PHILOSOPHICAL TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

Review

Cite this article: Matzinger T, Fitch WT. 2021 Voice modulatory cues to structure across languages and species. *Phil. Trans. R. Soc. B* **376**: 20200393.
https://doi.org/10.1098/rstb.2020.0393

Accepted: 5 July 2021

One contribution of 11 to a theme issue 'Voice modulation: from origin and mechanism to social impact (Part I)'.

Subject Areas:
behaviour, cognition, evolution

Keywords:
voice modulation, prosody, linguistic structure, cross-species comparisons, cross-linguistic comparisons, speech segmentation

Author for correspondence:
Theresa Matzinger
e-mail: theresa.matzinger@univie.ac.at

Voice modulatory cues to structure across languages and species

Theresa Matzinger[1,2] and W. Tecumseh Fitch[1]

[1]Department of Behavioral and Cognitive Biology, University of Vienna, 1030 Vienna, Austria
[2]Department of English, University of Vienna, 1090 Vienna, Austria

TM, 0000-0001-5414-7962; WTF, 0000-0003-1830-0928

Voice modulatory cues such as variations in fundamental frequency, duration and pauses are key factors for structuring vocal signals in human speech and vocal communication in other tetrapods. Voice modulation physiology is highly similar in humans and other tetrapods due to shared ancestry and shared functional pressures for efficient communication. This has led to similarly structured vocalizations across humans and other tetrapods. Nonetheless, in their details, structural characteristics may vary across species and languages. Because data concerning voice modulation in non-human tetrapod vocal production and especially perception are relatively scarce compared to human vocal production and perception, this review focuses on voice modulatory cues used for speech segmentation across human languages, highlighting comparative data where available. Cues that are used similarly across many languages may help indicate which cues may result from physiological or basic cognitive constraints, and which cues may be employed more flexibly and are shaped by cultural evolution. This suggests promising candidates for future investigation of cues to structure in non-human tetrapod vocalizations.

This article is part of the theme issue 'Voice modulation: from origin and mechanism to social impact (Part I)'.

1. Introduction

Although human speech is often thought to be categorically different from non-human animal vocal communication, many aspects of human acoustic communication are directly comparable with those of other land vertebrates. These include both the vocal apparatus itself and the main voice modulatory cues involved in vocal production.[1] In this review, we will argue that voice modulatory cues are similar in the vocal communication of humans and other tetrapods because of (i) shared ancestry, resulting in a similar voice modulation physiology, and (ii) shared functional bases, i.e. similar pressures for efficient communication, resulting in similar cognitive processing due to domain-general mechanisms shared among species.

Voice modulatory cues that are shared and have similar functions in human and non-human tetrapod vocalizations as well as cross-linguistically can be hypothesized to result from anatomical, physiological and cognitive mechanisms that are evolutionarily conserved [4–6]. These include vocal tract anatomy or respiratory constraints, along with domain-general learning constraints and/or cognitive production and perception constraints (e.g. attention and memory; [4,7,8]). By contrast, cues that are neither paralleled in other tetrapods' vocalizations nor cross-linguistically varied may rely on less evolutionarily conserved mechanisms and therefore have larger potential to be shaped by cultural evolutionary processes. For example, the learnability and transmissibility of vocal features to future generations of signallers may not only be influenced by general mechanisms such as how easily the vocal features can be processed, but also by the social environment [9–14]. Thus, factors such as group identity,

THE ROYAL SOCIETY PUBLISHING

Table 1. Voice modulatory cues in human and non-human tetrapod vocal signals, including the physiological factors that constrain them, and the specific ways in which they vary.

shared voice modulatory cues in human and non-human tetrapod vocal signals	constrained by	variation
pauses	lung capacities, respiration	number, duration, position
fundamental frequency (pitch)	subglottal pressure, length of vibrating tissue	magnitude; location of modulation
duration of syllables/units	lung capacities, respiration	magnitude; location of modulation
intensity/amplitude (loudness)	effort with which air is pushed from the lungs	magnitude; location of modulation
voice quality: formants, overtones and spectral envelope	physiology of the vocal tract, flexibility to move articulators	different sound qualities (timbre) and speech sounds (e.g. vowels)
voice quality: glottal pulses	shape of the vibrating tissue, effort with which air is pushed from the lungs	manner of vibration and shape of the glottal pulses (e.g. breathy voice)

community size or prestige may lead to different conventions of voice modulatory patterns in different communities [15,16]. In this review, we attempt to begin disentangling which voice modulatory cues are the result of physiological constraints, of domain-general cognitive mechanisms, and of species- or language-specific conventions and learning pressures, aiming to contribute to the understanding of voice modulation in general evolutionary and cognitive terms.

Because this is a very large research program, our review will cover only some specific aspects of voice modulation. In the first section, we compare different voice modulatory cues across human speech and tetrapod communication, including pauses, fundamental frequency and syllable/unit duration. We discuss similarities and differences in the physiological mechanisms underlying these cues, and then discuss how the effort of producing and perceiving them may be linked to functional pressures in the environment. In the second part of the review, we take a comparative approach across languages, comparing whether different voice modulatory cues used for speech segmentation are similar between or differ among various human languages. Especially regarding the many voice modulatory cues for which animal data remain scarce, comparisons between different human languages may provide valuable insights as to whether the physiological and cognitive mechanisms behind those cues are species-typical (and therefore may be evolutionarily conserved and domain-general), or more flexible language-specific. Finally, our review will identify research gaps and suggest avenues for further work that may help more clearly reveal the underlying physiological and cognitive mechanisms underlying the realization of different voice modulatory cues.

Overall, our comparison between voice modulatory cues in tetrapod vocalizations and across various human languages will show that biological evolution can constrain cultural evolution, and that many of the structures and cues widely used in human speech rely upon basic acoustic and cognitive mechanisms that humans share with other tetrapods.

2. Voice modulation physiology and constraints on vocal production

Humans and other tetrapods share many similarities in the physiological mechanisms used to produce vocal signals. Multiple similarities result from shared respiratory mechanisms, which in turn result from shared ancestry during biological evolution [17,18]. Most tetrapods, including humans, produce vocal signals in a two-stage process: first, a **source** generates acoustic energy using an airflow from the lungs. This source is the larynx in most tetrapods and the syrinx in birds, and consists of vibrating tissue that creates sound by oscillating at a particular rate termed the fundamental frequency (f_0 hereafter). This source signal is then **filtered** in the supralaryngeal vocal tract (upper respiratory tract) via multiple formant frequencies that act as a series of bandpass filters, attenuating or enhancing certain frequency ranges. The actual vocal output fuses these two components (source and filter), which are mostly independent, meaning that f_0 can freely vary independent of formants and vice versa. This process, summarized as the source-filter-theory of vocal production, is shared by humans and most other tetrapods [1,19–21], with the exception of toothed whales [22] and certain whistle vocalization (e.g. in rodents; [23]). This shared physiological basis of vocal production leads to many similarities in both the production and the acoustic output of humans and other tetrapods. Nonetheless, while constrained by physiological production mechanisms, voice modulatory cues can to a certain extent be flexible, and dynamic modifications of particular acoustic parameters can provide structure to the vocal output. Specific voice modulatory cues and the extent to which they can vary (table 1) are reviewed below. In particular, we focus mainly on three cues that are well-investigated with regard to speech segmentation across human languages and will therefore be most relevant for the later sections of this review: pauses, pitch and durational cues.

(a) The physiology of pauses

Among the most distinctive voice modulatory cues are pauses in the vocal signal, which often result from the need to breathe via alternating between exhaling and inhaling. Typically, tetrapods vocalize during exhalation, and vocalizations pause during inhalation. However, some non-human tetrapods vocalize during both exhalation and inhalation, and thus do not need to pause during vocalization (e.g. donkey braying, chimpanzee pant hoots or howler monkey howling, during which inhaling vocalizations are shorter than exhaling vocalizations, but similar in terms of structure and amplitude; [24]). Humans are also capable of ingressive vocalizations such as gasps and chuckles, but these usually

do not replace respiratory pauses and are less flexible in encoding meaning than egressive vocalizations [25–27]. While pauses in tetrapods result from the same physiological mechanism, i.e. respiratory pausing, and are thus constrained by the individuals' lung capacities, they can also vary in their specific realizations. For example, pauses can differ in their duration, number and their position in the vocal stream. Because of this flexibility, tetrapods, including humans, can use pauses to structure the vocal signal in many different ways [28]. For example, birdsong is structured into units commonly termed 'syllables' that are separated by short pauses during which rapid inhalation—'mini-breaths'—occur [29].

