White, L., & Malisz, Z. (2020). Speech rhythm and timing.
In C. Gussenhoven & A. Chen (Eds.), *Oxford Handbook of Language Prosody* (pp. 167-182). Oxford: Oxford University Press.

Speech Rhythm and Timing

Laurence White[1] & Zofia Malisz[2]

[1]Newcastle University; [2]KTH Royal Institute of Technology, Stockholm

Chapter 11, Oxford Handbook of Language Prosody
(editors: Carlos Gussenhoven & Aoju Chen)

Abstract

Speech events do not typically exhibit the temporal regularity conspicuous in many musical rhythms. In the absence of such surface periodicity, hierarchical approaches to speech timing propose that nested prosodic domains, such as syllables and stress-delimited feet, can be modelled as coupled oscillators and that surface timing patterns reflect variation in the relative weights of oscillators. Localized approaches argue, by contrast, that speech timing is largely organized bottom-up, based on segmental identity and subsyllabic organization, with prosodic lengthening effects locally associated with domain heads and edges. We weigh the claims of the two speech timing approaches against empirical data. We also review attempts to develop quantitative indices ('rhythm metrics') of cross-linguistic variations in surface timing, in particular in the degree of contrast between stronger and weaker syllables. We further reflect on the shortcomings of categorical 'rhythm class' typologies in the face of cross-linguistic evidence from speech production and speech perception.

White, L., & Malisz, Z. (2020). Speech rhythm and timing.
In C. Gussenhoven & A. Chen (Eds.), *Oxford Handbook of Language Prosody* (pp. 167-182). Oxford: Oxford University Press.

11.1 Introduction

Rhythm is a temporal phenomenon, but how far speech rhythm and speech timing are commensurable is a perennial debate. Distinct prosodic perspectives echo two Ancient Greek conceptions of time. First, *chronos* (χρόνος) signified time's linear flow, measured in seconds, days, years, and so on. Temporal linearity implicitly informs much prosody research, wherein phonetic events are interpreted with respect to external clocks and surface timing patterns are expressible through quantitative measures such as milliseconds. Second and by contrast, *kairos* (καιρός) was a more subjective notion of time as providing occasions for action, situating events in the context of their prompting circumstances. *Kairos* was invoked in Greek rhetoric: what is spoken must be appropriate to the particular moment and audience. Rhythmic approaches to speech that might be broadly classified as 'dynamical' reflect – to varying degrees – this view of timing as emerging from the intrinsic affordances occasioned by spoken interaction.

Interpretation of observable timing patterns is complicated by the fact that vowel and consonant durations are only approximate indicators of the temporal coordination of articulatory gestures, although there is evidence that speakers do manipulate local surface durations for communicative goals (e.g., signalling boundaries and phonological length; reviewed by Turk and Shattuck-Hufnagel 2014). Furthermore, perception of speech's temporal flow is not wholly linear. For example, Morton, Marcus, and Frankish (1976) found that a syllable's perceived moment of occurrence ('p-centre') is affected by the nature of its sub-constituents. Moreover, variation in speech rate can affect the perception of a syllable's presence or absence (Dilley and Pitt 2010) and the placement of prosodic boundaries (Reinisch, Jesse, and McQueen 2011). Thus, surface timing patterns may have non-linear relationships to both underlying control structures and to listeners' perceptions of prominence and grouping.

More generally, the term 'speech rhythm', without qualification, can cause potentially serious misunderstandings because: ''rhythm' carries with it implicit assumptions about the way speech works, and about how (if at all) it involves periodicity' (Turk and Shattuck-Hufnagel 2013, p. 93). Various definitions of rhythm applied to speech, and the timing thereof, are considered by Turk and Shattuck-Hufnagel: periodicity (surface, underlying, perceptual), phonological/metrical structure, and surface timing patterns. In this chapter, we do not attempt a single definition of

speech rhythm, but review some of these diverse perspectives and consider whether it is appropriate to characterise the speech signal as rhythmical (for other definitions, see e.g., Allen 1975; Cummins and Port 1998; Gibbon 2006; Wagner 2010; Nolan and Jeon 2014).

With such caveats in mind, the remainder of this section reviews four aspects of speech that may influence perceptions of rhythmicity: periodicity, alternation between strong and weak elements, hierarchical coordination of timing, and articulation rate. Section 11.2 discusses attempts to derive quantitative indices of rhythm typology. Section 11.3 contrasts two approaches to speech timing, one based on linguistic structure and localized lengthening effects and the other on hierarchically-coupled metrical units, and Section 11.4 considers the prospects for a synthesis of such approaches. We do not attempt a definitive summary of empirical work on speech rhythm and timing (for reviews, see, e.g., Klatt 1976; Arvaniti 2009; Fletcher 2010; White 2014), but aim to highlight some key theoretical concepts and debates informing such research.

11.1.1 Periodicity in surface timing

Before technology made large-scale analyses of acoustic data tractable, descriptions of speech timing were often impressionistic, with terminology arrogated from traditional poetics. In particular, the assumption that metrical structure imposes global timing constraints has a long history (Steele 1779). A specific timing constraint that proved pervasively influential was 'isochrony', the periodic recurrence of equally-timed metrical units such as syllables or stress-delimited feet. Classe (1939), while maintaining that isochrony is an underlying principle of English speech, concluded from his data that 'normal speech [is] on the whole, rather irregular and arrhythmic' (p. 89), due to variation in the syllable number and phonetic composition of stress-delimited phrasal groups, as well as to grammatical structure. Pike (1945) contrasted typical 'stress-timed' English rhythm and 'syllable-timed' Spanish rhythms, while asserting that stylistic variation could produce 'syllable-timed' rhythm in English. Abercrombie (1967) formalized 'rhythm class' typology, asserting that all languages were either syllable-timed (e.g., French, Telugu, Yoruba) or stress-timed (e.g., Arabic, English, Russian).

Isochronous mora timing has been claimed for Japanese, among other languages (Ladefoged 1975). The mora is a subsyllabic constituent (e.g., consonant plus short vowel), with somewhat

language-specific definitions, and is important in Japanese poetics (e.g., haiku comprise 17 morae), where syllables with long vowels or consonantal rhymes constitute two morae. Apparently by extension from poetry (*cf* syllables in French and Spanish, stress feet in English and German), spoken Japanese morae were assumed to be isochronous (e.g., Bloch 1950). Some data suggested approximate mora-timing but with deviations due to the mora's internal structure (Han 1962) and utterance position (longer morae phrase-finally, Kaiki and Sagisaka 1992). Warner and Arai's (2001) review concluded that Japanese mora duration is not isochronous, and that relatively regular mora-timing – when observed – is due to contingent features such as syllable phonotactics.

The 'rhythm class' concept persisted despite much evidence (e.g., Bertinetto 1989; Eriksson 1991) demonstrating the lack of isochrony of syllables or stress-delimited feet in surface speech timing. In a proleptic challenge to the syllable-timing hypothesis, Gili Gaya (1940; cited in Pointon 1980) observed that Spanish syllable duration is strongly affected by structural complexity, stress and utterance position. Pointon (1980), reviewing Spanish timing studies, concluded that syllable duration is determined bottom-up – what he called an 'antirhythmic' or 'segment-timed' pattern – and found further support in a study of six Spanish speakers (Pointon 1995; see also Hoequist 1983, contra Spanish syllable-timing).

Roach (1982) found similar correlations between interstress interval duration and syllable counts in Abercrombie's (1967) 'stress-timed' and 'syllable-timed' languages, with variance measures of syllable and interstress interval duration failing to support the categorical typology. Although the elementary design and single speaker per language limits interpretation of Roach's study, it proved influential for the use of variance measures of interval duration, later adopted in 'rhythm metrics', and for challenging the rhythm class hypothesis.

11.1.2 Contrastive rhythm

Brown (1911) distinguished 'temporal rhythm' – the regular recurrence of structural elements (here termed 'periodicity') – from 'accentual rhythm', the relative prominence of certain structural elements (for similar distinctions, see, *inter alia*: Allen 1975; Nolan and Jeon 2014; White 2014). As discussed above, speech usually lacks periodicity in surface timing, but many

languages contrast stronger and weaker elements through lexical stress (relative within-word syllable prominence) and phrasal accent (relative within-phrase word prominence, also called sentence stress). Here we use the term 'contrastive rhythm' rather than 'accentual rhythm' (to avoid confusion with the nature of the contrast: lexical stress or phrasal accent).

