20 The Hierarchical Temporal Structure of Prosody

CHRISTOPHER T. KELLO

20.1 Introduction

Languages are layered with patterns and processes spanning multiple scales (Patel, 2003). Theories of phonology, morphology, and syntax address levels of linguistic analysis, and each one contains layers in terms of nested representations that correspond with smaller to larger linguistic structures across several scales. Phonemes nest within syllables, morphemes within words, and both within phrases; phrases nest within clauses and clauses within sentences.

Prosody is also theorized in terms of multiple levels of linguistic structure that are often expressed in terms of hierarchical structure (Beckman, 1996; Goldsmith, 1990; Krivokapić, 2014; O'Dell and Nieminen, 1999; Shattuck-Hufnagel and Turk, 1996; Tilsen and Arvaniti, 2013). Starting with the smaller timescales of prosody, the amplitude, duration, and pitch of vocalic and syllabic sounds may be varied for purposes of lexical stress, for example (Fry, 1958; Gordon and Roettger, 2017). At longer timescales, the same acoustic dimensions may be varied more slowly over phrases, clauses, and sentences as an expression of grammar, emphasis, emotion, and discourse (Bolinger, 1989). The rhythm and pace of speech may vary with longer-term changes in discourse (Wennerstrom, 2001), and overall patterns of intonation may vary with changes in speech register (Yaeger-Dror, 2002). All these timescales are layered together and combined to yield prosodic effects on overt speech.

While prosodic hierarchies are well established as linguistic representations, less is known about the patterns and relationships across the timescales of naturally produced speech that encode prosodic information. For instance, coupled oscillators have been formulated to model rhythmic patterns of activity hypothesized to encode prosodic information (Goswami, 2019; Meyer, 2018; Tilsen, 2009). Consistent with this hypothesis, speech can be overtly rhythmic, as when items are listed aloud (Schiffrin, 1994) or when words and phrases are repeated (Cummins and Port, 1998). Rhythmic patterning can also be found when nursery rhymes are read aloud (Leong and Goswami, 2015) and in infant-directed speech (Leong et al., 2017).

The studies by Leong et al. illustrate some of the advances, and complexities, in measuring prosodic hierarchies in the speech signal. They computed the modulation spectrum for acoustic

https://doi.org/10.1017/9781009295888.024 Published online by Cambridge University Press

recordings of speech and divided spectra into frequency bands corresponding to established bands of brain rhythms – delta, theta, and beta/low-gamma bands (also see Chapter 10). They found phase relations between peaks in the amplitude envelopes across different frequency bands, indicating a hierarchical rhythmic organization of the speech signal. Evidence indicates that brain rhythms can entrain to amplitude modulations in these frequency bands (Poeppel and Assaneo, 2020). While these specialized speech conditions demonstrate the capacity for speech to exhibit regular periodicities (see Tilsen and Arvaniti, 2013), the results also indicate that most natural speech conditions do not appear to be so rhythmic (Cummins, 2012) – for instance, Leong et al. (2017) found that adult-directed speech did not exhibit a hierarchical rhythmic organization like infant-directed speech (also see Chapter 14).

The lack of prominent periodicity in most naturally occurring speech signals does not mean there is a lack of temporal structure to carry prosodic information. Some researchers refer to the speech signal as being "quasi-regular," "quasi-periodic," or "quasi-rhythmic" (Giraud and Poeppel, 2012; Peelle and Davis, 2012), but a concentration in energy within a given frequency band does not necessarily entail any amount of regularity or periodicity. Another possibility of temporal structure takes the form of clustering or so-called burstiness in time (Goh and Barabási, 2008), which means more clustering than expected by a random process. Clustering is common in many human activities, and rhythmic activity can be a special case of temporal clustering at regular intervals, such as the beats of a drum. Nonrhythmic clustering is found in eye movements, for example, where irregular but nonrandom jumps to locations in the visual field are interspersed with periods of fixation composed of shorter microsaccades (Salvucci and Goldberg, 2000). Similarly, foraging behaviors consist of irregular bouts of exploratory locomotive activities interspersed with relatively stationary periods of intensive local search (Garg and Kello, 2021). Clustering can be completely irregular and aperiodic and still manifest as concentrated energy in a frequency band, the same as a comparable regular, periodic signal.