(b) The physiology of duration

The duration of phonation at the source can induce durational and rate variations in the vocal output. These durational variations can extend over different domains of the vocal output, such as individual sounds, individual syllables/units or larger stretches of vocalizations [30,31]. This can, for example, lead to different rhythmic patterns, to differences in vocalization tempo, or to distinctive vowel sounds in human speech, where phonemic distinctions between long and short durations are frequent. Duration of one syllable can also disambiguate neighbouring phonemes, as exemplified in the American English words *ladder* (/æ/ longer) and *latter* (/æ/ shorter), which only differ in their vowel length [32]. Human speech sounds that differ in their vowel quality (determined by formants), such as the vowels in the English words *feet* and *fit*, may also have distinctive lengths. Again, physiologically, durational variations are limited by the individuals' breathing capacities, but below that capacious limit, the duration can be varied more or less flexibly to give structure to the vocal output of humans and most non-human tetrapods alike.

(c) The physiology of pitch

Vocal signals are further characterized by the vibration rate of the vibrating tissue, which determines the signals' f_0, often termed pitch in the speech literature [21]. Typically, in tetrapods, f_0 is influenced both by subglottal air pressure and by muscles that regulate the length and tension of the vibrating tissues, i.e. the vocal folds in non-avian tetrapods and the syringeal membranes in birds [33–35]. By modulating these two factors, a pitch can vary within and between vocal signals. To increase pitch, individuals can either increase the subglottal air pressure or the tension of the vibrating tissues. Both of these options require increased effort (see §3) and can provide diversity and structure to vocal signals. For example, typically, on the level of syllables, an increase in pitch signals emphasis ('stress' in the speech literature), whereas pitch modulation on the phrase level can function as a boundary signal [36–39]. Again, the effort required for pitch modulation, and physiology such as the dimensions of the vibrating tissues, limit the pitch range that can be realized. However, within that range, the pitch can be employed flexibly to structure the vocal signal differently, as evidenced by different stress patterns observed in different languages [40].

Fundamentally, tetrapods share these voice modulatory cues because of their shared vocal production physiology, which in turn results from their shared ancestry. Nonetheless, the specific uses and manifestations of these cues can vary considerably across species and languages. For example, species, languages and individuals may differ in when and where they make pauses, when and where pitch rises and falls, or which segments they lengthen or shorten. One useful principle for categorizing and understanding this variation in vocal signals is based on the effort it takes to place emphasis in the vocal signal, using various voice modulatory cues. Thus, the following section will address emphasis and effort in the production of vocal signals, how they are influenced by functional pressures and how this can lead to the cultural evolution of prosodic patterns.

3. Emphasis and effort

It seems intuitively obvious that vocal signals can carry emphasis, and that this requires effort. In particular, it takes more effort to produce emphasized or stressed, i.e. louder, longer and higher-pitched syllables than non-emphasized or unstressed ones. However, despite a common assertion that producing certain voice modulatory cues is more 'energetically efficient' than producing others [41–45], the exact metabolic costs needed to produce and process these cues have rarely been systematically compared. In fact, several studies have shown that vocalizing is not very costly in terms of oxygen, glucose or ATP needed [46–50]. Thus, although it is clear that tensing muscles requires energy consumption, the costs involved in contracting the tiny muscles controlling source characteristics like f_0 may not be appreciable relative to an organism's overall energy budget. Respiratory muscles are larger and potentially more energy-consuming, but they need to be constantly working to serve respiratory functions, independent of vocalization. The relative cost of increased versus decreased pitch or duration during normal speech and frequently produced animal vocalizations will represent an even smaller proportion of net energy expenditure.[2] Finally, the cost of neuronal firing involved in producing or perceiving vocalizations is real, but also very difficult to quantify using current methods. Therefore, at present, we have little choice but to adopt an intuitive definition of 'effort', which can manifest in dynamic effort, i.e. muscular effort for moving the articulators, and neural control effort, i.e. cognitive effort for planning, producing and processing voice modulatory cues. The term 'stress' is used in phonology essentially as a catch-all term, connoting effort and emphasis, but not grounded in detailed syllable-by-syllable measures of expended effort.

How much effort senders will invest in emphasizing vocalizations is largely driven by an interplay of the functional pressure for successful versus efficient communication [42,51]. These pressures may also influence which parts of the signal are emphasized. Emphasis can either extend over the whole signal (e.g. louder vocalizations in noisy environments) or be specific to certain elements of the signal (e.g. stressing certain phrases or syllables); the latter should be more energetically efficient, so we may expect organisms to vary cues across a vocal stream in many cases, as humans do with speech.

One well-studied example where signals are emphasized in their entirety is the so-called Lombard effect: both humans and other tetrapods, including non-human primates, birds and whales tend to vocalize louder and with a higher pitch, i.e. with an increased effort, when there is more background noise [52–56]. When background noise in the environment is reduced, signallers return to vocalizations that need less

effort and decrease their pitch and intensity. A recent example in birdsong occurred when traffic reductions during the Covid-19 pandemic resulted in lower-frequency bird vocalizations, showing that signallers can flexibly adapt their vocalizations to functional pressures in the environment [57].

Further examples of signals with emphasized elements include rhythmic vocalizations and stress or intonation patterns. This kind of emphasis needs both dynamic and cognitive effort on the side of the sender, but creates structure in the signal, which may reduce error, combat habituation or facilitate meaning encoding and processing on the side of the listener [58]. The complex interplay of pressures acting on the sender and receiver may lead to variation in vocal signals that is not fixed genetically but influenced by current properties of the environment [9,10] and shows that once individuals begin to produce vocal cues, there is an opportunity to modulate them. Furthermore, in species that learn their vocalizations (e.g. birdsong or human speech), small production or perception biases for or against certain voice modulatory structural patterns in a certain environment may be amplified over generations of speakers [9]. This may lead to a process of cultural evolution and can result in within-species variation in structural patterns of vocalizations as exemplified by different human languages or different dialects in other tetrapods' vocalizations [59,60].

Thus, overall, how exactly the different voice modulatory cues are used varies within physiological constraints and results from a balancing act between communicating successfully, but with low effort. This in turn depends on functional pressures of listeners and environment, which can vary between different species and languages, and may include factors such as cultural evolution. How exactly different species and different linguistic communities deal with different functional pressures depends both on domain-specific factors such as auditory salience, domain-general cognitive constraints such as memory and attention, but also on more flexible constraints such as social factors. All of these factors will combine to constrain the range within which the different voice modulatory cues can be realized and determine the actual vocal output seen in a language or a species.

4. What we can learn from comparing voice modulatory cues across human languages

Different realizations of voice modulatory cues have been heavily investigated in human languages, but similar investigations in non-human tetrapod vocalizations are comparatively scarce and less systematic. Over the past decades, bioacoustics has made considerable advances in the investigation of non-human tetrapod vocal production, but research on the perception of voice modulatory cues in non-human tetrapods is still in its infancy [61,62]. It is especially difficult to reach firm conclusions about the communicative meaning of voice modulatory structures found in non-human tetrapod vocal signals, given how few cues and species have been systematically investigated.

Therefore, the remaining sections of this review will mainly focus on the comparison of voice modulatory cues across human languages, and specifically the voice modulatory cues that help listeners to segment continuous speech into words. When voice modulatory cues are realized similarly across human languages, this suggests that fundamental physiological constraints or basic cognitive mechanisms may be responsible for these patterns [4–6], and that therefore, due to their shared ancestry, similar cues may also be prevalent in non-human tetrapod vocalizations. We suggest that such patterns may provide starting points for investigating modulation in tetrapod vocal signals. By contrast, cues that differ across different linguistic communities may be largely influenced by different functional pressures in the environment and by cultural evolutionary processes and therefore are more likely to also differ across tetrapod vocalizations.

Comparing voice modulation across human languages and non-human animal vocalizations, and using similarities and differences between them to draw conclusions about the evolutionary roots of vocal communication, is not new [2,63–66]. Similar approaches have already been proposed, for example, by Morton [64,65], who suggested that high and low pitch vocalizations signal similar emotions and attitudes across languages and species. Across species, a low pitch signals largeness, dominance and self-confidence, whereas a high pitch signals smallness, submissiveness and prosociality. Ohala [67] suggests that this biological grounding helps to explain prosodic patterns that are consistent across human languages, such as a final pitch decrease in declarative statements (i.e. utterances signalling dominance and self-assurance) and final pitch increase in questions (i.e. utterances signalling insecurity, submissiveness and need).

Past approaches typically either avoid detailing the specific acoustic cues [66], or treat these cues as fixed for a particular sound class (e.g. low-pitched growls and high-pitched whines). Our goal below is to call attention to how dynamics *within* a call can play a role in structuring acoustic signals, and to investigate the specific acoustic parameters varied. Furthermore, our approach extends previous proposals by highlighting the importance of listener-associated cognitive factors, such as perceptual salience, memory, attention and learnability of prosodic patterns, for biological and cultural evolution. Finally, our proposal captures a more diverse range of prosodic patterns than previous accounts. In contrast with Ohala [67], who explained prosodic patterns by primarily drawing on emotional communication, our account attempts to explain a more diverse set of linguistic structures and meanings.

