14 Rhythm Is a Timescale

AVIAD ALBERT AND MARTINE GRICE

14.1 Introduction

Rhythm is perhaps one of the most perplexing notions in linguistic theory, a theory that has been struggling to accommodate it as a clear construct of language systems, with a relatively coherent function and definition. Rhythm is of major interest in prosodic analysis of speech, but in the mainstream linguistic literature, it has proven to be a less fruitful auditory dimension than pitch, which has long been the focus of intonation studies, aided by the relatively straightforward display of the fundamental frequency (F0) contour.

Rhythm and pitch are major aspects of auditory perception, and are exploited in both music and speech. However, while pitch has relatively similar cognitive goals in both domains (i.e., making use of our specialization in detecting fundamental frequencies in complex harmonic structures), rhythm seems to have different cognitive goals. In music, rhythm is most effectively used to couple oscillations internally with the motor system, as well as externally between agents. In speech, however, rhythm is most effectively used to construct an internal representation of timing effects in the prosody of language systems. This discrepancy between domains has likely contributed to the impression in the linguistic literature that speech rhythm is a highly elusive concept (Turk and Shattuck-Hufnagel, 2013; Nolan and Jeon, 2014).

In this chapter, we offer a new synthesis of existing theories and findings that relate to rhythm. We take into account evolutionary, physical, cognitive, neurological, musicological, and linguistic aspects of this question in order to paint a holistic picture of rhythm. This synthesis offers a framework for understanding rhythm in a manner that can be shared more coherently across disciplines. The immediate contribution is in explicating rhythm within linguistic contexts, where the notion of musical rhythm seems to have led many studies astray. We present a clear and principled delineation of rhythmic goals in music and in speech. Essentially, we claim that the shared behavior that makes those different goals rhythmic is not the adherence to metronome-like equal intervals (isochrony), which is so characteristic of musical signals, but the shared timescale of temporal integration that both music and speech exploit to different ends.

https://doi.org/10.1017/9781009295888.017 Published online by Cambridge University Press

14.1.1 Speech Rhythm: A Brief Overview

There is a general consensus that some aspects of speech are rhythmic. However, after many decades of research (e.g., Brown, 1911; Pike, 1945; Abercrombie, 1967; Allen, 1975; Dauer, 1983; Cummins and Port, 1998; Arvaniti, 2009; Nolan and Jeon, 2014; Inbar et al., 2020), it is still unclear which units of speech actually play a role in this rhythmicality, the most commonly reported being phonemes, syllables, and stressed syllables. Moreover, it is unclear across the different studies which type of rhythmicality is intended when the notion of rhythm is invoked.

One type of rhythmicality is often referred to with terms such as isochronous, periodic, temporal, coordinative rhythm or beat. It resembles the workings of a clock, characterized by equal intervals between successive events. Another type of rhythmicality is often referred to with terms such as meter, prominence, accentual or contrastive rhythm. It is less bound to an abstract external clock, and it is mostly based on the distinction between weak and strong events in a sequence, promoting the grouping of these events in different ways. While the former type (beat) is based on temporal relations, the latter type (meter) incorporates various dimensions that can contribute to strength. On top of a duration-based distinction, strength also includes the acoustic power, the spectral quality, and the F0 of the events in question (see overviews in Gordon and Roettger, 2017, and Baumann and Winter, 2018).

These two types of rhythmicality are often conflated. This is due to some long-standing assumptions that isochronous relations may hold between the strong members of successive events, rather than between all members in the string. Such ideas gained traction during the second half of the twentieth century, with the widely adopted distinction between syllable-timed and stress-timed languages (e.g., Pike, 1945; Abercrombie, 1967; Dauer, 1983).

The distinction between stress-timing and syllable-timing has been widely investigated using rhythm metrics (e.g., Low, 1998; Ramus et al., 1999; Dellwo, 2006). Although they have failed to adequately characterize different language types in a systematic manner (e.g., Nolan and Asu, 2009; Arvaniti, 2009, 2012; Barry, 2010; Lowit, 2014), they have not been abandoned. We return to this classic distinction in our discussion in Section 14.5.

14.1.2 Scope of Current Synthesis

In the following sections we lay the foundations for a holistic proposal. In Section 14.2 we discuss the notion of timescales in perception, and we introduce a theoretical framework for the principled reduction of the perceptual auditory space based on timescales. In Section 14.3 we provide a brief overview of the role of rhythm in prosody, and in Section 14.4 we extend the discussion with theories and findings from the literature on neural oscillations, considering both speech and music, and how they relate to the notion of entrainment. We end with a discussion in Section 14.5.

14.2 Timescales of Perception

The ability to construct a stable and useful representation of the external physical reality is a critical aspect of survival. An inevitable outcome is that perceptual and cognitive

https://doi.org/10.1017/9781009295888.017 Published online by Cambridge University Press

systems evolve to optimally capture physical phenomena that can be beneficial to survival. In that sense, the type of events that different species can see and hear were selected in evolution to support each species' occupation of a specialized niche in a shared ecosystem (Krause, 2012).

There are two major conclusions that can be drawn. The first is that the information that travels via sound waves (as well as light waves) is very beneficial to constructing useful representations of reality on earth's atmosphere. The second ensuing conclusion is that different species focus on the ranges of the spectra of energy waves that are most beneficial to them. Evolution selects the ranges of the spectra that support each species' successful occupation of a certain niche. Good examples can be gleaned from the auditory system of bats or the visual system of bees, both covering ranges different from humans.

Crucially for us, the spectra of sound waves are temporally distributed in ranges that can be characterized in terms of timescales. All of this is akin to saying that the competition for survival is a determinant force on the timescales in which living brains optimally operate. Humans' auditory perception and cognition are therefore a manifestation of the spectra of sound waves that we can temporally integrate. The timescales within which we temporally integrate acoustic events were most likely selected at a very early stage in the evolution of our species, to support the most basic needs of survival, such as detecting danger and locating food. Our higher cognitive abilities, such as the capacity for establishing intricate communication systems (i.e., spoken language) and the capacity for creative endeavors (e.g., music), must therefore piggyback on those previously selected timescales of temporal integration that were likely "hardwired" at a prior stage of evolution (see also Meyer, 2018).

14.2.1 Defining Human Auditory Timescales with PRiORS

In this chapter, we make use of a theoretical framework that was introduced in Albert (2023) to provide building blocks for linguistic models that are based on auditory perception. Perceptual Regimes of Repetitive Sound (PRiORS) offers a reduction of auditory perception into its basic primitives, based on distinct behavioral effects of temporal integration in human cognition. It reveals two types of perceptual regimes that operate at different timescales: the temporal regime and the spectral regime. These observations are not entirely new. Similar and related observations have been previously suggested from perspectives that include musicology (Stockhausen, 1959), acoustics (Flanagan and Guttman, 1960), cognitive psychology (Warren, 1982:80), linguistics (Rosen, 1992; Gibbon, 2023), and neuroscience (Zatorre et al., 2002; Chi et al., 2005). PRiORS borrows ideas from all of the above in order to present a simplified perception-based synthesis.

The temporal regime in PRiORS dominates the timescale within which we perceive successive acoustic events as isolated events that can be rhythmically related. We can thus sense whether successive discrete events within this timescale can be related, via one of the following scenarios: Events that occur at (relatively) regular (quasi-)isochronous intervals within the timescale of the temporal regime will yield the sensation of a steady beat; we can also perceive whether successive acoustic events within this timescale display dynamic patterns of deceleration or acceleration, that is, whether the interval between successive events decreases or

https://doi.org/10.1017/9781009295888.017 Published online by Cambridge University Press

increases; and finally, if intervals between events are randomly distributed within this timescale, these successive acoustic events will be perceived as meandering.

The spectral regime in PRiORS is the timescale within which auditory repetition is too fast to support the perception of isolated events. Instead, we perceive successive acoustic events as a continuous sound. The differences in the rates of repetition at this timescale result in the perception of different spectral details (i.e., different frequencies). We can sense if successive events within this timescale occur in (quasi-)isochronous succession, which will result in a complex harmonic sound, and we can also perceive whether successive events exhibit deceleration or acceleration patterns, such as falling or rising pitch. If successive acoustic events are randomly distributed within this timescale, they will be perceived as aperiodic noise.

The two distinct perceptual regimes in the PRiORS framework should not be confused with the more familiar notions of time and frequency domains in mathematical representations, which lack a perceptual component (see Section 14.2.2). Likewise, they should not be confused with suggestions in the mainstream neurolinguistic literature (e.g., Poeppel, 2003; Giraud and Poeppel, 2012; Doelling et al., 2014; see also Chapters 3 and 5) for different temporal integration windows, whereby spectral detail is linked to phoneme-size intervals (20–50 ms long) and is therefore related to rates of 20–50 Hz. Furthermore, the simple one-dimensional spectral regime in PRiORS should not be confused with the more complex notion of frequency modulation (FM) in speech research (e.g., Teoh et al., 2019; Chapter 23). FM carrier signals exhibit high-frequency oscillations (e.g., the F0 in human vocalization), but FM modulation signals, which are usually the signals of interest in FM configurations, are themselves in the low-frequency range (up to about 8 Hz), which is the rate at which they modulate the high-frequency spectral information. As we shall see below, the spectral regime in PRiORS is inherently distinct and above those rates of oscillation – mostly above 50 Hz.