The present chapter considers that while speech signals may have some quasi-regularity/periodicity, prosodic information may also be carried by nonrhythmic, aperiodic clustering in the speech signal across a hierarchy of timescales. There are multiple possible features of the speech signal that might cluster to carry prosodic information. In terms of acoustic energy, for instance, there may be vocalic clusters of peak intensity or peak pitch values interspersed with irregular periods of relative silence or nonpeak/nonvocalic values. Rather than being characterized by specific timescales of activity, speech seems better characterized by activity that might cluster across a range of timescales, from tens of milliseconds to tens of seconds and even longer depending on the speaking conditions. Prosodic information in the form of irregular clustering might be available anywhere in this range, up to discourse-level timescales.

The intensity of acoustic energy is well established as a measure of speech signals that carries prosodic information (Shattuck-Hufnagel and Turk, 1996), and this measure is generally captured by the amplitude envelope that the modulation spectrum is based upon. Periods of intense acoustic energy can be irregularly interspersed with less intense periods and silence, and these periods correspond to rises and falls in the amplitude envelope, respectively. Fluctuations in the amplitude envelope can unfold both quickly and slowly at the same time, reflecting temporal structure across multiple timescales. For instance, Tilsen and Arvaniti (2013) used

https://doi.org/10.1017/9781009295888.024 Published online by Cambridge University Press

empirical mode decomposition to analyze temporal structure in the amplitude envelope for syllabic and stress-driven timescales in the speech signal (also see Chapter 16).

The present chapter examines how a measure of hierarchical temporal structure in the amplitude envelope (Falk and Kello, 2017; Kello et al., 2017) appears to reflect an overall degree of prosodic composition in the speech signal. While similar measures have been advanced in recent years (Ding et al., 2017; Goswami, 2019), this chapter focuses on an adaptation of Allan Factor analysis (Allan, 1966) designed to distill temporal structure across and applied over a relatively wide range of timescales, from phonemic effects on the order of tens of milliseconds to speech rate and turn-taking effects on the order of tens of seconds. The resulting Allan Factor function is agnostic to whether temporal structure is periodic or aperiodic, and it excludes phase information, which means that it does not capture nesting of linguistic units located in specific time intervals. Instead, the function measures the temporal structure around a given timescale in terms of Allan Factor variance (a type of coefficient of variation explained later), and a given signal is hierarchical when temporal structure grows with timescale.

20.2 Hierarchical Temporal Structure

Perhaps the most direct way to realize hierarchical temporal structure is to relate measures of speech activity to the hypothesized nesting of specific phonemes, syllables, and so on being produced. While this approach will yield hierarchical structures, it requires mapping between a physical space of measurement and the symbolic space of linguistic representation. In speech sounds, the physical space of prosody is typically defined in terms of amplitude and duration of periods of acoustic energy, plus durations of periods lacking energy, as well as measures of pitch and spectral frequency. Sounds and silent periods corresponding to specific phonemes and other phonological units may be produced with variations in amplitude, duration, and spectral measures that correspond to prosodic boundaries, contrasts, and features (Hayes, 1989; Langus et al., 2012). Also, electrophysiological activity corresponding to speech may similarly express variations in amplitude, duration, and spectral measures of voltage, for instance, that reflect hierarchical processing of language structures (Giraud and Poeppel, 2012).

Research on hierarchically nested structures has been fruitful in advancing our understanding of speech and prosody. That said, it can be difficult to identify phonological units and their occurrence in ongoing speech and brain activity (Port, 2008). Alternatively, one can quantify the shape and degree of hierarchical temporal structure in acoustic signals without knowledge of the underlying processes or representations (Singh and Theunissen, 2003). Perhaps the first study to analyze hierarchical temporal structure in speech was conducted by Voss and Clarke (1975). They computed power spectra of amplitude envelopes for radio station signals containing various kinds of speech and music. The envelope captures fluctuations in the amount of acoustic energy recorded over time, and the spectrum quantifies the degree (power) of fluctuation at different frequencies. In a signal with hierarchical temporal structure, power builds up from higher to lower frequencies because smaller units of speech that vary at higher frequencies are nested within larger units of speech that vary at lower frequencies, which are nested within even larger units at even lower frequencies, and so on.

https://doi.org/10.1017/9781009295888.024 Published online by Cambridge University Press