5. Structure in human languages: the speech segmentation problem and cues to solving it

One crucial first step in the acquisition of linguistic structure is the segmentation of fluent speech into words, before the words' meaning is known. This so-called speech segmentation problem is most acute for infants learning their first language, but also concerns second language learners. For adults, the challenge is particularly evident when they try to identify distinct words while listening to an unfamiliar foreign language [68–70]. Nevertheless, language learners eventually master the speech segmentation problem easily. This is because they implicitly use various cues in the speech stream to identify patterns and regularities, which in turn help them to extract words. Such cues may also play a role in complex sequence learning in bird or whale song (e.g. [71]), but this possibility remains little explored.

Speech segmentation is a challenge that speakers of all human languages have to face and that is therefore well

suited for cross-linguistic comparisons. Over the past decades, cues used in human speech segmentation have been the subject of a large body of research in a variety of different languages such as English [72–76], German [77–80], Italian [78,79,81], French [74], Dutch [74], Spanish [79,82], Portuguese [83], Basque [79], Japanese [73], Cantonese, Mandarin and Russian [84]. This makes it possible to compare the characteristics of speech segmentation cues across languages, answer questions about more general physiological and cognitive mechanisms that are necessary to create and process linguistic structure and identify functional pressures in the respective environments. Among the cues that have been identified to be very important for speech segmentation and creating linguistic structure are transitional probability cues (statistical learning) and the voice modulatory cues that are our focus (e.g. [68,73–75,85–92]).

Transitional probability cues are based on listeners tracking the co-occurrence frequencies of syllables in vocal input ([75,93]; see [94] for a meta-analysis). For example, when hearing the sound sequence *pretty#baby*, listeners can infer that *pretty* and *baby* are distinct words because the syllables *pre* and *ty* as well as *ba* and *by* also co-occur in other sequences such as *pretty#girl* or *lovely#baby*. By contrast, *ty* and *ba* co-occur less frequently and can therefore be assumed to span a word boundary [95]. Speakers of a wide variety of languages have been demonstrated to use such transitional probability cues for language acquisition in similar ways (English: e.g. [72–76]; German: [77–79]; Italian: [78,79,81]; French: [74]; Dutch: [74]; Spanish: [79,82]; Portuguese: [83]; Basque: [79]; Japanese: [73]). Notably, producing different speech sounds and syllable identities is itself a form of voice modulation and is a prerequisite for syllable creation and thus for tracking transitional probabilities. Specifically, individual vowels and consonants are created by moving the articulators, which leads to different formant frequency patterns (see table 1; [96]). While different languages have different speech sounds [40,97], the cross-linguistic ability to modulate the voice in a way that produces different speech sounds is crucial for the cross-linguistic use of transitional probabilities for speech segmentation.

Using transitional probabilities to infer characteristics of a signal appears to be a very general behaviour since in basically any domain of action, including animal vocalizations, certain events are more likely to follow each other than others [98,99]. In humans, the identification of transitional probability cues appears to be based on a domain-general cognitive mechanism, namely statistical learning [100–103]. Furthermore, statistical learning is not a uniquely human cognitive mechanism, and also other species have been demonstrated to use it to deduce signal structure [104]. These can even apply across species; for example, many non-human animals form associations between heterospecific alarm calls and the presence of a predator [105,106]. Also, vocal learning in non-human animals, most notably in birds, is suggested to be supported by statistical computations, although the precise mechanisms behind it are not yet fully understood [104]. It thus seems likely that both humans and many non-human tetrapods rely on a combination of statistical learning and acoustic modulations when learning the structure of their species-specific sound sequences.

Statistical learning is a very general and prominent perceptual and cognitive skill. However, in human languages, voice modulatory cues in the speech stream, such as pauses, or variations in fundamental frequency, syllable duration or intensity (which create word stress, speech rhythm or intonation), can be processed more easily than statistical cues and therefore have more significant effects on speech segmentation [68,76,80,81,91]. However, since voice modulatory cues come in many different realizations and can have many different functions [107], their overall role in signalling linguistic structure, and the cognitive mechanisms needed for processing them, are less understood. While some voice modulatory cues are realized and processed similarly across languages (e.g. [73]), others are subject to cross-linguistic variation (e.g. [74,79]). This raises the question how much the realization and processing of voice modulatory cues are determined by domain-general cognitive or physiological constraints, and how much these cues may be shaped by cultural evolution.

6. Cues to speech perception: when voice modulatory cues count more than transitional probability cues

The efficiency of different voice modulatory cues for speech segmentation has traditionally been tested in artificial language learning experiments [75]. In these experiments, participants are exposed to several minutes of a continuous stream of nonsense speech, consisting of randomly concatenated invented pseudo-words. Listeners can infer from the transition probabilities between syllables which syllable combinations are 'words' of the artificial language and can segment these items from the stream. To test the influence of voice modulatory cues on listeners' segmentation performance, voice modulatory cues are added at different positions to the speech stream and it is measured how this changes listeners' perception of words in the stream.

In such artificial language learning experiments, voice modulatory cues added to continuous speech on the word (e.g. [73,74,80]) and phrase level (e.g. [108–111]) typically enhance speech segmentation compared to transitional probability cues only. Crucially, these cues facilitate speech segmentation most effectively when they converge with the transitional probability cues in the speech stream, i.e. when the voice modulatory cues sound as 'natural' to the listeners as they do in natural speech. By contrast, when voice modulatory cues are designed to conflict with the transitional probability cues in experimental settings and sound 'unnatural' to the listeners, voice modulatory cues disrupt speech segmentation or even override the transitional probability cues [68,76,80,81,91]. Whether voice modulatory cues at certain positions in the speech stream sound natural or unnatural with respect to the transitional probability cues depends both on language-universal cognitive predispositions such as attention, perception or preferences in pattern recognition, and on language-specific word stress patterns typical of the listeners' native languages [73,74,81].

Crucially, many artificial language learning studies tested the influence of language-specific word stress on speech segmentation by using a combination of different voice modulatory cues [74,78,81]. For example, stress cues dominated transitional probability cues when they were implemented as a combination of longer-duration, higher-pitch and higher-intensity of stressed syllables [68,76,91]. While using a combination of different voice modulatory cues closely

simulates natural languages [70,91,92], it does not tell anything about the effects of the individual voice modulatory cues in isolation. However, since different voice modulatory cues have different physiological origins and may be cognitively processed and culturally transmitted differently, investigating them separately can reveal more about the functional pressures acting on linguistic structure [81,88].

Several studies have already addressed the role of voice modulatory cues in isolation. These studies suggest that pauses and lengthening serve as language-universal signals for word-finality (e.g. [73,74,78,79,85,88,112]; but also: [81,113]). By contrast, pitch increase is suggested to be the main perceptual correlate of word stress and is therefore processed differently by speakers of different languages [68,74,78,114]. Speech segmentation studies investigating other prosodic cues such as intensity or voice quality are comparatively rare [88,115], which is why our review below focuses on pauses, durational and pitch modifications.

7. Pauses

Pause cues typically result from the physiological necessity to breathe, but pauses could in principle be expressed at different positions in a vocal signal, or differ in number and duration. Still, in practice, pauses are realized in strikingly similar ways across human languages. Language-universally, pauses are realized at the end of sentences or phrases but hardly ever occur within phrases or within words [28,116]. This is further supported by second language learning studies finding that second language learners have hardly any problems acquiring pause characteristics typical of their second language [117,118]. Thus, while in principle, pauses could occur anywhere within the breathing range, it is most probable that domain-general cognitive processing mechanisms constrain them to occur at specific positions in the vocal output—namely at those positions where they structure the vocal output most efficiently and with the least processing effort.

This and their perceptual salience may explain why pauses are very effective for speech segmentation and outrank other cues in speech segmentation experiments [80].

In animal vocal signals, it is challenging to determine whether pauses occur between or within phrases because units and phrases in animal vocalizations are less clearly defined [119]. Still, because of their shared ancestry with humans, it can be expected that pauses manifest similarly in non-human tetrapods' vocalizations, i.e. at the end of phrases or units. This is why pauses are often used by researchers to determine units in non-human tetrapod vocalizations [120].

8. Final lengthening as a cross-linguistic segmentation cue

One reason why final lengthening may serve as a language-independent speech segmentation cue is that—language-universally—sentence-final or phrase-final elements are lengthened in everyday speech production [28,74,121–124]. The evolutionary origins of final lengthening are that at sentence or phrase boundaries, speakers need to switch from exhaling to inhaling, leading to a pause, and that it takes less effort to slow articulators down before a pause than to stop them abruptly [125–129]. Similar patterns can also be observed in movements in other domains than vocalization. For example, runners also decelerate their movements before stopping [130]. This mechanistic factor seems like a good candidate for a factor that could play a role across languages and in other species' vocal communication systems: a potential universal in vocal communication.

