Speech rhythms and their neural foundations

David Poeppel[1,2] ✉ *and M. Florencia Assaneo*[2,3]

Abstract | The recognition of spoken language has typically been studied by focusing on either words or their constituent elements (for example, low-level features or phonemes). More recently, the 'temporal mesoscale' of speech has been explored, specifically regularities in the envelope of the acoustic signal that correlate with syllabic information and that play a central role in production and perception processes. The temporal structure of speech at this scale is remarkably stable across languages, with a preferred range of rhythmicity of 2–8 Hz. Importantly, this rhythmicity is required by the processes underlying the construction of intelligible speech. A lot of current work focuses on audio-motor interactions in speech, highlighting behavioural and neural evidence that demonstrates how properties of perceptual and motor systems, and their relation, can underlie the mesoscale speech rhythms. The data invite the hypothesis that the speech motor cortex is best modelled as a neural oscillator, a conjecture that aligns well with current proposals highlighting the fundamental role of neural oscillations in perception and cognition. The findings also show motor theories (of speech) in a different light, placing new mechanistic constraints on accounts of the action–perception interface.

Distinctive features
Stable auditory and/or articulatory patterns that distinguish phonemes, for example 'voicing' in /b/ versus /p/.

Phones
Brief segments of speech that have characteristic physical or perceptual attributes.

Phonemes
The speech elements of a language (vowels and consonants) that encode words.

[1]*Department of Neuroscience, Max Planck Institute, Frankfurt, Germany.*

[2]*Department of Psychology, New York University, New York, NY, USA.*

[3]*Instituto de Neurobiologia, Universidad Nacional Autónoma de México Juriquilla, Querétaro, México.*

✉*e-mail: david.poeppel@nyu.edu*

https://doi.org/10.1038/s41583-020-0304-4

Research on speech processing has typically focused on 'the little bits', that is to say on how 'distinctive features', 'phones' or 'phonemes' constitute the elements that must be identified and decoded in recognizing and producing speech (FIG. 1a,b). This approach has been very successful, forming the basis of our understanding from the perspectives of acoustics, psychology, linguistics and neuroscience[1–3] — as well as, more recently, engineering, where automatic speech recognition systems are yielding remarkable performance. The critical role of the constituent elements (constitutive of, informally speaking, 'words') in perception and production as well as lexical processing is widely appreciated and investigated[4,5]. In research that has proceeded somewhat independently, a different attribute of speech has begun to be highlighted — the slower signal modulations more characteristic of 'intermediate bits' or chunks, namely syllables (FIG. 1c). In contrast to the consideration paid to elemental acoustic–phonetic features (FIG. 1b), this 'mesoscale of speech' has received less attention (FIG. 1c). One of the surprising recent discoveries is that speech quantified at this timescale has a temporal structure of high regularity, an attribute likely to be a consequence of the organization of brain circuitry and the biomechanics of the speech motor system[6,7]. This temporal, rhythmic regularity is exploited by recognition systems, as well. There is now a growing body of work (from psychophysics to physiology to modelling) that builds on these observations to develop models that incorporate 'segmentation', 'decoding' and 'audio-motor integration' of speech. Here, we review the latter concept and outline a linking hypothesis that, we suggest, deepens our mechanistic understanding of speech processing.

The statement that speech is rhythmic leads in equal measures to fascination and frustration. On the one hand, the conjecture has led to a range of novel research that we review. On the other, scepticism about rhythmicity has resulted in vigorous debate. Speech is, to be sure, not periodic (in the formal sense of an isochronous signal) — and, to our knowledge, this has never been claimed. However, both informal observation (for example, the rate of speech across languages seems clearly bounded) and quantitative characterization lead to the conclusion that speech is rhythmic to the degree that it has a regular temporal structure that presumably reflects deep properties of perception and production systems.

The data, models and arguments we review proceed from the empirically well-justified position that there is sufficient temporal regularity in the mesoscale of speech signals (FIG. 1c) that this attribute merits consideration from various perspectives. Although mindful of the remarkable variety and diversity of human languages, we are equally moved by the degree of similarity across them. We review here those aspects of speech production, perception and sensorimotor interaction that

Segmentation
The process of chunking the continuous acoustic stream of spoken language into units.

Decoding
Mapping the segmented acoustic chunks into linguistic units (phonemes, syllables or words) stored in the mental dictionary.

point to fundamental and shared principles of cerebral organization and computation.

We focus on the literature on the temporal structure of speech at the syllabic scale. First, we summarize the data revealing that speech production and perception exhibit rhythmicity across languages. Second, we describe research exploring the interaction between perception and production rhythms. Third, we introduce a biophysical model for the sensorimotor integration of speech in the time domain. Finally, we discuss the implications in the broader context of biology.

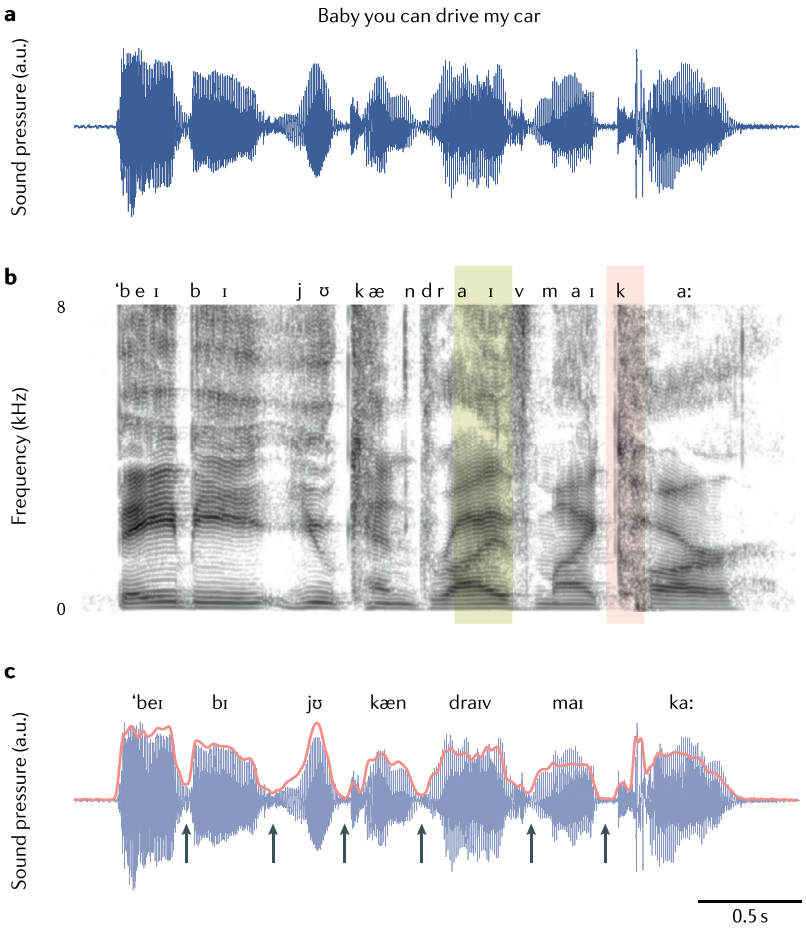


Fig. 1 | **Different timescale representations of a generic acoustic speech signal.** **a** | Acoustic waveform of the utterance 'Baby you can drive my car' produced by a male English speaker. The *y* axis represents the amplitude of the sound pressure level in arbitrary units (a.u.); same in panel **c**. **b** | Spectrogram of the signal, emphasizing short-scale dynamics. Typical spectrogram values to analyse speech use time windows ranging from 10 to 25 ms (here 25 ms), short scale when compared with the typical syllable duration (~200 ms). Spectrogram representations allow the visualization of phonemic features over time. For example, vowels are defined by their formant patterns (enhanced frequency bands of energy; for example, dark stripes in the yellow-shadowed region define the /a/ and /I/ vowels), and consonants, such as /k/ (red-shadowed region), by a noise burst with no structured frequency pattern. **c** | Mesoscale representation: evolution of the waveform amplitude over time, the so-called speech envelope (pale red trace). Although the envelope — as any other natural signal — is not perfectly periodic, it displays certain temporal regularities. Troughs in the envelope (grey arrows), which typically correspond to syllable boundaries, are roughly regularly distributed in time. ˈbeɪbi juː kæn drav maɪ kɑ represents the International Phonetic Alphabet transcription for the corresponding speech chunk immediately below each sign.

Speech production exhibits rhythmicity

Rhythmicity in the acoustic domain. The waveform of an acoustic speech signal reveals a sequence of increments and decrements in signal amplitude (FIG. 1a,c). This modulation, commonly referred to as the speech envelope, has received considerable attention. Many studies show that the speech envelope exhibits compelling temporal regularity[8–12]. Rather than being merely an incidental feature of speech signals, this regularity is suggested to play a key role in speech comprehension (reviewed in the next section). As envelope modulation is so clear a feature and so strong a cue in the signal, not capitalizing on it for processing would be akin to not using luminance cues in visual perception. It is possible to see with few luminance cues (and even at isoluminance), but onerous. Likewise, minimal audio envelope information makes speech recognition hard; it can be done, but it is not fun[13,14].

Because of the link between intelligibility and speech temporal structure, the envelope has mostly been explored from the perspective of perception. Instead of characterizing the broadband acoustic waveform that enters the ear (FIG. 1c), researchers typically first decompose the signal into frequency bands and explore the amplitude modulation within those segregated bands — a narrowband analysis. Various such decompositions can be found in the literature, representing different properties of the human auditory system; for example, cochlear-derived frequency decompositions (compare with critical band filtering of the auditory system) or modulation tuning of cortical neurons. Regardless of the type of analysis employed, all studies converge on the same conclusion: the speech envelope possesses an overall 1/f noise spectrum[15], which, when removed, reveals an increase in power for frequencies between 2 and 8 Hz, with a notable peak between 4 and 5 Hz[8–12,16]. Critically, these features are preserved across speakers, languages and speaking conditions (for example, interviews, telephone conversations or audiobooks) (FIG. 2a).

What is the signal like prior to the frequency decomposition that is characteristic of the auditory system? When computed over the original (straight out of the mouth, broadband) acoustic signal, the speech envelope spectrum displays the same pattern as the one computed on the narrowband decomposition but containing more power at lower frequencies[17]. To quantify this, we reanalysed a corpus previously explored with a narrowband approach using a broadband analysis (FIG. 2b). As with the narrowband analysis, the modulation spectrum is remarkably robust across languages and speakers. However, the peak frequency is shifted to a slightly lower range. The narrowband approach (applying critical band filtering before computing the envelope) shows a peak range from 4.3 to 5.4 Hz[10]; the broadband analysis shows a range from 3.5 to 4.5 Hz (FIG. 2b).

To summarize, speech is not only perceived as quasi-rhythmic but the produced acoustic signal itself shows striking temporal regularities. Interestingly, it seems that the perceived rhythm is slightly faster than the one carried by the physical signal. Further work should clarify the origin and consequences of this slight misalignment between rhythms.