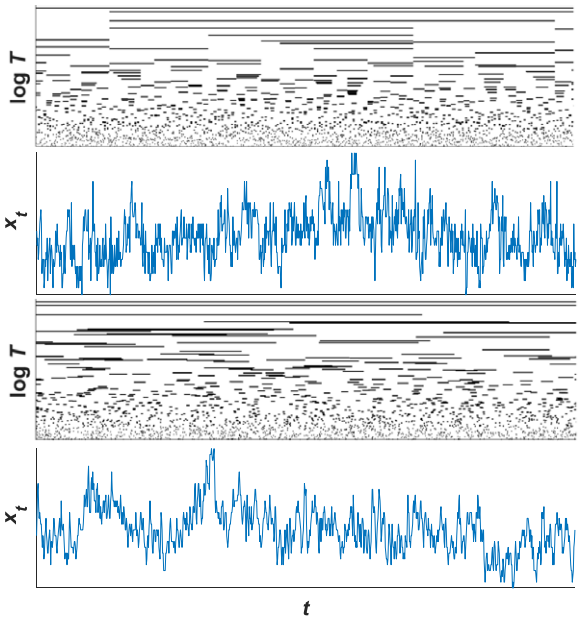


Figure 20.1 An illustrative model of hierarchically nested intervals and their summed fluctuations.
In the top graph, one interval spans the entire length of time (x-axis), represented as a horizontal line located at the longest timescale (the highest point on the logarithmic y-axis). The interval is copied and divided at a random point along its length (towards the left in this case). The two resulting intervals remain at their x-axis positions and are plotted at their new corresponding timescales (lengths). The process is repeated recursively until all intervals reach a minimal length. The second graph down plots a series of sums across the nested intervals at each time point. The third graph down shows the same hierarchical intervals but with their locations randomized along the time axis, and the resulting sums are plotted below them.

Voss and Clarke (1975) found that an inverse relationship between power and frequency known as 1/f scaling (aka 1/f noise), which since has been found across a wide range of neural and behavioral activity (see Kello et al., 2010). Like hierarchical temporal structure, 1/f scaling is agnostic to whether or not units of speech or music are nested in time (i.e., phase information is removed from the analysis). To illustrate this point, a simple model of nested temporal structure is shown in Figure 20.1 (Kello et al., 2024). Nesting is created by starting with a long unit interval and then copying and splitting the interval at a random point along its length to create two new intervals. The process is repeated recursively until units reach an arbitrary minimum length.

In the top half of the figure, one can see smaller unit intervals repeatedly nested within larger unit intervals across timescales to form hierarchical structures. The structures in this example are aperiodic because intervals are split at randomized points, but structures could instead be made periodic by splitting intervals at their midpoints instead. In the bottom half of the figure, the same unit intervals from above are shown with their phases randomized. Amplitude envelopes are created for nested and phase-randomized intervals by summing the values (all equal to one) of intervals present at each successive time point.

The nesting of structures is destroyed by phase randomization, which changes the resulting fluctuations in their amplitude envelopes. However, both envelopes converge on 1/f scaling because hierarchical temporal structure is preserved, as it does not capture information about specific clusters or nested structures. The same would be true for a periodic version of the model because 1/f scaling is also agnostic to periodicity – it essentially reflects the clustering of amplitudes (power) as a function of frequency.

This simple model creates "pure" hierarchical temporal structure that is devoid of the many other factors and influences that combine to create the richness of real speech and music

https://doi.org/10.1017/9781009295888.024 Published online by Cambridge University Press