Because kinematic articulatory constraints result in lengthened syllables before sentence or phrase boundaries, listeners may have learned to associate lengthening with boundaries and to exploit it as a cue for speech segmentation [131]. In turn, speakers may have started to intentionally use lengthening to indicate boundaries in the speech stream, also at positions where they did not pause [132]. Via cultural transmission, this may have resulted in final lengthening becoming a conventionalized but still language-universal boundary signal [133]. Because final lengthening is used as a convention for indicating boundaries cross-linguistically, it can be assumed that besides the articulatory constraints that speakers of all languages face equally, its transmission and processing is based on domain-general cognitive constraints.

This notion is supported by the putatively language-independent Iambic/Trochaic Law (=ITL; [134–138]), which states that cross-linguistically, listeners group sounds with longer duration as sequence-final (iambic grouping). Although the ITL focuses on disyllabic words, it can also be generalized to trisyllabic words, suggesting that domain-general cognitive mechanisms may be responsible for this flexibility [73,80]. Still, recently, there has also been evidence that the perceptual groupings of sequences of syllables with variable duration may be shaped more by cultural variation than previously assumed [81,139–141]. Interestingly, the ITL not only applies to linguistic stimuli, but also to tone sequences [115,137] or visual patterns [142]. This further supports the idea that final lengthening as a signal to linguistic structure and thus to low-effort communication results from general cognitive processing mechanisms that also apply to non-linguistic stimuli.

Since deceleration before pauses occurs across various human movements [130] and final lengthening is perceived as a boundary signal across different sensory domains, the mechanisms behind it seem likely to be evolutionarily old. Because of their shared ancestry with humans, a similar vocal tract physiology and similar energetic constraints, final lengthening and its perception as a boundary signal are promising targets for investigation in non-human tetrapods, and there is already some evidence for final lengthening in birdsong [143,144]. Such a cue could play an important role, for example, in structuring turn-taking exchanges between individuals [145,146]. However, to our knowledge, there is no current evidence that non-human tetrapods use final lengthening as a boundary cue at a perceptual level, and when listening to human speech, rats do not appear to group syllables varying in duration according to the ITL [138]. Research with other tetrapods is badly needed to further examine this potential universal.

9. Pitch cues as language-specific segmentation cues

In multiple speech segmentation experiments, similar pitch modifications led to different segmentation patterns in speakers of different native languages [74,78]. For example, word-initial pitch increase facilitated speech segmentation for native speakers of English, whereas word-final pitch increase

facilitated speech segmentation for native speakers of French. These patterns are consistent with the typical stress placements of these languages [74,147].

One explanation why duration and pitch are used differently for speech segmentation is that, potentially, pitch is used as a more reliable cue for the perception of word stress than duration. In speech *production*, stressed syllables are characterized by a co-occurrence of higher pitch and longer duration, and interestingly, cross-linguistically, duration seems to be a more consistent marker of word stress than pitch ([81,148]; but also: [74] for French and English). Still, while being an important acoustic correlate of word stress, lengthening at the same time occurs at boundaries (as discussed in the previous section) and most likely, this durational increase is larger and more consistently applied than that at stressed syllables [125]. As a result, during *perception*, to avoid ambiguities, listeners may rely on lengthening for perceiving boundaries, but rather focus on the pitch for perceiving word stress [74,80].

In general, listeners may need to be more flexible in the perception and cognitive processing of pitch variations compared to durational variations. In natural speech, pitch as a signal for word stress varies more than duration as a signal for sentence or phrase finality, for example, because of loan words with non-typical stress patterns [149–151]. In addition, intonation patterns are variable and depend for example on speaker emotions, attitudes, grammatical structure and focus [152]. Also, while sentence-final pitch decrease in declarative sentences is common across languages [110,123,153], listeners may equally encounter sentence-final pitch increase in yes–no questions. Therefore, overall, the pitch may be a less consistent [41,154–156] and less informative cue during speech segmentation than lengthening. This may explain why neither word-final pitch decrease [80] nor increase facilitated speech segmentation [74,78,157] in artificial language learning experiments, unless for speakers of languages with word-final stress [74,147].

According to the ITL [136,138,158–160], listeners perceive sounds with a higher pitch in sequence-initial positions (trochaic grouping). Interestingly, rats similarly group sequences that vary in pitch as trochees [138]. However, apparently, this perceptual grouping does not play a big role for speech segmentation, since cross-linguistically, a word-initial higher pitch has facilitated speech segmentation in artificial language learning experiments only inconsistently [74,80,81,157]. It can therefore be inferred that the ITL for pitch does not systematically generalize from disyllabic to trisyllabic words, but pitch is instead processed more flexibly.

The apparently rather flexible processing of pitch may result in weak production, perception or learning biases amplifying pitch cues in different directions during the cultural transmission of languages. This may in turn lead to different stress patterns in different languages, making pitch a less reliable signal for speech segmentation than duration. While still originating from basic cognitive processing mechanisms, the cognitive and physiological structures responsible for pitch processing are therefore suggested to be less conserved than those responsible for duration processing. This may have constrained the cultural evolution of pitch cues to linguistic structure less than that of durational cues. Thus, functional pressures for structured signals may hold equally across languages, but how exactly this linguistic structure is archieved, can vary cross-linguistically.

While lexical stress patterns vary across languages and it can be assumed that similar variation should be expected in other tetrapod vocalizations, utterance-final pitch decrease in declarative statements is common across many languages [39,110,123,153]. One reason for this declination may be that the articulators, in this case the vibrating tissues, are slowed down before being brought to a halt, and this lower vibration rate of the tissues leads to a lower pitch [161]. A functional reason may be that pitch declination facilitates turn-taking and thus decreases communicative effort.[3] These physiological and functional constraints are shared across species, which is why pitch declination may be an interesting target for investigation in non-human tetrapod vocal signals. Indeed, there are some indications for final pitch declination and turn-taking in vervet monkeys and rhesus macaques [38]. Investigating other species for final pitch declination could further corroborate the hypothesis that a shared ancestry drives similarities in pitch realization and processing in humans and non-human tetrapods.

10. Conclusion and outlook

Summarizing, our review of human speech modulation shows that f_0, duration and pauses are typically used in systematic ways across languages to help structure the speech signal, but that there is nonetheless considerable variation across languages in the details. Voice modulation can, in many cases, provide cues to structure that are more salient and effective to listeners and learners than statistical measures over the vocal units (e.g. sequential transition probabilities), and can work together with such statistical information or in some cases override it. Thus, although such statistical cues are important (and can be readily computed in animal signals like bird or whale song), they obscure the importance of voice modulation as a key factor in structuring animal communication signals.

How language- or species-specific and cross-linguistic and cross-species cues interact certainly warrants further research. In those cases where comparative information is available, it suggests that the cues used to indicate a structure in the speech signal are both present in vocalizations of other species (unsurprising given their fundamentally similar production mechanisms) and also can be used in similar ways (e.g. phrase-final lengthening in speech and birdsong). Nonetheless, there is currently far too little comparative data to allow any clear conclusions about the degree to which human-typical cues to structure are also used by other species. More research in this area—what we might term 'animal phonology'—is needed to evaluate whether there are broad phylogenetic generalizations to be made, as we have hypothesized here. A rich comparative analysis of these issues could be expected to shed light not just on the evolution of communication across vertebrates, but also about the phylogenetic origins of universals in human speech production and perception.

Data accessibility. This article has no additional data.

Authors' contributions. T.M. was involved in conceptualization and writing the original draft; W.T.F. was involved in conceptualization, writing the review and editing, and supervision.

Competing interests. We declare we have no competing interests.

Funding. This work was supported by the Austrian Science Fund (FWF) DK Grant Cognition and Communication (grant no. W1262-B29) to W.T.F.

Acknowledgements. We would like to thank Andrey Anikin and an anonymous reviewer for helpful comments on an earlier version of this manuscript.

Endnotes

[1]The terms 'voice modulation' and 'prosody' essentially describe the same concept, namely all kinds of vocal dynamic modifications of acoustic parameters during production in humans and non-human tetrapods [1–3]. For the sake of consistency, we will use the term 'voice modulation' throughout this review.

[2]Note that respiratory muscles may induce higher energetic costs in very loud, high or long vocalizations such as during human singing and oratory, or mammalian roaring contests or infrasonic long-distance calls. Because subglottal pressure is an important factor determining both f_0 and sound intensity, very loud and high-pitched vocalizations may require more respiratory effort than normal breathing and vocalization. In addition, very long syllables may disrupt the natural respiratory rhythm.