Dauer (1983), in an influential discussion, elaborated upon Roach's (1982) suggestion that cross-linguistic rhythmic differences may inhere in structural regularities such as vowel reduction and syllable complexity, and their relation with syllable stress. In particular, Dauer observed that the phonetic realization of stressed syllables and their (lexically/syntactically determined) distribution conspire to make (for example) English and Spanish seem rhythmically distinct. Most Spanish syllables have consonant-vowel (CV) structure, whereas the predominant English syllable structure is CVC and up to three onset consonants and four coda consonants are permissible. Moreover, the association between lexical stress and syllable weight (related to coda cluster complexity) is stronger for English, and also for Arabic and Thai, than Spanish. Additionally, unstressed syllables undergo minimal vowel reduction in Spanish, but most English unstressed syllables contain a reduced vowel, predominantly schwa (Dauer noted unstressed vowel reduction also for Swedish and Russian). All these patterns converge towards a high durational contrast between English strong and weak syllables. Furthermore, English stressed syllables tend to recur with relative regularity, particularly given the existence of secondary lexical stress, while long unstressed syllable sequences are more likely in Greek, Italian and Spanish (Dauer 1983). Structural trends do not converge onto a high-low contrast gradient for all languages, however: for example, Polish has high syllable complexity, but limited vowel reduction, while Catalan has low syllable complexity, but significant vowel reduction (Nespor 1990).

In part due to the language's recent status as a scientific *lingua franca*, analytical concepts derived from English linguistics have sometimes guided the characterization of other languages. Thus, much early comparative field linguistics had a guiding assumption that 'stress' was universally meaningful. In fact, English – particularly standard southern British English (SSBE) – seems a conspicuously 'high-contrast' language in terms of lexical stress and also phrasal accent. Comparisons of global timing properties between selected languages often show English

with the highest variation in vowel duration (e.g., Ramus, Nespor, and Mehler 1999; White and Mattys 2007a). In Nolan and Asu (2009)'s terminology, English has a markedly steep 'prominence gradient'. Even other Germanic languages, such as Dutch, have sparser occurrence of reduced vowels in unstressed syllables (Cutler and Van Donselaar 2001). However, stress is – manifestly – linguistically important in many languages lacking such marked stress cues as English: thus, Dauer (1983) observed that while stress-related duration contrasts are substantially greater in English than Spanish, combinations of cues make stressed syllables in Spanish, Greek or Italian salient to native listeners (in contrast with French, which lacks lexical-stress). Indeed, Cumming (2011) suggested that languages may appear less rhythmically distinct once prosodic perceptual integration is taken into account (see also Arvaniti 2009).

On the other hand, it is also becoming clear that many languages may lack metrically-contrasting elements (e.g., Korean: Jun 2005; Ambonese Malay: Maskikit-Essed and Gussenhoven 2016; see Nolan and Jeon 2014, for references questioning the status of stress in several languages). Tabain, Fletcher and Butcher (2014) suggested the term 'stress ghosting' to highlight how Germanic language speakers' native intuitions may induce stress perception in languages unfamiliar to them. Stress ghosting arises due to misinterpretation of phonetic or structural patterns that would be associated with prominence in languages – like Dutch, English or German – with unambiguous lexical stress contrast (e.g., English ˈ*insight* vs *in*ˈ*cite*). By contrast, native speakers of languages without variable stress placement as a cue to lexical identity have been characterized as having 'stress deafness' (Dupoux, Peperkamp, and Sebastián-Gallés 2001). Specifically, speakers of languages that either lack lexical stress (e.g., French) or have non-contrastive stress (e.g., Finnish or Hungarian fixed word-initial stress) do not appear to retain stress patterns of nonwords in short-term memory, suggesting that their phonological representations do not include stress (Peperkamp and Dupoux 2002; see also Rahmani et al. 2015). Thus, the notion of contrastive rhythm, while pertinent for some widely-studied linguistic systems, may be inapplicable for many languages (Nolan and Jeon 2014).

11.1.3 Hierarchical timing

Unlike typologies based on isochrony of morae, syllables or stressed syllables (reviewed above), hierarchical timing approaches do not identify a single privileged unit strictly governing any

language's surface timing. They describe relative timing dependencies between at least two hierarchically-nested constituents, for example, the syllable and the stress-delimited foot (e.g., O'Dell and Nieminen 1999). The syllable (or syllable-sized unit, e.g. vowel-to-vowel interval) is regarded as a basic cyclic event in speech perception/production (Fowler 1983) and the smallest beat-induction speech unit (Morton et al. 1976). With regard to the stress-delimited foot, various definitions are proposed, sometimes related to the metrical structure of particular languages, with a key distinction being whether or not the foot respects word boundaries (e.g., Eriksson 1991; Beckman 1992; Bouzon and Hirst 2004).

Analysis of timing relationships between hierarchically-nested constituents were developed from Dauer's (1983) findings that, in various languages, stress foot duration is neither independent of syllable number (the expectation based on foot isochrony) nor an additive function of syllable number (the expectation based on syllable isochrony). Eriksson (1991) further explored Dauer's data on the positive relationship between total foot duration and syllable number. The durational effect of adding a syllable to the foot (i.e., the slope of the regression line) was similar for all five of Dauer (1983)'s languages. However, the intercept differed between putative 'rhythm classes' ('syllable-timed' Greek, Italian, Spanish: ~100 ms; 'stress-timed' English, Thai: ~200 ms). Eriksson claimed that the natural interpretation of the intercept variation was that the durational difference between stressed and unstressed syllables is greater in English and Thai than in Greek, Italian or Spanish. However, as Eriksson observed (also O'Dell and Nieminen 1999), the positive intercept does not itself indicate where the durational variation takes place. Eriksson further noted an inverse relationship between the number of syllables in the foot and the average duration of those syllables. Similarly, Bouzon and Hirst (2004) found sub-additive relationships between several levels of structure in British English: syllables in a foot; phones in a syllable; feet in an intonational unit.

These linear relationship between foot duration and the number of sub-constituents, with positive slope and intercept coefficients of the linear function, can – as described in more detail in Section 11.3.2 – be modelled as systems of coupled oscillators (e.g., O'Dell and Nieminen 1999, at the syllable level and foot level). Other approaches that relate surface timing to the coupled interaction of hierarchically-nested constituents include work on the coordination of articulatory

gestures within syllables and prosodic phrases (e.g., Byrd and Choi 2010) and on the coordination of articulatory gestures within syllables and feet (Tilsen 2009).

11.1.4 Articulation rate

Cross-linguistic variations in predominant syllable structures (Dauer 1983) are associated with systematic differences in 'articulation rate', defined as syllables per second excluding pauses (given that pause frequency and duration significantly affect overall speech rate, Goldman-Eisler 1956). Estimated rates vary between studies due to the spoken materials, the accents chosen for each language and speaker idiosyncrasies. Stylistic and idiosyncratic effects notwithstanding, languages with predominantly simple syllable structures, such as Spanish, tend to be spoken at a higher syllables-per-second rate than languages, such as English, with more complex syllable structures (White and Mattys 2007a; Pellegrino, Coupé, and Marsico 2011). Of course, such differences in syllable rates do not imply that Spanish speakers articulate more quickly than English speakers, rather that more syllables are produced per unit of time when those syllables contain fewer segments. Additionally, Pellegrino et al. (2011) pointed to an effect of information density on rate: for example, Mandarin Chinese has lower syllable-per-second rates than Spanish, but more informationally-rich syllables when taking lexical tone into account, hence their information density is roughly similar.

Listeners' linguistic experience may strongly affect rate judgements, particularly with unfamiliar languages. Thus, where Japanese and German utterances were assessed by native speakers of both languages, there was overestimation of the unfamiliar language's rate compared to the first language (Pfitzinger and Tamashima 2006). This has been described as the 'gabbling foreigner illusion' (Cutler 2012): when confronted with speech that we cannot understand, we tend to perceive it as spoken faster (see also Bosker and Reinisch 2017, regarding effects of second language proficiency). This illusion may, in part, be due to difficulties segmenting individual words in unfamiliar languages (Snijders, Kooijman, Cutler, and Hagoort 2007). Conversely, when judging non-native accents, listeners generally interpret faster speech rate as evidence of more native-like production (e.g., White and Mattys 2007b; see Hayes-Harb 2014, for a review of rate influences on accentedness judgements). Moreover, the perception of cross-linguistic 'rhythm' contrasts is influenced by structurally-based rate differences (Dellwo 2010). For

example, when hearing delexicalized Spanish and English *sasasa* stimuli (all vowels replaced by /a/, all consonants by /s/, but with the original segment durations preserved), English speakers were more likely to correctly classify faster Spanish but slower English utterances (White, Mattys, and Wiget 2012; Polyanskaya, Ordin, and Busa 2016). Thus, some perceptions of linguistic differences typically described as 'rhythmic' may be associated with systematic variations in rate (Dellwo 2010).