14.2.2 Time and Frequency Domains

The notion of temporal and spectral regimes echoes the time and frequency domains in mathematical representations, going back to the Fourier transform (Fourier, 1822), which decomposes a time series into a sum of finite series of sine or cosine functions. Note that in most contexts of acoustic analysis, the procedure is referred to as fast Fourier transform (FFT), which is the name given to a wide range of algorithms that perform quick calculations of the Fourier transform (see the overview in Brigham, 1988). FFT allows us to switch the representation of complex natural sounds from the time domain (where all the subcomponents are bundled together) to the frequency domain (where we can represent the distribution of acoustic power of different frequencies at different points in time).

In principle, the temporal regime is congruent with the notion of the time domain, while the spectral regime is congruent with the notion of the frequency domain. However, while the time and frequency domains can independently describe the same event in mathematical terms, the perceptual regimes in PRiORS operate at two mostly distinct and mutually exclusive timescales. The non-overlapping qualia in PRiORS are therefore referred to as temporal and spectral regimes to set them apart from the overlapping time and frequency domains.

https://doi.org/10.1017/9781009295888.017 Published online by Cambridge University Press

14.2.3 Visual FFT-Based Simulations

It is useful to illustrate the distinction between perceptual regimes with a Band-Limited Impulse Train (BLIT) synthesis that produces a train of transient acoustic bursts at adjustable rates (in contrast to continuous sine waves). Each burst is a single impulse, which is the shortest burst a given system can produce (i.e., one sample in digital setups), with equal power across the frequency scale. A perfect impulse has acoustic power over an infinite frequency range, but the impulses in a BLIT are band-limited to human hearing ranges, between approximately 20 and 20k Hz. The BLIT signal can be effectively visualized with standard FFT-based tools that convert signals between time domain and frequency domain representations (see Section 14.2.2).

Table 14.1 presents a rough sketch of the relevant timescales of the two perceptual regimes. Within each regime the effects of repetition are termed differently in order to maintain a distinction (although note that they are largely interchangeable in many conventional uses). We use the term Rhythm for the main effect occurring within the timescale of the temporal regime versus Periodicity for the main effect occurring within the timescale of the spectral regime. Table 14.1 also shows the upper and lower boundaries of these temporal integration effects. Repetitions within the temporal regime may be too slow to be integrated as rhythmic (below 0.5 Hz for intervals longer than 2 seconds; see Fraisse, 1984; Repp, 2005; Ulbrich et al., 2007; Wittmann, 2011; Farbood et al., 2013). Likewise, repetitions within the spectral regime may be too fast to be perceived as periodic (above 5k Hz, given that our auditory system can typically discern pitch up to about 5k Hz; see Ward, 1954; Attneave and Olson, 1971). Furthermore, the switch between regimes does not occur at once. Table 14.1 displays a transitional range between the temporal and the spectral regimes, in which both effects are present but neither is clear enough, resulting in an indeterminate effective sensation.

Four examples are provided in Figure 14.1, each one with three corresponding visualization panels. The bottom white panel presents a one-second-long waveform (oscillogram) that shows the unipolar transient bursts produced by the BLIT synthesis in the time domain, going from left to right. The number of visible bursts within this one-second interval corresponds to the rate of the BLIT in Hz. The two upper dark panels show FFT-based analyses exhibiting the dispersion of acoustic power across the audible frequency range in the frequency domain.[1] The middle panel, often called a spectrum or a spectrograph, exhibits a two-dimensional representation of frequency (x-axis) and power (y-axis), while the top panel, which is typically called a spectrogram, exhibits a three-dimensional representation of frequency (x-axis), power (shade), and time (y-axis). The frequency x-axes of the spectrum and the spectrogram are perfectly aligned to facilitate the interpretation of the spectrum in the middle as a "slice," or a still image of the temporal representation in the spectrogram above it. This means that unlike the more

[1] The visual effects of the FFT-based representations presented here are calibrated to reflect human perception. We use a commercial metering application, SpectraFoo by Metric Halo (version 4.2.3), with an analyzer depth setting of 4,096 points (10 Hz). The BLIT synthesis and the oscillogram were produced with the sound design software Plogue Bidule (version 0.9766).

https://doi.org/10.1017/9781009295888.017 Published online by Cambridge University Press

Table 14.1 Perceptual regimes and their timescales

Perceptual regimes with corresponding effects and timescales (rough sketch). Hz = Hertz (repetitions per second); ms = millisecond (duration of intervals).

		Timescales	
Perceptual regimes	Effects	Hz	ms
Temporal	*Infra-rhythmic*	*0–0.5*	*∞–2k*
	Rhythm	0.5–12	2k–83.3
(Transitional)	*(indeterminate)*	*12–50*	*83.3–20*
Spectral	Periodicity	50–5k	20–0.2
	Ultra-periodic	*5k–20k*	*0.2–0.05*

typical configuration, the spectrogram here is moving in time from bottom to top, rather than from left to right.

Figure 14.1(a) (top left) shows a clear rhythmic effect at 4 Hz, indicated by four bursts in the bottom oscillogram panel. A single burst appears with equal power along the (band-limited) frequency range in the spectrum, indicated by the fairly straight horizontal line across the middle panel. Note that the still image shown here captured a moment in time in which the power graph of the spectrum was high. With rhythmic bursts, such as the 4 Hz BLIT in Figure 14.1(a), this graph goes visibly up and down over time. Above it, in the corresponding upper spectrogram panel, a succession of 10 impulses over a short period of time (about 2.5 seconds) is visible as isolated bursts, indicated by the horizontal lines going from bottom to top.[2]

In sharp contrast, Figure 14.1(d) (bottom right) clearly shows tonal behavior at 120 Hz. There are, indeed, 120 bursts in the time domain display of the bottom oscillogram panel, but the isolated bursts are no longer visible in the top spectrogram panel; that is, there are no horizontal lines going from bottom to top across the upper panel. The sensation of isolated discrete bursts transitions into one of continuous sound at these higher rates of repetition. This perceptual effect is reflected by the two FFT-based representations in Figure 14.1(d), which display a signal with the properties of a continuous sound that has a complex harmonic structure. The middle spectrum panel shows a series of "bumps" along the white curve, from left to right, corresponding to a series of continuous energy "poles" in the vertical representations of the upper spectrogram panel. This is a harmonic series in which the rate of repetition of the BLIT synthesis is mapped onto the F0 of the continuous sound (120 Hz in this case).[3] This demonstrates that at this faster timescale of the spectral regime, the sensation of repetition feeds perceptual effects of continuity and pitch, rather than of discreteness and rhythm.

[2] Note that the fading strength of white towards the left side of the spectrogram (horizontal lines in Figure 14.1a) is due to this analyzer using a pink noise accumulation mode, resulting in gradual attenuation of low frequencies in the display.

[3] The polarity of the impulses can also play a role with pitch perception at higher rates within the spectral regime, as demonstrated in Flanagan and Guttman (1960). For this reason, we used only unipolar impulses for which the relationship between impulse rate and frequency rate is kept stable.

https://doi.org/10.1017/9781009295888.017 Published online by Cambridge University Press