[3]However, potential analogies between turn taking in human and non-human animal vocalizations have to be interpreted with caution. Since it is difficult to assess the underlying meaning or the intentions behind non-human animal vocal signals, alternation of signals may not necessarily be the result of active turn-taking [146]. In such cases, the communicative benefit gained from alternating vocalizations may differ among species.

References

1. Pisanski K, Cartei V, McGettigan C, Raine J, Reby D. 2016 Voice modulation: a window into the origins of human vocal control? *Trends Cogn. Sci.* **20**, 304–318. (doi:10.1016/j.tics.2016.01.002)
2. Filippi P. 2016 Emotional and interactional prosody across animal communication systems: a comparative approach to the emergence of language. *Front. Psychol.* **7**, 1–19. (doi:10.3389/fpsyg.2016.01393)
3. Pisanski K, Oleszkiewicz A, Plachetka J, Gmiterek M, Reby D. 2018 Voice pitch modulation in human mate choice. *Proc. R. Soc. B* **285**, 1–8. (doi:10.1098/rspb.2018.1634)
4. Fitch WT. 2010 *The evolution of language*. Cambridge, UK: Cambridge University Press.
5. ten Cate C. 2017 Assessing the uniqueness of language: animal grammatical abilities take center stage. *Psychon. Bull. Rev.* **24**, 91–96. (doi:10.3758/s13423-016-1091-9)
6. Christiansen MH, Chater N. 2015 The language faculty that wasn't: a usage-based account of natural language recursion. *Front. Psychol.* **6**, 1–18. (doi:10.3389/fpsyg.2015.01182)
7. Evans N, Levinson SC. 2009 The myth of language universals: language diversity and its importance for cognitive science. *Behav. Brain Sci.* **32**, 429–492. (doi:10.1017/S0140525X0999094X)
8. Fitch WT, Boer Bd, Mathur N, Ghazanfar AA. 2016 Monkey vocal tracts are speech-ready. *Sci. Adv.* **2**, e1600723. (doi:10.1126/sciadv.1600723)
9. Smith K. 2011 Learning bias, cultural evolution of language, and the biological evolution of the language faculty. *Hum. Biol.* **83**, 261–278. (doi:10.3378/027.083.0207)
10. Smith K, Kirby S. 2008 Cultural evolution: implications for understanding the human language faculty and its evolution. *Phil. Trans. R. Soc. B* **363**, 3591–3603. (doi:10.1098/rstb.2008.0145)
11. Smith K, Kalish ML, Griffiths TL, Lewandowsky S. 2008 Introduction. Cultural transmission and the evolution of human behaviour. *Phil. Trans. R. Soc. B* **363**, 3469–3476. (doi:10.1098/rstb.2008.0147)
12. Watson SK, Townsend SW, Schel AM, Wilke C, Wallace EK, Cheng L, West V, Slocombe KE. 2015 Vocal learning in the functionally referential food grunts of chimpanzees. *Curr. Biol.* **25**, 495–499. (doi:10.1016/j.cub.2014.12.032)
13. Whiten A. 2019 Cultural evolution in animals. *Annu. Rev. Ecol. Evol. Syst.* **50**, 27–48. (doi:10.1146/annurev-ecolsys-110218-025040)
14. Williams H, Levin II, Norris DR, Newman AEM, Wheelwright NT. 2013 Three decades of cultural evolution in Savannah sparrow songs. *Anim. Behav.* **85**, 213–223. (doi:10.1016/j.anbehav.2012.10.028)
15. Smith K, Perfors A, Fehér O, Samara A, Swoboda K, Wonnacott E. 2017 Language learning, language use and the evolution of linguistic variation. *Phil. Trans. R. Soc. B* **372**, 1–20. (doi:10.1098/rstb.2016.0051)
16. Raviv L, de Heer Kloots M, Meyer A. 2021 What makes a language easy to learn? A preregistered study on how systematic structure and community size affect language learnability. *Cognition* **210**, 104620. (doi:10.1016/j.cognition.2021.104620)
17. Hoffman M, Taylor BE, Harris MB. 2016 Evolution of lung breathing from a lungless primitive vertebrate. *Respir. Physiol. Neurobiol.* **224**, 11–16. (doi:10.1016/j.resp.2015.09.016)
18. Perry SF, Sander M. 2004 Reconstructing the evolution of the respiratory apparatus in tetrapods. *Respir. Physiol. Neurobiol.* **144**, 125–139. (doi:10.1016/j.resp.2004.06.018)
19. Titze I. 1994 *Principles of voice production*. Englewood Cliffs, NJ: Prentice Hall.
20. Taylor AM, Reby D. 2010 The contribution of source-filter theory to mammal vocal communication research. *J. Zool.* **280**, 221–236. (doi:10.1111/j.1469-7998.2009.00661.x)
21. Fitch WT. 2000 The evolution of speech: a comparative review. *Trends Cogn. Sci.* **4**, 258–267. (doi:10.1016/S1364-6613(00)01494-7)
22. Reidenberg JS, Laitman JT. 2018 Anatomy of underwater sound production with a focus on ultrasonic vocalization in toothed whales including dolphins and porpoises. In *Handbook of behavioral neuroscience, volume 25—Handbook of ultrasonic vocalization: a window into the emotional brain* (ed. SM Brudzynski), pp. 509–519. Amsterdam, The Netherlands: Elsevier B.V. (doi:10.1016/B978-0-12-809600-0.00047-0)
23. Riede T, Borgard HL, Pasch B. 2017 Laryngeal airway reconstruction indicates that rodent ultrasonic vocalizations are produced by an edge-tone mechanism. *R. Soc. Open Sci.* **4**, 170976. (doi:10.1098/rsos.170976)
24. de Cunha RGT, de Oliveira DAG, Holzmann I, Kitchen DM. 2015 Production of loud and quiet calls in Howler Monkeys. In *Howler monkeys: adaptive radiation, systematics, and morphology* (eds MM Kowalewski, PA Garber, L Cortés-Ortiz, B Urbani, D Youlatos), pp. 337–368. New York, NY: Springer.
25. Eklund R. 2008 Pulmonic ingressive phonation: diachronic and synchronic characteristics, distribution and function in animal and human sound production and in human speech. *J. Int. Phon. Assoc.* **38**, 235–324. (doi:10.1017/S0025100308003563)
26. Eklund R. 2007 Pulmonic ingressive speech: a neglected universal? *Proc. Fonetik* **50**, 21–24.
27. Eklund R. 2019 Pulmonic ingressive speech. In *The SAGE encyclopedia of human communication sciences and disorders* (eds JS Damico, MJ Ball), pp. 1529–1532. Thousand Oaks, CA: Sage Publications.
28. Fletcher J. 2010 The prosody of speech: timing and rhythm. In *The handbook of phonetic sciences*, 2nd edn (eds WJ Hardcastle, J Laver, FE Gibbon), pp. 523–602. Hoboken, NJ: Wiley-Blackwell.
29. Hartley RS, Suthers RA. 1989 Airflow and pressure during canary song: direct evidence for mini-breaths. *J. Comp. Physiol. A* **165**, 15–26. (doi:10.1007/BF00613795)
30. Ravignani A, Dalla Bella S, Falk S, Kello CT, Noriega F, Kotz SA. 2019 Rhythm in speech and animal vocalizations: a cross-species perspective. *Ann. NY Acad. Sci.* **1453**, 79–98. (doi:10.1111/nyas.14166)
31. Jacewicz E, Fox RA, Wei L. 2010 Between-speaker and within-speaker variation in speech tempo of American English. *J. Acoust. Soc. Am.* **128**, 839–850. (doi:10.1121/1.3459842)
32. House AS. 1961 On vowel duration in English. *J. Acoust. Soc. Am.* **33**, 1174–1178. (doi:10.1121/1.1908941)

