporal intervals, we set off by discussing interval timing. (A thorough treatment of interval timing can be found elsewhere, e.g., [2-6].)

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. A general finding from these studies is that predictions from the so called scalar expectancy theory hold across species and domains (with some exceptions, see [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 estimate longer intervals with less accuracy, including humans.

Research in timing and time perception is necessary to understand rhythm, but not sufficient. As a parallel, a deep understanding of fundamental frequencies is necessary, but not sufficient, to understand the harmonies and timbres of sounds. Some perceptual phenomena go beyond fundamental frequency because perceptual effects arise when several frequencies are combined that are not predicted by perceptual data on individual frequencies.

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

Rhythm involves series of time intervals, often at multiple timescales, that can combine to produce hierarchical metrical structure [7]. The perception of rhythmic concepts, such as grouping and meter is usually studied in operant experiments [8]. 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)[9]. Zebra finches show similar discrimination capacities as rats [10]. Follow-up work showed that if thoroughly trained for a durational alternation, rats can indeed discriminate between iambs and trochees [11]. In a related experiment, although with different setup, budgerigars could distinguish between iambic and trochaic meter, but required more than one cue among e.g. pitch, duration, loudness, vowel quality, to succeed [12]. Testing rats with stimuli identical to those used for budgerigars held a very different result: unlike parrots rodents

need all 4 cues to discriminate between prosodic patterns. Of these 4 cues, one is purely (duration) and other two partly (loudness and pitch) rhythmic.

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

Several species have been found to produce spontaneous vocal rhythms, and are therefore particularly promising for human-animal comparisons [13], including (1) lab rodents, such as mice, because biomedical research has thoroughly studied their neurobiology; (2) non-human primates, because of their phylogenetic relatedness to humans; (3) songbirds, in particular zebra finches, because they are an established model species for avian vocal flexibility and learning; and (4) vocal learning mammals, such as seals and bats, because they represent the closest vocal learning animals to humans. Below, we will briefly discuss examples of vocal rhythms in rodents, songbirds and vocal learning mammals. (Primate perspectives on speech rhythm can be found elsewhere, e.g. [14].)

Ultrasonic vocalizations in mice exhibit quite stable transition probabilities in durations, especially for short-short and long-long transitions ([15], for more on transition probabilities see e.g. Figure 1). Mice vocalizations appear temporally organized in a hierarchical fashion [15], but more work is needed in this direction to test for the existence of real hierarchical organization, i.e. temporal events structured at different time scales, possibly embedding one level into the higher one, and bootstrapping learning of 'syntactic' rhythmic structure. Finally, in a developmental perspective, the rhythm of mice vocalizations as pups is predictive of the vocal rhythms in the same mice as adults [15].

Zebra finches have long been a model for vocal learning, but research in this species has historically focused on the spectral and combinatorial domains, rather than the temporal and rhythmic domains. Only recently, the temporal dimension of their songs has been explored. Zebra finches' rhythms are characterized by plasticity and inter-individual variability, which are connected to learning and often in contrast to stereotypical calling [16]. Past methods used in birdsong research concluded strong stereotypy in zebra finches' rhythms, but this may have been because of analytical methods only focusing on short time scales and ignoring organization at longer time scales [16]. 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 [17]. This interplay between plasticity and regularity makes intuitive sense: an underlying isochronous grid can provide anchor points in time in order to create and sing plastically.

The isochrony-detection technique used in zebra finches has also been applied to a bat species capable of vocal production learning. Surprisingly, the neo-tropical 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 [18]. In addition, the tempo of a hypothetical superimposed metronomic grid (e.g. see Figure 2) perfectly matches the wing-beat of the animals [18], potentially a reminder of the cross-modality of rhythm.

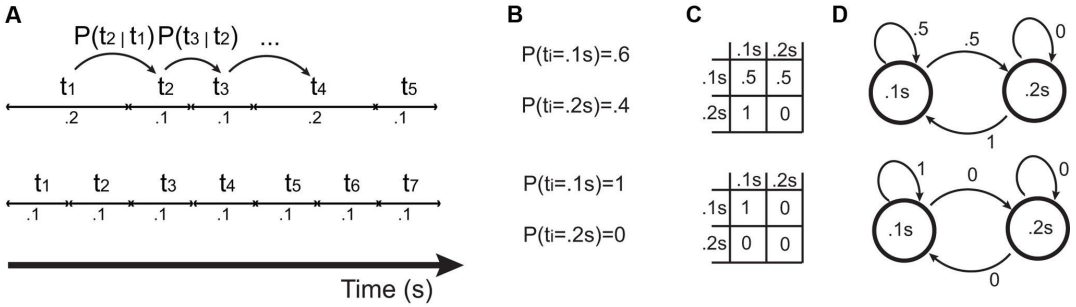

Figure 1. Some of the possible ways of representing temporal patterns (bottom: isochronous, top: non-isochronous). (A) Time series of intervals, inducing transition probabilities such as P(t2 | t1), which means the probability that the interval t2 of length x msec follows an interval t1 of length y msec. (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 [19], an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY).

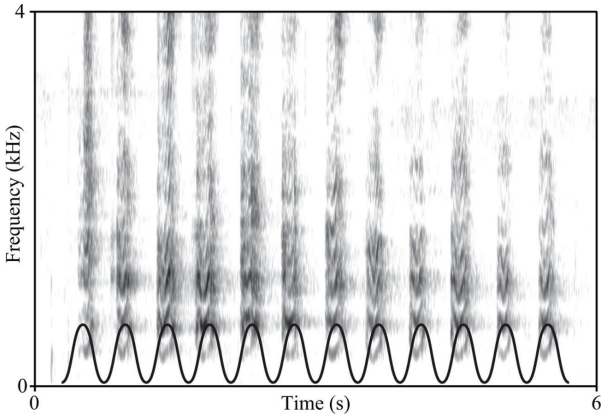

Figure 2. Spectrogram showing the isochronous barking of a California sea lion. To visually detect isochronous regularity, one can superimpose a metronomic grid, like the regular sinusoidal function, to the spectrogram. Figure reproduced verbatim from [19], an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY).

When the study species is difficult to sample, as in adult seal males who sing underwater [20], rhythmicity can be indirectly hypothesized by testing whether temporal structures of different song elements covary [21]. This, however, must be complemented with more rigorous and technically challenging fieldwork, where each vocalization can be uniquely attributed to one individual. 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 (Figure 3; [22]).

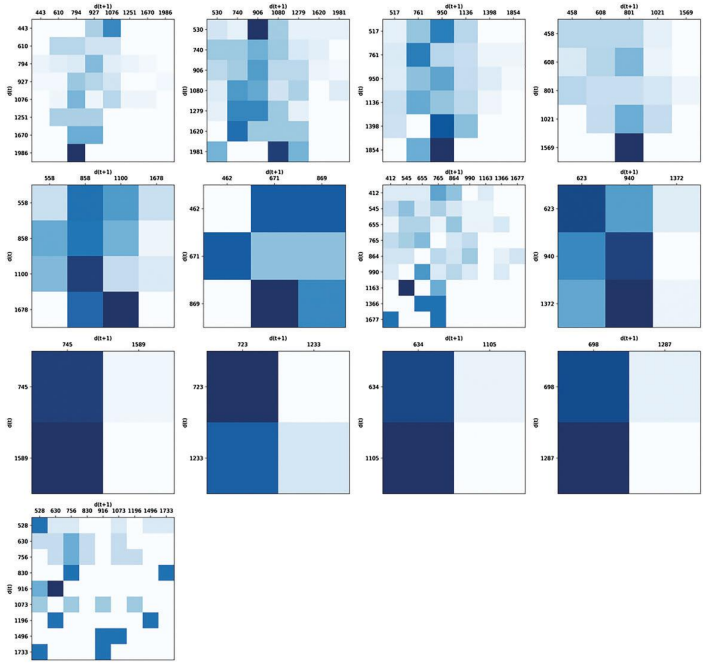

Figure 3. Transition matrices for one individual seal pup across days. Each matrix represents 1 day, with calendar days progressing from left to right and top to bottom. Each row and column of one matrix represents the centroid of a durational category: leftmost columns and upmost rows are shorter (400-700 msec) categories; the further down and right, the longer the category. Shades of blue represent transition probabilities; i.e. 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 last matrix, the dark square means that a durational category centered at 916 msec is very likely to be followed by a durational category centered at 630 msec, but very unlikely by one centered at 756 msec. Notice how, over days, the number of categories shrink and the transitions from one to the other become more predictable. Figure reproduced verbatim from [22], an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License; additional details in the original paper.