11.2 'Rhythm metrics' and prosodic typology

Informed, in particular, by Dauer's (1983) re-evaluation of rhythmic typology, various studies under the 'rhythm metrics' umbrella have attempted to empirically capture cross-linguistic differences in 'rhythm' (often loosely defined, see below and Turk and Shattuck-Hufnagel 2013). These studies employed diverse metrics of durational variation (*cf* Roach 1982), notably in vocalic and consonantal intervals. Some studies were premised on the validity of 'rhythm class' distinctions (e.g., Ramus, Nespor, and Mehler 1999), raising a potential circularity problem where the primary test of a metric's worth is whether it evidences the hypothesized class distinctions (Arvaniti 2009), although studies of perceptual discrimination between languages (e.g., Nazzi and Ramus 2003) were sometimes cited as external corroboration. However, the accumulated evidence from speech production and perception – reviewed below – strongly questions the validity and usefulness of categorical rhythmic distinctions.

Some evaluative studies have highlighted empirical strengths and limitations of different rhythm metrics, observing that while certain metrics might provide data about cross-linguistic variation in the durational marking of stress contrast, they neglect much else that might be relevant to 'rhythm', notably distributional information (White and Mattys 2007a; Wiget et al. 2010). More trenchantly, other researchers have argued that the 'rhythm metrics' enterprise was compromised by a lack of consistency regarding which languages were distinguished (Loukina, Kochanski, Rosner, Keane, and Shih 2011), for example, when comparing read and spontaneous speech (Arvaniti 2012). Indeed, the term 'rhythm metrics' is a misnomer: aggregating surface timing features does not capture the essence of 'speech rhythm', however defined (e.g., Cummins 2002; Arvaniti 2009). We next consider some lessons from the 'rhythm metrics' approach.

11.2.1 Acoustically-based metrics of speech rhythm: lessons and limitations

In the development of so-called 'rhythm metrics' for typological studies, there was a threefold rationale for quantifying durational variation based on vowels and consonants, rather than syllables or stress feet. First, languages such as Spanish typically have less vowel reduction and less complex consonant clusters than, for example, English (Dauer 1983). Second, Mehler, Dupoux, Nazzi and Dehaene-Lambertz (1996), assuming early sensitivity to vowel/consonant contrasts, proposed that young infants use variation in vowel duration and intensity to determine their native language 'rhythm class.' Third, syllabification rules vary cross-linguistically and are not uncontroversial even within languages, while applying heuristics to identify vowel/consonant boundaries is (comparatively) straightforward (Low, Grabe, and Nolan 2000).

Thus, Ramus et al. (1999) proposed the standard deviation of vocalic and consonantal interval duration ('ΔV' and 'ΔC' respectively), along with the percentage of utterance duration that is vocalic rather than consonantal (%V). They found that a combination of ΔC and %V statistically reflected their – *predefined* – rhythm classification of, in increasing %V order: Dutch/English/Polish *vs* Catalan/French/Italian/Spanish *vs* Japanese.

Seeking to capture syntagmatic contrast within an utterance as well as global variation, pairwise variability indices average the durational differences between successive intervals – primarily, vocalic/consonantal – over an utterance (see Nolan and Asu's 2009, account of this development). PVI-based measures showed differences between a Singaporean and a British dialect of English that had been claimed to be rhythmically distinct (Low et al. 2000), and gradient variation between languages previously categorized as 'stress-timed' or 'syllable-timed' (Grabe and Low 2002, based on one speaker per language). While PVIs were intended to capture sequential durational variation more directly than global measures, Gibbon (2006) noted that PVIs do not necessarily discriminate between alternating vs geometrically increasing sequences (although the latter are implausible in speech, Nolan and Jeon 2014).

Variance-based measures of interval duration tend to show high correlation with speech rate: as overall intervals lengthen with slower rate, so – other things being equal – do standard deviations

(Barry, Andreeva, Russo, Dimitrova, and Kostadinova 2003; Dellwo and Wagner 2003; White and Mattys 2007a). With normalized PVI (nPVI)-based metrics, interval durations were normalized to take account of speech rate variation (Low et al. 2000). With standard deviation measures (ΔV, ΔC), speech rate normalization was implemented through coefficients of variation for consonantal intervals (VarcoC, Dellwo and Wagner 2003) and vocalic intervals (VarcoV, Ferragne and Pellegrino 2004). In the case of consonants, however, the coefficient of variation (VarcoC) lacked discriminative power (White and Mattys 2007a): as noted by Grabe and Low (2002), mean consonantal interval duration varies substantially due to language-specific phonotactics, so using the mean as a normalising denominator also eliminates linguistically-relevant variation.

Comparing the power of various metrics, White and Mattys (2007a) suggested that rate-normalized metrics of vowel duration (VarcoV, nPVI-V) are more effective in capturing cross-linguistic variation, alongside %V to represent differences in consonant cluster complexity. (For broadly similar conclusions about the relative efficacy of the normalized vocalic metrics, see Loukina et al. 2011; Prieto, Vanrell, Astruc, Payne, and Post 2012.) In contrast with Ramus et al. (1999), cross-linguistic studies employing such metrics often found variation in scores within hypothesized rhythm classes to be as great as those between classes (Grabe and Low 2002; White and Mattys 2007a; Arvaniti 2012). While conclusions about prosodic typology based only on rhythm metrics should be treated with circumspection, these data generally align with recent perceptual studies (White et al. 2012; Arvaniti and Rodriquez 2013; White, Delle Luche, and Floccia 2016) in refuting categorical notions of rhythm class.

Several studies emphasise the limitations of even the more reliable metrics for capturing language-specific durational characteristics, given their susceptibility to variation in utterance composition and idiosyncratic differences between speakers (e.g., Wiget et al. 2010; Loukina et al. 2011; Arvaniti 2012; Prieto, Vanrell, Astruc, Payne, and Post 2012;). Given that %V, for example, is designed to reflect variation in the preponderance of syllable structures between languages, it is unsurprising to find that sentences constructed to represent language-*atypical* structures elicit anomalous scores (Arvaniti 2012; Prieto et al. 2012). Moreover, the sensitivity of rhythm metrics to speaker-specific variation, a potential problem for typological studies, has

been exploited in forensic phonetics and speaker recognition (Leemann, Kolly, and Dellwo 2014; Dellwo, Leemann, and Kolly 2015) and in discriminating motor speech disorders (Liss et al. 2009).

It is clear, however, that large sample sizes and a variety of materials are needed to represent languages in typological studies, a major limitation given the laborious nature of manual measurement of segment duration (and the potential for unconscious language-specific biases in application of acoustic segmentation criteria, Loukina et al. 2011). While automated approaches have potential (Wiget et al. 2010), data-trained models for recognition and forced alignment may not be available for many languages; furthermore, Loukina et al. (2011) indicated drawbacks with forced alignment that they addressed using purely acoustic-based automated segmentation.

Also problematic for 'rhythm metrics' is that relationships between sampled languages vary according to elicitation methods (for comparison of read and spontaneous speech, see Barry et al. 2003; Arvaniti 2012) and that no small set of metrics, even the more reliable, consistently distinguishes all languages (Loukina et al. 2011). Furthermore, articulation rates should be also reported, as the more reliable metrics are rate-normalized (VarcoV and nPVI, although not %V), but perceptual evidence shows the importance for language discrimination of syllable-per-second rate differences (Dellwo 2010; White et al. 2012; Arvaniti and Rodriquez 2013).

At best, metrics such as VarcoV and %V are approximate indicators of broad phonetic and phonotactic patterns. Questions about cross-linguistic timing differences – for example, comparing the durational marking of prominences and boundaries – could often be better addressed by more direct methods (Turk and Shattuck-Hufnagel 2013). Moreover, duration-based metrics neglect other perceptually-important prosodic dimensions (Cumming 2011). From a theoretical perspective, the need to declare one's assumptions about the nature of speech rhythm is paramount (e.g., Wiget et al.'s 2010, specific use of the term '*contrastive* rhythm metrics'). Indeed, there is usually a more directly appropriate term for one's object of phonetic or phonological study than 'rhythm' (Turk and Shattuck-Hufnagel 2013).

11.2.2 The fall of the rhythm class hypothesis

Rhythm classes based on isochronous units – syllable-timed, stress-timed, mora-timed – have long been undermined by durational evidence, as discussed above. The multi-faceted nature of prominence provides further counter-arguments to the rhythm class hypothesis. Two languages characterized as 'syllable-timed' are illustrative. Spanish and French both have limited consonant clustering, minimal prominence-related vowel reduction and relatively transparent syllabification. As Pointon (1995) observed, however, French lacks lexical stress and has phrase-final prominence, while Spanish has predominantly word-penultimate stress but lexically-contrastive exceptions, minimally distinguishing many word pairs (e.g., *tomo* 'I take' vs *tomó* 'she took').