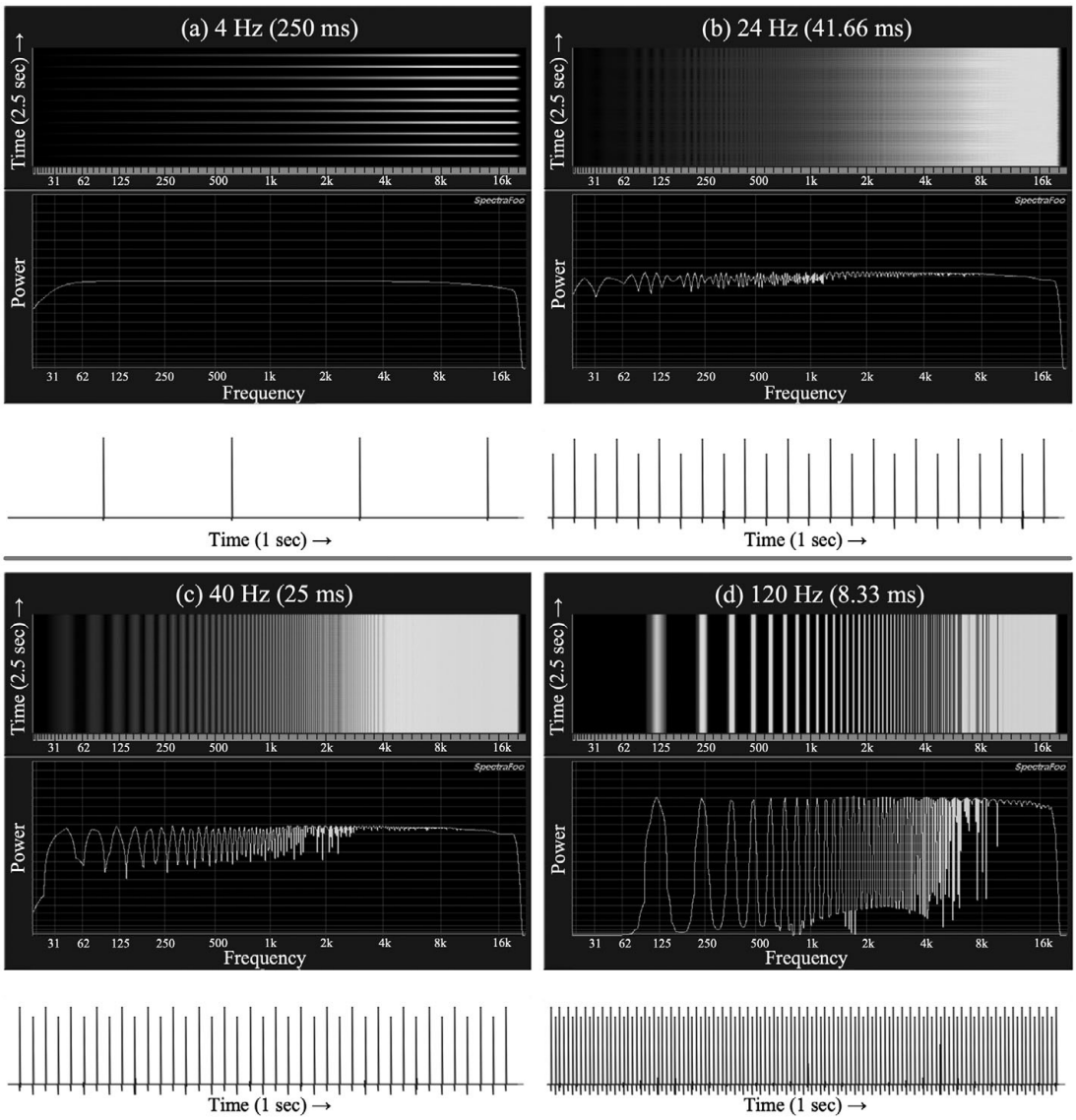


Figure 14.1 A BLIT demonstration.
Illustration of perceptual regimes with visual analyses of acoustic impulse trains (BLITs) at different rates and different domains (see text for details).
Image taken from Albert (2023).

Between the two regimes, we can observe a transitional range in which effects of both rhythm and periodicity are present, but neither is strong enough to be sufficiently clear. Figures 14.1 (b)–14.1(c) demonstrate this transitional range between the two distinct regimes. Figure 14.1(b) (top right) is especially well suited for illustrating the indeterminacy of the transitional range. At a BLIT rate of 24 Hz, the impulses seem to be too fast to support a rhythmic perception of discrete bursts, and, at the same time, too slow to support the perception of a continuous

https://doi.org/10.1017/9781009295888.017 Published online by Cambridge University Press

harmonic (pitch-bearing) sound. The upper spectrogram panel of Figure 14.1(b) reflects that by showing a combination of both faint horizontal lines that reflect isolated events in time, as well as faint vertical lines that reflect the emerging harmonic structure of a continuous complex tone (visible also as small corresponding energy fluctuations in the middle spectrum panel).

14.2.4 PRiORS-Derived Hypotheses

14.2.4.1 Universal Aspects of Syllabic Structure

Syllables are abstract units of phonological systems, and they do not easily lend themselves to consistent and straightforward phonetic descriptions in terms of perception.[4] The PRiORS framework can do a lot of heavy lifting in this regard, by providing the conditions that can explain the evolutionary trajectory of syllables from a perceptual perspective. According to this analysis, syllables were shaped by selection to optimally take advantage of the two perceptual regimes: carrying pitch in the spectral regime and giving rise to dynamic timing relations in the temporal regime (see also Strauß and Schwartz, 2017, and Räsänen et al., 2018, for proposals that suggest somewhat similar divisions of labor). In other words, syllables universally exploit the spectral regime with an internal segmental makeup that is optimized to carry pitch (namely by the requirement for sonorous nuclei; see Albert and Nicenboim, 2022). At the same time, syllables universally exploit the temporal regime with sizes that give rise to dynamic speech rate effects.[5] Figure 14.2 illustrates this.

14.2.4.2 Prosodic Effects Are Dynamic

Pitch contours in speech signals are not static but dynamic. They are constantly changing in order to achieve communicative goals. Consider, for example, the periods during a rising pitch contour, in which every period is shorter than the previous one. These degrees of change do not hinder the perception of a coherent rising pitch contour, demonstrating our specialized ability to perceive gradually changing dynamic pitch (Temperley, 2008; Morgan et al., 2019). As long as these communicatively relevant pitch changes occur within the timescale of the spectral regime – and follow basic Gestalt principles – they invoke a reliable effect in perception.

A similar behavior can be observed for rhythm (e.g., Cope et al., 2012). Rather than exhibiting static isochrony, speech units within the temporal regime exhibit mostly dynamic changes in terms of acceleration and deceleration patterns within certain ranges. These patterns are exploited for communicative goals via prosody, such as chunking the message into phrases,

[4] For articulation-based descriptions of syllables, see the *frame/content* theory (MacNeilage, 1998), which takes an evolutionary perspective based on prior capacities for chewing, sucking, and licking, and see *articulatory phonology* (e.g., Browman and Goldstein, 1992; Gafos et al., 2014), which takes a dynamical systems approach, where syllables are units of motor gesture coordination.

[5] Pitch perception is considered to require a minimum of three periods to detect the rate of repetition (Josephs, 1967; Fyk, 1987), such that typical syllabic size also ensures the sufficient length for pitch perception: 60 ms will be the minimum size required for a F0 of 50 Hz (the bottom of the range for an adult male) – most syllables are much longer than 60 ms and most of the F0 data in speech are higher than 50 Hz, making pitch perception all the more feasible.

https://doi.org/10.1017/9781009295888.017 Published online by Cambridge University Press

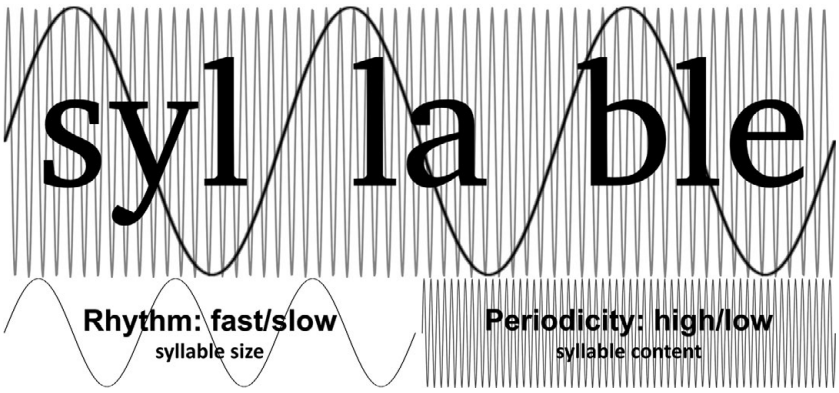


Figure 14.2 Perceptual regimes and syllables.
Schematic illustration of the relationship between perceptual regimes and syllabic units. Segmental makeup in terms of sonority is related to the spectral regime with high-frequency oscillations within syllables, while syllabic size is related to the temporal regime with low-frequency oscillations between syllables. The ratio between the low- and high-frequency oscillations in this illustration is arbitrarily set to be 1:20. This is a realistic ratio such that if syllables are taken to have a typical average duration of 200 ms (5 Hz), the high-frequency oscillation within it would reflect a typical F0 for adult males at 100 Hz. For simplicity, this generalized illustration shows a single rate at each timescale using a steady phase (isochronous repetitions). Image taken from Albert (2023).

highlighting important information and turn-taking management (see a more detailed non-exhaustive overview in the following Section 14.3).

From a functional linguistic perspective, isochrony is not a useful effect of temporal integration, as it is not immediately clear what purpose this would serve in speech. There are no behaviorally observable isochronous responses to (non-isochronous) spontaneous speech, or, in other words, we do not – and likely cannot – dance to spontaneous speech (see the discussion in Section 14.5.3). This can be contrasted with music, which much more clearly exploits isochrony to achieve certain goals (see Section 14.4.1).

A truly steady isochronous signal – or a quasi-isochronous one, given perceptually negligible jitter – would not be very useful for prosody. Within the spectral regime, that would entail that all the syllables have the same static F0 rate, and within the temporal regime that would entail that all the syllables have the same duration. In fact, it is the constant state of flux in the prosody of spontaneous speech that is critical to effectively exploit the sensations of rhythm and pitch in their respective timescales in speech (see Section 14.3).