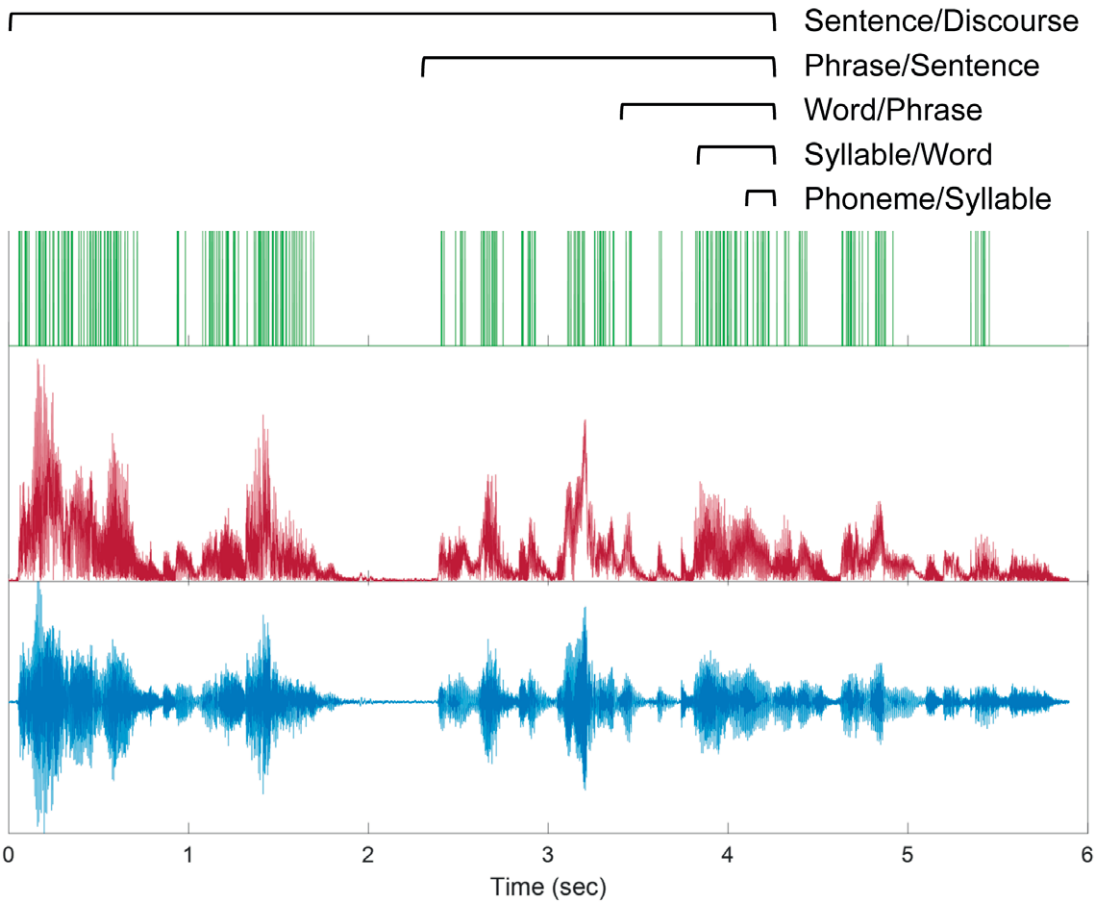


Figure 20.2 Illustration of Allan Factor analysis.
A speech example of hierarchical temporal structure is shown for the utterance "all we have to decide is what to do with the time that is given to us." The acoustic waveform is at the bottom, the amplitude envelope in the middle, and peaks of the envelope above the threshold used for Allan Factor analysis. Brackets above show example windows over which peaks are counted and differenced over time to measure clustering. Window sizes roughly correspond to different overlapping linguistic timescales of prosodic temporal structure.

signals. To distill hierarchical temporal structure from real signals, one can convert their amplitude envelopes into series of peaks defined only by their timing – a time series of zeros and ones that is called a point process (see Figure 20.2). Converting amplitude envelopes to point processes focuses analyses on temporal structure per se, and the threshold for identifying peaks can be set to achieve a specified overall rate of peaks that facilitates comparisons across recordings. The rate of peaks is a free parameter that is set to allow for a large dynamic range of peaks per unit time, so that hierarchical temporal clustering in the signal may be expressed as variability in the clustering of peaks. Peaks sometimes roughly correspond to glottal pulses, but they also correspond to aperiodic sounds of sufficient energy and generally any significant

https://doi.org/10.1017/9781009295888.024 Published online by Cambridge University Press

concentration of acoustic energy. Also, individual peaks are not intended to capture perceptual events (e.g., they may occur more rapidly than the auditory flicker fusion rate). Instead, clusters of peaks are the basic unit of analysis.

The resulting point processes are submitted to Allan Factor analysis, which quantifies the degree of clustering across the range of available timescales given the time series length and resolution, without identifying individual clusters at any given timescale. Hierarchical temporal structure is expressed in Allan Factor analysis as a positive relationship between timescale and average amount of clustering at that timescale, in logarithmic coordinates. The degree of hierarchical temporal structure can be measured by the logarithmic rate of increase in average clustering.

20.3 Hierarchical Temporal Structure in Speech

Allan Factor analysis has been applied to many hours of speech recordings from several different languages under a range of speaking conditions, including talks and other monologues, rapping, singing, infant-directed speech, computer-generated speech, and conversations in one and two languages (Alviar et al., 2018; Falk and Kello, 2017; Kello et al., 2017; Schneider et al., 2020). The first and foremost result is that speech exhibits a characteristic pattern of hierarchical temporal structure that holds true in all cases, regardless of language spoken or idiosyncrasies of recording conditions.

The pattern is characterized by a relatively steep rise in average clustering as the Allan Factor function goes from timescales of tens of milliseconds to about one second, followed by a tapering in the rate of increase in clustering as timescales extend to tens of seconds (longer timescales require longer recordings of 10+ minutes that are difficult to find or collect). This general pattern of steep-then-tapering suggests that speech exhibits strong hierarchical structure for the timescales over which much of linguistic processing occurs – phonemes, syllables, words, phrases, and short sentences. Clustering subsides as sentences become longer and timescales stretch out to multiple sentences.