Despite such 'within-class' structural distinctions, some studies have suggested that initial speech processing depends upon speakers' native 'rhythm class.' Thus, French speakers were quicker to spot targets corresponding to exactly one syllable of a carrier word: for example, *ba* in *ba.lance*, *bal* in *bal.con* vs (slower) *bal* in *ba.lance*, *ba* in *bal.con* (Mehler, Dommergues, Frauenfelder, and Segui 1981). This 'syllable effect' was contrasted with metrical segmentation, wherein speakers of Germanic languages with predominant word-initial stress (e.g., Dutch and English) were held to infer word boundaries preceding stressed (full) syllables (Cutler and Norris 1988; although Mattys and Melhorn 2005, argued that stressed-syllable-based segmentation implies, additionally, a syllabic representation).

These different segmentation strategies were explicitly associated with 'rhythm class' (Cutler 1990), which Cutler and Otake (1994) extended to Japanese, the 'mora-timed' archetype. Furthermore, the importance of early childhood experience was emphasized, suggesting that infants detect their native 'rhythm class' to select a (lifelong) segmentation strategy (Cutler and Mehler 1993). It is questionable, however, whether Spanish and French speakers would share rhythmical segmentation strategies, given differences in prominence distribution and function. Indeed, the 'syllable effect' subsequently appeared elusive in Spanish, Catalan and Italian, all with variable, lexically-contrastive stress placement (Sebastián-Gallés, Dupoux, Segui, and Mehler 1992; Tabossi, Collina, Mazzetti, and Zoppello 2000). Moreover, Zwitserlood, Schriefers, Lahiri and Van Donselaar (1993) found that speakers of ('stress-timed') Dutch

showed syllabic matching effects comparable to those Mehler et al. (1981) reported for French (for syllabic effects in native English speakers, see Bruck, Treiman, and Caravolas 1995; Mattys and Melhorn 2005). It appears that syllabic and metrical effects are heavily influenced by linguistic materials and task demands, rather than fixed by listeners' earliest linguistic experiences (for a review, see White 2018).

Some perceptual studies have shown that listeners can distinguish two languages from distinct 'rhythm classes', but not two languages within a class. For example, American English-learning five-month-olds distinguished Japanese utterances from – separately – British English and Italian utterances, but did not distinguish Italian and Spanish, or Dutch and German (Nazzi, Jusczyk, and Johnson 2000). Using monotone delexicalized *sasasa* speech preserving natural utterance timing (as described above), Ramus, Nespor and Mehler (2003) found between-class, but not within-class, discrimination by French adult listeners (but postulated a fourth 'rhythm class' to account for discrimination of Polish from – separately – Catalan, Spanish and English). However, subsequent similar studies found discrimination within 'rhythm classes': for five-month-olds hearing intact speech (White et al. 2016) and for adults with delexicalized speech (White et al. 2012; Arvaniti and Rodriquez 2013). Discrimination patterns can be explained by cross-linguistic similarity on salient prosodic dimensions, including speech rate and utterance-final lengthening, without requiring categorical distinctions (White et al. 2012).

In her influential paper 'Isochrony Reconsidered,' Lehiste (1977) argued that support for isochrony-based theories was primarily perceptual; indeed, data from perception studies have since been invoked to buttress the rhythm class hypothesis. It now seems clear, however, that responses to speech stimuli are not determined by listeners' native 'rhythm class' (segmentation studies) nor by categorical prosodic classes of language materials (discrimination studies). Languages clearly vary in their exploitation of temporal information to indicate speech structure, notably prominences and boundaries, but this variation is gradient and – integrating other prosodic features – multi-dimensional. There remain typological rhythm-based proposals, such as the 'control vs compensation hypothesis' (Bertinetto and Bertini 2008), but these assume *gradient* between-language variation in key parameters. The concept of *categorical* rhythm class seems superfluous, indeed misleading, for theories of speech production and perception.

11.3 Models of prosodic speech timing

Factors affecting speech duration patterns are diverse and not wholly predictable, including – beyond this linguistically-oriented survey's scope – word frequencies, emotional states and performance idiosyncrasies. At the segmental level, voicing and place/manner of articulation influence consonant duration, while high vowels tend to be shorter than low vowels (for reviews see Klatt 1976; van Santen 1992). Some languages signal consonant or vowel identity by length distinctions, sometimes with concomitant quality contrasts (for a review see Ladefoged 1975). Connected speech structure also has durational consequences: e.g., vowels are shorter preceding voiceless obstruents than voiced obstruents (Delattre 1962). This consonant-vowel duration trade-off ('pre-fortis clipping') is amplified phrase-finally (Klatt 1975, hinting at the importance of considering prosodic structure when interpreting durational data, e.g., White and Turk 2010).

Beyond segmental and intersegmental durational effects, an ongoing discussion concerns the nature of the higher-level structures that are important for describing speech timing, and the mechanisms through which these structures influence observed durational patterns. Here we review two of the many extant approaches to these problems (see also, *inter alia*, Byrd and Saltzman 2003; Aylett and Turk 2004; Barbosa 2007).

Section 11.3.1 considers approaches based on localized lengthening effects associated with linguistic constituents. Section 11.3.2 considers dynamical systems models based on hierarchical coupling of oscillators. For each, we briefly highlight key features and consider their accounts of some observed timing effects.

11.3.1 Localized approaches to prosodic timing

The fundamental claim of 'localized' approaches to prosodic timing is that no speech units impose temporal constraints on their sub-constituents throughout the utterance (van Santen 1997). Timing is primarily determined bottom-up, based on segmental identity (echoing Pointon's 1980 description of Spanish as 'segment-timed') and processes of accommodation and coarticulation between neighbouring segments. Higher-level structure influences timing via localized lengthening effects at linguistically important positions (White 2002, 2014).

The most well-attested lengthening effects are at prosodic domain edges and, for some languages, at prosodic heads (see Beckman 1992, regarding edge-effect universality vs head-effect language-specificity). Final ('pre-boundary') lengthening is widely observed at various levels of linguistic structure: (e.g., English: Oller 1973; Dutch: Gussenhoven and Rietveld 1992; Hebrew: Berkovits 1994; Czech: Dankovičová 1997; see Fletcher 2010 for an extensive review). Lengthening (and gestural strengthening) of word-initial consonants is also reported cross-linguistically (e.g., Oller 1973; Cho, McQueen, and Cox 2007), with greater lengthening after higher-level boundaries (e.g., Fougeron and Keating 1997). In many languages, lexically-stressed syllables are lengthened relative to unstressed syllables (e.g., Crystal and House 1988), although the magnitude of lengthening varies (e.g., Dauer 1983; Hoequist 1983) and, as discussed above, some languages may lack lexical stress (e.g., Jun, 2005; Maskikit-Essed and Gussenhoven, 2016). Additionally, stressed and other syllables are lengthened in phrasally-accented words (e.g., Sluijter and van Heuven 1995).

White's (2002, 2014) prosodic timing framework proposed that lengthening is the durational means by which speakers signal structure for listeners. The distribution of lengthening depends on the particular (edge or head) structural influence: for example, the syllable onset is the locus of word-initial lengthening (Oller 1973), while the pre-boundary syllable rhyme is lengthened phrase-finally (as well as syllable rhymes preceding a final *unstressed* syllable, Turk and Shattuck-Hufnagel 2007). Thus, the distribution ('locus') of lengthening disambiguates the nature of the structural cue (e.g., Monaghan, White, and Merkx 2013).

This emphasis on localized lengthening affords a reinterpretation of 'compensatory' timing processes, inverse relationships between constituent length and the duration of sub-constituents. For example, Lehiste (1972) reported 'polysyllabic shortening', an inverse relationship between a word's syllable count and its primary stressed syllable's duration.As observed by White and Turk (2010), however, many duration studies have only measured phrasally-accented words, such as in fixed frame sentences (e.g., 'Say WORD again'). The primary stressed syllables are lengthened in these phrasally-accented words, as – to a lesser extent – are unstressed syllables; moreover, the greater the number of unstressed syllables, the smaller the accentual lengthening on the primary stressed syllable (Turk and White 1999). Hence, pitch-accented words appear to

demonstrate polysyllabic shortening (e.g., *cap* is progressively shorter in *cap*, *captain*, *captaincy*; likewise *mend* in *mend*, *commend*, *recommend*); however, in the absence of pitch accent, there is no consistent relationship between word length and stressed syllable duration (White 2002; White and Turk 2010).