Vocal learners such as harbor seals and *Saccopteryx bilineata* bats [23, 24] are useful model species to study the potential interplay between rhythm and vocal learning ontogeny [16, 22]. 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., [25, 26]). Babbling features native language sound production and imitation of prosodic aspects of adult speech (e.g., [26, 27]). 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 [28], in sac-winged bat pups [18, 29], and in zebra finches [30]. According to the Frame & Content theory [31], babbling in human infants is a rhythmic motor training laying the grounds for basic syllable structure. Infants learn that vocalizing at different times during quasi-periodic cycles of mandibular opening and closure results in vowels at maximal mandibular opening and consonants at maximal mandibular closure. However, is is still unclear how these early syllable rhythms in babbling contribute to later adult rhythms or language acquisition in general [32]. Results from non-human animals may help understand some of these issues. For example, female baby bats, during the babbling period, also produce adult male songs and trills without producing them as adults [29]. In zebra finches, different brain

circuits were active during juvenile babbling and later adult song production [30]. These results suggest that animal babbling is not linked in a simple way to later adult vocal production, which may inspire future research on human infant babbling in relation to later speech and language acquisition as well.

Finally, potential parallels between the animal world and humans can be investigated in disorders of the rhythm of early vocal production. Stuttering, for example, is a speech fluency disorder typically emerging between the 2nd and 4th year of life in humans [33]. Children show untypical disfluencies during speech production such as silent blocks, syllable and sound repetitions and prolongations. Stuttering-like behavior can be observed in songbirds such as zebra finches as well. 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 [34-36]. Interestingly, animal research points to a prominent role of basal ganglia dysfunction in stuttering zebra finches [37], paralleling findings of impaired basal ganglia functioning in human children and adults who stutter [38, 39]. More research is needed to unravel similarities in how rhythm contributes to develop skilled speech motor control across species.

Uniquely human? Interactive rhythms during speech development

As there are some parallels in human and animal rhythmic vocalizations during development, the question arises to what extent vocal rhythms in interaction are also comparable across species. Animal studies tend to find that “solo” vocalizations produced by non-human infants and adults have similar acoustic characteristics to those produced during infant-adult interactions. Only certain animals, though, utter specific pup-directed vocalizations by making them shorter, more repetitive, or more specialized than adult-directed vocalizations (see [40], for male zebra finches; [41], for free-ranging female rhesus macaque; [42], 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, 2) *temporal predictions* about linguistic structure derived from rhythmic cues in infant-directed communication. These aspects could further be studied in the animal domain.

Whether female, male, parent, sibling or stranger, older interlocutors across cultures display a distinct infant-directed speech register. Their utterances are shorter, higher pitched and contain distinct melodic contours, and more repetition [43-45]. These salient alterations in speech, as well as songs, chants, and rhythmic vocal play [46, 47] contribute to an overall highly musical, and, thereby 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 to each other and to develop mutual understanding and experience sharing beyond symbolic communication [48]. In line with this idea, Jaffe and colleagues [49] 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 two months

on, turn-taking structure between mother and infant vocalizations is already observable with only a 30-40 % overlap between reciprocal vocalizations. Two to 3 turns are the most frequent exchange structure, and pause gaps are under 1s [50, 51].

Mutual alignment is considered a key aspect of adult verbal interaction [52]. 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-years longitudinal study on mother-infant coordination, Abney and colleagues [53] identified 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 extracting hierarchically nested bouts of temporal clusters across timescales. They found that mothers particularly align with their infants in terms of the hierarchical temporal structure of speech which is generally emphasized in infant-directed speech and singing compared to adult-directed communication [54]. Supporting the idea of a precursor to linguistic skills, preverbal vocalizations of infants (e.g., vocalic and syllabic sequences) were overall temporally better coordinated with their mother's vocalizations than any non-verbal vocalization (e.g., laughter, cries).

Adult listeners use temporal predictions in order to better attend to and process phonological, lexical, semantic, and syntactic structure in their interlocutor's speech [55-58]. Higher repetitiveness, greater metrical stability, shorter utterances, and enhanced utterance-final lengthening in infant-directed speech are all temporal aspects 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 pre-boundary lengthening and longer post-boundary pauses [44, 59]. 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 as phrase-final vs. medial positions in speech [60]. Infants are also able to generate temporal predictions from a regular beat structure such as found in music [61-63]. As a musical stimulus, infant-directed singing may particularly support beat-related predictions in caregiver-infant communication. Infant-directed singing can be discriminated from infant-directed speech by infants as young as 6 months, adults [64], and even by non-human avian species (i.e., zebra finches [65]). Infant-directed singing features clearer metrical structure than speech [46], and therefore may better direct infants' attention towards words associated with a beat. First results showed a trend that infants at 11 months of age process word-related information in song better than in speech [66]. Yet, unique contributions of the rhythm of singing to infants' language skills still await further investigation.

Rhythm as Temporal Hierarchy in Human and Non-Human Vocalization

Rhythm and timing in speech, as in complex animal vocalizations, has hierarchical temporal structure. We know where this structure comes from in speech: Units of perception and production are built up hierarchically [67]. 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 [68-70].

Regardless of whether we know the units or not, we can observe hierarchical temporal structure directly in the acoustic signal that results from vocalization. This structure is importantly different than symbolic hierarchical expressions, as in linguistic research, because symbolic expressions do not specify timing or temporal durations. Linguistic hierarchies must be elaborated to include temporal structure, which is often done either implicitly or explicitly. 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 (see Figure 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, rather than try to identify the particular units of perception or production. 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. While we do not have a corresponding symbolic hierarchy as we do in speech, we can nonetheless directly compare the hierarchical temporal 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 [71], which quantifies the bursts and lulls in acoustic energy that are characteristic of speech and animal vocalizations. The timing and duration of the bursts are captured by clustering in peak events in the amplitude envelope across a wide range of timescales. Smaller clusters are nested within larger clusters across timescales, and nesting is quantified using Allan Factor (AF) variance [72]. The result is an AF function over timescales that is analyzed in log-log coordinates because AF variance is often a power law function of timescales, indicating self-similar (fractal) nesting. AF functions can be compared using various metrics such as correlating their log-log slopes [73] or computing the Euclidean distance between them [74].

Falk and Kello [54] were the first to submit peak amplitude events to AF analysis of speech recordings. They 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 ten 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 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.

With this result in hand, Kello and colleagues [75] applied AF analysis to a wide range of speech, music, and animal vocalization recordings. Results further supported the relationship between hierarchical temporal structure and prosodic exaggeration, in that synthesized speech with impoverished prosodic cues contained less nested clustered compared with natural speech. Results also showed that nested clustering is enhanced by musical composition compared with improvisation or speech. But perhaps most interesting, machine learning analyses of AF functions revealed a natural taxonomy of complex acoustic signals, where recordings within a given category yielded AF functions that closely followed a pattern specific to that category (see Figure 4).

Figure 4. Scatter plot with each point representing the curvature (x-axis) and slope (y-axis) of the AF function for each of 10 recordings per category shown in the legend. The four main categories are represented by color (blue = animal vocalization, red = human vocalization, green = classical music, and cyan = popular music). Large symbols are placed at the centroid of each subcategory. **Figure 4 is omitted because it is a copy of the non-OA article [75].**

The AF function category most relevant to the current discussion corresponds to *conversational interactions*. In [75] and two subsequent studies [76, 77], 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. While it is possible that the same AF function shape could be generated in other ways, observations to date establish a diagnostic relationship between a particular shape and conversational interaction. The relationship was further established by Kello et al. who found that jazz improvisations, which have been likened to conversations [78], also yield AF functions with the same particular shape. Most notably, recordings of *communicative* animal vocalizations from killer whales 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 [22] 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 non-human communicative interactions of all kinds may manifest the same, unique kind of hierarchical temporal structure depending on the particular communicative function but irrespective of the species. 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 of incoming communicative stimuli, and learning.