14.3 Speech Prosody and Dynamic Speech Rate

More often than not, when the notion of speech rate is invoked, it is meant in the sense of the global speech rate, looking at the ratio between a certain linguistic unit and a given unit of time. This is often measured in terms of the average syllable duration (e.g., Miller et al., 1984) or

https://doi.org/10.1017/9781009295888.017 Published online by Cambridge University Press

other similar measurements (see Tilsen and Tiede, 2023, for a recent proposal for global speech rate measurements). We explicitly refer here to dynamic speech rate, to express the idea that the tempo of speech is in a constant state of flux (explained in Section 14.2.4.2; also, see Gibbon, 2023, and Chapter 23 for a related perspective on speech rate dynamics). Unlike global speech rate, dynamic speech rate should be more adequately characterized in terms of a time series trajectory. Few studies have used this type of dynamic speech rate trajectory representation thus far, making the perceptual local speech rate in Pfitzinger (2001) a notable exception.

One of the most important goals achieved by dynamic speech rate in the prosody of speech is the division of the message into chunks (see Christiansen and Chater, 2016, for a functional cognitive account of chunks as crucial units in models of speech processing). Chunking can be achieved by various prosodic effects (e.g., Gee and Grosjean, 1984; Price et al., 1991; Dilley and Pitt, 2010; Reinisch et al., 2011), which play a major role in comprehension of speech in real time (see Chapters 17 and 18, and see the debate in Kazanina and Tavano, 2023, and Lo et al., 2023, about the syntactic and neurological aspects of chunking). Additionally, chunking is essential in turn-taking management to maintain the conversational flow (e.g., Sacks et al., 1978; Wilson and Wilson, 2005; Levinson and Torreira, 2015; Ogden and Hawkins, 2015; Roberts et al., 2015).

The chunking function of dynamic speech rate is often studied in terms of prosodic boundary phenomena (Schubö et al., 2023), which are generally assumed to be present in all languages, regardless of their typology (Fletcher, 2010). One of the most commonly researched aspects is (progressive) domain final lengthening (Klatt, 1976; Cummins, 1999; White, 2002, 2014; Kohler, 2003; Paschen et al., 2022). Domain final lengthening involves deceleration in articulation when approaching a prosodic boundary, something that can be found in movement in general (Turk and Shattuck-Hufnagel, 2007). Despite the mechanical explanations (along with others relating to planning of upcoming phrases), domain final lengthening interacts with the phonology of the language it operates on, leading to language-specific differences in the domain, scope, and execution of this lengthening (Paschen et al., 2022).

Another, less investigated aspect of chunking is often studied in terms of acceleration at the beginning of a prosodic domain, referred to as initial rush, anacrusis, or phrase-initial acceleration (Cruttenden, 1997; Fletcher, 2010). This is particularly commonly (but not exclusively) reported in languages with (final or right-branching) lexical stress, with unstressed syllables being rapidly produced at the beginning of the domain.

The temporal integration of events at the timescale of the temporal regime can also be exploited via silent gaps in the sequence, which are usually considered as pauses in prosodic analyses. Pauses serve as strong cues to demarcation of phrasal units, which can also promote effective chunking of speech (e.g., Grosjean, 1979; Duez, 1982, 1985; Gee and Grosjean, 1984; Heldner and Edlund, 2010), and they appear to be more common in slow speech (Trouvain and Grice, 1999).

Furthermore, dynamic speech rate is sometimes used to mark the distinctive status of a certain phrase in the stream of speech. Phrases that are part of a self-repair strategy (Schegloff et al., 1977; Levelt and Cutler, 1983; Dingemanse and Floyd, 2014) make a good case in point as they have already been found to be systematically employing speech rate cues (Plug, 2016).

Dynamic speech rate can also be used to highlight informative parts of the message: Slowing down allows for more precise articulation of the individual sounds (hyperarticulation)

https://doi.org/10.1017/9781009295888.017 Published online by Cambridge University Press

(Lindblom, 1990) and for an intonation contour to be produced in full (Grice et al., 2018). Conversely, less informative parts can be speeded up with reduction in the segmental domain (Cohen Priva, 2017; Hall et al., 2018) as well as a possible truncation of intonation contours (Rathcke, 2013). Domain final lengthening and accentual lengthening can coexist, indicating a cumulative effect (Turk and White, 1999).

14.4 Neural Perspectives

14.4.1 Entrainment in Music

Unlike spontaneous speech, music tends to exhibit a steady (quasi-)isochronous beat (e.g., Bolton, 1894; Fraisse, 1963; Repp, 2005; Bispham, 2006, 2021; Fitch, 2012; Grahn, 2012). In this respect, the notion of entrainment is often invoked to describe how listening to music may involve the phase-locking of neural activity to the external musical signal (e.g., Merker et al., 2009; Phillips-Silver et al., 2010; Nozaradan et al., 2012; Doelling et al., 2019). Entrainment makes perfect sense in cases where an external isochronous beat promotes motor coordination in synchrony, such as dancing, tapping, nodding, and so on (see Repp, 2005; Ravignani et al., 2014; Kotz et al., 2018), even in the absence of actual motor response (Chen et al., 2008). Entrainment to a musical rhythm is not only useful to internally couple one's motor system with an external clock; it also allows coupling oscillations across different agents who share the same space. The goal of isochrony in this case is to achieve this type of system(s)-wide phase-locking, which many researchers equate with rewarding goals such as rapid social bonding, social cohesion, and even pain relief (Wiltermuth and Heath, 2009; Cohen et al., 2010; Kokal et al., 2011; Bowling et al., 2013; Tarr et al., 2016; Savage et al., 2021).

14.4.2 Entrainment in Speech

Recent decades have seen an influx of studies that link language processing with neural activity within a range of low-frequency oscillations (see reviews in Meyer, 2018; Myers et al., 2019; Poeppel and Assaneo, 2020). This is largely in line with the entire effective rhythmic range of the temporal regime in PRiORS, at about 0.5–12 Hz (Keitel et al., 2018). Central among these are Theta band oscillations (about 4–8 Hz) that correspond to syllable-size events. Frequencies above and below the Theta band are likewise often linked with units below and above the syllable, respectively (e.g., Delta band oscillations at about 0.5–4 Hz are often linked with words/phrases; see Inbar et al., 2020; Rimmele et al., 2021; Chapter 15).

The majority of the literature on the link between brain oscillations and speech makes two (implicit or explicit) assumptions: (i) that the mechanism at play is entrainment, whereby the rate of internal oscillations phase-lock to the external speech stimulus (exogenous neural oscillation); and (ii) that this phase-locking procedure facilitates and improves comprehension in general, under the assumption that the rate of activity in the brain can be equated with the rate of attention allocation when processing incoming information (see, for example, Large and Jones, 1999; Ghitza, 2011; Peelle and Davis, 2012; Doelling et al., 2014; Goswami, 2018).

https://doi.org/10.1017/9781009295888.017 Published online by Cambridge University Press

With regards to the first assumption, a growing number of voices among the researchers in the field have been calling for a revised understanding of the notion of entrainment in the case of brain oscillations that respond to speech (see Cummins, 2012; Breska and Deouell, 2017; Haegens and Golumbic, 2018; Rimmele et al., 2018; Kotz et al., 2018; Meyer et al., 2019, 2020). For example, Cummins (2012) claims that entrainment and phase-locking are not adequate descriptions for the tracking of spontaneous speech, which is essentially non-isochronous. Meyer et al. (2019) suggest the term *intrinsic synchronicity* as a separate process from classic entrainment, to cover cases in which the external signal is non-isochronous yet linked to relatively isochronous internal oscillations (endogenous neural oscillations).

The second assumption regarding the direct link between brain oscillations and comprehension should also be questioned. Such a link between oscillations and comprehension via rate of attention seems to reflect a computer metaphor of the mind (e.g., Searle, 1990; Spivey, 2007): The frequency of the oscillations in this type of explanation is likened to the sample rate in digital systems, where the resolution of the sample rate determines the resolution of the obtained signal. While this is not implausible, there seem to be other links to speech comprehension via prosody that would benefit from the time-keeping capability of internal oscillations, without the need to resort to new theoretical entities.

As briefly summarized above in Section 14.4.1, music tends much more than speech to exploit the timescale of the temporal regime to produce isochronous signals that promote the phase-locking of motor systems. This is arguably one of the major effects of the musical experience as a social phenomenon. Non-isochronous beats in music destroy this effect. Likewise, speech tends much more than music to exploit the timescale of the temporal regime to produce non-isochronous (yet dynamically changing) signals that are designed to use changes in speech rate to create prosodic effects. Isochronous rhythm in speech is devoid of this communicative dimension of the language sound system (but see the discussion in Section 14.5.3 for specific roles isochrony may play in the prosodic repertoire of languages).

In our current proposal, we argue along with others that entrainment in music cannot be the same as entrainment in speech perception. Speech does not have the same goal that music has in terms of social bonding via syncing of the motor systems to external signals (see Section 14.5.3). We add to that a skeptical view towards the second assumption in its simplistic form. When processing speech, brain oscillations should allow us to perceive the dynamic timing patterns in speech rate, which, in turn, enrich and facilitate comprehension via prosody (see Glushko et al., 2022). This seems like a stronger explanation than the currently dominant attention-based explanation, which is modeled in terms of digital sampling.