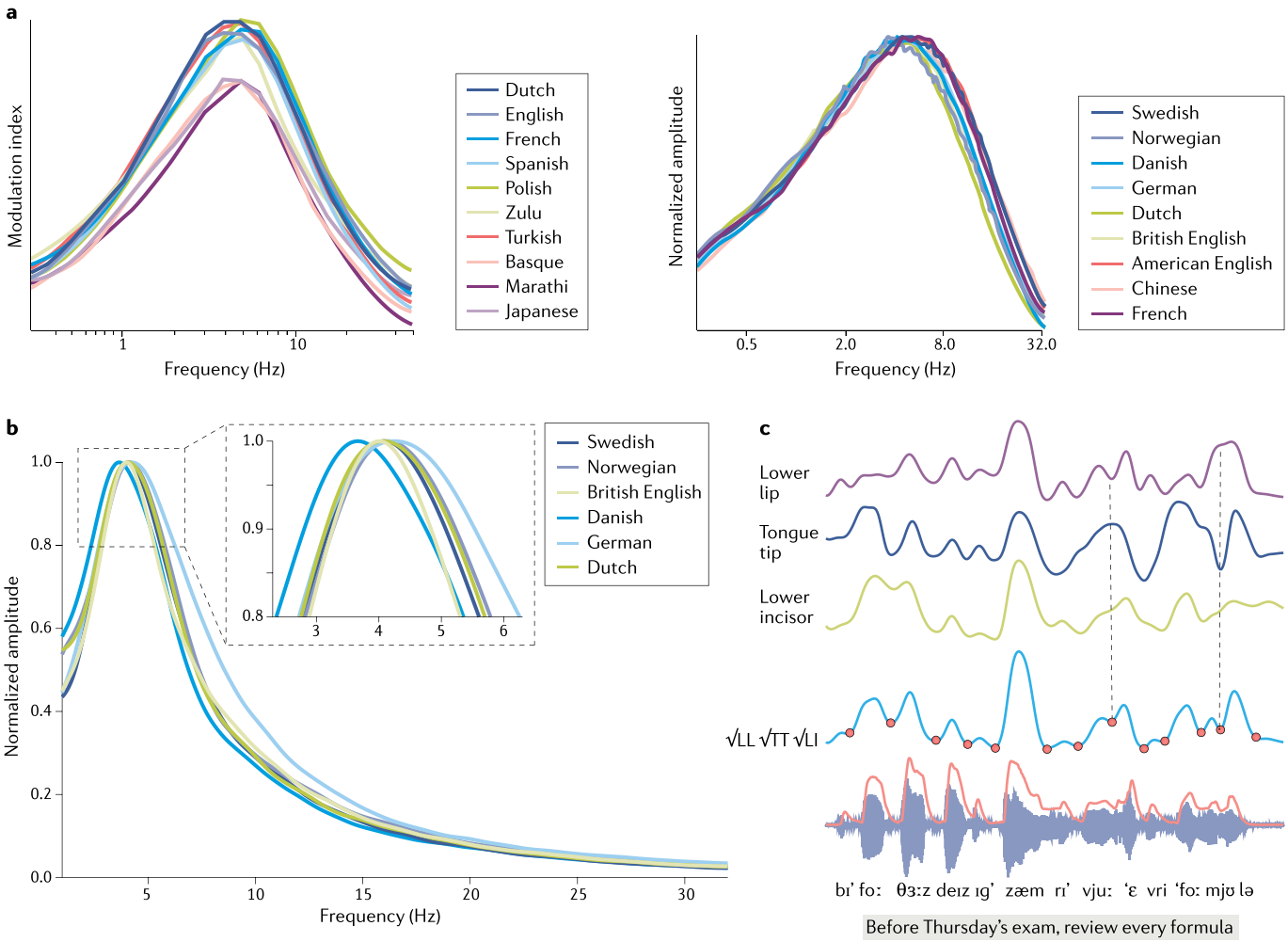


Fig. 2 | **Speech production exhibits rhythmicity. a** | Temporal structure of the speech envelope across languages. Two (analytically differently derived) narrowband analyses applied over different speech samples (different speakers, under different speaking conditions; that is, reading, having a conversation or telling a story) from various languages. Independently of the language and analysis filter/pipeline used for narrowband decomposition, the spectra display clear peaks between 2 and 10 Hz. **b** | Speech envelope spectrum computed over the original acoustic signal; that is, without applying any filter — the broadband modulation spectrum. Again, the modulation spectrum peak shows small variations across language, and is restricted between 3.5 and 4.5 Hz. **c** | Speech rhythmicity in different domains (articulatory, acoustic, linguistic). Three upper traces: generic vertical displacement of the lower lip (LL), tongue tip (TT) and lower incisor (LI) for one subject, during the production of the sentence shown[132]. Light blue trace: multiplication of the square root of the articulatory measurements. The selected articulatory measures and the multiplication of square roots are merely used to exemplify how to construct a signal that captures the cooperative dynamics between articulators; further work is needed to identify the exact functional form. During natural speech production, most of the time the articulators' movements are highly correlated. However, sometimes (dashed lines) a constriction is caused just by one of the articulators. At this time point, the rhythmicity is broken in some of the articulators' trajectories but not in the cooperative dynamics (light blue trace). Lower traces: speech acoustic waveform with its corresponding envelope. Under each cycle of the envelope is the International Phonetic Alphabet transcription for the corresponding speech chunk. Panel **a** adapted with permission from REF.[12], AIP Publishing and REF.[10], Elsevier.

Audio-motor integration
The alignment or merging of information computed in the auditory and (speech) motor systems.

Spectrogram
A visualization of how the frequency composition of a signal evolves over time.

Rhythmicity in the articulatory domain. The speech signal arises from complex motor gestures involving precise movements of the upper vocal tract articulators — the velum, lips, tongue and jaw — and activation of the vocal folds[3,18]. From a biomechanical perspective, speech production represents a high-dimensional problem, given the number of effectors, and especially the many degrees of freedom of the tongue[19].

However, experimental evidence shows a dimensionality reduction of the problem: the temporal dynamics of the main articulators during speech can be described by just seven parameters[20]; the vocal tract shape defining (English) vowels can be defined as a linear combination of two canonical 'deformation patterns'[21]; and one can synthesize intelligible speech from the measurements of the position of a discrete number of points in the oral cavity and lips[22,23].

This dimensionality reduction can be seen as a consequence of the vocal tract articulators not having independent kinematics. Instead, their movements are coordinated to achieve a common vocal tract goal[24,25]. For example, during the pronunciation of /b/, the goal is to occlude the front of the vocal tract, thus there is a synergy between the movements of the lips and the jaw in order to achieve complete closure. The trajectory of each individual articulator is not unequivocally defined by the

Critical band filtering
Decomposing a signal into different frequency bands defined according to the frequency response of the relevant biophysical system.

1/f noise spectrum
The power spectrum of noise decreases with frequency, an attribute of many biological signals.

Spectrum
A representation of how much energy a signal carries in each frequency band.

Vocal tract
The set of anatomical cavities above the larynx that shape the production of speech.

Velum
Part of the roof of the oral cavity comprising connective tissue and muscle, also called the soft palate.

Syllable
A basic unit of spoken language, typically comprising a vowel (energy peak) with adjoining consonants (for example, /bar/), and thus a short sequence of speech sounds.

acoustic goal; instead, it is the relationship between the different articulators' dynamics that is preserved — for example, /b/ can be pronounced by small displacement of the lips compensated by a larger closure of the jaw, or vice versa[26]. Moreover, at the neural level, the motor cortex does not encode individual articulator movements but, rather, the coordinated articulatory pattern necessary to produce a vocal tract constriction[27].

The features of motor gestures in the time domain reveal temporal regularity. For instance, jaw and lip displacement during repetition of a sentence at a normal rate — sentences with an over-representation of bilabials such as 'buy Bobby a puppy' or 'mommy bakes pot pies' — displays oscillatory behaviour in the trajectories of both articulators between 4 and 5 Hz[28–30]. Also, the tongue's dynamics during syllable repetition at a normal rate exhibits rhythmicity within the same range[31]. Lindblad et al. measured 12 subjects' jaw movements while pronouncing two phrases at a natural tempo, finding quasi-regular behaviour with a frequency close to 5 Hz across participants[32]. Furthermore, rhythmicity has also been reported under more natural conditions. Ohala measured the time interval between successive jaw openings for one subject during 1.5 h of natural reading. The obtained histogram displays a clear peak close to 250 ms (that is, 4 Hz)[33]. More recently, the area of the mouth opening was measured for different subjects under different speaking conditions, in unstructured conversation or full sentence production. Again, the mouth area shows a rhythmic modulation between 2 and 7 Hz[17].

Summarizing, rhythmicity appears under different speaking conditions for independent articulator movements. However, speech does not rely on the kinematics of individual articulators (FIG. 2c, upper three traces); it emerges from the cooperative dynamics between articulators (FIG. 2c, lower trace) in order to achieve a common goal. We suggest that during natural speech production the rhythm of the interplay between the articulators is preserved instead of the rhythm of the isolated elements.

Rhythmicity in the linguistic domain. Although we have intuitive notions of a syllabic unit, formal phonetic definitions are debated[1,34–36]. The syllabification of some English words, for example, is speaker dependent (for example, 'predatory' can be pronounced with three or four syllables); and for some, even when pronounced equally, listeners will differ in the estimated number of syllables, including as a function of literacy (for example, 'communism' can be described as a three or four-syllabic word). Such disagreements, however, represent isolated cases, and the majority of linguistic utterances have relatively clear syllabifications[1].

Moreover, one cannot neglect the crucial role being played by syllables from the point of view of perception and production. Many studies demonstrate that syllables have relevant cognitive implications[37–40] and are intended to be optimal articulatory motor units[1,7]. Speech motor programming has been proposed to consist of serially ordered coordinative structures (articulatory motor units), each representing the motor command to produce a syllable. This hypothesis is computationally motivated and experimentally supported[41–43]. Guenther and colleagues developed a neural network model for speech acquisition and production assuming that unique 'speech sound map cells' encode frequently used syllables of the language, and such map cells are sequentially activated during fluent speech[6]. This model successfully accounts for many speech production phenomena[44]. Thus, the syllable appears as the natural motor unit for fluent speakers.

In the phonetics literature, the syllable rate is a common measurement to assess a speaker's speech tempo. Consequently, many studies estimate this parameter under different experimental conditions[8,45,46]. Note that the syllable rate is an average value across spoken items and that sequences of syllables, as any other biologically grounded signal, are not perfectly isochronous. Furthermore, systematic variation in syllable duration is well established. For example, in English, whereas the mean syllable duration is ~200 ms, unstressed syllables tend to be shorter (<150 ms) than stressed ones (>300 ms)[8]. Despite variable syllable duration, the syllable rate displays a rather restricted range of values. Although significant differences in syllable rate have been reported between languages[47], dialects[48], speaking conditions[49], age[50] and gender[51], the variation of this value is always restricted between 2.5 and 8 Hz (corresponding to durations of ~125–400 ms). Moreover, part of the variation in the reported values derives from using different definitions of the syllable rate. For example, syllable rate can been computed as the number of syllables divided by the total length of the vocalization – the raw syllable rate — or by the total length of the vocalization minus any silent gaps — the articulation rate. Studies using the raw syllable rate report values between 3 and 5.5 Hz, whereas the range is 5–8 Hz in those employing the articulation rate. For either measurement approach, the results reflect an underlying rhythmic regularity, suggesting that, during speaking, syllables are produced sequentially with a relatively consistent rate across speakers, conditions and languages.