The pattern of tapered clustering at longer timescales may be partly due to the lengths of recordings being limited to around four–five minutes. Recordings were generally intended to contain uninterrupted speech signals throughout, which inhibits clustering at the longest timescales measured. Breaks in activity lasting tens of seconds are generally unwanted in audio recordings because there is an expectation of something being recorded. Such breaks may come about naturally under certain conditions, perhaps during casual or intimate conversation for example, but further work is needed to investigate.

While it is possible that tapered clustering may be partly due to recording conditions, results indicate that tapering also varies systematically across individuals, even when they are speaking different languages. Schneider et al. (2020) recorded pairs of Spanish-English bilingual speakers having conversations in the three different language conditions: Spanish, English, and a mixed condition in which one partner spoke Spanish while the other spoke English. Audio recordings were individual to each speaker, which allowed for within-speaker comparisons of hierarchical temporal structure when speaking the same language or different languages. Results showed

https://doi.org/10.1017/9781009295888.024 Published online by Cambridge University Press

that variations around the general steep-then-tapered pattern were highly correlated (mean B = 0.77) within speakers, regardless of whether correlations were within Spanish or English, or between the two languages.

These results show that hierarchical temporal structure, as measured by the adaptation of Allan Factor analysis, reflects the speaking style of individuals rather than the specific language or words being spoken. Consistent with this conclusion, Kello et al. (2017) found that TED talks all had highly similar patterns of hierarchical temporal structure across seven different languages. When combined with the results from Schneider et al. (2020), the TED talk pattern seems to reflect the consistent style and manner in which TED talks are given. The pattern consists of a specific tilt and bend to the increase in clustering with timescale, which is difficult to interpret on its own. Next, recent studies are reviewed that begin to shed light on how to interpret systematic patterns and variability in hierarchical temporal structure.

20.4 Interpretable Variability in Hierarchical Temporal Structure

Abney et al. (2014) reported the first study to employ Allan Factor analysis adapted for measuring hierarchical temporal structure in speech acoustics. They analyzed conversations from a study by Paxton and Dale (2013) in which pairs of individuals engaged in either affiliative conversations (e.g., shared interests in entertainment) or argumentative conversations. The latter were created by pairing individuals who took opposing positions on "hot button" issues such as abortion based on a prior survey, and then instructing them to talk about the issue. The authors analyzed audio recordings of individual speakers as in Schneider et al. (2020), and results showed a similar steep-then-taper pattern, but there was greater hierarchical temporal structure in the longer timescales for conversations versus monologues such as TED talks. The reason is that conversations include more pauses for turn-taking and interacting, which creates clustering, whereas speaking is more continuous in monologues such as TED talks.

Abney et al. (2014) also found more tapering for affiliative compared with argumentative conversations. This difference indicates that speakers can alter the degree of hierarchical temporal structure based on the mode of conversation. The authors conjectured that the argumentative conditions triggered more formal speech registers that, among other factors, caused speakers to enunciate more carefully and clearly compared with affiliative conversations in which speaking can be looser and more fluid. This conjecture is supported by the finding that hierarchical temporal structure changed in the prosodic timescales of hundreds of milliseconds and longer, rather than more fine-grained phonetic features that speakers may not be able to control. Since this first study, three other studies of speech under varying conditions have found similar effects that corroborate the original conjecture.

First, Kello et al. (2017) compared hierarchical temporal structure measured in original TED talk recordings with those created by Google speech synthesis when reading transcripts of the talks. Allan Factor analysis showed that prosodic timescales for synthesized speech were devoid of temporal structure compared with original recordings, in that the Allan Factor function tapered off to a flat line. Given that the same words were spoken in both conditions,

https://doi.org/10.1017/9781009295888.024 Published online by Cambridge University Press

this result can be attributed to the failure of speech synthesis (at the time) to use prosody as a channel for communicating meaning and emotion.

Second, Ramirez-Aristizabal et al. (2018) found that hierarchical temporal structure in prosodic timescales (and not the faster phonetic timescales) is more tapered for fast versus slow rates of speech. They instructed participants to read off a teleprompter excerpt from a speech by Barack Obama. The speed of the teleprompter was manipulated to be either slower or faster than the pace of the original speech to induce relatively slow versus fast speaking. Also, the original recording of Obama giving the speech was slowed down or sped up using an algorithm for manipulating speech rate without affecting pitch so that the result is relatively natural sounding. For both the algorithm and the teleprompter, tapering was less pronounced for slow speaking and more pronounced for fast speaking. Changing the pace of speaking has several effects of prosody (Jun, 2003), but giving a slow and deliberate speech entails greater enunciation compared with "cutting corners" to get through a hasty speech. Indeed, hastened speech is more likely to sound flat, akin to a computerized voice.