14.5 Discussion

14.5.1 Isochrony, Where Are You?

In music, the isochronous element is typically in the external signal such that the internal brain oscillations can entrain to it (exogenous neural oscillations). This is essential for achieving some of the most powerful goals that music can achieve – rapid social bonding and group coherence via coupled oscillations.

https://doi.org/10.1017/9781009295888.017 Published online by Cambridge University Press

In speech processing, the relatively isochronous element comes from the brain oscillations themselves (endogenous neural oscillations), such that the external and temporally dynamic speech can be measured internally. This is essential for achieving communicative goals via prosody (see Section 14.3) by obtaining a mental representation of timing patterns (for models of internal clocks, see Treisman, 1963; Church, 1984; Wittmann, 2013; Allman et al., 2014; Paton and Buonomano, 2018).

The proposed description that assumes that (quasi-)isochrony in speech is internal rather than external fits well with some prominent critical reviews on the topic of rhythm in the linguistic literature. Having found no convincing case for isochrony in speech signals, Lehiste (1977) suggested that isochrony in speech may be projected from perception. Likewise, Nolan and Jeon (2014) suggested that isochrony may be a metaphor that we project onto speech in perception. These intuitive descriptions are in line with the current proposal.

Rhythm, as we understand it here, is the main effect of temporal integration at the timescale of the temporal regime (roughly 0.5–12 Hz).[6] Isochrony is not what defines rhythm – it is one of the goals that rhythm can achieve. In music, this goal requires external sources, while in speech, isochrony needs to be sourced internally, since it is likely serving as the baseline measurement for mental representation of dynamic speech rate effects in prosodic perception.

14.5.2 Meter Is Independent of Isochrony

Meter (or contrastive rhythm) is therefore not a relationship that needs to be tied to isochrony. Any sequence of elements that can be grouped via temporal integration (i.e., that fall within the relevant timescale of the temporal regime) can be in a strong–weak relationship, regardless of whether they are temporally equidistant or not. This more nuanced understanding can also provide an explanation for the continued prevalence of one of the most contested notions in the literature on rhythm and speech, namely the notions of stress-timing versus syllable-timing (Pike, 1945; Dauer, 1983). In our proposed understanding of rhythm, it is possible to extract timing patterns from all successive syllabic units as much as it is possible to extract such patterns more selectively from the string of strong syllables only. Again, this should not entail that the units (either all syllables or just the stressed ones) are equidistant, only that they are in a rhythmically relevant relationship within the timescale of the temporal regime.

For example, languages such as English, that include many weak syllables next to strong ones, can plausibly use the strong syllables (either all stressed syllables or just the primary stressed ones) to signal dynamic change in speech rate. Languages such as French, with phrase-final prominence, may tune to this phrasal position for similar effects, while languages with no apparent prominence asymmetries may likely signal speech rate patterns using all the syllables (or moras) in the stream of speech.

[6] Interestingly, Ding et al. (2017) found that the rate of speech tends to center around 5 Hz (reflecting the tendency of syllable durations towards 200 ms-long intervals), while music tends to center around slower rates of 2 Hz (reflecting the tendency of musical beats towards 120 beats per minute).

https://doi.org/10.1017/9781009295888.017 Published online by Cambridge University Press

14.5.3 Isochrony between Music and Speech

We make a distinction between speech in spontaneous communication scenarios and speech types that incorporate musical aspects, which may include prayers, mantras, chants, poems, nursery rhymes, freestyle rap, and many other types of belief-based and/or artistic expressions (see Leong and Goswami, 2015; Fuchs and Reichel, 2016; Davis, 2017; Cummins, 2018; Danner et al., 2021). All the latter tend to incorporate quasi-isochrony that should be related to the musical aspect of these complex fusions between music and speech. The nature of the compromise in the vast majority of music and speech fusions is such that speech gives way to musical isochrony.

Speaking intentionally in a quasi-isochronous rhythm can be attested as part of the repertoire of intonation patterns of at least some languages. In English, for example, a tendency to make strong syllables equidistant can often characterize the intonation pattern of lists (e.g., alignment of p-centers; see Morton et al., 1976; Couper-Kuhlen, 1993). Likewise, quasi-isochrony can be attested in declarations in public speeches and statements (e.g., see White's 2014:45 analysis of Bill Clinton's statement). Here, again, the task of the listener does not appear to be entrainment to the rhythm in the speech signal in order to couple oscillations between their own motor system and the external auditory signal. It seems more likely that the task of the listener is to infer the timing relations such that a "sing-song" effect can arise (which in turn might make the message more appealing and/or long-lasting in memory, due to its atypical "musicality").

The speech-to-song illusion (Deutsch et al., 2011) shows that stretches of spontaneous speech can be perceived as music when a certain portion is repeated in a loop (various factors, such as the size of the loop, the number of repetitions, and the segmental makeup of the looped speech, can affect the strength of this illusion; see Rathcke et al., 2021a). This is of interest in the context of the current synthesis because a looped auditory signal is, in fact, isochronous at the level of the entire loop size, and it can become "musical" as soon as the structure of the repeating pattern is revealed in perception (see tapping to speech in Rathcke et al., 2021b). The invisible boundaries between music and speech are brought to light in this illusion and they imply that isochrony is at the heart of the distinction between music and speech in perception (rather than being what they both share).

Seen this way, speech rhythm is not such an elusive concept after all. We suggest that rhythm is the timescale within which temporal relationships between isolated events are perceived. While music tends to use this timescale to promote phase-locking to an external clock, speech rather exploits it to achieve a mental representation of dynamic speech rate in prosody.

Box 14.1 Chapter Overview

Summary

Rhythm is the timescale within which temporal relationships between isolated events are perceived. While music tends to use this timescale in terms of isochrony to promote phase-locking to an external clock, speech rather exploits the rhythmic timescale to achieve a mental representation of dynamic speech rate.

https://doi.org/10.1017/9781009295888.017 Published online by Cambridge University Press

Box 14.1 (cont.)

Implications

Speech rhythm is too irregular to be described using isochrony and entrainment. Instead, speech rhythm should be understood in terms of dynamic speech rate, which is an important aspect of speech prosody, serving major communicative goals such as chunking the signal, highlighting newsworthy information, and turn-taking.

Gains

We provided a principled framework for understanding temporal relations in the auditory domain in relation to music and speech. For speech, this framework can help linguists focus on the rhythmic properties employed in communication, and helps neuroscientists appreciate how prosody, in particular speech rate, facilitates comprehension via neural oscillations.

REFERENCES

Abercrombie, D. (1967). *Elements of general phonetics*. Edinburgh: Edinburgh University Press.

Albert, A. (2023). *A model of sonority based on pitch intelligibility*. Studies in Laboratory Phonology: Vol. 13. Berlin: Language Science Press.

Albert, A., and Nicenboim, B. (2022). Modeling sonority in terms of pitch intelligibility with the nucleus attraction principle. *Cognitive Science*, 46(7), e13161. https://doi.org/10.1111/cogs.13161

Allen, G. D. (1975). Speech rhythm: Its relation to performance universals and articulatory timing. *Journal of Phonetics*, 3(2), 75–86. https://doi.org/10.1016/S0095-4470(19)31351-8

Allman, M. J., Teki, S., Griffiths, T. D., and Meck, W. H. (2014). Properties of the internal clock: First- and second-order principles of subjective time. *Annual Review of Psychology*, 65, 743–771. https://doi.org/10.1146/annurev-psych-010213-115117

Arvaniti, A. (2009). Rhythm, timing and the timing of rhythm. *Phonetica*, 66(1–2), 46–63. https://doi.org/10.1159/000208930

Arvaniti, A. (2012). The usefulness of metrics in the quantification of speech rhythm. *Journal of Phonetics*, 40(3), 351–373. https://doi.org/10.1016/j.wocn.2012.02.003

Attneave, F., and Olson, R. K. (1971). Pitch as a medium: A new approach to psychophysical scaling. *American Journal of Psychology*, 84(2), 147–166. https://doi.org/10.2307/1421351

Barry, W. J. (2010). Rhythm measures in retrospect: Reflections on the nature of spoken language rhythm. Proceedings of the Fifth AISV National Conference, pp. 3–12.

Baumann, S., and Winter, B. (2018). What makes a word prominent? Predicting untrained German listeners' perceptual judgments. *Journal of Phonetics*, 70, 20–38. https://doi.org/10.1016/j.wocn.2018.05.004

Bispham, J. C. (2006). Rhythm in music: What is it? Who has it? And why? *Music Perception*, 24(2), 125–134. https://doi.org/10.1525/mp.2006.24.2.125

Bispham, J. C. (2021). Music, evolution, and the experience of time. In M. Doffman, E. Payne, and T. Young (Eds.), The oxford handbook of time in music (pp. 149–168). Oxford: Oxford University Press.

https://doi.org/10.1017/9781009295888.017 Published online by Cambridge University Press

Bolton, T. L. (1894). Rhythm. *American Journal of Psychology*, 6(2), 145–238. https://doi.org/10.2307/1410948

Bowling, D. L., Herbst, C. T., and Fitch, W. T. (2013). Social origins of rhythm? Synchrony and temporal regularity in human vocalization. *PLoS ONE*, 8(11), e80402. https://doi.org/10.1371/journal.pone.0080402

Breska, A., and Deouell, L. Y. (2017). Neural mechanisms of rhythm-based temporal prediction: Delta phase-locking reflects temporal predictability but not rhythmic entrainment. *PLoS Biology*, 15(2), e2001665. https://doi.org/10.1371/journal.pbio.2001665

Brigham, E. O. (1988). *The fast Fourier transform and its applications*. Engelwood Cliffs: Prentice-Hall.