Speech across domains. The three domains of characterizing speech — the envelope of the signal (from acoustics), movements of the vocal tract (from articulation) and the syllable duration and rate (from linguistics) — are highly interconnected. At one end, the articulators' dynamics — a consequence of the interaction between neural motor control and biomechanical properties of the anatomy — largely determine the modulation in the amplitude of speech acoustics[17,52] and the produced syllable rate[32]. At the other end, local minima of the envelope approximate syllable boundaries[53,54]. Furthermore, it is plausible to link the cyclic properties of the different domains by linking the following facts: the sequence of cooperative movements between articulators to achieve a common goal occurs at a quasi-rhythmic pace, creating cycles of opening and narrowing of the vocal tract; from an acoustic perspective, the vocal tract acts as a filter for the sound generated by the vocal folds[55] — thus, the amplitude of the emitted acoustic signal is minimal during an occlusion or constriction of the tract; and during fluent speech comprising elements with a

Entrainment
The synchronization of brain activity to the temporal structure of a stimulus or between the activity of neural elements.

clear syllabification, the syllable boundaries are defined by local minima in the speech envelope[7,53].

The experimental evidence shows that there is a correlation between the different speech domains, all of them revealing quasi-rhythmic features within the same narrow frequency range. This match in rhythms is not accidental. Instead, the rhythmicity across the domains of description derives from the same cause: the motor gestures of speech are sequentially executed at a relatively regular time interval. It is noteworthy that similar ideas have been advanced in the field of birdsong[56,57], which represents an animal model for speech.

Speech perception exhibits rhythmicity

Next, we turn to two obvious questions. Is the speech rhythm retrieved by the perceptual system? If so, does the perceptual system rely on this rhythmic structure to facilitate comprehension?

Auditory entrainment to the envelope. When presented with an acoustic stimulus — whether speech[58,59] or modulated white noise[60] — the auditory cortex faithfully tracks the amplitude modulation of the input. This effect is often called 'entrainment', although that particular locution comes with certain assumptions about the neural mechanisms[61]. In the case of speech, this stimulus–brain interaction becomes quite complex, relying on combined bottom-up (that is, feedforward) and top-down (that is, feedback) processes. The data show that while perceiving unintelligible signals, for example a foreign language[62] or backward speech[63], auditory areas are entrained by the acoustic envelope. In addition, when the processes constituting comprehension are also in play, this entrainment is top-down modulated, for example by attention[64,65] or semantic context[66]. The precise mechanisms supporting the brain-to-speech entrainment remain a topic of intensive investigation. Specifically, how the top-down and bottom-up processes interact[67] and whether the entrainment reflects the resetting of ongoing oscillations in the auditory cortex[68] or additive brain responses to the physical attributes of the acoustic signal[69], or a combination of both[70], are debated. However, beyond the nature of the underlying mechanisms — an issue that merits discussion but lies outside the scope of this Review — evidence for the entrainment of auditory cortex activity to the perceived speech envelope is ubiquitous. The data show that the rhythmic structure of speech is conveyed as an input to the early stages of speech neural processing.

It has been conjectured that speech envelope tracking plays a causal role for speech comprehension. Specifically, one hypothesis posits that the rhythmic structure recovered by auditory areas allows the listener to transform the continuous input speech signal into segmented, discrete units, which form the input for subsequent decoding steps[71]. Neurophysiological data show that auditory entrainment to the speech envelope, specifically in the theta band from 4 to 8 Hz, correlates with intelligibility[72–74]. In addition, from a related perspective, entrainment has been shown to be abnormal in poor readers[75] and dyslexic children[76]. Although such findings reveal a link between envelope tracking and comprehension, they do not cement a causal relationship. Causality has recently been assessed more directly in studies that applied electrical stimulation over temporal regions during speech perception, interfering with neural entrainment to the speech envelope in the theta band. The results demonstrated that compromised entrainment led to intelligibility decrements[77–79].

Cumulatively, the evidence points to the rhythmic structure of speech being recovered by the perceptual system and playing a critical role for spoken language comprehension.

The speech rhythm enhances perception. Insofar as the perceptual system tracks and relies on the rhythmic structure of speech, is the system equally efficient for any frequency range or is it tuned to the natural rhythm of speech; that is, is perception optimal for syllable rates between 2 and 8 Hz.

The perception of amplitude modulation, in the absence of spectral cues, has been widely investigated as a measurement of the temporal resolution of the auditory system. In such psychophysical studies, a carrier signal is sinusoidally modulated, and a threshold is defined as the minimum modulation amplitude required for a listener to discriminate between modulated and unmodulated waveforms. Despite results showing an interaction between the spectral content of a carrier and the frequency of the modulation signal[80], a common finding is that the threshold remains relatively stable for modulation frequencies between 2 and 8 Hz, and increases – that is, perceptual performance decreases — outside this range[81,82].

In line with these behavioural results, the auditory cortex response is better tuned to temporal modulation rates within this range. Across studies, and regardless of the spectral content of the signal being modulated, auditory cortex activity shows an enhancement, mainly in the right hemisphere, for modulation frequencies of 2–8 Hz (REFS[83,84]), and better tracking of the modulation phase within the same range[85,86]. The heightened sensitivity of auditory cortex to these modulation rates is also seen in functional MRI data that explicitly probed along the auditory pathway[87].

Turning to speech, intelligibility is enhanced when the rhythmic structure of the signal lies within a certain range of frequencies. Behavioural studies assessing the comprehension of compressed speech showed that performance remains stable when speech is speeded up to ~8 syllables per second and significantly decreases at 10 syllables per second[88,89]. Furthermore, when periodic silent 'gaps' are inserted into a compressed speech signal, restoring the overall temporal structure of the original signal, intelligibility is partially recovered[90].

The experimental evidence shows that the rhythmic structure of the speech envelope is not only a descriptive feature of the acoustic signal but a crucial attribute facilitating comprehension.

Rhythmic action–perception interaction

Speech output is rhythmic (first section) and our perceptual system appears to capitalize on this attribute (second section). We turn now to an aspect of processing that remains one of the foundational questions

regarding the cognitive and neural basis of speech, namely the extent to which the perception and action systems are coordinated or not. Broadly speaking, interaction between the speech perception and production systems, a highly contentious issue with significant implications, has been proposed and experimentally supported, both in functional terms (BOX 1) and structurally (BOX 2).

Over the last years, different studies have explored the interaction between perception and production brain areas (FIG. 3). In line with modified forms of a motor theory of speech perception (BOX 1), the experiments showed that passive listening to speech activates brain areas involved in speech production[91] and that transcranial magnetic stimulation over articulation areas interferes with speech perception[92]. Although the findings do not suggest a critical causal role for motor activation, they do advocate for a supportive role of the speech–motor systems, for example by aiding speech perception under adverse listening conditions[93].

Sensorimotor interaction also takes place in the other direction; that is, speech production systems recruiting perception. At the behavioural level, feedback modification studies show that real-time perturbations of acoustic parameters — such as pitch or vowel formants — of an ongoing utterance elicit automatic, unconscious behavioural compensation[94]. Relatedly, delayed auditory feedback induces dysfluencies and decreases the speech rate[95], providing evidence for online interaction. At the brain level, so-called speech-induced suppression has been reported: the responses of auditory areas to ongoing self-produced speech are smaller than those to similar but externally produced speech[96]. Compelling evidence for sensorimotor interaction also comes from studies on speech motor planning that implicate 'efference copies' during production[97]. These experiments build on the fact that the speech motor system sends a copy of the intended target to sensory regions — somatosensory and auditory — where the alignment between planned and executed targets is evaluated.

Summarizing, the experimental evidence consistently points to a bidirectional interaction between the perception and production systems. However, the mechanisms and causal role remain poorly understood.

Despite the fact that the link has been explored, only in the last few years has the temporal dynamics of the sensorimotor integration of speech been assessed. Two different approaches have been pursued. One explores the synchronization between brain regions in the service of speech comprehension. The second line of research aims to characterize the sensorimotor integration of speech in phase space, regardless of any functional role.

Synchrony helps speech processing. As already outlined, listening to speech induces the entrainment of auditory low-frequency brain activity within the temporal lobe to the quasi-rhythmic structure of the acoustic signal's envelope. This entrainment is required for the correct segmentation and subsequent decoding of the signal. Building on this result, recent experiments explored how this brain-to-stimulus synchronization is modulated by interaction between the temporal cortex and the rhythms of other brain regions during the processing of intelligible speech.

Park et al. assessed the connectivity between the auditory cortex and other brain areas while participants listened to natural (intelligible) speech and backwards (unintelligible) speech[98]. Their data showed that oscillatory activity generated in the left inferior frontal (frequencies between 1 and 3 Hz) and the precentral gyri (frequencies between 4 and 8 Hz) modulated the phase of low-frequency activity in left auditory regions, significantly more in the intelligible than the unintelligible condition. This top-down control yielded better tracking of the speech envelope in the auditory cortex. In line with this finding, it has been shown that the phase of the ongoing slow oscillations in bilateral motor and supplementary motor regions modulate reaction times during a word recognition task[99]. New experiments testing the extent of feedforward and feedback processing in syllable and word-level perception reveal both phase–amplitude and phase–phase coupling effects in the relevant low-frequency ranges between frontal and temporal regions that may be (part of) the substrate to support bottom-up and top-down online sensorimotor alignment[100].

Patient data also suggest that the rhythmic interaction between temporal and frontal regions enhances spoken language comprehension. Cope et al. explored the brain activity of patients with frontal neurodegeneration and healthy subjects when they were presented with degraded speech samples (words) combined with matching or mismatching text cues[101]. They showed that neural responses related to prediction generated in the intact temporal cortex are delayed as a consequence of the neurodegeneration of frontal speech regions. Moreover, although fronto-temporal connectivity was enhanced for both groups for frequencies below 25 Hz, it was stronger for patients in the beta band, 13–23 Hz. Using Granger causality analyses they showed that the interaction is top-down, meaning frontal areas influenced temporal areas, for frequencies within the beta band, whereas the 5-Hz activity reflected bottom-up, temporal to frontal modulation. Relatedly, the activity

Box 1 | **Motor theory of speech perception**

In the 1950s, researchers observed that the sounds comprising different languages — phonemes — cannot be unequivocally defined in the acoustic space; the problem derives from the related challenges of 'linearity' (that is, sequences of sound do not map cleanly to sequences of phonemes), 'invariance' (that is, the mapping from tokens of spoken phoneme realizations to types of phonemes is not unique) and 'perceptual constancy' (that is, both within-speaker and across-speaker variability is difficult to identify in the acoustics). Consequently, it was proposed that phonemic perception occurs in motor rather than acoustic space. To comprehend speech, listeners retrieve the intended articulatory gestures of speakers. The incoming speech is thus mapped to its corresponding neuromotor command, which in turn is invariant[140]. This hypothesis, the motor theory of speech perception, continues to elicit vigorous debate[141,142]. The original version has been strongly criticized[143], for example, because: infants exhibit exquisite sensitivity to speech despite having very little to no speech — motor control; patient data show that listeners with dramatic lesions to speech — motor areas perform well on speech discrimination; and data and models demonstrate that speech perception can be approached as an auditory task for which no motoric support is required. Nevertheless, there is a body of evidence that one can take to be consistent with a modified, weaker version of the theory.