Similar arguments apply to apparent foot-level compression effects. Beckman (1992, p. 458) noted the difficulty in distinguishing 'rhythmic compression of the stressed syllable in a polysyllabic foot from the absence of a final lengthening for the prosodic word.' Likewise, Hirst (2009) considered the durational consequences of the length of the English 'narrow rhythm unit' (NRU) (or 'within-word foot', from a stressed syllable to a subsequent word boundary). He found the expected linear relationship between syllable number and NRU duration, not the negative acceleration expected for a cross-foot compression tendency (Nakatani, O'Connor, and Aston 1981; Beckman 1992). Furthermore, Hirst (2009) attributed the 'residual' extra duration within each NRU (the intercept of the regression line for NRU length vs duration) to localized lengthening effects at the beginning and end of the NRU (*cf* White 2002, 2014; White and Turk 2010). Similarly, Fant, Kruckenberg and Nord (1991, p. 84), considering Swedish, French and English, suggested that the primary (but 'marginal') durational consequence of foot-level structure was in 'the step from none to one following unstressed syllables in the foot.' This localized lengthening of the first of two successive stressed syllables (e.g., Fourakis and Monahan 1988; Rakerd, Sennett, and Fowler 1987, called 'stress-adjacent lengthening' by White 2002) may relate to accentual lengthening variation in cases of stress class.

Generalising from these observations, White (2002 2014) reinterpreted apparent compensatory timing as being due to variation in the distribution of localized prosodic lengthening effects at domain heads and domain edges (e.g., phrasal-accent lengthening or phrase-final lengthening). The localized lengthening framework further argues that, outside the loci of prosodic lengthening effects, there is little evidence for relationships between constituent length and sub-constituent duration (see also, e.g., Suomi 2009; Windmann, Šimko, and Wagner 2015). Beyond such localized lengthening, the primary determiner of a syllable's duration is its segmental composition (van Santen and Shih 2000).

11.3.2 Coupled oscillator approaches to prosodic timing

Coupled oscillator models posit at least two cyclical processes that are coupled, that is, influence each other's evolution in time. O'Dell and Nieminen's (1999, 2009) model of prosodic timing considers the hierarchically-coupled syllable and stress-delimited foot oscillators (but see Malisz, O'Dell, Nieminen, and Wagner 2016, regarding other units). Some models additionally include non-hierarchically coupled subsystems: Barbosa's (2007) complex model also contains a coupled syntax and articulation module, the syntactic component being controlled by a probabilistic model as well as a coupled prosody-segmental interaction, and generates abstract vowel-to-vowel durations tested on a corpus of Brazilian Portuguese. (For overviews of dynamical approaches to speech, including timing, see Van Lieshout 2004; Tilsen 2009).

Empirical support for coupled oscillator models on the surface timing level has been found in the linear relationship between the number of syllables in a foot and the foot's duration, discussed in Section 11.1.3, with non-zero coefficients. This relationship naturally emerges from O'Dell and Nieminen's (2009) mathematical modelling of foot and syllable oscillator coupling. Importantly, there is variable asymmetry in the coupling strengths of the two oscillators, between and within languages (see also Cummins 2002). If one process wholly dominated, isochrony of syllables or stress feet would be observed: in strict foot-level isochrony, foot duration would be independent of syllable count; in strict syllable-level isochrony, foot duration would be additively proportional to syllable count. That such invariance is rarely observed is, of course, not evidence against oscillator models. Surface regularity of temporal units is not a prerequisite; rather, it is the underlying cyclicity of coupled units that is assumed (for a discussion see Turk and Shattuck-Hufnagel 2013; Malisz et al. 2016). Indeed, periodic control mechanisms, if coupled, should not typically produce static surface isochrony on any subsystem level (e.g., syllable or foot): hierarchical coupling promotes variability in temporal units (Barbosa 2007; Malisz et al. 2016), and only under specific functional conditions is surface periodicity achieved.

Regression of unit duration against number of sub-constituents cannot, however, distinguish *where* local expansion or compression may take place. Indeed, coupled oscillator models are neutral about where durational effects are allocated within temporal domains (Malisz, O'Dell, Nieminen and Wagner 2016), ranging from extreme centralization to equal allocation throughout

the domain. By contrast, localized approaches, e.g., White (2014) suggest strict binding of lengthening effects to specific loci (e.g. syllable onset or rhyme) within the domain (e.g. a word) while other syllables are predicted to remain unaffected as they are added to the domain outside of this locus. Effects on surface timing predicted by the coupled oscillator model are thus less constrained than those of localized approaches, which specifically argue for the absence of compression effects beyond the locus (see above).

Dynamical models, such as O'Dell and Nieminen (2009), updated in Malisz et al., (2016), depend rather on evidence of hierarchical coupling, such as that provided by Cummins and Port (1998) in the specific case of speech-cycling tasks. While repeating a phrase to a uniformly-varying metronome target beat, English speakers tended to phase-lock stressed syllables to the simple ratios (1:3, 1:2, 2:3) of the repetition cycle. Furthermore, other periodicities emerge at harmonic fractions of the phrase repetition cycle (Port 2003), who relates these observations to periodic attentional mechanisms (Jones and Boltz 1989; see also McAuley and Fromboluti 2014).

There is also suggestive empirical support for metrical influences on speech production in findings that speakers may prefer words and word combinations that maintain language-specific metrical structures (Lee and Gibbons 2007; Schlüter 2009; Temperley 2009; Shih 2014) although Temperley (2009) found that contextually-driven variations from canonical form (e.g., stress-clash avoidance) actually increase interval irregularity in English.

In dynamical theories, coupling is evident within hierarchical speech structures, between speakers in dialogue and within language communities (Port and Leary 2005; Cummins 2009). Periodic behaviour is understood to be one of the mechanisms of coordination within complex systems (Turvey 1990), mathematically modelled by oscillators. Furthermore, coupled oscillatory activity behaviour is a control mechanism that *spontaneously* arises in complex systems where at least two subsystems interact, without necessarily requiring a periodic referent, such as a regular beat (Cummins 2011).

Whether the undoubted human ability to dynamically entrain our actions is mirrored in the entrainment of metrical speech units remains debatable, as discussed here. Evidence of the entrainment of endogenous neural oscillators (e.g., theta waves) to the amplitude envelope of speech (e.g., Peelle and Davis 2012) suggests a possible neural substrate for oscillator-based speech behaviour, potentially important in listeners' generation of durational predictions based on speech rate (e.g., Dilley and Pitt 2010). Theories of neural entrainment need, however, to address the lack of surface periodicity in most speech, as well as the imprecise mapping between the amplitude envelope and linguistic units (Cummins 2012). More generally, oscillator models of timing may find a challenge in evidence that many languages lack levels of prominence, such as lexical stress, that were once thought universal (e.g., Jun, 2005; Maskikit-Essed and Gussenhoven, 2016).

11.4 Conclusions and prospects

The hypothesis that speech is consistently characterized by isochrony succumbed to the weight of counterevidence, and the associated hypothesis about categorical 'rhythm class' has, at best, scant support. The accumulated production and perception data do, however, support a continuing diversity of approaches to speech timing, varying in their balance between *chronos* and *kairos*, notably the degree to which surface timing patterns or hierarchical control structures are emphasized.

Regarding the two approaches sketched here, there may appear superficial contrasts between dynamical timing models, emphasising underlying coupling between hierarchically-organized levels of metrical structure (e.g., Cummins and Port 1998; O'Dell and Nieminen 2009; Malisz et al. 2016), and localized approaches, emphasising the irregularity of surface timing and the information about structure and phonology provided for listeners by this temporal unpredictability (e.g., Cauldwell 2002; Nolan and Jeon 2014; White 2014). A synthesis of dynamical and localized models may, however, emerge from a deeper understanding of the complex interaction between the information transmission imperative in language and the affordance that speech offers for multi-level entrainment of interlocutors' gestural, prosodic, linguistic and social behaviour (Tilsen 2009; Pickering and Garrod 2013; Mücke, Grice, and Cho 2014; Fusaroli and Tylén 2016).