Browman, C. P., and Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica*, 49(3–4), 155–180. https://doi.org/10.1159/000261913

Brown, W. (1911). Studies from the psychological laboratory of the university of California: Temporal and accentual rhythm. *Psychological Review*, 18(5), 336–346. https://doi.org/10.1037/h0074259

Chen, J. L., Penhune, V. B., and Zatorre, R. J. (2008). Listening to musical rhythms recruits motor regions of the brain. *Cerebral Cortex*, 18(12), 2844–2854. https://doi.org/10.1093/cercor/bhn042

Chi, T., Ru, P., and Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *Journal of the Acoustical Society of America*, 118(2), 887–906. https://doi.org/10.1121/1.1945807

Christiansen, M. H., and Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, 39, e62. https://doi.org/10.1017/S0140525X1500031X

Church, R. M. (1984). Properties of the internal clock. *Annals of the New York Academy of Sciences*, 423, 566–582. https://doi.org/10.1111/j.1749-6632.1984.tb23459.x

Cohen, E. E., Ejsmond-Frey, R., Knight, N., and Dunbar, R. I. (2010). Rowers' high: Behavioural synchrony is correlated with elevated pain thresholds. *Biology Letters*, 6(1), 106–108. https://doi.org/10.1098/rsbl.2009.0670

Cohen Priva, U. (2017). Informativity and the actuation of lenition. *Language*, 93(3), 569–597. https://doi.org/10.1353/lan.2017.0037

Cope, T. E., Grube, M., and Griffiths, T. D. (2012). Temporal predictions based on a gradual change in tempo. *Journal of the Acoustical Society of America*, 131(5), 4013–4022. https://doi.org/10.1121/1.3699266

Couper-Kuhlen, E. (1993). *English speech rhythm: Form and function in everyday verbal interaction*. Amsterdam: John Benjamins.

Cruttenden, A. (1997). *Intonation*. Cambridge: Cambridge University Press.

Cummins, F. (1999). Some lengthening factors in English speech combine additively at most rates. *Journal of the Acoustical Society of America*, 105(1), 476–480. https://doi.org/10.1121/1.424576

Cummins, F. (2012). Oscillators and syllables: A cautionary note. *Frontiers in Psychology*, 3, 364. https://doi.org/10.3389/fpsyg.2012.00364

Cummins, F. (2018). *The ground from which we speak: Joint speech and the collective subject*. Newcastle-upon-Tyne, UK: Cambridge Scholars Publishing.

Cummins, F., and Port, R. (1998). Rhythmic constraints on stress timing in English. *Journal of Phonetics*, 26(2), 145–171. https://doi.org/10.1006/jpho.1998.0070

Danner, S. G., Krivokapić, J., and Byrd, D. (2021). Co-speech movement in conversational turn-taking. *Frontiers in Communication*, 6, 779814. https://doi.org/10.3389/fcomm.2021.779814

Dauer, R. M. (1983). Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics*, 11(1), 51–62. https://doi.org/10.1016/S0095-4470(19)30776-4

Davis, I. R. (2017). Rhythmic analysis of rap: What can we learn from "flow"? Master's thesis, University of Canterbury.

Dellwo, V. (2006). Rhythm and speech rate: A variation coefficient for deltac. In P. Karnowski and I. Szigeti (Eds.), *Language and language-processing: Proceedings of the 38th linguistics colloquium, piliscsaba* (pp. 231–241). Frankfurt/Main: Peter Lang.

https://doi.org/10.1017/9781009295888.017 Published online by Cambridge University Press

Deutsch, D., Henthorn, T., and Lapidis, R. (2011). Illusory transformation from speech to song. *Journal of the Acoustical Society of America*, 129(4), 2245–2252. https://doi.org/10.1121/1.3562174

Dilley, L. C., and Pitt, M. A. (2010). Altering context speech rate can cause words to appear or disappear. *Psychological Science*, 21(11), 1664–1670. https://doi.org/10.1177/0956797610384743

Ding, N., Patel, A. D., Chen, L., et al. (2017). Temporal modulations in speech and music. *Neuroscience & Biobehavioral Reviews*, 81, 181–187. https://doi.org/10.1016/j.neubiorev.2017.02.011

Dingemanse, M., and Floyd, S. (2014). Conversation across cultures. In N. J. Enfield, P. Kockelman, and J. Sidnell (Eds.), *The Cambridge handbook of linguistic anthropology* (pp. 447–480). Cambridge: Cambridge University Press.

Doelling, K. B., Arnal, L. H., Ghitza, O., and Poeppel, D. (2014). Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing. *NeuroImage*, 85, 761–768. https://doi.org/10.1016/j.neuroimage.2013.06.035

Doelling, K. B., Assaneo, M. F., Bevilacqua, D., Pesaran, B., and Poeppel, D. (2019). An oscillator model better predicts cortical entrainment to music. *Proceedings of the National Academy of Sciences*, 116(20), 10113–10121. https://doi.org/10.1073/pnas.1816414116

Duez, D. (1982). Silent and non-silent pauses in three speech styles. *Language and Speech*, 25(1), 11–28.

Duez, D. (1985). Perception of silent pauses in continuous speech. *Language and Speech*, 28(4), 377–389. https://doi.org/10.1177/002383098502800403

Farbood, M. M., Marcus, G., and Poeppel, D. (2013). Temporal dynamics and the identification of musical key. *Journal of Experimental Psychology: Human Perception and Performance*, 39(4), 911. https://doi.org/10.1037/a0031087

Fitch, W. T. (2012). The biology and evolution of rhythm: Unraveling a paradox. In P. Rebuschat, M. Rohrmeier, J. A. Hawkins, and I. Cross (Eds.), *Language and music as cognitive systems* (pp. 73–95). Oxford: Oxford University Press.

Flanagan, J. L., and Guttman, N. (1960). On the pitch of periodic pulses. *Journal of the Acoustical Society of America*, 32(10), 1308–1319. https://doi.org/10.1121/1.1907900

Fletcher, J. (2010). The prosody of speech: Timing and rhythm. In W. J. Hardcastle, J. Laver, and F. E. Gibbon (Eds.), *The handbook of phonetic sciences* (second edition, pp. 521–602). New York: John Wiley & Sons. https://doi.org/10.1002/9781444317251.ch15

Fourier, J.-B.-J. (1822). *Théorie analytique de la chaleur*. Paris: F. Didot.

Fraisse, P. (1963). *The psychology of time*. New York: Harper & Row.

Fraisse, P. (1984). Perception and estimation of time. *Annual Review of Psychology*, 35(1), 1–37.

Fuchs, S., and Reichel, U. D. (2016). On the relationship between pointing gestures and speech production in german counting out rhymes: Evidence from motion capture data and speech acoustics. Proceedings of Tagung Phonetik und Phonologie im deutschsprachigen Raum, pp. 50–53.

Fyk, J. (1987). Duration of tones required for satisfactory precision of pitch matching. *Bulletin of the Council for Research in Music Education*, 91, 38–44.

Gafos, A. I., Charlow, S., Shaw, J., and Hoole, P. (2014). Stochastic time analysis of syllable-referential intervals and simplex onsets. *Journal of Phonetics*, 44(1), 152–166. https://doi.org/10.1016/j.wocn.2013.11.007

Gee, J. P., and Grosjean, F. (1984). Empirical evidence for narrative structure. *Cognitive Science*, 8(1), 59–85. https://doi.org/10.1016/S0364-0213(84)80025-7

Ghitza, O. (2011). Linking speech perception and neurophysiology: Speech decoding guided by cascaded oscillators locked to the input rhythm. *Frontiers in Psychology*, 2, 130. https://doi.org/10.3389/fpsyg.2011.00130

Gibbon, D. (2023). The rhythms of rhythm. *Journal of the International Phonetic Association*, 53(1), 233–265. https://doi.org/10.1017/S002510032100008

https://doi.org/10.1017/9781009295888.017 Published online by Cambridge University Press

Giraud, A.-L., and Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience*, 15(4), 511–517.

Glushko, A., Poeppel, D., and Steinhauer, K. (2022). Overt and implicit prosody contribute to neurophysiological responses previously attributed to grammatical processing. *Scientific Reports*, 12(1), 1–18.