Box 2 | **Spoken language pathways**

Current models for spoken language processing assume a dual-stream architecture (FIG. 3a) wherein sound-to-meaning mapping is mostly performed by ventral brain regions and sound-to-articulation transformation by dorsal ones[114,144]. In line with this approach, two main routes of long-range white matter connections have been described: the dorsal and ventral pathways. The precise number of fibre bundles composing each pathway, their origin and termination points, and the functions that each of them support are under debate[145,146]. Here, we focus on the broad features on which reasonable consensus has been achieved. Ventral and dorsal streams are composed of more than one pathway, running in parallel, connecting — directly and/or indirectly — temporal and frontal regions[139,147]. The dorsal pathways connect the peri-Sylvian cortex of the temporal, parietal and frontal lobes; the ventral pathways run along the temporal lobe and connect it to the orbitofrontal cortex (FIG. 3b). Regarding their functional roles, the ventral pathways subserve, preferentially, semantic-level computations[148], whereas dorsal pathways support phonological processing[147] and, perhaps, syntactic structure processing[149].

In addition to these two principal long-range routes, the frontal cortex and the temporal cortex are traversed by shorter fibres enabling the flow of information within regions[150]. Furthermore, the speech and language processing networks are tightly connected with speech motor production regions through several cortico-subcortical loops comprising the cerebellum and the basal ganglia[146]. Overall, brain areas underpinning spoken language perception and production display dense and redundant connections.

in the left inferior frontal region[102] as well as its synchronization with auditory areas[103] is reduced in dyslexic readers.

Speech envelope tracking during language processing has also been investigated at the whole brain level[104]. Keitel et al. explored brain-to-envelope entrainment while participants listened to meaningful but unpredictable sentences. Comparing correctly comprehended trials with incorrect ones, they found that entrainment in a frequency band aligned with the phrasal timescale (here, 0.6–1.3 Hz) was enhanced not only in the temporal cortex but also in the premotor cortex. Moreover, the phase of this low-frequency band was coupled to the beta power in left motor areas. As the phrasal structure was consistent across stimuli, the authors hypothesized that the premotor and motor cortex exploited this temporal regularity to produce top-down temporal predictions.

Consistent with this hypothesis, it has been proposed that predictive timing processes rely on the functional coupling between beta and delta–theta rhythms; although delta–theta oscillations in sensory areas are primarily entrained by the temporal structure of external stimuli, this entrainment can be top-down modulated through an interaction with beta-frequency activity generated in frontal regions[105].

The interplay between the temporal dynamics of auditory regions and frontal brain areas reveals a complex landscape of coexisting top-down and bottom-up connections operating in different frequency bands. Even though the precise role of these interactions is unclear, the existing data suggest that the role of motor activity in speech perception in the time domain is to enhance auditory temporal prediction for speech processing.

Speech production: an oscillatory view. Next, we review a set of studies that characterize the synchronization between the speech perception and production systems regardless of any specific functional role, with the aim of elucidating a more mechanistic account.

Recently, Assaneo and Poeppel[106] measured the synchronization between speech motor and auditory cortices while participants passively listened to trains of isochronous syllables presented at different rates, uniformly distributed within the delta–theta range, from 2.5 to 6.5 syllables per second. Based on neurophysiological (magnetoencephalography) recordings, they found that the coupling between areas is restricted to the lower stimulus rates and is significantly enhanced for the condition of 4.5 syllables per second (FIG. 4b). This result invites the hypothesis that speech motor cortex has its own preferred rhythm, motivating the proposal that it behaves like a neural oscillator receiving auditory activity as an input (FIG. 4a). An oscillator is a system capable of generating oscillations at its own characteristic frequency and showing entrainment to a rhythmic input only if the external frequency is near its characteristic one[107]. To evaluate this proposal, numerical simulations of neurophysiological data based on such a model were run. The simulated data reproduced the experimental pattern (FIG. 4b), providing initial support for the model. Note that such a model converges with previous research showing an over-representation of frequencies within the theta band in the speech motor cortex during the resting state[108,109].

Given such neural coupling in the phase domain, there should be testable behavioural effects. At that level, two protocols have been used to examine the temporal aspects of perception–production interaction: the classic delayed auditory feedback[110] and a new Spontaneous Speech Synchronization test[111]. In a delayed auditory feedback experiment, participants wearing headphones speak into a microphone while their own speech is played back with a time delay. Delays in the range of hundreds of milliseconds decrease the speaker's speech rate and/or produce dysfluencies[95,112]. A first-order approximation to model delayed auditory feedback results assumes that the speech envelope is proportional to speech motor cortex activity. Given that the speech envelope entrains the auditory cortex, it follows that the auditory activity is proportional to the motor output with a given delay, T. Interestingly, one can show that numerical simulations obtained by applying this approximation to the previously introduced model[106] — the speech motor cortex represented by an oscillator coupled to auditory activity — reproduce the well-known syllabic temporal elongation reported in the literature[95] (FIG. 4c). This type of evidence is thus consistent with conceptualizing the speech motor cortex as an oscillator.

The Spontaneous Speech Synchronization test explores how a speaker's produced syllable rate is spontaneously modulated by the perceived one[111]. In this test, participants listen to a rhythmic syllable train, presented at 4.5 syllables per second, while continuously whispering the syllable /ta/. Importantly, auditory feedback from their own vocalization is masked: participants wear headphones and whisper, and thus they do not listen to their own production, allowing experimenters to control the auditory input precisely. The spontaneous nature of the test derives from the fact that there is no instruction to synchronize to the external audio input; participants are instructed to recall target syllables.

By means of this simple protocol, an unexpected phenomenon was revealed: the population of speaker-listeners is segregated into two groups according to robust individual differences in speech-to-speech synchronization. Some subjects are spontaneously compelled to adjust their produced syllable rate to the perceived one (high synchronizers), whereas others remain unmoved by the external rhythm (low synchronizers; FIG. 4d). Examining these two subject groups neurally (using magnetoencephalography and MRI), clear functional and structural differences between participant types are visible. In terms of neurophysiological function, when participants passively listen to isochronous streams of syllables, the brain-to-envelope entrainment in, remarkably, the left inferior frontal cortex is stronger for high synchronizers compared with low synchronizers. Relatedly, high synchronizers show more white matter volume in the left dorsal pathway (that is, the red tracts depicted in FIG. 3b are thicker for high synchronizers). Furthermore, the results are correlated: subjects with greater volume in the dorsal pathways display stronger entrainment in the left inferior frontal cortex to the perceived syllable rate. Also, in line with the previous section, wherein fronto-temporal synchrony was shown to positively facilitate language processing, high synchronizers also perform better than low synchronizers in a classic statistical word-form learning task[113].

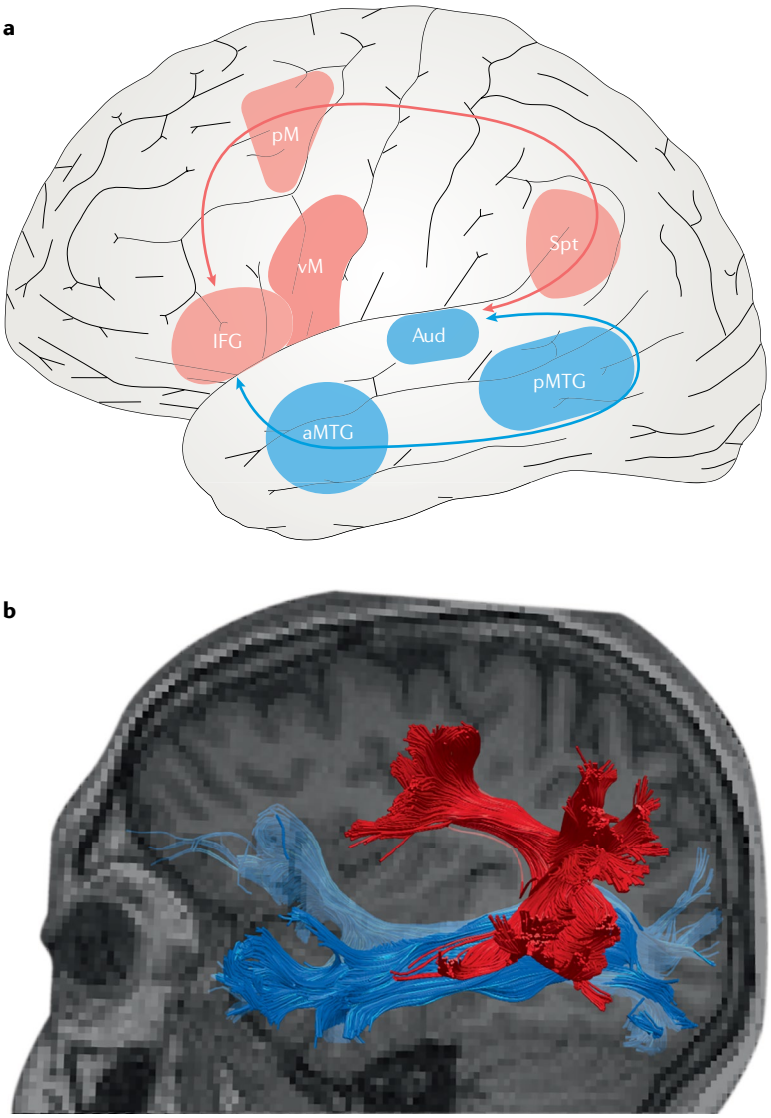
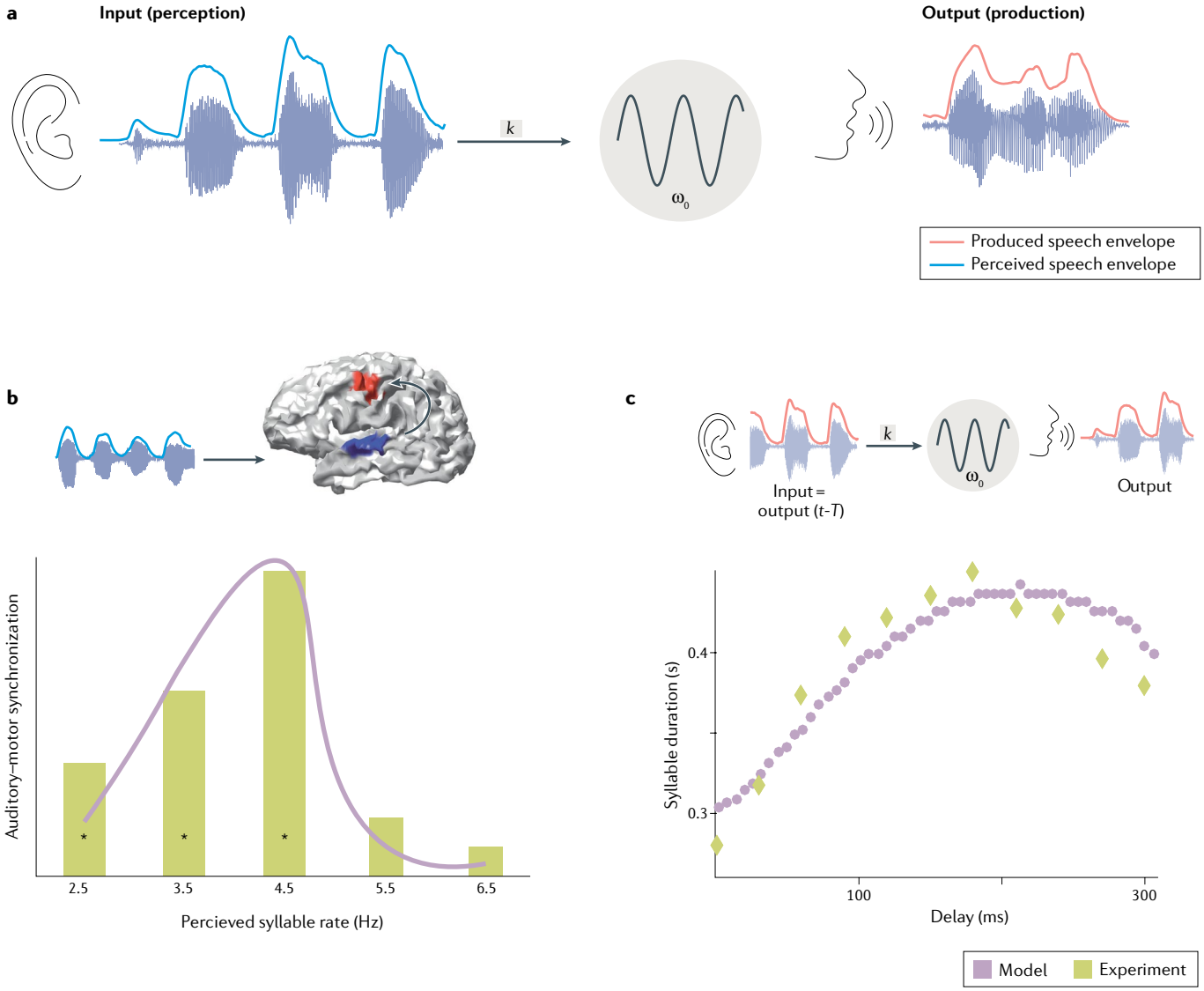