Some degree of broad predictability is a prerequisite for humans interacting in conversation or other joint action. More specifically, local unpredictability in speech timing cannot be interpreted as structurally or prosodically motivated unless listeners have a foundation on which to base temporal predictions and the ability to spot violations of predictions (e.g., Baese-Berk et al. 2014; Morrill, Dilley, McAuley, and Pitt 2014). Where mutual understanding confers predictability – for example, via a common social framework or foregoing linguistic context – then the surface timing of speech may be freer to vary unpredictably, towards maximising encoding of information. When interlocutors lack shared ground and predictability is consequently elusive, then relative underlying periodicity may dominate, supporting mutual coordination and ease of processing, but with potential loss of redundancy in information encoding (see Wagner, Malisz, Inden, and Wachsmuth 2013). This proposal, which we tentatively call the 'periodicity modulation hypothesis', lends itself to ecologically-embedded studies of infant and adult spoken interactions and their relationship to neurophysiological indices of perception and understanding.

References

Abercrombie, D. (1967). Elements of General Phonetics. Edinburgh: Edinburgh University Press.

Allen, G. D. (1975). Speech rhythm: Its relation to performance universals and articulatory timing. Journal of Phonetics, 3(75–86).

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

Arvaniti, A., & Rodriquez, T. (2013). The role of rhythm class, speaking rate, and F0 in language discrimination. Laboratory Phonology, 4(1), 7–38.

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. Language and Speech, 47(1), 31–56.

Baese-Berk, M. M., Heffner, C. C., Dilley, L. C., Pitt, M. A., Morrill, T. H., & McAuley, J. D. (2014). Long-term temporal tracking of speech rate affects spoken-word recognition. Psychological Science, 25(8), 1546–1553. https://doi.org/10.1177/0956797614533705

Barbosa, P. A. (2007). From syntax to acoustic duration: A dynamical model of speech rhythm production. Speech Communication, 49(9), 725–742. https://doi.org/10.1016/j.specom.2007.04.013

Barry, W. J., Andreeva, B., Russo, M., Dimitrova, S., & Kostadinova, T. (2003). Do rhythm measures tell us anything about language type? 2693–2696. Barcelona.

Beckman, M. E. (1992). Evidence for speech rhythms across languages. In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (Eds.), Speech Perception, Production and Linguistic Structure (pp. 457–463). Tokyo: OHM Publishing Co.

Berkovits, R. (1994). Durational effects in final lengthening, gapping, and contrastive stress. Language and Speech, 37(3), 237–250.

Bertinetto, P. M. (1989). Reflections on the dichotomy 'stress' vs.'syllable-timing'. Revue De Phonétique Appliquée, 91(93), 99–130.

Bertinetto, P. M., & Bertini, C. (2008). On modeling the rhythm of natural languages. 5. Campinas, Brazil.

Bloch, B. (1950). Studies in Colloquial Japanese IV Phonemics. Language, 26(1), 86–125.

Bosker, H. R., & Reinisch, E. (2017). Foreign languages sound fast: evidence from implicit rate normalization. Frontiers in Psychology, 8. https://doi.org/10.3389/fpsyg.2017.01063

Bouzon, C., & Hirst, D. (2004). Isochrony and prosodic structure in British English. Speech Prosody 2004, International Conference.

Brown, W. (1911). Studies from the psychological laboratory of the University of California. Temporal and accentual rhythm. Psychological Review, 18(5), 336–346.

Bruck, M., Treiman, R., & Caravolas, M. (1995). Role of the syllable in the processing of spoken English: Evidence from a nonword comparison task. Journal of Experimental Psychology: Human Perception and Performance, 21(3), 469.

Byrd, D., & Choi, S. (2010). At the juncture of prosody, phonology, and phonetics—The interaction of phrasal and syllable structure in shaping the timing of consonant gestures. Laboratory Phonology, 10, 31–59.

Byrd, D., & Saltzman, E. (2003). The elastic phrase: Modeling the dynamics of boundary-

adjacent lengthening. Journal of Phonetics, 31(2), 149–180.

Cauldwell, R. (2002). The functional irrhythmicality of spontaneous speech : A discourse view of speech rhythms. Apples - Journal of Applied Language Studies. Retrieved from https://jyx.jyu.fi/handle/123456789/22698

Cho, T., McQueen, J. M., & Cox, E. A. (2007). Prosodically driven phonetic detail in speech processing: The case of domain-initial strengthening in English. Journal of Phonetics, 35(2), 210–243.

Classe, A. (1939). The rhythm of English prose. B. Blackwell.

Crystal, T. H., & House, A. S. (1988). Segmental durations in connected-speech signals: Syllabic stress. The Journal of the Acoustical Society of America, 83(4), 1574–1585. https://doi.org/10.1121/1.395912

Cumming, R. E. (2011). Perceptually informed quantification of speech rhythm in pairwise variability indices. Phonetica, 68(4), 256–277. https://doi.org/10.1159/000335416

Cummins, F. (2002). Speech rhythm and rhythmic taxonomy. Proceedings of Speech Prosody 2002, 121–126. Aix-en-Provence.

Cummins, F. (2009). Rhythm as entrainment: The case of synchronous speech. Journal of Phonetics, 37(1), 16–28.

Cummins, F. (2011). Periodic and aperiodic synchronization in skilled action. Frontiers in Human Neuroscience, 170(5).

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

Cummins, F., & Port, R. (1998). Rhythmic constraints on stress timing in English. Journal of Phonetics, 26(2), 145–171.

Cutler, A. (1990). Exploiting prosodic probabilities in speech segmentation. Retrieved from http://psycnet.apa.org/psycinfo/1991-97555-004

Cutler, A. (2012). Native Listening: Language Experience and the Recognition of Spoken Words. MIT Press.

Cutler, A., & Mehler, J. (1993). The periodicity bias. Journal of Phonetics, 21(1–2), 103–108.

Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. Journal of Experimental Psychology: Human Perception and Performance, 14(1), 113.

Cutler, A., & Otake, T. (1994). Mora or phoneme? Further evidence for language-specific

listening. Journal of Memory and Language, 33(6), 824.

Cutler, A., & Van Donselaar, W. (2001). Voornaam is not (really) a homophone: Lexical prosody and lexical access in Dutch. Language and Speech, 44(2), 171–195.

Dankovičová, J. (1997). The domain of articulation rate variation in Czech. Journal of Phonetics, 25(3), 287–312. https://doi.org/10.1006/jpho.1997.0045

Dauer, R. M. (1983). Stress-timing and syllable-timing reanalyzed. Journal of Phonetics, 11, 51–62.

Delattre, P. (1962). Some factors of vowel duration and their cross-linguistic validity. The Journal of the Acoustical Society of America, 34(8), 1141–1143.

Dellwo, V. (2010). Influences of speech rate on the acoustic correlates of speech rhythm: An experimental phonetic study based on acoustic and perceptual evidence. Bonn.

Dellwo, V., Leemann, A., & Kolly, M.-J. (2015). Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors. The Journal of the Acoustical Society of America, 137(3), 1513–1528.

Dellwo, V., & Wagner, P. (2003). Relationships between rhythm and speech rate. In M.-J. Solé & D. Recasens i Vives (Eds.), Proceedings of the 15th International Congress of Phonetic Sciences [also] 15th IcPhS, Barcelona 3-9 August 2003 (pp. 471–474). Barcelona: Universitat Autónoma de Barcelona.

Dilley, L. C., & 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

Dupoux, E., Peperkamp, S., & Sebastián-Gallés, N. (2001). A robust method to study stress "deafness". The Journal of the Acoustical Society of America, 110(3), 1606–1618. https://doi.org/10.1121/1.1380437

Eriksson, A. (1991). Aspects of Swedish Speech Rhythm (PhD). Göteborg.

Fant, G., Kruckenberg, A., & Nord, L. (1991). Durational correlates of stress in Swedish, French, and English. Journal of Phonetics, 19, 351–365.

Ferragne, E., & Pellegrino, F. (2004). A comparative account of the suprasegmental and rhythmic features of British English dialects. Presented at the Modelisations pour l'Identification des Langues, Paris.

Fletcher, J. (2010). The prosody of speech: Timing and rhythm. The Handbook of Phonetic

Sciences, Second Edition, 521–602.

Fougeron, C., & Keating, P. A. (1997). Articulatory strengthening at edges of prosodic domains. The Journal of the Acoustical Society of America, 101(6), 3728–3740.

Fourakis, M., & Monahan, C. B. (1988). Effects of metrical foot structure on syllable timing. Language and Speech, 31(3), 283–306.

Fowler, C. A. (1983). Converging sources of evidence on spoken and perceived rhythms of speech: Cyclic production of vowels in monosyllabic stress feet. Journal of Experimental Psychology: General, 112(3), 386.

Fusaroli, R., & Tylén, K. (2016). Investigating conversational dynamics: Interactive alignment, Interpersonal synergy, and collective task performance. Cognitive Science, 40(1), 145–171.

Gibbon, D. (2006). Time types and time trees: Prosodic mining and alignment of temporally annotated data. Methods in Empirical Prosody Research, 281–209.