Third, Falk and Kello (2017) analyzed hierarchical temporal structure in recordings of mothers singing a song or telling a story in German to either their infants or adult confederates. Singing or speaking to their infants induced infant-directed speech, whereas doing the same for adults induced adult-directed speech. As one might expect given the prior two studies, hierarchical temporal structure was more pronounced in prosodic timescales (and again not the faster phonetic timescales) for infant-directed versus adult-directed speech (also see Chapter 38). The authors interpreted this result as indicating that mothers exaggerated prosodic information for their infants. Convergent evidence for purposeful exaggeration also came from Boorom et al. (2022) who found that adult speech directed at children diagnosed with autism spectrum disorder had greater hierarchical temporal structure compared with typically developing controls. In both cases, exaggeration may serve to grab the child's attention, and it could also serve to aid speech development and comprehension.

The latter possibility leads to the hypothesis that hierarchical temporal structure may vary as a function of the degree of prosodic composition – that is, the presence and distinctiveness of prosodic boundaries and units as expressed in the speech signal. As with prosody in general, hierarchical temporal structure is hypothesized to carry information about meaning, intention, emotion, mode of interaction, and other possible channels of communication and coordination. On this hypothesis, tapering in hierarchical temporal structure occurs when these layers are less composed.

As a further test of the composition hypothesis, Kello et al. (2017) tested whether different kinds of music might also vary in their degree of layering of temporal structure. The authors applied the same Allan Factor method of analysis to four different categories of music – contemporary pop, rap, jazz improvisation, and classical symphonies. The composition hypothesis predicts that symphonies should have the greatest degree of nested clustering because composers explicitly add layers of structure for each of the instruments. Results bore out the prediction, in that nested clustering increased steadily across all measured timescales, as in 1/f scaling, but only for classical music. By contrast, rap and popular music had bends in their Allan Factor functions that reflected their beat structure. More interestingly, jazz showed

https://doi.org/10.1017/9781009295888.024 Published online by Cambridge University Press

a degree of tapering in the longer timescales that was remarkably similar to the tapering for recordings of conversations. Jazz improvisation has been likened to a conversation among musicians (Sawyer, 2005), but heretofore there was little if any quantitative evidence for the comparison. The Allan Factor results support an underlying connection between prosody and music hypothesized in previous studies (Patel et al., 1998).

20.5 Conclusions

This chapter reviewed a growing body of evidence that speakers can control the rate of growth of temporal structure with timescales, referred to as the degree of hierarchical temporal structure. Evidence indicates that this control governs the amount of prosodic structure and composition in speech, possibly through a relative sharpening or blurring of prosodic units and boundaries. Studies show that the degree of hierarchical temporal structure reflects the difference between more hastened versus measured rates of speech (Ramirez-Aristizabal et al., 2018), more informal versus formal speech registers (Abney et al., 2014), adult-directed versus infant-directed speech (Falk and Kello, 2017), and monologues versus dialogues (Kello et al., 2017).

A similar control parameter also appears in analyses of musical recordings in which the more layered compositions of classical music have greater degrees of hierarchical temporal structure compared with contemporary popular music and other less layered forms of music (Kello et al., 2017). Taken together, these studies suggest that listeners can distinguish degrees of hierarchical temporal structure, and speakers may vary this control parameter to signal different modes of communication to interlocutors (Alviar et al., 2020), and aid speech development and comprehension (Abney et al., 2017; Boorom et al., 2022).

Hierarchical temporal structure such as classical music has been found in recordings of thunderstorms as well, and Allan Factor functions for human conversations were like those for improvisational jazz as well as vocalizations of killer whales communicating in groups. These results indicate that prosodic and hierarchical temporal structure can be composed by people, animals, or natural causes (Ravignani et al., 2019). Hierarchical temporal structure serves to situate prosodic composition and structure in the broader context of language, music, and vocal communication.

Box 20.1 Chapter Overview

Summary

Prosody is defined over several layers of hierarchical temporal structure. A growing number of studies indicate that variations in prosodic information correspond to variations in temporal clustering of acoustic energy as a function of timescale. Speakers can control hierarchical temporal structure in prosodic timescales for various reasons, including who is being addressed and under what circumstances.

https://doi.org/10.1017/9781009295888.024 Published online by Cambridge University Press

Box 20.1 (cont.)