Fig. 3 | **Cortical structures supporting spoken language and sensorimotor interaction. a** | Schematic display of the main areas comprising the dual-stream model[114]. Red and blue shading represent areas within the dorsal and ventral streams, respectively. Red and blue arrows represent the dorsal and ventral white matter connections, respectively. The inferior frontal gyrus (IFG) is associated, among many other functions, with phonological encoding[133] and temporal organization[116] during overt speech production. Activity in the ventral motor cortex (vM) is likely to encode the coordination between the movements of the upper vocal tract articulators that are generating speech[134]. The dorsal premotor cortex (pM) performs functions related to selection of motor responses[135] and the Sylvian parieto-temporal area (Spt) supports motor to auditory mapping[136]. Whereas the anterior middle temporal gyrus (aMTG) is possibly involved in processing of sentence-level intelligibility[137], the posterior middle temporal gyrus (pMTG) is involved in accessing lexical semantic information[138]. The primary auditory cortex (Aud) is involved in the early stages of sound processing. The broad functional description of the regions shown in the figure acknowledges that the precise computational roles remain underspecified. As reviewed, data suggest that the computational processes required to perceive and produce spoken language emerge from a cooperative interaction between all of these brain areas. **b** | Single-subject dissections of peri-Sylvian ventral (blue) and dorsal (red) pathways, superimposed on the subject's structural brain image. Dissections performed following Catani et al.[139].

Finally, the simple model introduced above — production areas behave as an oscillator coupled to the perception regions — can be adapted to predict the surprising bimodal distribution for speech-to-speech synchronization. The model (FIG. 4a) has two basic parameters: the characteristic production frequency, ω_0, and the strength of the perception-to-production coupling, k. When subjects passively listen to syllables, ω_0 corresponds to the natural frequency of the speech motor system, and its value approximates 4.5 Hz[106]. However, we know that we can volitionally speak faster or slower. Thus, we hypothesize that during overt speech production this value can be adjusted, within a range, presumably through top-down signals coming from other brain regions. Also, based on previously described results[111], we hypothesize that k is a structural variable whose magnitude is proportional to the volume of the left dorsal pathway(s). Pursuant to these assumptions, an individual subject can be modelled as a combination of parameters (ω_0, k), where ω_0 is the intended syllable rate and k depends on the individual's brain structure. According to the model, when an external auditory stimulus is presented at a frequency f_{ext}, speech production will synchronize to it only for some combination of parameters (ω_0, k). Thus, parameters within the synchronization region (the shaded region in the lower panel of FIG. 4d) would represent 'high synchronizers'. Conversely, low synchronizers correspond to sets of parameters (ω_0, k) outside the shaded region for which no synchrony is predicted (FIG. 4d).

How, then, might one extend and test this model? A non-trivial prediction is that synchronization will occur for any combination of the intended syllable rate (ω_0) and the external syllable rate (f_{ext}, rate of the auditory stimulus) fulfilling the constraint $m\omega_0 = nf_{ext}$, where m and n are integers[107]. Thus, for example, if participants are cued to speak at 3 Hz and the external rate is set to 6 Hz ($m = 2$ and $n = 1$), a bimodal outcome is expected, but at a 2:1 synchrony ratio. Some participants


a
Input (perception)
Output (production)
k
ω0
Produced speech envelope
Perceived speech envelope
b
Auditory–motor synchronization
2.5
3.5
4.5
5.5
6.5
Percieved syllable rate (Hz)
c
k
ω0
Input = output (t-T)
Output
Syllable duration (s)
0.4
0.3
100
300
Delay (ms)
Model
Experiment
d
Low synchronizer
High synchronizer
foo
beh
lah
dee
tah
k∞
fext
ω0∝

◀ Fig. 4 | **Speech production system can be modelled as an oscillator. a** | Schematic representation of the proposed model: the speech production system is described as an oscillator coupled to the perception system, which in turn follows the envelope of perceived speech. The parameters of this model are *k*, the strength of the perception–production coupling, and ω_0, the characteristic frequency of the production system oscillator, which is represented by the grey circle. **b** | Synchronization between auditory and motor regions during passive listening to rhythmic syllable trains. Upper panel: sketch of the experimental magnetoencephalography measurements, synchronization between brain activity originating in the auditory (blue) and motor (red) regions. Lower panel: green bars, experimental pattern, synchronization between regions is restricted to the lower syllable rates and enhanced at 4.5 syllables per second; purple curve, synchronization pattern obtained from numerical simulations generated by the proposed model. As depicted in the figure, numerical simulations closely match the experimental pattern. Asterisks, significant increments from baseline. **c** | Numerical simulations replicating a delayed auditory feedback experiment. Upper panel: sketch of the modification applied to the model to reproduce delayed auditory feedback conditions. The oscillator's activity at time *t* is fed back with a time delay, *T*. Lower panel: green diamonds, experimental mean syllable duration as a function of *T*[95]; purple dots, data simulated with the adapted version of the model. **d** | The proposed model explains the spontaneous speech synchronization test bimodal outcome[111]. Left panel: schematic representation of the experiment; low synchronizer on the left, high synchronizer on the right. Coloured lines display the amplitude variations of the sound waves, illustrating the syllable rate. Light blue, perceived syllables; dark blue and pale red, produced syllables. Right panel: schematic representation of participants in the parameter space — each (ω_0, *k*) defines a subject. The model predicts that production will synchronize to a rhythmic input presented at the rate of the auditory stimulus, f_{ext}, only for combinations of parameters within the grey-shaded region. Accordingly, dark blue/pale red dots represent low/high synchronizers, respectively. Based on Assaneo et al.[111], it is assumed that *k* is a structural parameter proportional to the subject's white matter volume in the left dorsal pathway, whereas ω_0 is a volitional parameter given by the intended syllable rate, that is the rate at which a subject would speak without interaction with the external auditory input. Parts **b** and **c** adapted with permission from REF.[106], AAAS.

(high synchronizers) would whisper one syllable every two syllables of the external audio, and others (low synchronizers) would not keep any fixed ratio between produced and perceived syllables. This hypothesis represents a natural follow-up psychophysical experiment to further examine the model.

Although the model accounts for the currently available experimental evidence, it represents an idealization. For example, the speech perception and production substrates comprise extended, densely interconnected brain regions (BOXES 1,2) underwriting various complex processes[6,114]. The biophysical model advanced here simply outlines the basis for a mechanistic description of the sensorimotor integration of speech in the time domain. Next, we point to limitations of the model that require further research.

First, the model lacks spatial precision. In describing the relation between auditory and speech production systems in the time domain, one must address which brain areas mediate the interactions. Which is the brain region within the speech production network that behaves as an oscillator coupled to auditory activity? How does the dynamics of this region spread to the rest of the network? On the one hand, as discussed, it has been shown that primary motor regions for speech synchronize to auditory activity for frequencies within a range[106]. On the other, synchrony within the inferior frontal gyrus (IFG) to the external stimulus distinguishes between high and low synchronizers[111]. We hypothesize that the main region acting as an oscillator coupled to auditory areas is the IFG. The synchrony pattern observed in speech motor areas could derive from their dense connection to IFG. Existing evidence aligns with this hypothesis: the activity in the IFG has been shown to anticipate the envelope of produced speech[115]; and although intrasurgically cooling the speech motor cortex interferes with articulation, the same procedure over the IFG leads to changes in the speech rate[116]. Furthermore, the auditory cortex is directly connected with the IFG through the arcuate fasciculus (BOX 2).

Second, as already stated, during overt speech production it is possible to volitionally adjust one's produced syllable rate. Although we assume that the natural frequency of the oscillator matches the produced syllable rate, this frequency adjustment requires a mechanistic account. For example, in the Wilson–Cowan approach adopted here, varying the basal input activity coming from other brain areas can shift the frequency of the oscillator. Further research should explore whether oscillatory activity in the IFG is modulated through top-down signals coming from other brain regions and include this interaction in the model.

Last, an oversimplification of the model is the unidirectional perception-to-production interaction, as evidence for the reverse direction is abundant (for example, REFS[96,98,105]). A natural next step would be to include a bidirectional auditory–motor interaction as well as non-trivial dynamics for the auditory activity. One possibility would be, based on previous evidence[68,70], representing auditory areas as a second oscillator coupled to the speech envelope. The acquisition of new experimental data is required to guide ongoing research in the process of adding such additional, necessary layers of complexity to the model.