Quantitative methods for characterizing temporal structure in speech and animal vocalizations

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 unsupervised techniques acting directly over the waveform. An alternative in this situation is to annotate the recordings with the onset and or offset times and investigate the temporal structure of these events [22, 79, 80]. In this section, we discuss five more quantitative methods (in addition to the AF analyses above) for characterizing temporal patterns in series of events. The discussed methods can be divided into two categories depending on the sort of temporal data they deal with. The first kind uses times series as input data and includes the power spectral density and autocorrelation techniques. The second kind uses the inter-event intervals (IEI) as input data and includes the normalized pairwise variability index (nPVI) [81], distributions of IEI intervals [82], and phase portraits [83], techniques. We evaluate these methods based on their ability to characterize temporal structures in four datasets: random, isochronous, hierarchical, and speech. We briefly describe the datasets before discussing what happens when the different methods are used on them.

The datasets consist of time series of events represented by pulses. The isochronous series has a pulse every 0.2s. The random series is a Poisson process with a rate $\lambda = 12$ pulses per second. The hierarchical series composed of hierarchically grouped pulses. All artificial sets — random, isochronous, hierarchical — are 10 s longs with a sampling rate of one kilohertz (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 s. 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 [84]. This annotated speech dataset contains the syllable centers of all languages. However, because there are language differences, we also analyze the syllable centers of each language separately, thereby obtaining 18 additional datasets.

We start the discussion of the first kind of methods with what is perhaps the best-known approach for investigating time series: Fourier analysis. By projecting a signal into a basis of sinusoids, Fourier's power spectral density reveals the periodicities within the signal. The structure of the isochronous signal is well captured by this method, as shown (Figure 5) by the 5 Hz peaks of its power spectral density. Fourier analysis also captures well the structures in the hierarchical series as indicated by the hierarchical distances of the spectrum's peaks. However, Fourier analysis is of little help for characterizing the temporal structure of the random and the speech datasets. Similarly, the autocorrelation function captures well the structure of the isochronous and the hierarchical signals but is of limited insight for the random and speech signals. With this in mind, autocorrelation and Fourier analysis are useful for characterizing

temporal structure in vocal sequences with a high level of periodicity but of limited help otherwise.

Another alternative for investigating timing is by looking at the intervals between consecutive events, or IEI [85-87].

The normalized pairwise variability index (**nPVI)** measures the relative variance of consecutive inter-event intervals (IEI). Expectedly, the isochronous signal has the smallest nPVI. The random signal has an nPVI value of 104 in our dataset, only exceeded by the one of the hierarchical signal. Speech nPVI's range between 33 and 77. As the nPVI is a single number its insight power is limited; for instance, it is hard to distinguish variability due to randomness from variability due hierarchy.

The **distribution of the logarithm of the IEI** (log-IEI) highlights the typical IEI in the time series (Figure 5). This method cannot resolve high order temporal structure, as randomizing the IEI would yield the same distributions. On the other hand, distributions are easy to interpret and can be compared using the symmetric Kullback-Leibler divergence (Figure 6) [82]. 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.

Like the IEI distributions, **phase portraits** also highlight the typical scales of the IEIs and further reveal structure within consecutive IEIs. Phase portraits are an excellent alternative for visual inspection, as a structured portrait indicates structured timing [88, 89].

Both, distributions of IEI and phase portraits, can be employed on either the IEI or the log-IEI — as we do here. Taking the logarithm is advantageous because it scales the IEI according to their magnitude thereby allowing to deal with different time scales simultaneously. This logarhythmic scaling may also be quite plausible neuro-biologically [3, 90]. 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 [86, 91]. The fact that these methods can work with both IEIs and their logarithm makes them flexible to work with different types of data sets.

We discussed five frequently used methods for characterizing temporal structure. The first kind of methods, acting on the time series, proved to be insightful for signals with periodicities, but of limited insight otherwise. As for the second kind of methods, acting on the IEI, the nPVIs reduces the series to a single number thereby being of little insight. Distributions of log-IEIs are easy to interpret and combining them with metrics like the symmetric Kullback-Leibler divergence comparison can be automated. Higher order structures are missed by the distributions of log-IEI but well captured by the phase portraits. Certainly, this list of methods is not extensive and the reader may refer to [89] for other methods. Our focus here was in characterizing temporal structures in a time series of events. Other questions related to timing such as how signalers interact vocally over time [86, 91, 92] can be addressed with alternative computational methods [93-95].

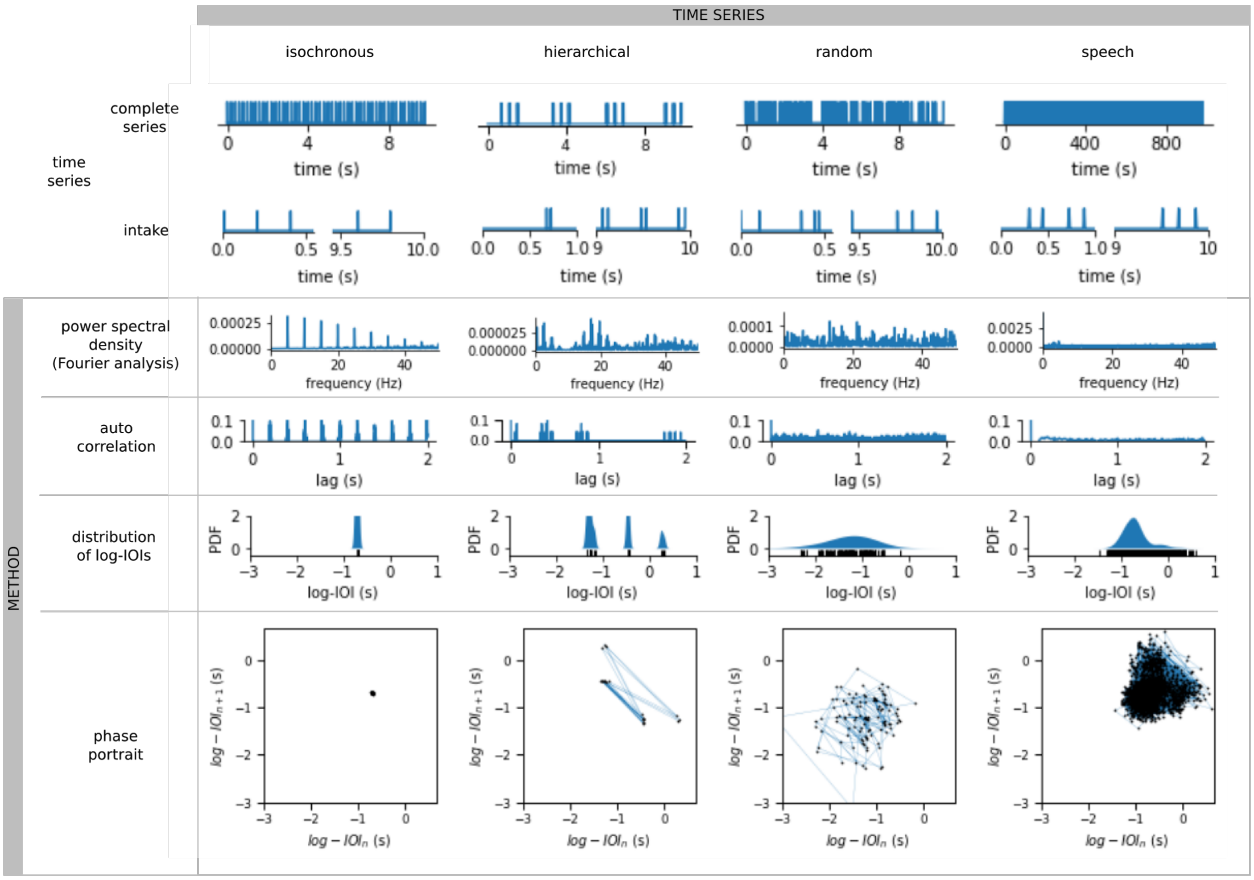

Figure 5. Characterization of the temporal structure of four time series (columns) — isochronous, hierarchical, random and speech — with four techniques (rows): Fourier analysis, autocorrelation, distribution of inter-event intervals (IEIs) and phase portraits. Top two rows show the full time-series and an intake of the beginning and end of the time series. Power spectral density is shown in the range 0 to 50 Hz, computed with the same window size of 2^{17} samples and by zero-padding signals, so that all densities vary in the same range. Autocorrelation function is computed for up to a two seconds lag in the range 0 to 0.1. Distribution of the logarithm base 10 of the IEI (log-IEI). Phase portraits of the log-IEI.