33. Riede T. 2011 Subglottal pressure, tracheal airflow, and intrinsic laryngeal muscle activity during rat ultrasound vocalization. *J. Neurophysiol.* **106**, 2580–2592. (doi:10.1152/jn.00478.2011)
34. Riede T, Tokuda IT, Farmer CG. 2011 Subglottal pressure and fundamental frequency control in contact calls of juvenile *Alligator mississippiensis*. *J. Exp. Biol.* **214**, 3082–3095. (doi:10.1242/jeb.051110)
35. Riede T, Goller F. 2010 Functional morphology of the sound-generating labia in the syrinx of two songbird species. *J. Anat.* **216**, 23–36. (doi:10.1111/j.1469-7580.2009.01161.x)
36. Lehiste I. 1970 *Suprasegmentals*. Cambridge, UK: MIT Press.
37. Adams C, Munro RR. 1978 In search of the acoustic correlates of stress: fundamental frequency.*Phonetica* **35**, 125–156. (doi:10.1159/000259926)
38. Hauser MD, Fowler CA. 1992 Fundamental frequency declination is not unique to human speech: evidence from nonhuman primates. *J. Acoust. Soc. Am.* **91**, 363–369. (doi:10.1121/1.402779)
39. Pierrehumbert J. 1979 The perception of fundamental frequency declination. *J. Acoust. Soc. Am.* **66**, 363–369. (doi:10.1121/1.383670)
40. Dryer MS, Haspelmath M (eds). 2013 *The world atlas of language structures online [internet]*. Munich, Germany: Max Planck Digital Library. See https://wals.info.
41. Bybee J. 2007 *Frequency of use and the organization of language*. Oxford, UK: Oxford University Press.
42. Fedzechkina M, Jaeger TF, Newport EL. 2012 Language learners restructure their input to facilitate efficient communication. *Proc. Natl Acad. Sci. USA* **109**, 17 897–17 902. (doi:10.1073/pnas.1215776109)
43. Gibson E, Futrell R, Piandadosi ST, Dautriche I, Mahowald K, Bergen L, Levy R. 2019 How efficiency shapes human language. *Trends Cogn. Sci.* **23**, 389–407. (doi:10.1016/j.tics.2019.02.003)
44. Blevins J. 2004 *Evolutionary phonology - The emergence of sound patterns*, vol. 112. Cambridge, UK: Cambridge University Press.
45. Zipf GK. 1949 *Human behavior and the principle of least effort*. Cambridge, MA: Addison-Wesley Press.
46. Moon SJ, Lindblom B. 2003 Two experiments on oxygen consumption during speech production: vocal effort and speaking tempo. In *15th ICPhS.*, pp. 3129–3132.
47. Horn AG, Leonard ML, Weary DM. 1995 Oxygen consumption during crowing by roosters: talk is cheap. *Anim. Behav.* **50**, 1171–1175. (doi:10.1016/0003-3472(95)80033-6)
48. Oberweger K, Goller F. 2001 The metabolic cost of birdsong production. *J. Exp. Biol.* **204**, 3379–3388. (doi:10.1242/jeb.204.19.3379)
49. Speakman JR, Racey PA. 1991 No cost of echolocation for bats in flight. *Nature* **350**, 421–423. (doi:10.1038/350421a0)
50. Foskolos I, Aguilar de Soto N, Madsen PT, Johnson M. 2019 Deep-diving pilot whales make cheap, but powerful, echolocation clicks with 50 µl of air. *Sci. Rep.* **9**, 1–9. (doi:10.1038/s41598-019-51619-6)
51. Grice HP. 1975 Logic and conversation. In *Syntax and semantics* (eds P Cole, J Morgan), pp. 41–58. New York, NY: Academic Press.
52. Brumm H, Naguib M. 2009 Chapter 1 Environmental acoustics and the evolution of bird song. *Adv. Study Behav.* **40**, 1–33. (doi:10.1016/s0065-3454(09)40001-9)
53. Brumm H, Zollinger A. 2011 The evolution of the Lombard effect: 100 years of psychoacoustic research. *Behaviour* **148**, 1173–1198. (doi:10.1163/000579511X605759)
54. Egnor SER, Hauser MD. 2006 Noise-induced vocal modulation in cotton-top tamarins (*Saguinus oedipus*). *Am. J. Primatol.* **68**, 1183–1190. (doi:10.1002/ajp.20317)
55. Nemeth E, Pieretti N, Zollinger SA, Geberzahn N, Partecke J, Mirand AC, Brumm H. 2013 Bird song and anthropogenic noise: vocal constraints may explain why birds sing higher-frequency songs in cities. *Proc. R. Soc. B* **280**, 20122798. (doi:10.1098/rspb.2012.2798)
56. Manabe K, Sadr EI, Dooling RJ. 1998 Control of vocal intensity in budgerigars (*Melopsittacus undulatus*): differential reinforcement of vocal intensity and the Lombard effect. *J. Acoust. Soc. Am.* **103**, 1190–1198. (doi:10.1121/1.421227)
57. Derryberry EP, Phillips JN, Derryberry GE, Blum MJ, Luther D. 2020 Singing in a silent spring: birds respond to a half-century soundscape reversion during the COVID-19 shutdown. *Science* **370**, 575–579. (doi:10.1126/science.abd5777)
58. Jaeger TF, Tily H. 2011 On language 'utility': processing complexity and communicative efficiency. *Wiley Interdiscip. Rev. Cogn. Sci.* **2**, 323–335. (doi:10.1002/wcs.126)
59. Garland EC, Goldizen AW, Rekdahl ML, Constantine R, Garrigue C, Hauser ND, Poole MM, Robbins J, Noad MJ. 2011 Dynamic horizontal cultural transmission of humpback whale song at the ocean basin scale. *Curr. Biol.* **21**, 687–691. (doi:10.1016/j.cub.2011.03.019)
60. Laland KN, Galef B. (eds) 2009 *The question of animal culture*. Cambridge, MA: Harvard University Press.
61. Bradbury JW, Vehrencamp SL. 2011 *Principles of animal communication*, 2nd edn. Sunderland, MA: Sinauer Associates.
62. Erbe C, Dent ML. 2017 Animal bioacoustics. *Acoust Today* **13**, 65–67.
63. Filippi P, Hoeschele M, Spierings M, Bowling DL. 2019 Temporal modulation in speech, music, and animal vocal communication: evidence of conserved function. *Ann. N Y Acad. Sci.* **1453**, 99–113. (doi:10.1111/nyas.14228)
64. Morton ES. 1977 On the occurrence and significance of motivation–structural rules in some bird and mammal sounds. *Am. Nat.* **111**, 855–869. (doi:10.1086/283219)
65. Morton ES. 1982 Grading, discretenesss, redundancy, and motivational–structural rules. In *Acoustic communication in birds* (eds D Kroodsma, EH Miller), pp. 183–212. New York, NY: Academic Press.
66. Filippi P *et al.* 2017 Humans recognize emotional arousal in vocalizations across all classes of terrestrial vertebrates: evidence for acoustic universals. *Proc. R. Soc. B* **284**, 20170990. (doi:10.1098/rspb.2017.0990)
67. Ohala JJ. 1983 Cross-language use of pitch: an ethological view. *Phonetica* **40**, 1–18. (doi:10.1159/000261678)
68. Johnson EK, Jusczyk PW. 2001 Word segmentation by 8-month-olds: when speech cues count more than statistics. *J. Mem. Lang.* **44**, 548–567. (doi:10.1006/jmla.2000.2755)
69. Endress AD, Hauser MD. 2010 Word segmentation with universal prosodic cues. *Cogn. Psychol.* **61**, 177–199. (doi:10.1016/j.cogpsych.2010.05.001)
70. Erickson LC, Thiessen ED. 2015 Statistical learning of language: theory, validity, and predictions of a statistical learning account of language acquisition. *Dev. Rev.* **37**, 66–108. (doi:10.1016/j.dr.2015.05.002)
71. Fehér O, Ljubičić I, Suzuki K, Okanoya K, Tchernichovski O. 2017 Statistical learning in songbirds: from self-tutoring to song culture. *Phil. Trans. R. Soc. B* **372**, 20160053. (doi:10.1098/rstb.2016.0053)
72. Mattys SL, Jusczyk PW, Luce PA, Morgan JL. 1999 Phonotactic and prosodic effects on word segmentation in infants. *Cogn. Psychol.* **38**, 465–494. (doi:10.1006/cogp.1999.0721)
73. Frost RLA, Monaghan P, Tatsumi T. 2017 Domain-general mechanisms for speech segmentation: the role of duration information in language learning. *J. Exp. Psychol. Hum. Percept. Perform.* **43**, 466–476. (doi:10.1037/xhp0000325)
74. Tyler MD, Cutler A. 2009 Cross-language differences in cue use for speech segmentation. *J. Acoust. Soc. Am.* **126**, 367–376. (doi:10.1121/1.3129127)
75. Saffran JR, Aslin RN, Newport EL. 1996 Statistical learning by 8-month-old infants. *Science* **274**, 1926–1928. (doi:10.1126/science.274.5294.1926)
76. Thiessen ED, Saffran JR. 2003 When cues collide: use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Dev. Psychol.* **39**, 706–716. (doi:10.1037/0012-1649.39.4.706)
77. Marimon Tarter M. 2019 *Word segmentation in German-learning infants and German-speaking adults: prosodic and statistical cues*. Potsdam, Germany: University of Potsdam.
78. Ordin M, Nespor M. 2016 Native language influence in the segmentation of a novel language. *Lang. Learn Dev.* **12**, 461–481. (doi:10.1080/15475441.2016.1154858)
79. Ordin M, Polyanskaya L, Laka I, Nespor M. 2017 Cross-linguistic differences in the use of durational cues for the segmentation of a novel language. *Mem. Cognit.* **45**, 863–876. (doi:10.3758/s13421-017-0700-9)
80. Matzinger T, Ritt N, Fitch WT. 2021 The influence of different prosodic cues on word segmentation. *Front. Psychol.* **12**, 458. (doi:10.3389/fpsyg.2021.622042)