Gili Gaya, S. (1940). La cantidad silábica en la frase. Castilla, 1, 287–298.

Goldman-Eisler, F. (1956). The determinants of the rate of speech output and their mutual relations. Journal of Psychosomatic Research, 1(2), 137–143.

Grabe, E., & Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. Papers in Laboratory Phonology, 7(515–546). Retrieved from http://wwwhomes.uni-bielefeld.de/~gibbon/AK-Phon/Rhythmus/Grabe/Grabe_Low-reformatted.pdf

Gussenhoven, C., & Rietveld, A. C. M. (1992). Intonation contours, prosodic structure and preboundary lengthening. Journal of Phonetics, 20(3), 283–303.

Han, M. (1962). The feature of duration in Japanese. Onsei No Kenkyuu, 10, 65–80.

Hayes-Harb, R. (2014). Acoustic-phonetic parameters in the perception of an accent. In J. M. Levis & A. Moyer (Eds.), Social Dynamics in Second Language Accent. Berlin: De Gruyter Mouton.

Hirst, D. (2009). The rhythm of text and the rhythm of utterances: from metrics to models. 1519–1522. Brighton.

Hoequist, C. (1983). Syllable duration in stress-, syllable-and mora-timed languages. Phonetica, 40(3), 203–237.

Jones, M. R., & Boltz, M. (1989). Dynamic attending and responses to time. Psychological Review, 96(3), 459.

Jun, S.-A. (2005). Korean intonational phonology and prosodic transcription. Prosodic

Typology: The Phonology of Intonation and Phrasing, 1, 201.

Kaiki, N., & Sagisaka, Y. (1992). The control of segmental duration in speech synthesis using statistical methods. In Speech Perception, Production and Linguistic Structure (pp. 391–402). Tokyo: Ohmsha.

Klatt, D. H. (1975). Vowel lengthening is syntactically determined in a connected discourse. Journal of Phonetics, 3(3), 129–140.

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. The Journal of the Acoustical Society of America, 59(5), 1208–1221.

Ladefoged, P. (1975). A Course in Phonetics. New York: Harcourt Brace Jovanovich.

Lee, M.-W., & Gibbons, J. (2007). Rhythmic alternation and the optional complementiser in English: New evidence of phonological influence on grammatical encoding. Cognition, 105(2), 446–456.

Leemann, A., Kolly, M.-J., & Dellwo, V. (2014). Speaker-individuality in suprasegmental temporal features: Implications for forensic voice comparison. Forensic Science International, 238, 59–67.

Lehiste, I. (1972). The timing of utterances and linguistic boundaries. The Journal of the Acoustical Society of America, 51(6B), 2018–2024.

Lehiste, I. (1977). Isochrony reconsidered. Journal of Phonetics, 5, 253–263.

Liss, J. M., White, L., Mattys, S. L., Lansford, K., Lotto, A. J., Spitzer, S. M., & Caviness, J. N. (2009). Quantifying speech rhythm abnormalities in the dysarthrias. Journal of Speech, Language, and Hearing Research, 52(5), 1334–1352. https://doi.org/10.1044/1092-4388(2009/08-0208)

Loukina, A., Kochanski, G., Rosner, B., Keane, E., & Shih, C. (2011). Rhythm measures and dimensions of durational variation in speech. The Journal of the Acoustical Society of America, 129(5), 3258–3270.

Low, E. L., Grabe, E., & Nolan, F. (2000). Quantitative characterizations of speech rhythm: syllable-timing in Singapore English. Language and Speech, 43(4), 377–401.

Malisz, Z., O'Dell, M., Nieminen, T., & Wagner, P. (2016). Perspectives on speech timing: coupled oscillator modeling of Polish and Finnish. Phonetica, 73(3–4), 229–255.

Maskikit-Essed, R., & Gussenhoven, C. (2016). No stress, no pitch accent, no prosodic focus: the case of Ambonese Malay. Phonology, 33(2), 353–389.

https://doi.org/10.1017/S0952675716000154

Mattys, S. L., & Melhorn, J. F. (2005). How do syllables contribute to the perception of spoken English? Insight from the migration paradigm. Language and Speech, 48(2), 223–252.

McAuley, J. D., & Fromboluti, E. K. (2014). Attentional entrainment and perceived event duration. Philosophical Transactions of the Royal Society B: Biological Sciences, 369(1658), 20130401–20130401. https://doi.org/10.1098/rstb.2013.0401

Mehler, J., Dommergues, J. Y., Frauenfelder, U., & Segui, J. (1981). The syllable's role in speech segmentation. Journal of Verbal Learning and Verbal Behavior, 20(3), 298–305.

Mehler, J., Dupoux, E., Nazzi, T., & Dehaene-Lambertz, G. (1996). Coping with linguistic diversity: The infant's viewpoint. Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition, 101–116.

Monaghan, P., White, L., & Merkx, M. M. (2013). Disambiguating durational cues for speech segmentation. The Journal of the Acoustical Society of America, 134(1), EL45–EL51. https://doi.org/10.1121/1.4809775

Morrill, T. H., Dilley, L. C., McAuley, J. D., & Pitt, M. A. (2014). Distal rhythm influences whether or not listeners hear a word in continuous speech: Support for a perceptual grouping hypothesis. Cognition, 131(1), 69–74. https://doi.org/10.1016/j.cognition.2013.12.006

Morton, J., Marcus, S., & Frankish, C. (1976). Perceptual centers (P-centers). Psychological Review, 83(5), 405.

Mücke, D., Grice, M., & Cho, T. (2014). More than a magic moment–Paving the way for dynamics of articulation and prosodic structure. Journal of Phonetics, 44, 1–7.

Nakatani, L. H., O'Connor, K. D., & Aston, C. H. (1981). Prosodic aspects of American English speech rhythm. Phonetica, 38(1–3), 84–105. https://doi.org/10.1159/000260016

Nazzi, T., Jusczyk, P. W., & Johnson, E. K. (2000). Language discrimination by English-learning 5-month-olds: Effects of rhythm and familiarity. Journal of Memory and Language, 43(1), 1–19.

Nazzi, T., & Ramus, F. (2003). Perception and acquisition of linguistic rhythm by infants. Speech Communication, 41(1), 233–243.

Nespor, M. (1990). On the rhythm parameter in phonology. In I. M. Roca (Ed.), Logical Issues in Language Acquisition (pp. 157–175). Dordrecht.

Nolan, F., & Asu, E. L. (2009). The pairwise variability index and coexisting rhythms in

language. Phonetica, 66(1–2), 64–77.

Nolan, F., & Jeon, H.-S. (2014). Speech rhythm: a metaphor? Phil. Trans. R. Soc. B, 369(1658), 20130396. https://doi.org/10.1098/rstb.2013.0396

O'Dell, M. L., & Nieminen, T. (1999). Coupled oscillator model of speech rhythm. Proceedings of the XIVth International Congress of Phonetic Sciences, 2, 1075–1078.

O'Dell, M. L., & Nieminen, T. (2009). Coupled oscillator model for speech timing: Overview and examples. Nordic Prosody: Proceedings of the 10th Conference, Helsinki, 179–190. Retrieved from http://www.academia.edu/download/3247213/odell-nieminen2008.pdf

Oller, D. K. (1973). The effect of position in utterance on speech segment duration in English. The Journal of the Acoustical Society of America, 54(5), 1235–1247.

Peelle, J. E., & 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

Pellegrino, F., Coupé, C., & Marsico, E. (2011). Across-language perspective on speech information rate. Language, 87(3), 539–558. https://doi.org/10.1353/lan.2011.0057

Peperkamp, S., & Dupoux, E. (2002). A typological study of stress 'deafness'. In C. Gussenhoven & N. Warner (Eds.), Laboratory Phonology (Vol. 7, pp. 203–240). Berlin: Mouton de Gruyter.

Pfitzinger, H. R., & Tamashima, M. (2006). Comparing perceptual local speech rate of German and Japanese speech. Proceedings of the Third International Conference on Speech Prosody, 1, 105–108. TUD Press Dresden.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. Behavioral and Brain Sciences, 36(4), 329–347. https://doi.org/10.1017/S0140525X12001495

Pike, K. L. (1945). The Intonation of American English. Retrieved from http://eric.ed.gov/?id=ED077259

Pointon, G. E. (1980). Is Spanish really syllable-timed? Journal of Phonetics, 8(3), 293–304.

Pointon, G. E. (1995). Rhythm and duration in Spanish. In J. Windsor-Lewis (Ed.), Studies in General and English Phonetics: Essays in Honour of Professor J.D. O'Connor (pp. 266–269). London: Routledge.