Gordon, M., and Roettger, T. (2017). Acoustic correlates of word stress: A cross-linguistic survey. *Linguistics Vanguard*, 3(1), 20170007. https://doi.org/10.1515/lingvan-2017-0007

Goswami, U. (2018). A neural basis for phonological awareness? An oscillatory temporal-sampling perspective. *Current Directions in Psychological Science*, 27(1), 56–63. https://doi.org/10.1177/0963721417727520

Grahn, J. A. (2012). Neural mechanisms of rhythm perception: Current findings and future perspectives. *Topics in Cognitive Science*, 4(4), 585–606. https://doi.org/10.1111/j.1756-8765.2012.01213.x

Grice, M., Savino, M., and Roettger, T. B. (2018). Word final schwa is driven by intonation: The case of Bari Italian. *Journal of the Acoustical Society of America*, 143(4), 2474–2486. https://doi.org/10.1121/1.5030923

Grosjean, F. (1979). A study of timing in a manual and a spoken language: American Sign Language and English. *Journal of Psycholinguistic Research*, 8(4), 379–405. https://doi.org/10.1007/BF01067141

Haegens, S., and Golumbic, E. Z. (2018). Rhythmic facilitation of sensory processing: A critical review. *Neuroscience & Biobehavioral Reviews*, 86, 150–165. https://doi.org/10.1016/j.neubiorev.2017.12.002

Hall, K. C., Hume, E., Jaeger, T. F., and Wedel, A. (2018). The role of predictability in shaping phonological patterns. *Linguistics Vanguard*, 4(s2), 20170027. https://doi.org/10.1515/lingvan-2017-0027

Heldner, M., and Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4), 555–568. https://doi.org/10.1016/j.wocn.2010.08.002

Inbar, M., Grossman, E., and Landau, A. N. (2020). Sequences of intonation units form a ~1 hz rhythm. *Scientific Reports*, 10(1), 1–9. https://doi.org/10.1038/s41598-020-72739-4

Josephs, J. J. (1967). *The physics of musical sound*. Princeton, NJ: D. Van Nostrand Co.

Kazanina, N., and Tavano, A. (2023). What neural oscillations can and cannot do for syntactic structure building. *Nature Reviews Neuroscience*, 24(2), 113–128.

Keitel, A., Gross, J., and Kayser, C. (2018). Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. *PLoS Biology*, 16(3), e2004473. https://doi.org/10.1371/journal.pbio.2004473

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, 59(5), 1208–1221. https://doi.org/10.1121/1.380986

Kohler, K. J. (2003). Domains of temporal control in speech and language: From utterance to segment. Proceedings of the 16th International Congress of Phonetic Sciences, Barcelona, Spain, pp. 7–10.

Kokal, I., Engel, A., Kirschner, S., and Keysers, C. (2011). Synchronized drumming enhances activity in the caudate and facilitates prosocial commitment: If the rhythm comes easily. *PLoS ONE*, 6(11), e27272. https://doi.org/10.1371/journal.pone.0027272

Kotz, S. A., Ravignani, A., and Fitch, W. T. (2018). The evolution of rhythm processing. *Trends in Cognitive Sciences*, 22(10), 896–910. https://doi.org/10.1016/j.tics.2018.08.002

Krause, B. (2012) *The great animal orchestra: Finding the origins of music in the world's wild places*. New York: Little, Brown and Co.

Large, E. W., and Jones, M. R. (1999). The dynamics of attending: How people track time-varying events. *Psychological Review*, 106(1), 119–159. https://doi.org/10.1037/0033-295x.106.1.119

Lehiste, I. (1977). Isochrony reconsidered. *Journal of Phonetics*, 5(3), 253–263. https://doi.org/10.1016/S0095-4470(19)31139-8

Leong, V., and Goswami, U. (2015). Acoustic-emergent phonology in the amplitude envelope of child-directed speech. *PLoS ONE*, 10(12), e0144411. https://doi.org/10.1371/journal.pone.0144411

https://doi.org/10.1017/9781009295888.017 Published online by Cambridge University Press

Levelt, W. J., and Cutler, A. (1983). Prosodic marking in speech repair. *Journal of Semantics*, 2(2), 205–218. https://doi.org/10.1093/semant/2.2.205

Levinson, S. C., and Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, 6, 731. https://doi.org/10.3389/fpsyg.2015.00731

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In W. J. Hardcastle and A. Marchal (Eds.), *Speech production and speech modelling* (pp. 403–439). Dordrecht: Kluwer Academic Publishers.

Lo, C. -W., Henke, L., Martorell, J., and Meyer, L. (2023). When linguistic dogma rejects a neuroscientific hypothesis. *Nature Reviews Neuroscience*, 24, 725. https://doi.org/10.1038/s41583-023-00738-1

Low, E. L. (1998). Prosodic prominence in Singapore English. Doctoral dissertation, University of Cambridge.

Lowit, A. (2014). Quantification of rhythm problems in disordered speech: A re-evaluation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1658), 20130404. https://doi.org/10.1098/rstb.2013.0404

MacNeilage, P. F. (1998). The frame/content theory of evolution of speech production. *Behavioral and Brain Sciences*, 21(4), 499–546. https://doi.org/10.1017/S0140525X98001265

Merker, B. H., Madison, G. S., and Eckerdal, P. (2009). On the role and origin of isochrony in human rhythmic entrainment. *Cortex*, 45(1), 4–17. https://doi.org/10.1016/j.cortex.2008.06.011

Meyer, L. (2018). The neural oscillations of speech processing and language comprehension: State of the art and emerging mechanisms. *European Journal of Neuroscience*, 48(7), 2609–2621. https://doi.org/10.1111/ejn.1374

Meyer, L., Sun, Y., and Martin, A. E. (2019). Synchronous, but not entrained: Exogenous and endogenous cortical rhythms of speech and language processing. *Language, Cognition and Neuroscience*, 35(9), 1089–1099. https://doi.org/10.1080/23273798.2019.1693050

Meyer, L., Sun, Y., and Martin, A. E. (2020). "Entraining" to speech, generating language? *Language, Cognition and Neuroscience*, 35(9), 1138–1148. https://doi.org/10.1080/23273798.2020.1827155

Miller, J. L., Grosjean, F., and Lomanto, C. (1984). Articulation rate and its variability in spontaneous speech: A reanalysis and some implications. *Phonetica*, 41(4), 215–225. https://doi.org/10.1159/000261728

Morgan, E., Fogel, A., Nair, A., and Patel, A. D. (2019). Statistical learning and gestalt-like principles predict melodic expectations. *Cognition*, 189, 23–34. https://doi.org/10.1016/j.cognition.2018.12.015

Morton, J., Marcus, S., and Frankish, C. (1976). Perceptual centers (p-centers). *Psychological Review*, 83(5), 405–408. https://doi.org/10.1037/0033-295X.83.5.405

Myers, B. R., Lense, M. D., and Gordon, R. L. (2019). Pushing the envelope: Developments in neural entrainment to speech and the biological underpinnings of prosody perception. *Brain Sciences*, 9(3), 70. https://doi.org/10.3390/brainsci9030070

Nolan, F., and Asu, E. L. (2009). The pairwise variability index and coexisting rhythms in language. *Phonetica*, 66(1–2), 64–77. https://doi.org/10.1159/000208931

Nolan, F., and Jeon, H.-S. (2014). Speech rhythm: A metaphor? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1658), 20130396. https://doi.org/10.1098/rstb.2013.0396

Nozaradan, S., Peretz, I., and Mouraux, A. (2012). Selective neuronal entrainment to the beat and meter embedded in a musical rhythm. *Journal of Neuroscience*, 32(49), 17572–17581. https://doi.org/10.1523/JNEUROSCI.3203-12.2012

Ogden, R., and Hawkins, S. (2015). Entrainment as a basis for co-ordinated actions in speech. Proceedings of the 18th International Congress of the Phonetic Sciences, York, UK.

Paschen, L., Fuchs, S., and Seifart, F. (2022). Final lengthening and vowel length in 25 languages. *Journal of Phonetics*, 94, 101179. https://doi.org/10.1016/j.wocn.2022.101179

https://doi.org/10.1017/9781009295888.017 Published online by Cambridge University Press

Paton, J. J., and Buonomano, D. V. (2018). The neural basis of timing: Distributed mechanisms for diverse functions. *Neuron*, 98(4), 687–705. https://doi.org/10.1016/j.neuron.2018.03.045

Peelle, J. E., and Davis, M. H. (2012). Neural oscillations carry speech rhythm through to comprehension. *Frontiers in Psychology*, 3, 320. https://doi.org/10.3389/fpsyg.2012.00320

Pfitzinger, H. R. (2001). Phonetische Analyse der Sprechgeschwindigkeit. Doctoral dissertation, Universität München.

Phillips-Silver, J., Aktipis, C. A., and Gregory, G. (2010). The ecology of entrainment: Foundations of coordinated rhythmic movement. *Music Perception*, 28(1), 3–14. https://doi.org/10.1525/mp.2010.28.1.3

Pike, K. L. (1945). *The intonation of American English*. Ann Arbor, MI: University of Michigan Press.