Contextualizing speech rhythms

The proposal that interconnected populations of inhibitory and excitatory neurons within a brain region can give rise to a larger neurocomputational 'unit' which behaves as an oscillator has already been discussed in the literature[107,117]. Furthermore, brain rhythms — including the theta rhythm of 4–8 Hz in focus here — are evident and consistent across brain regions and across species, suggesting that such temporal motifs emerge as a consequence of a preserved underlying neural architecture that is probably necessary to perform basic computational subroutines in the time domain[118]. Birdsong is an animal model widely studied in the literature, mostly because of similarities between that system and speech. For example, as in speech, birdsong comprises sequences of highly stereotyped acoustic units, and the tutor (the adult providing a template for the learned vocalizations) plays a fundamental role during development. The idea that different bird brain nuclei can be modelled as a set of coupled neural oscillators has been theorized and experimentally assessed[56,119]. These established lines of research lend further plausibility to the proposed position considered here.

Given that there is reason to assume that the cortical infrastructure underpinning speech perception and production derives from mechanisms that are neurobiologically elemental, ubiquitous and preserved, evolutionary speculations lead to interesting hypotheses. First, from this perspective, the 'frame/content' theory postulates

that speech is a sequence of concatenated units — syllables — originating from the natural cycle of the jaw, which itself evolved from ingestion-related rhythms[120]. This hypothesis is inspired by a set of behavioural observations: first, the stability of the syllable rate, described in the first section of this Review; second, infant babbling that displays rhythmicity with a frequency of 2–6 Hz[121], and presumably at this early developmental stage the natural rhythm of the mouth movements is exposed as the speech motor control system is not yet developed; and third, non-human primates' lip-smacking, orofacial gestures facilitating social interaction that could be one precursor of speech, involving cyclic mouth movements occurring at a rate between 4 and 5 Hz[122]. In line with the frame/content theory and these behavioural observations, we describe a plausible biophysical model for the temporal dynamics of the speech production brain regions, implying that the rhythm of speech emerges as a consequence of the underlying neural architecture. However, we believe, in accordance with others[122], that the central role given to the jaw needs to be revised. Evidence suggests (see Rhythmicity in the articulatory domain) that the coordination between the movements of the vocal tract articulators forms the basis of the speech rhythm, rather than individual articulator kinematics.

Second, we turn to the notion of 'active sensing' in relation to the reviewed data[123,124]. A common finding across sensory modalities and species is that perception relies on the motor system to scan the external word and that this scanning or sampling occurs at a quasi-rhythmic rate. It will come as no surprise that the typical sampling rate is around 2–10 Hz. This phenomenon has been widely reported in the literature, for example: the olfactory system uses sniffing (cyclic voluntary nose inhalations[125] and regular antenna sweeps in insects[126]); the somatosensory system in humans relies on periodic finger movements to determine the roughness of a surface[127], and in rodents the cyclic whisker movement for exploration of novel environments or object recognition has been described[128]; and in primates, the visual exploration of scenes is guided by the saccadic eye movements that occur at a restricted rate[129]. These data have inspired the proposal that rhythmic motor patterns allow perceptual systems to sample the environment at the appropriate temporal granularity for successful perception, discrimination, memory encoding and so forth[124]. The auditory system, although certainly having rich efferent projections, does not have a motor system to sample the world that way. But, as argued above (see Speech perception exhibits rhythmicity), auditory perception is enhanced for rhythmic acoustic signals within a range of frequencies, which happens to be the same range privileged by the speech production system. We suggest that this rhythmic alignment is no coincidence. Rather, it is a work-around for the absence of a motor system, to perform acoustic signal sampling at the same rate as other sensory systems. A motor system that produces speech at the correct temporal granularity pre-packages information for the perceptual auditory system: it is not necessary to impose any sampling on the stimulus, as it already carries the correct temporal pattern. The speech motor system, in its rhythmicity, obviates the need for a motor system for the ear.

Last, an intriguing relation exists between speech rhythms and their use for reading. In particular, when quantifying eye movements during the reading of real text across different orthographies, saccadic timing aligns well with typical syllabic rates, highlighting that the temporal structure of perceptual experience substantially overlaps across domains[130]. It has also been shown that the speech rate as well as audio-motor synchrony correlate with reading ability[131]. These data, jointly, suggest that the synchronization of rhythms described here extends beyond spoken language. Further experimentation is required to understand more thoroughly how reading acquisition could modify the properties of the model.

To summarize, the reviewed data arise from the literature on speech rhythmicity at the 'mesoscale', corresponding to the syllabic rate. The data reveal clear temporal patterns: speech production yields a remarkably stable rate that is faithfully recovered by the perception system, subserving the successful comprehension of spoken language. To characterize the mechanisms underlying the auditory–motor basis of speech in the time domain, we introduce a simple biophysical description of the speech production system as an oscillator coupled to a perceptual system that follows the envelope of the perceived speech signal. This offers a first explanation for several neurophysiological and behavioural findings. The model taps into the basic features of the temporal dynamics of the speech perception–production interaction.

Published online 6 May 2020

1. Ladefoged, P. *A Course in Phonetics* (Harcourt Brace, 1993).
2. Greenberg, S. & Ainsworth, W. A. (eds) *Listening to Speech: An Auditory Perspective* (Psychology Press, 2012).
3. Stevens, K. N. *Acoustic Phonetics* (MIT Press, 2000).
4. Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. Phonetic feature encoding in human superior temporal gyrus. *Science*, **343**, 1006–1010 (2014)
5. Marslen-Wilson, W. D. Functional parallelism in spoken word-recognition. *Cognition* **25**, 71–102 (1987).
6. Guenther, F. H. *Neural Control of Speech* (MIT Press, 2016).
7. Levelt, W. J. M. *Speaking: From Intention to Articulation* (MIT Press, 1993).
This foundational book describes in detail the many steps involved in spoken language production.
8. Greenberg, S., Carvey, H., Hitchcock, L., & Chang, S. Temporal properties of spontaneous speech — a syllable-centric perspective. *J. Phonetics* **31**, 465–485 (2003).
9. Goswami, U., & Leong, V. Speech rhythm and temporal structure: converging perspectives. *Lab. Phon.* **4**, 67–92 (2013).
10. Ding, N. et al. Temporal modulations in speech and music. *Neurosci. Biobehav. Rev.* **81**, 181–187 (2017).
This study includes an analysis of several large speech and music corpora demonstrating the acoustic regular modulation rate of these basic signal types.
11. Houtgast, T., & Steeneken, H. J. A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. *J. Acoust. Soc. Am.* **77**, 1069–1077 (1985).
12. Varnet, L., Ortiz-Barajas, M. C., Erra, R. G., Gervain, J. & Lorenzi, C. A cross-linguistic study of speech modulation spectra. *J. Acoust. Soc. Am.* **142**, 1976–1989 (2017).
Together with Ding et al. (2017), this paper reveals that signal processing for a wide variety of languages shows the temporal regularity of continuous speech.
13. Drullman, R., Festen, J. M., & Plomp, R. Effect of temporal envelope smearing on speech reception. *J. Acoust. Soc. Am.* **95**, 1053–1064 (1994).
14. Elliott, T. M., & Theunissen, F. E. The modulation transfer function for speech intelligibility. *PLOS Comput. Biol.* **5**, e1000302 (2009).
15. Clarke, J., & Voss, R. 1/f noise. music and speech. *Nature* **258**, 317–318 (1975).
16. Drullman, R. in *Listening to Speech: An Auditory Perspective* (eds Greenberg, S. & Ainsworth, W.) ch. 3 (Tailor & Francis, 2012).

17. Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A. & Ghazanfar, A. A. The natural statistics of audiovisual speech. *PLOS Comput. Biol.* **5**, e1000436 (2009).
18. Titze, I. R. *Principles of Voice Production* (Prentice Hall, 1994).
19. Sanders, I., & Mu, L. A three-dimensional atlas of human tongue muscles. *Anat. Rec.* **296**, 1102–1114 (2013).
20. Maeda, S. in *Speech Production and Speech Modelling* 2(eds Hardcastle, W. J. & Marchal, A.) 63–403 (Springer, 2012).
21. Story, B. & Titze, I. R. Parametrization of vocal tract area functions by empirical orthogonal modes. *Natl. Cent. Voice Speech Status Prog. Rep.* **10**, 9–23 (1996).
22. Assaneo, M. F., Ramirez Butavand, D., Trevisan, M. A., & Mindlin, G. B. Discrete anatomical coordinates for speech production and synthesis. *Front. Commun.* **4**, 13 (2019).
23. Bocquelet, F., Hueber, T., Girin, L., Savariaux, C. & Yvert, B. Real-time control of an articulatory-based speech synthesizer for brain computer interfaces. *PLOS Comput. Biol.* **12**, e1005119 (2016).
24. Abbs, J. H., Gracco, V. L., & Cole, K. J. Control of multimovement coordination: Sensorimotor mechansims in Speech motor programming. *J. Mot Behav.* **16**, 195–232 (1984).
25. Browman, C. P., & Goldstein, L. Articulatory phonology: An overview. *Phonetica* **49**, 155–180 (1992).
26. Hughes, O. M., & Abbs, J. H. Labial-mandibular coordination in the production of speech: Implications for the operation of motor equivalence. *Phonetica* **33**, 199–221 (1976).
27. Chartier, J., Anumanchipalli, G. K., Johnson, K., & Chang, E. F. Encoding of articulatory kinematic trajectories in human speech sensorimotor cortex. *Neuron* **98**, 1042–1054 (2018).
28. Walsh, B., & Smith, A. Articulatory movements in adolescents. *J. Speech Lang. Hear. R.* **45**, 1119–1133 (2002).
29. Chakraborty, R., Goffman, L., & Smith, A. Physiological indices of bilingualism: Oral–motor coordination and speech rate in Bengali–English speakers. *J. Speech Lang. Hear. R.* **51**, 321–332 (2008).
30. Riely, R. R., & Smith, A. Speech movements do not scale by orofacial structure size. *J. Appl. Physiol.* **94**, 2119–2126 (2003).
31. Bennett, J. W., Van Lieshout, P. H., & Steele, C. M. Tongue control for speech and swallowing in healthy younger and older adults *Int. J. Orofac. Myol.* **33**, 5–18.(2007).
32. Lindblad, P., Karlsson, S., & Heller, E. Mandibular movements in speech phrases — A syllabic quasiregular continuous oscillation. *Logop. Phoniatr. Vocol.* **16**, 36–42 (1991).
33. Ohala, J. J. The temporal regulation of speech. Auditory analysis and perception of speech, (eds. G. Fant, M. A. A. Tatham) 431–453 (Academic Press 1975).
34. Cummins, F. Oscillators and syllables: a cautionary note. *Front. Psychol.* **3**, 364 (2012).
35. Ghitza, O. The theta-syllable: a unit of speech information defined by cortical function. *Front. Psychol.* **4**, 138 (2013).
36. Strauβ, A., & Schwartz, J. L. The syllable in the light of motor skills and neural oscillations. *Lang. Cogn. Neurosci.* **32**, 562–569 (2017).
37. Mehler, J. The role of syllables in speech processing: Infant and adult data. *Philos. T. R. Soc. B. Biol. Sci.* **295**, 333–352 (1981).
38. Hooper, J. B. The syllable in phonological theory. *Language* **48**, 525–540 (1972).
39. Eimas, P. D. Segmental and syllabic representations in the perception of speech by young infants. *J. Acoust. Soc. Am.* **105**, 1901–1911 (1999).
40. Liberman, I. Y., Shankweiler, D., Fischer, F. W., & Carter, B. Explicit syllable and phoneme segmentation in the young child. *J. Exp. Child Psychol.* **18**, 201–212 (1974).
41. Ziegler, W., Aichert, I., & Staiger, A. Syllable-and rhythm-based approaches in the treatment of apraxia of speech. *Perspec. Neurophysiol. Neurogenic Speech Lang. Disord.* **20**, 59–66 (2010).
42. Carreiras, M., & Perea, M. Naming pseudowords in Spanish: Effects of syllable frequency. *Brain Lang.* **90**, 393–400 (2004).
43. Cholin, J., Levelt, W. J., & Schiller, N. O. Effects of syllable frequency in speech production. *Cognition* **99**, 205–235 (2006).
44. Guenther, F. H., Ghosh, S. S. & Tourville, J. A. Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain Lang.* **96**, 280–301 (2006).
This paper uses neuroimaging data and computational modelling to highlight the complex steps and brain regions implicated in syllable acquisition and production.
45. Jessen, M. Forensic reference data on articulation rate in German. *Sci. Justice* **47**, 50–67 (2007).
46. Fosler-Lussier, E., & Morgan, N. Effects of speaking rate and word frequency on pronunciations in convertional speech. *Speech Commun.* **29**, 137–158 (1999).
47. Pellegrino, F., Coupé, C. & Marsico, E. Across-language perspective on speech information rate. *Language* **87**, 539–558 (2011).
48. Jacewicz, E., Fox, R. A., O'Neill, C., & Salmons, J. Articulation rate across dialect, age, and gender. *Lang. Var. Change* **21**, 233–256 (2009).
49. Künzel, H. J. Some general phonetic and forensic aspects of speaking tempo. *Int. J. Speech Lang. Law* **4**, 48–83 (1997).
50. Ramig, L. A., & Ringel, R. L. Effects of physiological aging on selected acoustic characteristics of voice. *J. Speech Lang. Hear. R.* **26**, 22–30 (1983).
51. Clopper, C. G., & Smiljanic, R. Effects of gender and regional dialect on prosodic patterns in American English. *J. Phon.* **39**, 237–245 (2011).
52. He, L., & Dellwo, V. Amplitude envelope kinematics of speech: Parameter extraction and applications. *J. Acoust. Soc. Am.* **141**, 3582–3582 (2017).
53. Mermelstein, P. Automatic segmentation of speech into syllabic units. *J. Acoust. Soc. Am.* **58**, 880–883 (1975).
54. Tilsen, S., & Arvaniti, A. Speech rhythm analysis with decomposition of the amplitude envelope: characterizing rhythmic patterns within and across languages. *J. Acoust. Soc. Am.* **134**, 628–639 (2013).
55. Titze, I. R. Measurements for voice production: research and clinical applications. *J. Acoust. Soc. Am.* **104**, 1148 (1998).
56. Amador, A., Perl, Y. S., Mindlin, G. B. & Margoliash, D. Elemental gesture dynamics are encoded by song premotor cortical neurons. *Nature* **495**, 59–64 (2013).
57. Norton, P., & Scharff, C. "Bird song Metronomics": isochronous organization of zebra finch song rhythm. *Front. Neurosci.* **10**, 309 (2016).
58. Ahissar, E. et al. Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proc. Natl Acad. Sci. USA* **98**, 13367–13372 (2001).
59. Luo, H. & Poeppel, D. Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron* **54**, 1001–1010 (2007).
60. Henry, M. J., & Obleser, J. Frequency modulation entrains slow neural oscillations and optimizes human listening behavior. *Proc. Natl Acad. Sci. USA* **109**, 20095–20100 (2012).
61. Lakatos, P., Gross, J., & Thut, G. A new unifying account of the roles of neuronal entrainment. *Curr. Biol.* **29**, R890–R905 (2019).
62. Peña, M., & Melloni, L. Brain oscillations during spoken sentence processing. *J. Cogn. Neurosci.* **24**, 1149–1164 (2012).
63. Howard, M. F., & Poeppel, D. Discrimination of speech stimuli based on neuronal response phase patterns depends on acoustics but not comprehension. *J. Neurophysiol.* **104**, 2500–2511 (2010).
64. Golumbic, E. M. Z., et al. Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". *Neuron* **77**, 980–991 (2013).
65. Ding, N., & Simon, J. Z. Emergence of neural encoding of auditory objects while listening to competing speakers. *Proc. Natl Acad. Sci. USA* **109**, 11854–11859 (2012).
66. Broderick, M. P., Anderson, A. J., & Lalor, E. C. Semantic context enhances the early auditory encoding of natural speech. *J. Neurosci.* **39**, 7564–7575 (2019).
67. Assaneo, M. F. et al. The lateralization of speech–brain coupling is differentially modulated by intrinsic auditory and top-down mechanisms. *Front. Integr. Neurosci.* **13**, 28 (2019).
68. Peelle, J. E., & Davis, M. H. Neural oscillations carry speech rhythm through to comprehension. *Front. Psychol.* **3**, 320 (2012).
69. Capilla, A., Pazo-Alvarez, P., Darriba, A., Campo, P., & Gross, J. Steady-state visual evoked potentials can be explained by temporal superposition of transient event-related responses. *PLOS ONE* **6**, e0014543 (2011).
70. Doelling, K. B., Assaneo, M. F., Bevilacqua, D., Pesaran, B., & Poeppel, D. An oscillator model better predicts cortical entrainment to music. *Proc. Natl Acad. Sci. USA* **116**, 10113–10121 (2019).
71. Giraud, A. L., & Poeppel, D. Cortical oscillations and speech processing: emerging computational principles and operations. *Nat. Neurosci.* **15**, 511–517 (2012).
This article provides a perspective on how oscillatory neural activity may form the basis of segmenting speech to create units appropriate for cortical processing.
72. Gross, J. et al. Speech rhythms and multiplexed oscillatory sensory coding in the human brain. *PLOS Biol.* **11**, e1001752 (2013).
This work presents neurophysiological data revealing a nested hierarchy of entrained cortical oscillations underlying the segmentation and coding of spoken language.
73. Peelle, J. E., Gross, J. & Davis, M. H. Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cereb. Cortex* **23**, 1378–1387 (2013).
74. Doelling, K. B., Arnal, L. H., Ghitza, O. & Poeppel, D. Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing. *Neuroimage* **85**, 761–768 (2014).
75. Abrams, D. A., Nicol, T., Zecker, S., & Kraus, N. Abnormal cortical processing of the syllable rate of speech in poor readers. *J. Neurosci.* **29**, 7686–7693 (2009).
76. Cutini, S., Szűcs, D., Mead, N., Huss, M., & Goswami, U. Atypical right hemisphere response to slow temporal modulations in children with developmental dyslexia. *Neuroimage* **143**, 40–49 (2016).
77. Wilsch, A., Neuling, T., Obleser, J., & Herrmann, C. S. Transcranial alternating current stimulation with speech envelopes modulates speech comprehension. *Neuroimage* **172**, 766–774 (2018).
78. Zoefel, B., Archer-Boyd, A., & Davis, M. H. Phase entrainment of brain oscillations causally modulates neural responses to intelligible speech. *Curr. Biol.* **28**, 401–408 (2018).
79. Riecke, L., Formisano, E., Sorger, B., Başkent, D., & Gaudrain, E. Neural entrainment to speech modulates speech intelligibility. *Curr. Biol.* **28**, 161–169 (2018).
80. Luo, H., Wang, Y., Poeppel, D., & Simon, J. Z. Concurrent encoding of frequency and amplitude modulation in human auditory cortex: Encoding transition. *J. Neurophysiol.* **98**, 3473–3485 (2007).
81. Viemeister, N. F. Temporal modulation transfer functions based upon modulation thresholds. *J. Acoust. Soc. Am.* **66**, 1364–1380 (1979).
82. Zwicker, E. Die Grenzen der Hörbarkeit der Amplitudenmodulation und der Frequenzmodulation eines Tones [The limits of perceptibility of the amplitude-modulation and the frequency-modulation of a tone]. *Akust. Beih.* **2** (Suppl. 3), 125–133 (1952).
83. Giraud, A. L. et al. Representation of the temporal envelope of sounds in the human brain. *J. Neurophysiol.* **84**, 1588–1598 (2000).
84. Boemio, A., Fromm, S., Braun, A. & Poeppel, D. Hierarchical and asymmetric temporal sensitivity in human auditory cortices. *Nat. Neurosci.* **8**, 389–395 (2005).
85. Teng, X., Tian, X., Rowland, J., & Poeppel, D. Concurrent temporal channels for auditory processing: Oscillatory neural entrainment reveals segregation of function at different scales. *PLOS Biol.* **15**, e2000812 (2017).
86. Liégeois-Chauvel, C., Lorenzi, C., Trébuchon, A., Régis, J., & Chauvel, P. Temporal envelope processing in the human left and right auditory cortices. *Cereb. Cortex* **14**, 731–740 (2004).
87. Overath, T., Zhang, Y., Sanes, D. H. & Poeppel, D. Sensitivity to temporal modulation rate and spectral bandwidth in the human auditory system: fMRI evidence. *J. Neurophysiol.* **107**, 2042–2056 (2012).
88. Versfeld, N. J., & Dreschler, W. A. The relationship between the intelligibility of time-compressed speech and speech in noise in young and elderly listeners. *J. Acoust. Soc. Am.* **111**, 401–408 (2002).
89. Trouvain, J. On the comprehension of extremely fast synthetic speech. *Saarl. Work. Pap. Linguist.* **1**, 5–13 (2007).
90. Ghitza, O., & Greenberg, S. On the possible role of brain rhythms in speech perception: intelligibility of time-compressed speech with periodic and aperiodic insertions of silence. *Phonetica* **66**, 113–126 (2009).
This paper presents an innovative behavioural design using speech compression that highlights the relevance of syllable-sized units for intelligibility.