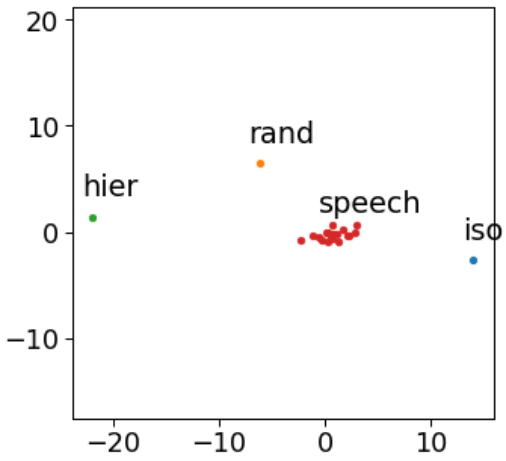

Figure 6. Comparison of distributions of log-IEI with the symmetric Kullback-Leibler divergence. Coordinates determined with a two dimensional scaling of the symmetric Kullback-Leibler divergence between all datasets: isochronous (**iso**), hierarchical (**hier**), random (**rand**) and the 18 **speech** datasets.

Moving to temporal patterns: motor entrainment to speech

Humans are generally highly skilled at processing complex temporal patterns such as music and speech. The majority can perceive the regular pulse of music (i.e., its *beat*), and detect stresses in spoken utterances. 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 [96]. Our proclivity to move to music manifests when we move, spontaneously or deliberately, to its beat by foot or hand tapping, in dance or synchronized walking. These skills are widespread in the general population [97, 98]. A compelling body of evidence from experimental psychology and cognitive neuroscience indicates that rhythm and movement are tightly linked [99-102]. Matching movements to a beat is possible because the temporal dynamics of rhythmic sound lead to the perception of the beat [103], a process linked to internal neurocognitive self-sustained oscillations [104, 105]. The underlying process, called *entrainment*, generates temporal expectancies which drive motor control, by allowing the alignment of movements to the anticipated event times.

The ubiquity of synchronization with music contrasts with the lack of spontaneous motor synchronization with spoken utterances. Yet, prominences in speech (stress patterns) can similarly 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 [106, 107]. For example, in English or German, rhythm is conveyed by accentual patterns whereby strong and weak positions are filled by prominent (i.e., stressed, [108]) and non-prominent (i.e., unstressed) syllables. Like in music, speech patterns evoke a subjective impression of isochrony [109]. This observation is striking, though, given that inter-stress intervals are typically quite variable in speech (coefficients of variations > 30% of the average inter-stress interval; [110, 111]), as compared to expressive music (around 10-30% for inter-beat intervals in performed expressive music; [112]). Moreover, speech meter in conversational speech is clearly less strict and regular than musical meter [113]. Higher regularity is found in metrical speech, however, such as poetry [114-117], and speech production in group such as prayers and chanting (i.e., choral speaking; Cummins, 2009).

In spite of the higher variability of speech temporal patterns, compared with music, the temporal dynamics of metrical speech can still induce expectancies about upcoming events [118, 119]. The substrate of this mechanism lays in the ability of quasi-rhythmic properties of the speech signal to engage oscillatory behavior in the brain [120]. Like music, speech patterns are thus capable of driving dynamic attending [103], underpinned by neurocognitive self-sustained oscillations [118, 121] which phase-lock to the temporal dynamics of syllabic nuclei in speech [5, 119, 122, 123]. 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 [111, 124, 125]. Interestingly, verbal expectancies can be enhanced by concurrent synchronized movement, as found in prosodically diverse languages such as German (a lexical stress-language) and French (a non-stress language) [124, 126]. For example, finger tapping

aligned to accented syllables of spoken utterances benefits the encoding and detection of subtle word changes [124, 126]. 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) is part of games supposed to enhance children's social and verbal skills [127]. 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 children [128] or autistic children [129].

The link between rhythm and movement, and the ability to couple movement to auditory prominences is ubiquitous in humans. The question as to whether other species are capable of synchronization to the beat has fueled research in the last decade. One intriguing hypothesis (the vocal learning - beat perception and synchronization hypothesis [113, 130]) postulates that synchronization to a beat is a by-product 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 [131]. This hypothesis received support by the finding that nonhuman animal species, namely sulfur-crested cockatoos [132, 133] and other bird species that are vocal learners [134, 135]. Motor synchronization in vocal learners is quite flexible (i.e., adapting to a wider range of tempos), occur with complex auditory signals, and is cross-modal [132, 133], thus displaying some of the properties of human synchronization. Recent evidence shows, however, that synchronization to a beat may extend to non-vocal learning species. There is evidence that a chimpanzee can tap above chance, thought quite inflexibly, with a 600-ms metronome [136, 137], a California sea lion can bob her head to the beat of a variety of auditory stimuli [138, 139], and horses do not seem to synchronize [140]. Thus, whether synchronization to beat is selectively associated with vocal learning across species is still an open question [141, 142].

Time and rhythm processing: evolutionary precursors of structural properties in speech, language, and music?

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 unique ability to process temporal and rhythmic structure emerged in humans [141-143]. One idea ties rhythm processing to social synchronization across a number of species (for a review see [7]). Other research exploring the neurocognitive function of time and rhythm processing also points towards similarities of rhythmic and structural properties in speech and music [144, 145] that are primarily denoted in vocal learners [130]. This co-evolution of properties might reside in and still relies on fronto-striatal brain circuitry [119, 146] ([147] for structure evolution), 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 temporal and structural cues found in speech and music (e.g. [5, 148]). However, it remains a mystery (i) how humans derived more complex structures in speech, language, and music from the temporal and sequencing properties of the

fronto-striatal system and (ii) where the structural and functional boundaries lie within this system that separate human and non-human species. Consequently, a comparative approach to evaluate the computational proximity and extent of temporal and rhythmic sequences in species relying on an extended fronto-striatal circuitry is called for [7].

The neurocognitive architecture of time and rhythm processing

The spatiotemporal properties of auditory signals reach the thalamus and cerebellum early on. While precise and 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. pre-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 fronto-striatal system for the encoding of temporal inter-event relations (intervals) that form the basic segmentation unit of sequences. The encoding and the evaluation of the temporal cohesion of sequences require working memory and rely on the prefrontal cortex [149], where temporal and memory information integrates [5].

In production, the generation of a sequence engages the prefrontal cortex. To start and continue this process, an interface of the pre-SMA and fronto-striatal circuitry acts as a “pacemaker” and stabilizes a temporal grid for auditory sequence processing. Sequences adhere to a sophisticated temporal architecture that integrates fast, short-range transitioning temporal events via the cerebellum and slower large-range intervals via the striatum (see also [150] 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. [151, 152]), followed by the premotor, and primary motor cortices for the execution of sequences.

In sum, the described temporal architecture (see Figure 7) composed of fast, short-range and slower, long-range temporal information contributes both to perception as well as the production of auditory-motor sequences such as found in human speech and music [5, 153, 154]. 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 fronto-striatal architecture as temporal processing per se [155, 156].