Implications

Prosodic structure is sometimes considered as periodic, and linguistic hierarchies are often represented in terms of nested structures, but hierarchical temporal structure in the range of prosodic timescales may also be aperiodic and not necessarily nested.

Gains

Studies of hierarchical temporal structure may help to link prosody with music, animal vocalizations, and other temporal structures in nature such as 1/f scaling. They may also illuminate mechanisms of control in communication and provide a broad basis for linking measures of cortical activity with speech activity.

REFERENCES

Abney, D. H., Paxton, A., Dale, R., and Kello, C. T. (2014). Complexity matching in dyadic conversation. *Journal of Experimental Psychology: General*, 143(6), 2304–2315. https://doi.org/10.1037/xge0000021

Abney, D. H., Warlaumont, A. S., Oller, D. K., Wallot, S., and Kello, C. T. (2017). Multiple coordination patterns in infant and adult vocalizations. *Infancy*, 22(4), 514–539. https://doi.org/doi:10.1111/infa.12165

Allan, D. W. (1966). Statistics of atomic frequency standards. *Proceedings of the IEEE*, 54(2), 221–230.

Alviar, C., Dale, R., and Kello, C. (2018). The fractal structure of extended communicative performance. In T. T. Rogers, M. Rau, X. Zhu, and C. W. Kalish (Eds.), *Proceedings of the 40th Annual Meeting of the Cognitive Science Society* (pp. 1292–1297). Cognitive Science Society.

Alviar, C., Dale, R., Dewitt, A., and Kello, C. (2020). Multimodal coordination of sound and movement in music and speech. *Discourse Processes*, 57(8), 682–702. https://doi.org/10.1080/0163853X.2020.1768500

Beckman, M. E. (1996). The parsing of prosody. *Language and Cognitive Processes*, 11(1–2), 17–68. https://doi.org/10.1080/016909696387213

Bolinger, D. (1989). *Intonation and its uses: Melody in grammar and discourse*. Stanford University Press.

Boorom, O., Alviar, C., Zhang, Y., et al. (2022). Child language and autism diagnosis impact hierarchical temporal structure of parent–child vocal interactions in early childhood. *Autism Research*, 15(11), 2099–2111. https://doi.org/10.1002/aur.2804

Cummins, F. (2012). Looking for rhythm in speech. *Empirical Musicology Review*, 7(1–2), 28–35. https://doi.org/10.18061/1811/52976

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

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

https://doi.org/10.1017/9781009295888.024 Published online by Cambridge University Press

Falk, S., and Kello, C. T. (2017). Hierarchical organization in the temporal structure of infant-direct speech and song. *Cognition*, 163, 80–86. https://doi.org/10.1016/j.cognition.2017.02.017

Fry, D. B. (1958). Experiments in the perception of stress. *Language and Speech*, 1(2), 126–152. https://doi.org/10.1177/002383095800100207

Garg, K., and Kello, C. T. (2021). Efficient Lévy walks in virtual human foraging. *Scientific Reports*, 11(1), 5242. https://doi.org/10.1038/s41598-021-84542-w

Giraud, A.-L., and Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience*, 15(4), 511–517. http://dx.doi.org/10.1038/nn.3063

Goh, K.-I., and Barabási, A.-L. (2008). Burstiness and memory in complex systems. *Europhysics Letters*, 81(4), 48002.

Goldsmith, J. A. (1990). *Autosegmental and metrical phonology*. Basil Blackwell.

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

Goswami, U. (2019). Speech rhythm and language acquisition: An amplitude modulation phase hierarchy perspective. *Annals of the New York Academy of Sciences*, 1453(1), 67–78. https://doi.org/10.1111/nyas.14137

Hayes, B. (1989). The prosodic hierarchy in meter. In P. Kiparsky and G. Youmans (Eds.), *Rhythm and meter* (pp. 201–260). Elsevier.

Jun, S.-A. (2003). The effect of phrase length and speech rate on prosodic phrasing. Proceedings of the XVth International Congress of Phonetic Sciences, ICPhS-15, pp. 483–486.

Kello, C. T., Bella, S. D., Médé, B., and Balasubramaniam, R. (2017). Hierarchical temporal structure in music, speech and animal vocalizations: Jazz is like a conversation, humpbacks sing like hermit thrushes. *Journal of The Royal Society Interface*, 14(135), 20170231. https://doi.org/doi:10.1098/rsif.2017.0231

Kello, C. T., Bhat, H., Turner, M., and Alviar, C. (2024). Hierarchical timescales and nested process composition. PsyArXiv Preprints. https://doi.org/10.31234/osf.io/tqyrz

Kello, C. T., Brown, G. D. A., Ferrer-i-Cancho, R., et al. (2010). Scaling laws in cognitive sciences. *Trends in Cognitive Sciences*, 14(5), 223–232.