91. Wilson, S. M., Saygin, A. P., Sereno, M. I. & Iacoboni, M. Listening to speech activates motor areas involved in speech production. *Nat. Neurosci.* **7**, 701–702 (2004).
92. D'Ausilio, A. et al. The motor somatotopy of speech perception. *Curr. Biol.* **19**, 381–385 (2009).
93. Du, Y., Buchsbaum, B. R., Grady, C. L. & Alain, C. Noise differentially impacts phoneme representations in the auditory and speech motor systems. *Proc. Natl Acad. Sci. USA* **111**, 7126–7131 (2014).
94. Houde, J. F. & Jordan, M. I. Sensorimotor adaptation in speech production. *Science* **279**, 1213–1216 (1998).
95. Black, J. W. The effect of delayed side-tone upon vocal rate and intensity. *J. Speech Disord.* **16**, 56–60 (1951).
This study is a first to demonstrate that delayed auditory feedback compromises and slows down speech production.
96. Flinker, A. et al. Single-trial speech suppression of auditory cortex activity in humans. *J. Neurosci.* **30**, 16643–16650 (2010).
97. Tian, X., & Poeppel, D. The effect of imagination on stimulation: the functional specificity of efference copies in speech processing. *J. Cogn. Neurosci.* **25**, 1020–1036 (2013).
98. Park, H., Ince, R. A. A., Schyns, P. G., Thut, G. & Gross, J. Frontal top-down signals increase coupling of auditory low-frequency oscillations to continuous speech in human listeners. *Curr. Biol.* **25**, 1649–1653 (2015).
99. Onojima, T., Kitajo, K., & Mizuhara, H. Ongoing slow oscillatory phase modulates speech intelligibility in cooperation with motor cortical activity. *PLOS ONE* **12**, e0183146 (2017).
100. Rimmele, J. M., Sun, Y., Michalareas, G., Ghitza, O. & Poeppel, D. Dynamics of functional networks for syllable and word-level processing. *BioRxiv* https://doi.org/10.1101/584375 (2019).
101. Cope, T. E. et al. Evidence for causal top-down frontal contributions to predictive processes in speech perception. *Nat. Commun.* **8**, 1–16 (2017).
102. Kovelman, I. et al. Brain basis of phonological awareness for spoken language in children and its disruption in dyslexia. *Cereb. Cortex* **22**, 754–764 (2012).
103. Molinaro, N., Lizarazu, M., Lallier, M., Bourguignon, M., & Carreiras, M. Out-of-synchrony speech entrainment in developmental dyslexia. *Hum. Brain Mapp.* **37**, 2767–2783 (2016).
104. Keitel, A., Gross, J., & Kayser, C. Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. *PLOS Biol.* **16**, e2004473 (2018).
105. Rimmele, J. M., Morillon, B., Poeppel, D., & Arnal, L. H. Proactive sensing of periodic and aperiodic auditory patterns. *Trends Cogn. Sci.* **22**, 870–882 (2018).
106. Assaneo, M. F. & Poeppel, D. The coupling between auditory and motor cortices is rate-restricted: evidence for an intrinsic speech–motor rhythm. *Sci. Adv.* **4**, eaao3842 (2018).
This study uses neural data and modelling to show how the auditory and speech–motor systems are coupled in phase most strongly at a time scale corresponding roughly to syllable duration.
107. Hoppensteadt, F. C. & Izhikevich, E. M. *Weakly Connected Neural Networks* (Springer, 1997).
108. Giraud, A. L. et al. Endogenous cortical rhythms determine cerebral specialization for speech perception and production. *Neuron* **56**, 1127–1134 (2007).
109. Keitel, A., & Gross, J. Individual human brain areas can be identified from their characteristic spectral activation fingerprints. *PLOS Biol.* **14**, e1002498 (2016).
110. Lee, B. S. Effects of delayed speech feedback. *J. Acoust. Soc. Am.* **22**, 824–826 (1950).
111. Assaneo, M. F. et al. Spontaneous synchronization to speech reveals neural mechanisms facilitating language learning. *Nature Neurosci.* **22**, 627–632 (2019).
This study uses an uncomplicated behavioural speech synchronization test to show how subjects differ anatomically and physiologically in their ability to align their sensorimotor systems.
112. Stuart, A., Kalinowski, J., Rastatter, M. P. & Lynch, K. Effect of delayed auditory feedback on normal speakers at two speech rates. *J. Acoust. Soc. Am.* **111**, 2237 (2002).
113. Saffran, J. R., Aslin, R. N. & Newport, E. L. Statistical learning by 8-month-old infants. *Science* **274**, 1926–1928 (1996).
114. Hickok, G. & Poeppel, D. The cortical organization of speech processing. *Nat. Rev. Neurosci.* **8**, 393–402 (2007).
115. Magrassi, L., Aromataris, G., Cabrini, A., Annovazzi-Lodi, V. & Moro, A. Sound representation in higher language areas during language generation. *Proc. Natl Acad. Sci. USA* **112**, 1868–1873 (2015).
116. Long, M. A. et al. Functional segregation of cortical regions underlying speech timing and articulation. *Neuron* **89**, 1187–1193 (2016).
117. Wilson, H. R., & Cowan, J. D. Excitatory and inhibitory interactions in localized populations of model neurons. *Biophys. J.* **12**, 1–24 (1972).
118. Buzsáki, G., Logothetis, N., & Singer, W. Scaling brain size, keeping timing: evolutionary preservation of brain rhythms. *Neuron* **80**, 751–764 (2013).
119. Laje, R. & Mindlin, G. B. *The Physics of Birdsong* (Springer-Verlag, 2005).
120. MacNeilage, P. F. The frame/content theory of evolution of speech production. *Behav. Brain Sci.* **21**, 499–511 (1998).
This paper describes an influential theory on how evolution privileged syllables as the basic units of spoken language.
121. De Boysson-Bardies, B., Bacri, N., Sagart, L., & Poizat, M. Timing in late babbling. *J. Child Lang.* **8**, 525–539 (1981).
122. Ghazanfar, A. A., Takahashi, D. Y., Mathur, N., & Fitch, W. T. Cineradiography of monkey lip-smacking reveals putative precursors of speech dynamics. *Curr. Biol.* **22**, 1176–1182 (2012).
123. Brooks, J. X., & Cullen, K. Predictive sensing: The role of motor signals in sensory processing. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* **4**, 842–850 (2019).
124. Schroeder, C. E., Wilson, D. A., Radman, T., Scharfman, H., & Lakatos, P. Dynamics of active sensing and perceptual selection. *Curr. Opin. Neurobiol.* **20**, 172–176 (2010).
This article advances the perspective that motor systems can play an integral role in shaping perceptual processes by sampling the input.
125. Wesson, D. W., Verhagen, J. V., & Wachowiak, M. Why sniff fast? The relationship between sniff frequency, odor discrimination, and receptor neuron activation in the rat. *J. Neurophysiol.* **101**, 1089–1102 (2009).
126. Huston, S. J., Stopfer, M., Cassenaer, S., Aldworth, Z. N., & Laurent, G. Neural encoding of odors during active sampling and in turbulent plumes. *Neuron* **88**, 403–418 (2015).
127. Lederman, S. J. Tactual roughness perception: spatial and temporal determinants. *Can. J. Psychol.* **37**, 498 (1983).
128. Deschênes, M., Moore, J., & Kleinfeld, D. Sniffing and whisking in rodents. *Curr. Opin. Neurobiol.* **22**, 243–250 (2012).
129. Fiebelkorn, I. C., & Kastner, S. A rhythmic theory of attention. *Trends Cogn. Sci.* **23**, 87–101 (2019).
130. Gagl, B. et al. Reading at the speed of speech: the rate of eye movements aligns with auditory language processing. *bioRxiv* https://doi.org/10.1101/391896 (2018).
131. Tierney, A., & Kraus, N. Auditory-motor entrainment and phonological skills: precise auditory timing hypothesis (PATH). *Front. Hum. Neurosci.* **8**, 949 (2014).
132. Wrench, A. MOCHA-TIMIT database (CSTR, Univ. of Edinburgh, 1999).
133. Indefrey, P. & Levelt, W. J. M. in *The New Cognitive Neurosciences* (ed. Gazzaniga, M. S.) 845–866 (MIT Press, 2000).
134. Bouchard, K. E., Mesgarani, N., Johnson, K. & Chang, E. F. Functional organization of human sensorimotor cortex for speech articulation. *Nature* **495**, 327–332 (2013).
135. Tremblay, P., & Small, S. L. Motor response selection in overt sentence production: a functional MRI study. *Front. Psychol.* **2**, 253 (2011).
136. Hickok, G., Buchsbaum, B., Humphries, C. & Muftuler, T. Auditory–motor interaction revealed by fMRI: speech, music, and working memory in area Spt. *J. Cogn. Neurosci.* **15**, 673–682 (2003).
137. Brennan, J., & Pylkkänen, L. The time-course and spatial distribution of brain activity associated with sentence processing. *Neuroimage* **60**, 1139–1148 (2012).
138. Lau, E. F., Phillips, C., & Poeppel, D. A cortical network for semantics:(de) constructing the N400. *Nat. Rev. Neurosci.* **9**, 920–933 (2008).
139. Catani, M., & De Schotten, M. T. A diffusion tensor imaging tractography atlas for virtual in vivo dissections. *Cortex* **44**, 1105–1132 (2008).
140. Liberman, A. M., & Mattingly, I. G. The motor theory of speech perception revised. *Cognition* **21**, 1–36 (1985).
141. Lotto, A. J., Hickok, G. S. & Holt, L. L. Reflections on mirror neurons and speech perception. *Trends Cogn. Sci.* **13**, 110–114 (2009).
142. Skipper, J. I., Devlin, J. T., & Lametti, D. R. The hearing ear is always found close to the speaking tongue: Review of the role of the motor system in speech perception. *Brain Lang.* **164**, 77–105 (2017).
143. Lane, H. The motor theory of speech perception: A critical review. *Psychol. Rev.* **72**, 275 (1965).
144. Rauschecker, J. P. & Scott, S. K. Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nat. Neurosci.* **12**, 718–724 (2009).
145. Friederici, A. D. Pathways to language: fiber tracts in the human brain. *Trends Cogn. Sci.* **13**, 175–181 (2009).
146. Dick, A. S., Bernal, B., & Tremblay, P. The language connectome: new pathways, new concepts. *Neuroscientist* **20**, 453–467 (2014).
147. Saur, D. et al. Ventral and dorsal pathways for language. *Proc. Natl Acad. Sci. USA* **105**, 18035–18040 (2008).
This study presents some of the first anatomical data to demonstrate that there are distinct ventral and dorsal pathways underpinning language processing.
148. Wong, F. C., Chandrasekaran, B., Garibaldi, K., & Wong, P. C. White matter anisotropy in the ventral language pathway predicts sound-to-word learning success. *J. Neurosci.* **31**, 8780–8785 (2011).
149. Brauer, J., Anwander, A., Perani, D., & Friederici, A. D. Dorsal and ventral pathways in language development. *Brain Lang.* **127**, 289–295 (2013).
150. Catani, M., & De Schotten, M. T. A diffusion tensor imaging tractography atlas for virtual in vivo dissections. *Cortex* **44**, 1105–1132 (2008).

Acknowledgements
The authors thank O. Ghitza and J. Orpella for valuable feedback. They acknowledge the support of the Max Planck Society and NIH R01DC05660.

Author contributions
Both authors contributed equally to all aspects of the manuscript.

Competing interests
The authors declare no competing interests.

Peer reviewer information
Nature Reviews Neuroscience thanks J. Gross, G. Mindlin and the other anonymous reviewer for their contribution to the peer review of this work.

Publisher's note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.