Plug, L. (2016). Informativeness, timing and tempo in lexical self-repair. *Language and Speech*, 59(4), 516–543. https://doi.org/10.1177/0023830915618427

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: Cerebral lateralization as "asymmetric sampling in time." *Speech Communication*, 41(1), 245–255.

Poeppel, D., and Assaneo, M. F. (2020). Speech rhythms and their neural foundations. *Nature Reviews Neuroscience*, 21(6), 322–334. https://doi.org/10.1038/s41583-020-0304-4

Price, P. J., Ostendorf, M., Shattuck-Hufnagel, S., and Fong, C. (1991). The use of prosody in syntactic disambiguation. *Journal of the Acoustical Society of America*, 90(6), 2956–2970. https://doi.org/10.1121/1.401770

Ramus, F., Nespor, M., and Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, 73(3), 265–292. https://doi.org/10.1016/S0010-0277(00)00101-3

Rathcke, T. (2013). On the neutralizing status of truncation in intonation: A perception study of boundary tones in German and Russian. *Journal of Phonetics*, 41(3–4), 172–185. https://doi.org/10.1016/j.wocn.2013.01.003

Rathcke, T., Falk, S., and Dalla Bella, S. (2021a). Music to your ears: Sentence sonority and listener background modulate the "speech-to-song illusion." *Music Perception*, 38(5), 499–508. https://doi.org/10.1525/mp.2021.38.5.499

Rathcke, T., Lin, C. -Y., Falk, S., and Bella, S. D. (2021b). Tapping into linguistic rhythm. *Laboratory Phonology*, 12(1), 1–32. https://doi.org/10.5334/labphon.248

Ravignani, A., Bowling, D. L., and Fitch, W. T. (2014). Chorusing, synchrony, and the evolutionary functions of rhythm. *Frontiers in Psychology*, 5, 1118. https://doi.org/10.3389/fpsyg.2014.01118

Räsänen, O., Doyle, G., and Frank, M. C. (2018). Pre-linguistic segmentation of speech into syllable-like units. *Cognition*, 171, 130–150. https://doi.org/10.1016/j.cognition.2017.11.003

Reinisch, E., Jesse, A., and McQueen, J. M. (2011). Speaking rate from proximal and distal contexts is used during word segmentation. *Journal of Experimental Psychology: Human Perception and Performance*, 37(3), 978–996. https://doi.org/10.1037/a0021923

Repp, B. H. (2005). Sensorimotor synchronization: A review of the tapping literature. *Psychonomic Bulletin & Review*, 12(6), 969–992.

Rimmele, J. M., Poeppel, D., and Ghitza, O. (2021). Acoustically driven cortical delta oscillations underpin perceptual chunking. *eNeuro*, 8(4), 1–15. https://doi.org/10.1523/ENEURO.0562-20.2021

Rimmele, J. M., Morillon, B., Poeppel, D., and Arnal, L. H. (2018). Proactive sensing of periodic and aperiodic auditory patterns. *Trends in Cognitive Sciences*, 22(10), 870–882. https://doi.org/10.1016/j.tics.2018.08.003

Roberts, S. G., Torreira, F., and Levinson, S. C. (2015). The effects of processing and sequence organization on the timing of turn taking: A corpus study. *Frontiers in Psychology*, 6, 509. https://doi.org/10.3389/fpsyg.2015.00509

Rosen, S. (1992). Temporal information in speech: Acoustic, auditory and linguistic aspects. *Philosophical Transactions: Biological Sciences*, 336(1278), 367–373.

https://doi.org/10.1017/9781009295888.017 Published online by Cambridge University Press

Sacks, H., Schegloff, E. A., and Jefferson, G. (1978). A simplest systematics for the organization of turn taking for conversation. In J. Schenkein (Ed.), *Studies in the organization of conversational interaction* (pp. 696–735). New York: Academic Press. https://doi.org/10.1016/B978-0-12-623550-0.50008-2

Savage, P. E., Loui, P., Tarr, B., et al. (2021). Music as a coevolved system for social bonding. *Behavioral and Brain Sciences*, 44(e59), 1–22. https://doi.org/10.1017/S0140525X20000333

Schegloff, E. A., Jefferson, G., and Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language*, 53(2), 361–382. https://doi.org/10.1353/lan.1977.0041

Schubö, F., Zerbian, S., Hanne, S., and Wartenburger, I. (Eds.) (2023). *Prosodic boundary phenomena*. Studies in Laboratory Phonology: Vol. 12. Berlin: Language Science Press.

Searle, J. R. (1990). Cognitive science and the computer metaphor. In K. S. Gill, B. Göranzon, and M. Florin (Eds.), *Artificial intelligence, culture and language: On education and work* (pp. 23–34). London: Springer.

Spivey, M. (2007). *The continuity of mind*. New York: Oxford University Press.

Stockhausen, K. (1959). How time passes by. *Die Reihe*, 3, 10–40.

Strauß, A., and Schwartz, J.-L. (2017). The syllable in the light of motor skills and neural oscillations. *Language, Cognition and Neuroscience*, 32(5), 562–569. https://doi.org/10.1080/23273798.2016.1253852

Tarr, B., Launay, J., and Dunbar, R. I. (2016). Silent disco: Dancing in synchrony leads to elevated pain thresholds and social closeness. *Evolution and Human Behavior*, 37(5), 343–349. https://doi.org/10.1016/j.evolhumbehav.2016.02.004

Temperley, D. (2008). A probabilistic model of melody perception. *Cognitive Science*, 32(2), 418–444. https://doi.org/10.1080/03640210701864089

Teoh, E. S., Cappelloni, M. S., and Lalor, E. C. (2019). Prosodic pitch processing is represented in delta-band EEG and is dissociable from the cortical tracking of other acoustic and phonetic features. *European Journal of Neuroscience*, 50(11), 3831–3842. https://doi.org/10.1111/ejn.14510

Tilsen, S., and Tiede, M. (2023). Parameters of unit-based measures of speech rate. *Speech Communication*, 150, 73–97. https://doi.org/10.1016/j.specom.2023.05.006

Treisman, M. (1963). Temporal discrimination and the indifference interval: Implications for a model of the "internal clock." *Psychological Monographs: General and Applied*, 77(13), 1–31. https://doi.org/10.1037/h0093864

Trouvain, J., and Grice, M. (1999). The effect of tempo on prosodic structure. Proceedings of the 14th International Congress of Phonetic Sciences, San Francisco, USA, pp. 1067–1070.

Turk, A. E., and Shattuck-Hufnagel, S. (2007). Multiple targets of phrase-final lengthening in American English words. *Journal of Phonetics*, 35(4), 445–472. https://doi.org/10.1016/j.wocn.2006.12.001

Turk, A. E., and Shattuck-Hufnagel, S. (2013). What is speech rhythm? A commentary on Arvaniti and Rodriquez, Krivokapić, and Goswami and Leong. *Laboratory Phonology*, 4(1), 93–118. https://doi.org/10.1515/lp-2013-0005

Turk, A. E., and White, L. (1999). Structural influences on accentual lengthening in English. *Journal of Phonetics*, 27(2), 171–206.

Ulbrich, P., Churan, J., Fink, M., and Wittmann, M. (2007). Temporal reproduction: Further evidence for two processes. *Acta Psychologica*, 125(1), 51–65. https://doi.org/10.1016/j.actpsy.2006.06.004

Ward, W. D. (1954). Subjective musical pitch. *Journal of the Acoustical Society of America*, 26(3), 369–380. https://doi.org/10.1121/1.1907344

Warren, R. M. (1982). *Auditory perception: A new synthesis*. New York: Pergamon Press.

White, L. (2014). Communicative function and prosodic form in speech timing. *Speech Communication*, 63–64, 38–54. https://doi.org/10.1016/j.specom.2014.04.003

White, L. S. (2002). English speech timing: A domain and locus approach. Doctoral dissertation, University of Edinburgh.

https://doi.org/10.1017/9781009295888.017 Published online by Cambridge University Press

Wilson, M., and Wilson, T. P. (2005). An oscillator model of the timing of turn-taking. *Psychonomic Bulletin & Review*, 12, 957–968.

Wiltermuth, S. S., and Heath, C. (2009). Synchrony and cooperation. *Psychological Science*, 20(1), 1–5. https://doi.org/10.1111/j.1467-9280.2008.02253.x

Wittmann, M. (2011). Moments in time. *Frontiers in Integrative Neuroscience*, 5, 66. https://doi.org/10.3389/fnint.2011.00066

Wittmann, M. (2013). The inner sense of time: How the brain creates a representation of duration. *Nature Reviews Neuroscience*, 14(3), 217–223. https://doi.org/10.1038/nrn3452

Zatorre, R. J., Belin, P., and Penhune, V. B. (2002). Structure and function of auditory cortex: Music and speech. *Trends in Cognitive Sciences*, 6(1), 37–46. https://doi.org/10.1016/S1364-6613(00)01816-7

https://doi.org/10.1017/9781009295888.017 Published online by Cambridge University Press

https://doi.org/10.1017/9781009295888.017 Published online by Cambridge University Press