Krivokapić, J. (2014). Gestural coordination at prosodic boundaries and its role for prosodic structure and speech planning processes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1658), 20130397. https://doi.org/doi:10.1098/rstb.2013.0397

Langus, A., Marchetto, E., Bion, R. A. H., and Nespor, M. (2012). Can prosody be used to discover hierarchical structure in continuous speech? *Journal of Memory and Language*, 66(1), 285–306. https://doi.org/10.1016/j.jml.2011.09.004

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

Leong, V., Kalashnikova, M., Burnham, D., and Goswami, U. (2017). The temporal modulation structure of infant-directed speech. *Open Mind*, 1(2), 78–90. https://doi.org/10.1162/OPMI_a_00008

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.13748

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

Patel, A. D. (2003). Language, music, syntax and the brain. *Nature Neuroscience*, 6(7), 674–681.

Patel, A. D., Peretz, I., Tramo, M., and Labreque, R. (1998). Processing prosodic and musical patterns: A neuropsychological investigation. *Brain and Language*, 61(1), 123–144. https://doi.org/10.1006/brln.1997.1862

https://doi.org/10.1017/9781009295888.024 Published online by Cambridge University Press

Paxton, A., and Dale, R. (2013). Argument disrupts interpersonal synchrony. *Quarterly Journal of Experimental Psychology*, 66(11), 2092–2102. https://doi.org/10.1080/17470218.2013.853089

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

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

Port, R. (2008). All is prosody: Phones and phonemes are the ghosts of letters. Proceedings of the Fourth International Conference of Speech Prosody 2008, pp. 7–13. https://doi.org/10.21437/SpeechProsody.2008-1

Ramirez-Aristizabal, A. G., Médé, B., and Kello, C. T. (2018). Complexity matching in speech: Effects of speaking rate and naturalness. *Chaos, Solitons & Fractals*, 111, 175–179. https://doi.org/10.1016/j.chaos.2018.04.021

Ravignani, A., Dalla Bella, S., Falk, S., et al. (2019). Rhythm in speech and animal vocalizations: A cross-species perspective. *Annals of the New York Academy of Sciences*, 1453(1), 79–98. https://doi.org/10.1111/nyas.14166

Salvucci, D. D., and Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. Proceedings of the 2000 Symposium on Eye Tracking Research & Applications, Palm Beach Gardens, Florida, USA, Association for Computing Machinery, pp. 71–78. https://doi.org/10.1145/355017.355028

Sawyer, R. K. (2005). Music and conversation. In D. Miell, R. MacDonald, and D. J. Hargreaves (Eds.), *Musical communication* (Vol. 45) (pp. 45–60). Oxford University Press.

Schiffrin, D. (1994). Making a list. *Discourse Processes*, 17(3), 377–406.

Schneider, S., Ramirez-Aristizabal, A. G., Gavilan, C., and Kello, C. T. (2020). Complexity matching and lexical matching in monolingual and bilingual conversations. *Bilingualism: Language and Cognition*, 23(4), 845–857. https://doi.org/10.1017/S1366728919000774

Shattuck-Hufnagel, S., and Turk, A. E. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research*, 25(2), 193–247. https://doi.org/10.1007/BF01708572

Singh, N. C., and Theunissen, F. E. (2003). Modulation spectra of natural sounds and ethological theories of auditory processing. *Journal of the Acoustical Society of America*, 114(6), 3394–3411. https://doi.org/10.1121/1.1624067

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

Tilsen, S., and Arvaniti, A. (2013). Speech rhythm analysis with decomposition of the amplitude envelope: Characterizing rhythmic patterns within and across languages. *Journal of the Acoustical Society of America*, 134(1), 628–639. https://doi.org/10.1121/1.4807565

Voss, R. F., and Clarke, J. (1975). "1/f" noise in music and speech. *Nature*, 258, 317–318.

Wennerstrom, A. (2001). *The music of everyday speech: Prosody and discourse analysis*. Oxford University Press.

Yaeger-Dror, M. (2002). Register and prosodic variation: A cross language comparison. *Journal of Pragmatics*, 34(10–11), 1495–1536.

https://doi.org/10.1017/9781009295888.024 Published online by Cambridge University Press