81. Ordin M, Nespor M. 2013 Transition probabilities and different levels of prominence in segmentation. *Lang. Learn.* **63**, 800–834. (doi:10.1111/lang.12024)
82. Cunillera T, Càmara E, Laine M, Rodríguez-Fornells A. 2009 Speech segmentation is facilitated by visual cues. *Q. J. Exp. Psychol.* **63**, 1–15. (doi:10.1080/17470210902888809)
83. Fernandes T, Ventura P, Kolinsky R. 2011 The relative weight of statistical and prosodic cues in speech segmentation: a matter of language-(in)dependency and of signal quality. *J. Port. Linguist.* **10**, 87. (doi:10.5334/jpl.102)
84. Gómez DM, Mok P, Ordin M, Mehler J, Nespor M. 2018 Statistical speech segmentation in tone languages: the role of lexical tones. *Lang. Speech* **61**, 84–96. (doi:10.1177/0023830917706529)
85. Saffran JR, Newport EL, Aslin RN. 1996 Word segmentation: the role of distributional cues. *J. Mem. Lang.* **35**, 606–621. (doi:10.1006/jmla.1996.0032)
86. Saffran JR, Johnson EK, Aslin RN, Newport EL. 1999 Statistical learning of tone sequences by human infants and adults. *Cognition* **70**, 27–52. [cited 2016 Dec 16]. (doi:10.1016/S0010-0277(98)00075-4)
87. Aslin RN, Saffran JR, Newport EL. 1998 Computation of conditional probability statistics by human infants. *Psychol. Sci.* **9**, 321–324. (doi:10.1111/1467-9280.00063)
88. Hay JSF, Saffran JR. 2012 Rhythmic grouping biases constrain infant statistical learning. *Infancy* **17**, 610–641. (doi:10.1111/j.1532-7078.2011.00110.x)
89. Johnson EK. 2008 Infants use prosodically conditioned acoustic–phonetic cues to extract words from speech. *J. Acoust. Soc. Am.* **123**, EL144–EL148. (doi:10.1121/1.2908407)
90. Johnson EK. 2012 Bootstrapping language: are infant statisticians up to the job? In *Statistical learning and language acquisition* (eds P Rebuschat, J Williams), pp. 55–90. Berlin, Germany: Mouton de Gruyter.
91. Johnson EK, Seidl AH. 2009 At 11 months, prosody still outranks statistics. *Dev. Sci.* **12**, 131–141. (doi:10.1111/j.1467-7687.2008.00740.x)
92. Johnson EK, Tyler MD. 2010 Testing the limits of statistical learning for word segmentation. *Dev. Sci.* **13**, 339–345. (doi:10.1111/j.1467-7687.2009.00886.x)
93. Romberg AR, Saffran JR. 2010 Statistical learning and language acquisition. *Wiley Interdiscip. Rev. Cogn. Sci.* **1**, 906–914. (doi:10.1002/wcs.78)
94. Black A, Bergmann C. 2017 Quantifying infants' statistical word segmentation: a meta-analysis. In *Proc. 39th Annual Conf. Cogn. Sci. Soc.*, pp. 124–129. See https://pdfs.semanticscholar.org/0807/41051b6e2b74d2a1fc2e568c3dd11224984b.pdf.
95. Saffran JR. 2003 Statistical language learning: mechanisms and constraints. *Curr. Dir. Psychol. Sci.* **12**, 110–114. (doi:10.1111/1467-8721.01243)
96. Redford MA (ed.). 2015 *The handbook of speech production*. Chichester, UK: Wiley Blackwell.
97. Moran S, McCloy D. (eds) 2019 PHOIBLE 2.0. Jena, Germany: Max Planck Institute for the Science of Human History. (Available online at http://phoible.org; accessed on 29-09-2021.).
98. Gagniuc PA. 2017 *Markov chains: from theory to implementation and experimentation*. Hoboken, NJ: John Wiley & Sons.
99. Shannon CE. 1948 A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 623–656. (doi:10.1002/j.1538-7305.1948.tb00917.x)
100. Bogaerts L, Frost R, Christiansen MH. 2020 Integrating statistical learning into cognitive science. *J. Mem. Lang.* **115**, 104167. (doi:10.1016/j.jml.2020.104167)
101. Newport EL. 2016 Statistical language learning: computational, maturational, and linguistic constraints. *Lang. Cogn.* **8**, 447–461. (doi:10.1017/langcog.2016.20)
102. Thiessen ED, Erickson LC. 2013 Beyond word segmentation: a two-process account of statistical learning. *Curr. Dir. Psychol. Sci.* **22**, 239–243. (doi:10.1177/0963721413476035)
103. Palmer SD, Mattys SL. 2016 Speech segmentation by statistical learning is supported by domain-general processes within working memory. *Q. J. Exp. Psychol.* **69**, 2390–2401. (doi:10.1080/17470218.2015.1112825)
104. Santolin C, Saffran JR. 2018 Constraints on statistical learning across species. *Trends Cogn. Sci.* **22**, 52–63. (doi:10.1016/j.tics.2017.10.003)
105. Hauser MD. 1988 How infant vervet monkeys learn to recognize starling alarm calls: the role of experience. *Behaviour* **105**, 187–201. (doi:10.1163/156853988X00016)
106. Rainey HJ, Zuberbühler K, Slater PJB. 2004 Hornbills can distinguish between primate alarm calls. *Proc. R. Soc. B* **271**, 755–759. (doi:10.1098/rspb.2003.2619)
107. Cole J. 2015 Prosody in context: a review. *Lang. Cogn. Neurosci.* **30**, 1–31. (doi:10.1080/23273798.2014.963130)
108. Christophe A, Peperkamp S, Pallier C, Block E, Mehler J. 2004 Phonological phrase boundaries constrain lexical access I. Adult data. *J. Mem. Lang.* **51**, 523–547. (doi:10.1016/j.jml.2004.07.001)
109. Gout A, Christophe A, Morgan JL. 2004 Phonological phrase boundaries constrain lexical access II. Infant data. *J. Mem. Lang.* **51**, 548–567. (doi:10.1016/j.jml.2004.07.002)
110. Langus A, Marchetto E, Bion RAH, Nespor M. 2012 Can prosody be used to discover hierarchical structure in continuous speech? *J. Mem. Lang.* **66**, 285–306. (doi:10.1016/j.jml.2011.09.004)
111. Shukla M, Nespor M, Mehler J. 2007 An interaction between prosody and statistics in the segmentation of fluent speech. *Cogn. Psychol.* **54**, 1–32. (doi:10.1016/j.cogpsych.2006.04.002)
112. Kim S, Broersma M, Cho T. 2012 The use of prosodic cues in learning new words in an unfamiliar language. *Stud. Second Lang. Acquis.* **34**, 415–444. (doi:10.1017/S0272263112000137)
113. White L, Benavides-Varela S, Mády K. 2020 Are initial-consonant lengthening and final-vowel lengthening both universal word segmentation cues? *J. Phon.* **81**, 100982. (doi:10.1016/j.wocn.2020.100982)
114. Morgan JL, Saffran JR. 1995 Emerging integration of sequential and suprasegmental information in preverbal speech segmentation. *Child Dev.* **66**, 911–936. (doi:10.2307/1131789)
115. Trainor LJ, Adams B. 2000 Infants' and adults' use of duration and intensity cues in the segmentation of tone patterns. *Percept. Psychophys.* **62**, 333–340. (doi:10.3758/BF03205553)
116. Zellner B. 1994 Pauses and the temporal structure of speech. In *Fundamentals of speech synthesis and speech recognition* (ed. E Keller), pp. 41–62. Chichester, UK: John Wiley.
117. Matzinger T, Ritt N, Fitch WT. 2020 Non-native speaker pause patterns closely correspond to those of native speakers at different speech rates. *PLoS ONE* **15**, 1–20. (doi:10.1371/journal.pone.0230710)
118. Derwing TM, Munro MJ, Thomson RI, Rossiter MJ. 2009 The relationship between L1 fluency and L2 fluency development. *Stud. Second Lang. Acquis.* **31**, 533–557. (doi:10.1017/S0272263109990015)
119. Kershenbaum A *et al.* 2016 Acoustic sequences in non-human animals: a tutorial review and prospectus. *Biol. Rev. Camb. Phil. Soc.* **91**, 13–52. (doi:10.1111/brv.12160)
120. Mann DC, Hoeschele M. 2020 Segmental units in nonhuman animal vocalization as a window into meaning, structure, and the evolution of language. *Anim. Behav. Cogn.* **7**, 151–158. (doi:10.26451/abc.07.02.09.2020)
121. Klatt DH. 1975 Linguistic uses of segmental duration in English: acoustic and perceptual evidence. *J. Acoust. Soc. Am.* **59**, 1208–1220. (doi:10.1121/1.380986)
122. Oller DK. 1973 The effect of position in utterance on speech segment duration in English. *J. Acoust. Soc. Am.* **54**, 1235–1247. (doi:10.1121/1.1914393)
123. Vaissière J. 1983 Language-independent prosodic features. In *Springer series in language and communication 14: prosody: models and measurements* (eds A Cutler, DR Ladd), pp. 53–66. Berlin, Germany: Springer.
124. Seifart F, Strunk J, Danielsen S, Hartmann I, Pakendorf B, Wichmann S, Witzlack-Makarevich A, Himmelmann NP, Bickel B. 2021 The extent and degree of utterance-final word lengthening in spontaneous speech from 10 languages. *Linguist. Vanguard* **7**, 1–14. (doi:10.1515/lingvan-2019-0063)
125. Berkovits R. 1994 Durational effects in final lengthening, gapping, and contrastive stress. *Lang. Speech* **37**, 237–250. (doi:10.1177/002383099403700302)
126. Byrd D. 2000 Articulatory vowel lengthening and coordination at phrasal junctures. *Phonetica* **57**, 3–16. (doi:10.1159/000028456)
127. Edwards J, Beckman ME, Fletcher J. 1991 The articulatory kinematics of final lengthening. *J. Acoust. Soc. Am.* **89**, 369–382. (doi:10.1121/1.400674)
128. Myers S, Hansen BB. 2007 The origin of vowel length neutralization in final position: evidence