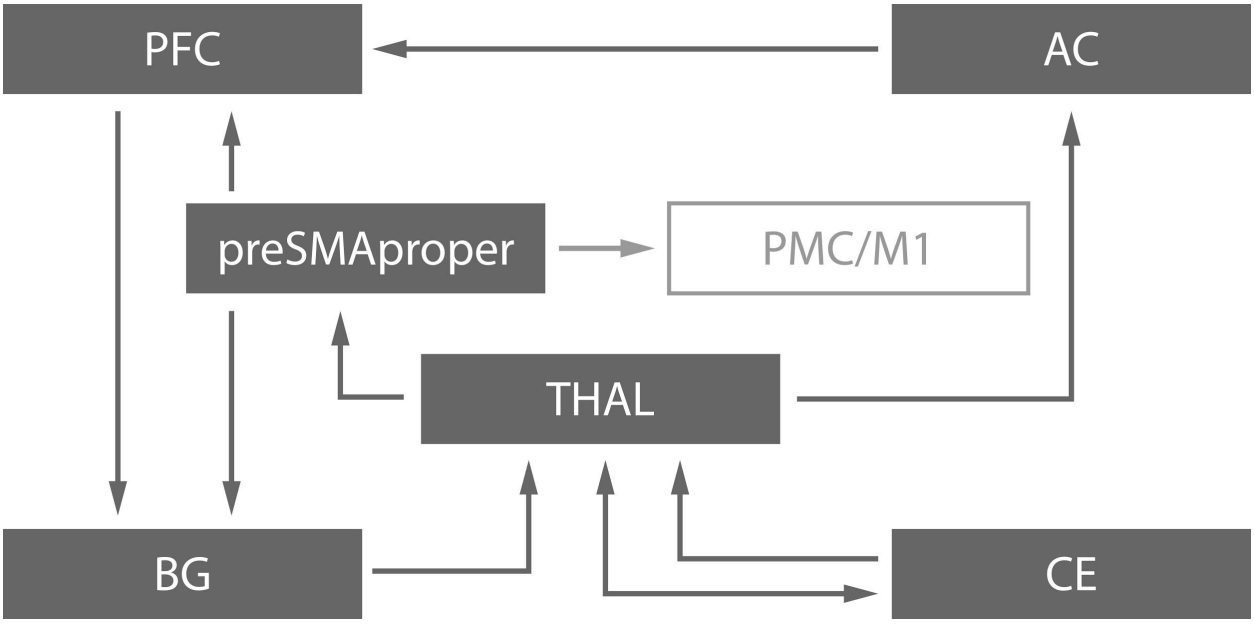

Figure 7. CE = cerebellum, THAL = thalamus, AC = auditory cortex, PFC = prefrontal cortex, BG = basal ganglia, PMC = premotor cortex, M1 = primary motor cortex, preSMA = presupplementary motor area. See also [157].

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

While there is now ample evidence that several species (birds, mammals, and some non-human primates) rely on comparable fronto-striatal circuitry (e.g. [158]) to acquire and produce simple and slightly more complex (grouped) temporally structured sequences, vocal learning alone does not suffice to acquire hierarchical temporal structures found in human speech and music [159]. For example, zebra finches produce temporally structured syllable sequences that align to isochronous click sequences [17] and can perceptually group auditory input [10]. Rhesus monkeys can produce single intervals and synchronize to a metronome [160], while macaques display auditory grouping [161, 162]. None of these species though display temporal structure beyond basic grouping while humans are capable to form simple and hierarchical metrical structure in 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 [119]. Second, complex temporal and rule structure building may rely on an intricate relationship between fronto-striatal and fronto-cerebellar circuitry, where the expansion of the neocerebellum reciprocally pushed the evolution of neocortex such as the prefrontal cortex [163, 164]. This latter structural development is considered crucial for hierarchical structure building. Consequently, investigations of this fronto-striato-cerbellar interface in species producing and perceiving basic temporal structure is required to understand the evolutionary gap between simple and hierarchical temporal structure building in humans and other species.

General discussion and conclusions

This paper is a first attempt at summarizing multiple comparative approaches to human speech rhythm evolution. We showed that 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 hierarchical temporal structure in vocalizations. We compared several techniques to measure temporal and rhythmic structure, both in human speech and in animal vocalizations. We concluded by discussing the neural circuitry underlying speech rhythm, and their relationship with non-vocal motoric actions.

Admittedly, however, there is a big divide among (1) what we know of human speech rhythm, especially from a developmental perspective, (2) speech-related work already performed in animal vocal production and perception, (3) techniques we can use to measure these rhythms behaviorally, (4) comparative work on rhythmic, non-vocal movement, and (5) how our knowledge of the human nervous system relates to that of other species with respect to speech rhythm. We suggest that future work should keep these issues in mind. This would translate into designing experiments which span at least 2 of the 5 still loosely connected areas discussed above.

In addition, some exciting areas of future research (not discussed here) include: the biology-culture interface and genetics. Studying the biology-culture interface can be used to reconcile old, unproductive nature-nurture debates by potentially showing how cognitive biases and cultural transmission interact to deliver the rhythmic structure of speech. Work along these lines has been done for linguistic morphology [165], poetry [166], and musical rhythm [90, 167-169]. An experimental design similar to these studies could be used to show how domain-general biases are amplified by cultural transmission resulting in rhythmic patterns of speech. Tools and methodologies from genetics can be used to map the population genotypes to behavioral variability in rhythmic traits [13, 170]. Initial work has been undertaken in special populations (e.g. those affected by Williams syndrome [171]), but could be extended to the whole population of one species, human or otherwise.

To conclude, the field of comparative rhythm research is rapidly growing, needs a multidisciplinary approach, and its low-hanging fruits are ready to be seized.

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). All authors wrote portions of the manuscript and edited it.

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. *Frontiers in human neuroscience*. **11**.

2. Buhusi, C.V. & W.H. Meck. 2005. What makes us tick? Functional and neural mechanisms of interval timing. *Nature reviews neuroscience*. **6**: 755.
3. Gibbon, J. 1977. Scalar expectancy theory and Weber's law in animal timing. *Psychological review*. **84**: 279.
4. Lejeune, H. & J. Wearden. 2006. Scalar properties in animal timing: Conformity and violations. *The Quarterly journal of experimental psychology*. **59**: 1875-1908.
5. Schwartze, M. & S.A. Kotz. 2013. A dual-pathway neural architecture for specific temporal prediction. *Neuroscience & Biobehavioral Reviews*. **37**: 2587-2596.
6. Teki, S. 2016. A citation-based analysis and review of significant papers on timing and time perception. *Frontiers in Neuroscience*. **10**.
7. Kotz, S., A. Ravignani & W.T. Fitch. 2018. The evolution of rhythm processing. *Trends in Cognitive Sciences*. **22**: 896-910.
8. ten Cate, C. & M. Spierings. 2018. Rules, rhythm and grouping: auditory pattern perception by birds. *Animal Behaviour*.
9. de la Mora, D.M., M. Nespor & J.M. Toro. 2013. Do humans and nonhuman animals share the grouping principles of the iambic‚Äìtrochaic law? *Attention, Perception, & Psychophysics*. **75**: 92-100.
10. Spierings, M., J. Hubert & C. ten Cate. 2017. Selective auditory grouping by zebra finches: testing the iambic–trochaic law. *Animal cognition*. **20**: 665-675.
11. Toro, J.M. & M. Nespor. 2015. Experience-dependent emergence of a grouping bias. *Biology letters*. **11**: 20150374.
12. Hoeschele, M. & W.T. Fitch. 2016. Phonological perception by birds: budgerigars can perceive lexical stress. *Animal cognition*. **19**: 643-654.
13. Ravignani, A. 2019. Rhythm and synchrony in animal movement and communication. *Current zoology*. **65**: 77.
14. Ghazanfar, A.A. 2013. Multisensory vocal communication in primates and the evolution of rhythmic speech. *Behavioral Ecology and Sociobiology*. **67**: 1441-1448.
15. Castellucci, G.A., D. Calbick & D. McCormick. 2018. The temporal organization of mouse ultrasonic vocalizations. *PloS one*. **13**: e0199929.
16. Hyland Bruno, J. & O. Tchernichovski. 2017. Regularities in zebra finch song beyond the repeated motif. *Behavioural processes*.
17. Norton, P. & C. Scharff. 2016. “Bird Song Metronomics”: Isochronous Organization of Zebra Finch Song Rhythm. *Frontiers in Neuroscience*. **10**.
18. Burchardt, L.S., P. Norton, O. Behr*, et al.* 2019. General isochronous rhythm in echolocation calls and social vocalizations of the bat Saccopteryx bilineata. *Royal Society Open Science*. **6**: 181076.
19. Ravignani, A. & G. Madison. 2017. The paradox of isochrony in the evolution of human rhythm. *Frontiers in psychology*.
20. Sabinsky, P.F., O.N. Larsen, M. Wahlberg*, et al.* 2017. Temporal and spatial variation in harbor seal (Phoca vitulina L.) roar calls from southern Scandinavia. *The Journal of the Acoustical Society of America*. **141**: 1824-1834.
21. 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)]. *Journal of the Acoustic Society of America*. **143**: 1-5.
22. Ravignani, A., C. Kello, K. de Reus*, et al.* 2018. Ontogeny of vocal rhythms in harbour seal pups: An exploratory study. *Current Zoology*.
23. 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. *Frontiers in Neuroscience*. **10**.
24. Vernes, S.C. 2017. What bats have to say about speech and language. *Psychonomic bulletin & review*. **24**: 111-117.
25. Oller, D.K. 2000. *The emergence of the speech capacity*. Psychology Press.