Polyanskaya, L., Ordin, M., & Busa, M. G. (2016). Relative salience of speech rhythm and speech rate on perceived foreign accent in a second language. Language and Speech,

0023830916648720.

Port, R. F. (2003). Meter and speech. Journal of Phonetics, 31(3–4), 599–611. https://doi.org/10.1016/j.wocn.2003.08.001

Port, R. F., & Leary, A. P. (2005). Against formal phonology. Language, 927–964.

Prieto, P., Vanrell, M. del M., Astruc, L., Payne, E., & Post, B. (2012). Phonotactic and phrasal properties of speech rhythm. Evidence from Catalan, English, and Spanish. Speech Communication, 54(6), 681–702. https://doi.org/10.1016/j.specom.2011.12.001

Rahmani, H., Rietveld, T., & Gussenhoven, C. (2015). Stress "deafness" reveals absence of lexical marking of stress or tone in the adult grammar. PLOS ONE, 10(12), e0143968. https://doi.org/10.1371/journal.pone.0143968

Rakerd, B., Sennett, W., & Fowler, C. A. (1987). Domain-final lengthening and foot-level shortening in spoken English. Phonetica, 44(3), 147–155.

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. Cognition, 73(3), 265–292.

Ramus, F., Nespor, M., & Mehler, J. (2003). The psychological reality of rhythm classes: perceptual studies. In M. J. Solé & D. Recasens i Vives (Eds.), Proceedings of the 15th International Congress of Phonetic Sciences [also] 15th IcPhS, Barcelona 3-9 August 2003. Barcelona: Universitat Autónoma de Barcelona.

Reinisch, E., Jesse, A., & McQueen, J. M. (2011). Speaking rate affects the perception of duration as a suprasegmental lexical-stress cue. Language and Speech, 54(2), 147–165. https://doi.org/10.1177/0023830910397489

Roach, P. (1982). On the distinction between 'stress-timed'and 'syllable-timed'languages. In D. Crystal (Ed.), Linguistic Controversies (pp. 73–79). London: Edward Arnold.

Schlüter, J. (2009). Rhythmic grammar: The influence of rhythm on grammatical variation and change in English (Vol. 46). Walter de Gruyter.

Sebastián-Gallés, N., Dupoux, E., Segui, J., & Mehler, J. (1992). Contrasting syllabic effects in Catalan and Spanish. Journal of Memory and Language, 31(1), 18–32.

Shih, S. S. (2014). Towards optimal rhythm (PhD Thesis). Stanford University Stanford, CA.

Sluijter, A. M. C., & van Heuven, V. J. (1995). Effects of focus distribution, pitch accent and lexical stress on the temporal organization of syllables in Dutch. Phonetica, 52(2), 71–89. https://doi.org/10.1159/000262061

Snijders, T. M., Kooijman, V., Cutler, A., & Hagoort, P. (2007). Neurophysiological evidence of delayed segmentation in a foreign language. Brain Research, 1178, 106–113.

Steele, J. (1779). Prosodia Rationalise: Or, an Essay Towards Establishing the Melody and Measure of Speech to Be Expressed and Perpetuated by Peculiar Symbols. London: J. Nichols.

Suomi, K. (2009). Durational elasticity for accentual purposes in Northern Finnish. Journal of Phonetics, 37(4), 397–416. https://doi.org/10.1016/j.wocn.2009.07.003

Tabain, M., Fletcher, J., & Butcher, A. (2014). Lexical stress in Pitjantjatjara. Journal of Phonetics, 42, 52–66. https://doi.org/10.1016/j.wocn.2013.11.005

Tabossi, P., Collina, S., Mazzetti, M., & Zoppello, M. (2000). Syllables in the processing of spoken Italian. Journal of Experimental Psychology: Human Perception and Performance, 26(2), 758.

Temperley, D. (2009). Distributional stress regularity: a corpus study. Journal of Psycholinguistic Research, 38(1), 75. https://doi.org/10.1007/s10936-008-9084-0

Tilsen, S. (2009). Multitimescale dynamical interactions between speech rhythm and gesture. Cognitive Science, 33(5), 839–879. https://doi.org/10.1111/j.1551-6709.2009.01037.x

Turk, A. E., & Shattuck-Hufnagel, S. (2007). Multiple targets of phrase-final lengthening in American English words. Journal of Phonetics, 35(4), 445–472.

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

Turk, A. E., & Shattuck-Hufnagel, S. (2014). Timing in talking: what is it used for, and how is it controlled? Philosophical Transactions of the Royal Society B: Biological Sciences, 369(1658), 20130395–20130395. https://doi.org/10.1098/rstb.2013.0395

Turvey, M. T. (1990). Coordination. American Psychologist, 45(8), 938–953.

Van Lieshout, P. (2004). Dynamical systems theory and its application in speech. In B. Maassen, R. Kent, H. Peters, P. van Lieshout, & W. Hulstijn (Eds.), Speech Motor Control in Normal and Disordered Speech (Vol. 3, pp. 51–82). Oxford: Oxford University Press.

van Santen, J. P. H. (1992). Contextual effects on vowel duration. Speech Communication, 11(6), 513–546. https://doi.org/10.1016/0167-6393(92)90027-5

van Santen, J. P. H. (1997). Segmental duration and speech timing. In Y. Sagisaka, N. Campbell,

& N. Higuchi (Eds.), Computing Prosody: Computational Models for Processing Spontaneous Speech (pp. 225–249). New York: Springer-Verlag.

van Santen, J. P. H., & Shih, C. (2000). Suprasegmental and segmental timing models in Mandarin Chinese and American English. The Journal of the Acoustical Society of America, 107(2), 1012–1026.

Wagner, P. (2010). A time-delay approach to speech rhythm visualization, modeling and measurement. Prosodic Universals: Comparative Studies in Rhythmic Modeling and Rhythm Typology.

Wagner, P., Malisz, Z., Inden, B., & Wachsmuth, I. (2013). Interaction phonology–A temporal co-ordination component enabling representational alignment within a model of communication. Alignment in Communication. Towards a New Theory of Communication, 109–132.

Warner, N., & Arai, T. (2001). Japanese mora-timing: a review. Phonetica, 58(1–2), 1–25. https://doi.org/10.1159/000028486

White, L. (2002). English speech timing: a domain and locus approach (University of Edinburgh). Retrieved from http://www.cstr.ed.ac.uk/projects/eustace/fulltext/lswdiss.pdf

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. (2018). Segmentation of speech. In S.-A. Rueschemeyer & G. Gaskell (Eds.), Oxford Handbook of Psycholinguistics. Oxford: Oxford University Press.

White, L., Delle Luche, C., & Floccia, C. (2016). Five-month-old infants' discrimination of unfamiliar languages does not accord with "rhythm class". In J. Barnes, A. Brugos, S. Shattuck-Hufnagel, & N. Veilleux (Eds.), Proceedings of Speech Prosody 2016, Boston (pp. 567–571). https://doi.org/10.21437/SpeechProsody.2016-116

White, L., & Mattys, S. L. (2007a). Calibrating rhythm: First language and second language studies. Journal of Phonetics, 35(4), 501–522. https://doi.org/10.1016/j.wocn.2007.02.003

White, L., & Mattys, S. L. (2007b). Rhythmic typology and variation in first and second languages. In P. Prieto, J. Mascaró, & M.-J. Solé (Eds.), Segmental and Prosodic Issues in Romance Phonology (pp. 237–257). Amsterdam: John Benjamins.

White, L., Mattys, S. L., & Wiget, L. (2012). Language categorization by adults is based on sensitivity to durational cues, not rhythm class. Journal of Memory and Language, 66(4),

665–679. https://doi.org/10.1016/j.jml.2011.12.010

White, L., & Turk, A. E. (2010). English words on the Procrustean bed: Polysyllabic shortening reconsidered. Journal of Phonetics, 38(3), 459–471. https://doi.org/10.1006/jpho.1999.0093

Wiget, L., White, L., Schuppler, B., Grenon, I., Rauch, O., & Mattys, S. L. (2010). How stable are acoustic metrics of contrastive speech rhythm? The Journal of the Acoustical Society of America, 127(3), 1559–1569. https://doi.org/10.1121/1.3293004

Windmann, A., Šimko, J., & Wagner, P. (2015). Optimization-based modeling account of speech timing. Speech Communication, 74, 76–92.

Zwitserlood, P., Schriefers, H., Lahiri, A., & Van Donselaar, W. (1993). The role of syllables in the perception of spoken Dutch. Journal of Experimental Psychology: Learning, Memory, and Cognition, 19(2), 260–271.