from Finnish speakers. *Nat. Lang. Linguist. Theory* **25**, 157–193. (doi:10.1007/s11049-006-0001-7)

129. Krivokapic J. 2014 Gestural coordination at prosodic boundaries and its role for prosodic structure and speech planning processes. *Phil. Trans. R. Soc. B* **369**, 20130397. (doi:10.1098/rstb.2013.0397)
130. Friberg A, Sundberg J. 1999 Does music performance allude to locomotion ? A model of final ritardandi derived from measurements of stopping runners. *J. Acoust. Soc. Am.* **105**, 1469–1484. (doi:10.1121/1.426687)
131. Scott DR. 1982 Duration as a cue to the perception of a phrase boundary. *J. Acoust. Soc. Am.* **71**, 996–1007. (doi:10.1121/1.387581)
132. Kachkovskaia T. 2014 Phrase-final lengthening in Russian: pre-boundary or pre-pausal? In *Speech and Computer. SPECOM 2014. Lecture Notes in Computer Science*, vol. 8773 (eds A Ronzhin, R Potapova, V Delic) pp. 353–359. Cham, Switzerland: Springer. (doi:10.1007/978-3-319-11581-8_44)
133. Christiansen MH, Kirby S. 2003 Language evolution: consensus and controversies. *Trends Cogn. Sci.* **7**, 300–307. (doi:10.1016/S1364-6613(03)00136-0)
134. Bolton TL. 1894 Rhythm. *Am. J. Psychol.* **6**, 145–238. (doi:10.2307/1410948)
135. Woodrow H. 1909 *A quantitative study of rhythm: the effect of variations in intensity, rate and duration*. New York, NY: The Science Press.
136. Hayes B. 1995 *Metrical stress theory: principles and case studies*. Chicago, IL: The University of Chicago Press.
137. Hay JSF, Diehl RL. 2007 Perception of rhythmic grouping: testing the iambic/trochaic law. *Perecept. Psychophys.* **69**, 113–122. (doi:10.3758/BF03194458)
138. De la Mora DM, Nespor M, Toro JM. 2013 Do humans and nonhuman animals share the grouping principles of the iambic – trochaic law? *Atten. Percept. Psychophys.* **75**, 92–100. (doi:10.3758/s13414-012-0371-3)
139. Iversen JR, Patel AD, Ohgushi K. 2008 Perception of rhythmic grouping depends on auditory experience. *J. Acoust. Soc. Am.* **124**, 2263–2271. (doi:10.1121/1.2973189)
140. Crowhurst M. 2016 Iambic–Trochaic law effects among native speakers of Spanish and English. *Lab. Phonol.* **7**, 12. (doi:10.5334/labphon.42)
141. Crowhurst M, Teodocio Olivares A. 2014 Beyond the Iambic–Trochaic law: the joint influence of duration and intensity on the perception of rhythmic speech. *Phonology* **31**, 51–94. (doi:10.1017/S0952675714000037)
142. Peña M, Bion RAH, Nespor M. 2011 How modality specific is the Iambic–Trochaic Law? Evidence from vision. *J. Exp. Psychol. Learn. Mem. Cogn.* **37**, 1199–1208. (doi:10.1037/a0023944)
143. Mann DC, Fitch WT, Tu HW, Hoeschele M. 2021 Universal principles underlying segmental structures in parrot song and human speech. *Sci. Rep.* **11**, 1–14. (doi:10.1038/s41598-020-79139-8)
144. Tierney AT, Russo FA, Patel AD. 2011 The motor origins of human and avian song structure. *Proc. Natl Acad. Sci. USA* **108**, 15 510–15 515. (doi:10.1073/pnas.1103882108)
145. Pika S, Wilkinson R, Kendrick KH, Vernes SC. 2018 Taking turns: bridging the gap between human and animal communication. *Proc. R. Soc. B* **285**, 20180598. (doi:10.1098/rspb.2018.0598)
146. Ravignani A, Verga L, Greenfield MD. 2019 Interactive rhythms across species: the evolutionary biology of animal chorusing and turn-taking. *Ann. NY Acad. Sci.* **1453**, 12–21. (doi:10.1111/nyas.14230)
147. Bagou O, Fougeron C, Frauenfelder UH. 2002 Contribution of prosody to the segmentation and storage of 'words' in the acquisition of a new mini-Language. In *Proc. Speech Prosody 2002*, pp. 159–162.
148. Gordon M, Roettger T. 2017 Acoustic correlates of word stress: a cross-linguistic survey. *Linguist. Vanguard* **3**, 1–11. (doi:10.1515/lingvan-2017-0007)
149. Andersson S, Sayeed O, Vaux B. 2017 The phonology of language contact. In *Oxford handbooks online*, pp. 1–33. Oxford, UK: Oxford University Press. (doi:10.1093/oxfordhb/9780199935345.013.55)
150. Broselow E. 2009 Stress adaptation in loanword phonology. In *Phonology in perception* (eds P Boersma, S Hamann), pp. 191–234. Berlin, Germany: De Gruyter Mouton.
151. Speyer A. 2009 On the change of word stress in the history of German. *Beitrage zur Geschichte der Dtsch Spr und Lit.* **131**, 413–441.
152. Nolan F. 2021 Intonation. In *The handbook of English linguistics* (eds B Aarts, AMS McMahon, L Hinrichs), pp. 385–405. Hoboken, NJ: John Wiley & Sons.
153. Hirst D, Di Cristo A (eds). 1998 *Intonation systems: a survey of twenty languages*. Cambridge, UK: Cambridge University Press.
154. Ellis NC. 2002 Frequency effects in language processing. *Stud. Second Lang. Acquis.* **24**, 143–188. (doi:10.1017/S0272263102002024)
155. Diessel H. 2007 Frequency effects in language acquisition, language use, and diachronic change. *New Ideas Psychol.* **25**, 108–127. (doi:10.1016/j.newideapsych.2007.02.002)
156. Ambridge B, Kidd E, Rowland CF, Theakston AL. 2015 The ubiquity of frequency effects in first language acquisition. *J. Child Lang.* **42**, 239–273. (doi:10.1017/S030500091400049X)
157. Toro JM, Sebastián-Gallés N, Mattys SL. 2009 The role of perceptual salience during the segmentation of connected speech. *Eur. J. Cogn. Psychol.* **21**, 786–800. (doi:10.1080/09541440802405584)
158. Abboub N, Boll-Avetisyan N, Bhatara A, Höhle B, Nazzi T. 2016 An exploration of rhythmic grouping of speech sequences by French- and German-learning infants. *Front. Hum. Neurosci.* **10**. 292. (doi:10.3389/fnhum.2016.00292)
159. Bion RAH, Benavides-Varela S, Nespor M. 2011 Acoustic markers of prominence influence infants' and adults' segmentation of speech sequences. *Lang. Speech* **54**, 123–140. (doi:10.1177/0023830910388018)
160. Nespor M, Shukla M, Van De Vijver R, Avesani C, Schraudolf H, Donati C. 2008 Different phrasal prominence realizations in VO and OV languages. *Lingue e Linguaggio* **7**, 139–167.
161. Hauser MD. 2000 A primate dictionary? Decoding the function and meaning of another species' vocalizations. *Cogn. Sci.* **24**, 445–475. (doi:10.1207/s15516709cog2403_5)