26. Vihman, M.M. 2014. *Phonological development: The first two years*. Wiley-Blackwell Boston, MA.
27. Esteve-Gibert, N. & P. Prieto. 2013. Prosody signals the emergence of intentional communication in the first year of life: Evidence from Catalan-babbling infants. *Journal of Child Language*. **40**: 919-944.
28. Snowdon, C.T. & A.M. Elowson. 2001. 'Babbling'in pygmy marmosets: Development after infancy. *Behaviour*. **138**: 1235-1248.
29. Knörnschild, M., O. Behr & O. von Helversen. 2006. Babbling behavior in the sac-winged bat (Saccopteryx bilineata). *Naturwissenschaften*. **93**: 451-454.
30. Aronov, D., A.S. Andalman & M.S. Fee. 2008. A specialized forebrain circuit for vocal babbling in the juvenile songbird. *Science*. **320**: 630-634.
31. MacNeilage, P.F. & B.L. Davis. 1990. "Acquisition of speech production: The achievement of segmental independence". In *Speech production and speech modelling*: 55-68. Springer.
32. 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 development*. **88**: 156-166.
33. Yairi, E., N.G. Ambrose, E.P. Paden*, et al.* 2005. *Early childhood stuttering for clinicians by clinicians*. Pro-ed Austin, TX.
34. Civier, O., D. Bullock, L. Max*, et al.* 2013. Computational modeling of stuttering caused by impairments in a basal ganglia thalamo-cortical circuit involved in syllable selection and initiation. *Brain and language*. **126**: 263-278.
35. 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. *Frontiers in human neuroscience*. **8**: 467.
36. Falk, S., T. Müller & S. Dalla Bella. 2015. Non-verbal sensorimotor timing deficits in children and adolescents who stutter. *Frontiers in psychology*. **6**: 847.
37. Kubikova, L., E. Bosikova, M. Cvikova*, et al.* 2014. Basal ganglia function, stuttering, sequencing, and repair in adult songbirds. *Scientific reports*. **4**: 6590.
38. Chang, S.-E. & D.C. Zhu. 2013. Neural network connectivity differences in children who stutter. *Brain*. **136**: 3709-3726.
39. 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 and language*. **104**: 190-199.
40. Chen, Y., L.E. Matheson & J.T. Sakata. 2016. Mechanisms underlying the social enhancement of vocal learning in songbirds. *Proceedings of the National Academy of Sciences*. **113**: 6641-6646.
41. 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.
42. Parks, S., L. Conger, D. Cusano*, et al.* 2014. Variation in the acoustic behavior of right whale mother-calf pairs. *The Journal of the Acoustical Society of America*. **135**: 2240-2240.
43. Fernald, A., T. Taeschner, J. Dunn*, et al.* 1989. A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of child language*. **16**: 477-501.
44. Martin, A., Y. Igarashi, N. Jincho*, et al.* 2016. Utterances in infant-directed speech are shorter, not slower. *Cognition*. **156**: 52-59.
45. Papoušek, M., H. Papoušek & D. Symmes. 1991. The meanings of melodies in motherese in tone and stress languages. *Infant behavior and development*. **14**: 415-440.
46. Bergeson, T.R. & S.E. Trehub. 2002. Absolute pitch and tempo in mothers' songs to infants. *Psychological Science*. **13**: 72-75.

47. Trehub, S.E. & L. Trainor. 1998. Singing to infants: Lullabies and play songs. *Advances in infancy research*. **12**: 43-78.
48. Dissanayake, E. 2000. *Art and intimacy: How the arts began*. University of Washington Press.
49. Jaffe, J., B. Beebe, S. Feldstein*, et al.* 2001. Rhythms of dialogue in infancy: Coordinated timing in development. *Monographs of the society for research in child development*. i-149.
50. Gratier, M., E. Devouche, B. Guellai*, et al.* 2015. Early development of turn-taking in vocal interaction between mothers and infants. *Front. Psychol*. **6**: 10.3389.
51. 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. *Frontiers in psychology*. **6**: 1492.
52. Pickering, M.J. & S. Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*. **27**: 169-190.
53. Abney, D.H., A.S. Warlaumont, D.K. Oller*, et al.* 2017. Multiple coordination patterns in infant and adult vocalizations. *Infancy*. **22**: 514-539.
54. Falk, S. & C.T. Kello. 2017. Hierarchical organization in the temporal structure of infant-direct speech and song. *Cognition*. **163**: 80-86.
55. Cason, N. & D. Schön. 2012. Rhythmic priming enhances the phonological processing of speech. *Neuropsychologia*. **50**: 2652-2658.
56. Quené, H. & R.F. Port. 2005. Effects of timing regularity and metrical expectancy on spoken-word perception. *Phonetica*. **62**: 1-13.
57. Roncaglia-Denissen, M.P., M. Schmidt-Kassow & S.A. Kotz. 2013. Speech rhythm facilitates syntactic ambiguity resolution: ERP evidence. *PloS one*. **8**: e56000.
58. Rothermich, K., M. Schmidt-Kassow & S.A. Kotz. 2012. Rhythm's gonna get you: regular meter facilitates semantic sentence processing. *Neuropsychologia*. **50**: 232-244.
59. Albin, D.D. & C.H. Echols. 1996. Stressed and word-final syllables in infant-directed speech. *Infant Behavior and Development*. **19**: 401-418.
60. Seidl, A. & E.K. Johnson. 2006. Infant word segmentation revisited: Edge alignment facilitates target extraction. *Developmental science*. **9**: 565-573.
61. Cirelli, L.K., C. Spinelli, S. Nozaradan*, et al.* 2016. Measuring neural entrainment to beat and meter in infants: effects of music background. *Frontiers in neuroscience*. **10**: 229.
62. Hannon, E.E. & S.E. Trehub. 2005. Tuning in to musical rhythms: Infants learn more readily than adults. *Proceedings of the National Academy of Sciences*. **102**: 12639-12643.
63. Winkler, I., G.P. Háden, O. Ladinig*, et al.* 2009. Newborn infants detect the beat in music. *Proceedings of the National Academy of Sciences*. **106**: 2468-2471.
64. Tsang, C.D., S. Falk & A. Hessel. 2017. Infants Prefer Infant-Directed Song Over Speech. *Child development*. **88**: 1207-1215.
65. Phillmore, L.S., J. Fisk, S. Falk*, et al.* 2017. Songbirds as objective listeners: Zebra finches (Taeniopygia guttata) can discriminate infant-directed song and speech in two languages. *International Journal of Comparative Psychology*. **30**.
66. 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 behavior and development*. **33**: 419-430.
67. Martin, J.G. 1972. Rhythmic (hierarchical) versus serial structure in speech and other behavior.
68. Kershenbaum, A., D.T. Blumstein, M.A. Roch*, et al.* 2016. Acoustic sequences in non-human animals: a tutorial review and prospectus. *Biological Reviews*. **91**: 13-52.
69. Kershenbaum, A., A.E. Bowles, T.M. Freeberg*, et al.* 2014. Animal vocal sequences: not the Markov chains we thought they were. *Proceedings of the Royal Society of London B: Biological Sciences*. **281**: 20141370.

70. Rohrmeier, M., W. Zuidema, G.A. Wiggins, *et al.* 2015. Principles of structure building in music, language and animal song. *Philosophical Transactions of the Royal Society B: Biological Sciences*. **370**: 20140097.
71. Singh, N.C. & F.E. Theunissen. 2003. Modulation spectra of natural sounds and ethological theories of auditory processing. *The Journal of the Acoustical Society of America*. **114**: 3394-3411.
72. Allan, D.W. 1966. Statistics of atomic frequency standards. *Proceedings of the IEEE*. **54**: 221-230.
73. Marmelat, V. & D. Delignières. 2012. Strong anticipation: complexity matching in interpersonal coordination. *Exp Brain Res*. **222**: 137-148.
74. Abney, D.H., A. Paxton, R. Dale, *et al.* 2014. Complexity matching in dyadic conversation. *Journal of Experimental Psychology: General*. **143**: 2304.
75. Kello, C.T., S. Dalla Bella, B. Médé, *et al.* 2017. Hierarchical temporal structure in music, speech and animal vocalizations: jazz is like a conversation, humpbacks sing like hermit thrushes. *Journal of The Royal Society Interface*. **14**: 20170231.
76. Ramirez-Aristizabal, A.G., B. Médé & C.T. Kello. 2018. Complexity matching in speech: Effects of speaking rate and naturalness. *Chaos, Solitons & Fractals*. **111**: 175-179.
77. Schneider, S., A.G. Ramirez-Aristizabal, C. Gavilan, *et al.* under review. Complexity matching and lexical matching in monolingual and bilingual conversations.
78. Sawyer, R.K. 2005. Music and conversation. *Musical communication*. **45**: 60.
79. Jadoul, Y., B. Thompson & B. de Boer. 2018. Introducing Parselmouth: A Python Interface to Praat. *Journal of Phonetics*.
80. Ravignani, A. 2018. Spontaneous rhythms in a harbor seal pup calls. *BMC Research Notes*. **11**: 1-4.
81. Grabe, E. & E.L. Low. 2002. Durational variability in speech and the rhythm class hypothesis. *Papers in laboratory phonology*. **7**.
82. Noriega, F. & e. al. 2019. Quantitative analysis of timing in animal vocal sequences. *Preprint*.
83. Rothenberg, D., T.C. Roeske, H.U. Voss, *et al.* 2014. Investigation of musicality in birdsong. *Hearing Research*. **308**: 71-83.
84. Jadoul, Y., A. Ravignani, B. Thompson, *et al.* 2016. Seeking Temporal Predictability in Speech: Comparing Statistical Approaches on 18 World Languages. *Frontiers in Human Neuroscience*. **10**.
85. Bohn, K.M., B. Schmidt-French, S.T. Ma, *et al.* 2008. Syllable acoustics, temporal patterns, and call composition vary with behavioral context in Mexican free-tailed bats. *The Journal of the Acoustical Society of America*. **124**: 1838-1848.
86. Demartsev, V., A. Strandburg-Peshkin, M. Ruffner, *et al.* 2018. Vocal turn-taking in meerkat group calling sessions. *Current Biology*. **28**: 3661-3666. e3663.
87. Frasier, K.E., M.A. Roch, M.S. Soldevilla, *et al.* 2017. Automated classification of dolphin echolocation click types from the Gulf of Mexico. *PLoS computational biology*. **13**: e1005823.
88. Ravignani, A. 2017. Visualizing and interpreting rhythmic patterns using phase space plots. *Music Perception*. **34**: 557-568.
89. Ravignani, A. & P. Norton. 2017. Measuring rhythmic complexity: A primer to quantify and compare temporal structure in speech, movement, and animal vocalizations. *Journal of Language Evolution*.
90. Ravignani, A., B. Thompson, M. Lumaca, *et al.* 2018. Why do durations in musical rhythms conform to small integer ratios? *Frontiers in computational neuroscience*. **12**.
91. Stivers, T., N.J. Enfield, P. Brown, *et al.* 2009. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*. pnas. 0903616106.
92. Ravignani, A. 2018. Timing of antisynchronous calling: A case study in a harbor seal pup (*Phoca vitulina*). *Journal of Comparative Psychology*.

93. Stowell, D., L. Gill & D. Clayton. 2016. Detailed temporal structure of communication networks in groups of songbirds. *Journal of the Royal Society Interface*. **13**: 20160296.
94. Ravignani, A., D.L. Bowling & W.T. Fitch. 2014. Chorusing, synchrony and the evolutionary functions of rhythm. *Frontiers in Psychology*. **5**: 1118.
95. Ravignani, A. & K. de Reus. 2019. Models of animal rhythmic signalling. *Evolutionary Bioinformatics*.
96. 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.
97. Sowiński, J. & S. Dalla Bella. 2013. Poor synchronization to the beat may result from deficient auditory-motor mapping. *Neuropsychologia*. **51**: 1952-1963.
98. 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.
99. 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. *Journal of cognitive neuroscience*. **20**: 226-239.
100. Grahn, J.A. & M. Brett. 2007. Rhythm and beat perception in motor areas of the brain. *Journal of cognitive neuroscience*. **19**: 893-906.
101. Janata, P., S.T. Tomic & J.M. Haberman. 2012. Sensorimotor coupling in music and the psychology of the groove. *Journal of Experimental Psychology: General*. **141**: 54.
102. Zatorre, R.J., J.L. Chen & V.B. Penhune. 2007. When the brain plays music: auditory–motor interactions in music perception and production. *Nature reviews neuroscience*. **8**: 547.
103. Large, E.W. & M.R. Jones. 1999. The dynamics of attending: How people track time-varying events. *Psychological review*. **106**: 119.
104. Fujioka, T., L.J. Trainor, E.W. Large*, et al.* 2012. Internalized timing of isochronous sounds is represented in neuromagnetic beta oscillations. *Journal of Neuroscience*. **32**: 1791-1802.
105. Nozaradan, S., I. Peretz, M. Missal*, et al.* 2011. Tagging the neuronal entrainment to beat and meter. *Journal of Neuroscience*. **31**: 10234-10240.
106. Kiparsky, P. & G. Youmans. 2014. *Rhythm and Meter: Phonetics and Phonology*. Academic Press.
107. Ong, W.J. 2002. 1982. *Orality and Literacy: The Technologizing of the Word*.
108. Beckman, M.E. 1986. Stress and Non-stress Accent. Netherlands Phonetic Archives. *Dordrecht: Foris*.
109. Lehiste, I. 1977. Isochrony reconsidered. *Journal of phonetics*.
110. Dauer, R.M. 1983. Stress-timing and syllable-timing reanalyzed. *Journal of phonetics*.
111. Lidji, P., C. Palmer, I. Peretz*, et al.* 2011. Listeners feel the beat: entrainment to English and French speech rhythms. *Psychonomic bulletin & review*. **18**: 1035-1041.
112. 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. *The Journal of the Acoustical Society of America*. **104**: 1085-1100.
113. Patel, A.D. 2010. *Music, language, and the brain*. Oxford University Press, USA.
114. Lerdahl, F. 2001. The sounds of poetry viewed as music. *Annals of the New York Academy of Sciences*. **930**: 337-354.
115. Obermeier, C., S.A. Kotz, S. Jessen*, et al.* 2016. Aesthetic appreciation of poetry correlates with ease of processing in event-related potentials. *Cognitive, Affective, & Behavioral Neuroscience*. **16**: 362-373.
116. Obermeier, C., W. Menninghaus, M. von Koppenfels*, et al.* 2013. Aesthetic and emotional effects of meter and rhyme in poetry. *Frontiers in psychology*. **4**: 10.
117. Tillmann, B. & W.J. Dowling. 2007. Memory decreases for prose, but not for poetry. *Memory & Cognition*. **35**: 628-639.
118. Jones, M.R. 2009. Musical time. *The handbook of music psychology*. 81-92.

119. Kotz, S.A. & M. Schwartze. 2010. Cortical speech processing unplugged: a timely subcortico-cortical framework. *Trends in cognitive sciences*. **14**: 392-399.
120. Giraud, A.-L. & D. Poeppel. 2012. Cortical oscillations and speech processing: emerging computational principles and operations. *Nature neuroscience*. **15**: 511.
121. Calderone, D.J., P. Lakatos, P.D. Butler*, et al.* 2014. Entrainment of neural oscillations as a modifiable substrate of attention. *Trends in cognitive sciences*. **18**: 300-309.
122. Nozaradan, S., I. Peretz & A. Mouraux. 2012. Selective neuronal entrainment to the beat and meter embedded in a musical rhythm. *Journal of Neuroscience*. **32**: 17572-17581.
123. Peelle, J.E. & M.H. Davis. 2012. Neural oscillations carry speech rhythm through to comprehension. *Frontiers in psychology*. **3**: 320.
124. Falk, S. & S. Dalla Bella. 2016. It is better when expected: aligning speech and motor rhythms enhances verbal processing. *Language, Cognition and Neuroscience*. **31**: 699-708.
125. Falk, S., T. Rathcke & S. Dalla Bella. 2014. When speech sounds like music. *Journal of Experimental Psychology: Human Perception and Performance*. **40**: 1491.
126. Falk, S., C. Volpi-Moncorger & S. Dalla Bella. 2017. Auditory-motor rhythms and speech processing in French and German listeners. *Frontiers in psychology*. **8**: 395.
127. Opie, P. & P. Opie. 1951. *The Oxford dictionary of nursery rhymes*. Clarendon Press Oxford.
128. Schön, D. & B. Tillmann. 2015. Short-and long-term rhythmic interventions: perspectives for language rehabilitation. *Annals of the New York Academy of Sciences*. **1337**: 32-39.
129. 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.
130. Patel, A.D. 2006. Musical rhythm, linguistic rhythm, and human evolution. *Music Perception: An Interdisciplinary Journal*. **24**: 99-104.
131. Dalla Bella, S., M. Berkowska & J. Sowiński. 2015. Moving to the Beat and Singing are Linked in Humans. *Frontiers in human neuroscience*. **9**: 663.
132. Patel, A.D., J.R. Iversen, M.R. Bregman*, et al.* 2009. Studying synchronization to a musical beat in nonhuman animals. *Annals of the New York Academy of Sciences*. **1169**: 459-469.
133. Patel, A.D., J.R. Iversen, M.R. Bregman*, et al.* 2009. Experimental evidence for synchronization to a musical beat in a nonhuman animal. *Current Biology*. **19**: 827-830.
134. Hasegawa, A., K. Okanoya, T. Hasegawa*, et al.* 2011. Rhythmic synchronization tapping to an audio-visual metronome in budgerigars. *Scientific reports*. **1**.
135. Schachner, A., T.F. Brady, I.M. Pepperberg*, et al.* 2009. Spontaneous motor entrainment to music in multiple vocal mimicking species. *Current Biology*. **19**: 831-836.
136. Hattori, Y., M. Tomonaga & T. Matsuzawa. 2013. Spontaneous synchronized tapping to an auditory rhythm in a chimpanzee. *Scientific Reports*. **3**: 1566.
137. Hattori, Y., M. Tomonaga & T. Matsuzawa. 2015. Distractor Effect of Auditory Rhythms on Self-Paced Tapping in Chimpanzees and Humans. *PloS one*. **10**: e0130682.
138. Cook, P., A. Rouse, M. Wilson*, et al.* 2013. A California Sea Lion (*Zalophus californianus*) Can Keep the Beat: Motor Entrainment to Rhythmic Auditory Stimuli in a Non Vocal Mimic. *Journal of Comparative Psychology*. **127**: 1-16.
139. Rouse, A.A., P.F. Cook, E.W. Large*, et al.* 2016. Beat keeping in a sea lion as coupled oscillation: implications for comparative understanding of human rhythm. *Frontiers in Neuroscience*. **10**.
140. Fitzroy, A.B., L. Lobdell, S. Norman*, et al.* 2018. "Horses do not spontaneously engage in tempo-flexible synchronization to a musical beat". In. ICMPC2018, Ed.
141. Ravignani, A. & P. Cook. 2016. The evolutionary biology of dance without frills. *Current Biology*. **26**: R878-R879.

142. 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. *Psychonomic bulletin & review*. **23**: 1647-1659.
143. 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. Cambridge University Press.
144. Kotz, S.A. & M. Schmidt-Kassow. 2015. Basal ganglia contribution to rule expectancy and temporal predictability in speech. *Cortex*. **68**: 48-60.
145. Harding, E.E., D. Sammler, M.J. Henry*, et al.* 2019. Cortical tracking of rhythm in music and speech. *NeuroImage*. **185**: 96-101.
146. 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.
147. Lieberman, P. 2009. FOXP2 and human cognition. *Cell*. **137**: 800-802.
148. Pastor, M.A., E. Macaluso, B. Day*, et al.* 2006. The neural basis of temporal auditory discrimination. *Neuroimage*. **30**: 512-520.
149. Fuster, J.M. 2001. The prefrontal cortex—an update: time is of the essence. *Neuron*. **30**: 319-333.
150. Teki, S., M. Grube, S. Kumar*, et al.* 2011. Distinct neural substrates of duration-based and beat-based auditory timing. *Journal of Neuroscience*. **31**: 3805-3812.
151. Hertrich, I., S. Dietrich & H. Ackermann. 2016. The role of the supplementary motor area for speech and language processing. *Neuroscience & Biobehavioral Reviews*. **68**: 602-610.
152. Tourville, J.A. & F.H. Guenther. 2011. The DIVA model: A neural theory of speech acquisition and production. *Language and cognitive processes*. **26**: 952-981.
153. Kotz, S.A., R.M. Brown & M. Schwartze. 2016. Cortico-striatal circuits and the timing of action and perception. *Current Opinion in Behavioral Sciences*. **8**: 42-45.
154. 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: Academic Press.
155. Grahn, J.A. 2009. The role of the basal ganglia in beat perception: neuroimaging and neuropsychological investigations. *Annals of the New York Academy of Sciences*. **1169**: 35-45.
156. Nozaradan, S., M. Schwartze, C. Obermeier*, et al.* 2017. Specific contributions of basal ganglia and cerebellum to the neural tracking of rhythm. *Cortex*. **95**: 156-168.
157. Schwartze, M. 2012. Adaptation to temporal structure. Max Planck Institute for Human Cognitive and Brain Sciences Leipzig.
158. Merchant, H., J. Grahn, L. Trainor*, et al.* 2015. Finding the beat: a neural perspective across humans and non-human primates. *Philosophical Transactions of The Royal Society B*. **370**: 20140093.
159. Fitch, W.T. 2009. "The biology and evolution of rhythm: Unraveling a paradox". In *Language and Music as Cognitive Systems*. Oxford University Press. Oxford, UK.
160. Zarco, W., H. Merchant, L. Prado*, et al.* 2009. Subsecond timing in primates: Comparison of interval production between human subjects and rhesus monkeys. *Journal of neurophysiology*. **102**: 3191-3202.
161. 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. *Scientific reports*. **7**: 16687.
162. Honing, H. & H. Merchant. 2014. Differences in auditory timing between human and nonhuman primates. *Behavioral and Brain Sciences*. **37**: 557-558.
163. MacLeod, C.E., K. Zilles, A. Schleicher*, et al.* 2003. Expansion of the neocerebellum in Hominoidea. *Journal of Human Evolution*. **44**: 401-429.

164. Weaver, A.H. 2005. Reciprocal evolution of the cerebellum and neocortex in fossil humans. *Proceedings of the National Academy of Sciences*. **102**: 3576-3580.
165. Kirby, S., H. Cornish & K. Smith. 2008. Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*. **105**: 10681-10686.
166. deCastro-Arrazola, V. & S. Kirby. 2019. The emergence of verse templates through iterated learning. *Journal of Language Evolution*.
167. Jacoby, N. & J.H. McDermott. 2017. Integer Ratio Priors on Musical Rhythm Revealed Cross-culturally by Iterated Reproduction. *Current Biology*.
168. Ravignani, A., T. Delgado & S. Kirby. 2016. Musical evolution in the lab exhibits rhythmic universals. *Nature Human Behaviour*. **1**: 0007.
169. Ravignani, A., B. Thompson, T. Grossi*, et al.* 2018. Evolving building blocks of rhythm: How human cognition creates music via cultural transmission. *Annals of the New York Academy of Sciences*.
170. Gingras, B., H. Honing, I. Peretz*, et al.* 2015. Defining the biological bases of individual differences in musicality. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*. **370**: 20140092.
171. Nazzi, T., S. Paterson & A. Karmiloff-Smith. 2003. Early word segmentation by infants and toddlers with Williams syndrome. *Infancy*. **4**: 251-271.