Speech, Rhythm, and the Brain

Steven Greenberg

Introduction

Think of "rhythm" and what most likely comes to mind are music and dance. We intuitively "know" what good rhythm is, especially when it comes to entertainment. Indeed, rhythm is vital for the expression of emotion in the arts. But what often goes unappreciated is that rhythm also plays an important role in various forms of acoustic signaling, including spoken language and non-human communication (Kotz et al., 2018).

What is it about rhythm that accounts for its prevalence across the animal kingdom (Ravignani et al., 2019)? And why is it especially important for human communication?

Although definitive answers lie outside the scope of the present discussion, several of these issues are examined here through the lens of speech acoustics, perception, and neuroscience. It is argued that rhythm lies at the very heart of what makes humans especially adept at communication, binding sensory signals across modalities and linking such input with internal, often rhythmic, neural activity in the brain.

What Rhythm Is

For illustrative purposes, I begin our survey by examining rhythm from a *musical* perspective, distinguishing between two "flavors" of rhythm, the "cognitive" and the "physical." Cognitive rhythm is associated with musical elements like notes, accents, beats, measures, and phrases. Physical rhythmic elements are intensity, duration, interval, and modulation.

Musicologists have traditionally viewed rhythm as operating on a sequence of perceptual elements: "Rhythm may be defined as the way in which one or more unaccented beats are grouped in relation to an accented one... A rhythmic group can be apprehended only when its elements are distinguished from one another, rhythm... always involves an interrelationship between a single, accented (strong) beat and either one or two unaccented (weak) beats" (Cooper and Meyer, 1960, p. 6).

Within this cognitive framework, rhythm is deemed a *relational* property, one that governs how elements (e.g., musical notes, measures, phrases) interact with each other perceptually and cognitively. Such operations likely involve widespread communication across a constellation of brain centers associated with the senses, memory, and movement.

But rhythm doesn't function simply as a relational quality: "…rhythm is the one indispensable element of all music. Rhythm can exist without melody, as in the drumbeats of so-called primitive music, but *melody cannot exist without rhythm*. In music that has both harmony and melody, the rhythmic structure cannot be separated from them" (emphasis added) (Crossley-Holland, 1998; 2002; 2020).

In other words, rhythm serves as a unifying, global function, integrating different musical elements into a perceptual experience greater than the sum of its constituent parts. Precisely how rhythm performs this cognitive "magic" is not well understood (Ding et al., 2017). One possibility is that certain key physical elements of musical and speech rhythm "trigger" endogenous synchronous activity, or neural "oscillations," associated with the encoding and retrieval of information pertinent to a variety of sensory and cognitive experiences discussed in **Speech Rhythms in the Brain**.

Speech Rhythm and Linguistic Representations

What pertains to music also applies to speech; however, the specifics differ. Talkers do not generally speak in musical notes or measures, although poetic rhythm can reenforce emotion or be used to conjure imagery and scenarios (see youtu.be/S0mwhkv9ves for an online discussion) (Obermeier et al., 2013).

https://doi.org/10.1121/AT.2022.18.3.22

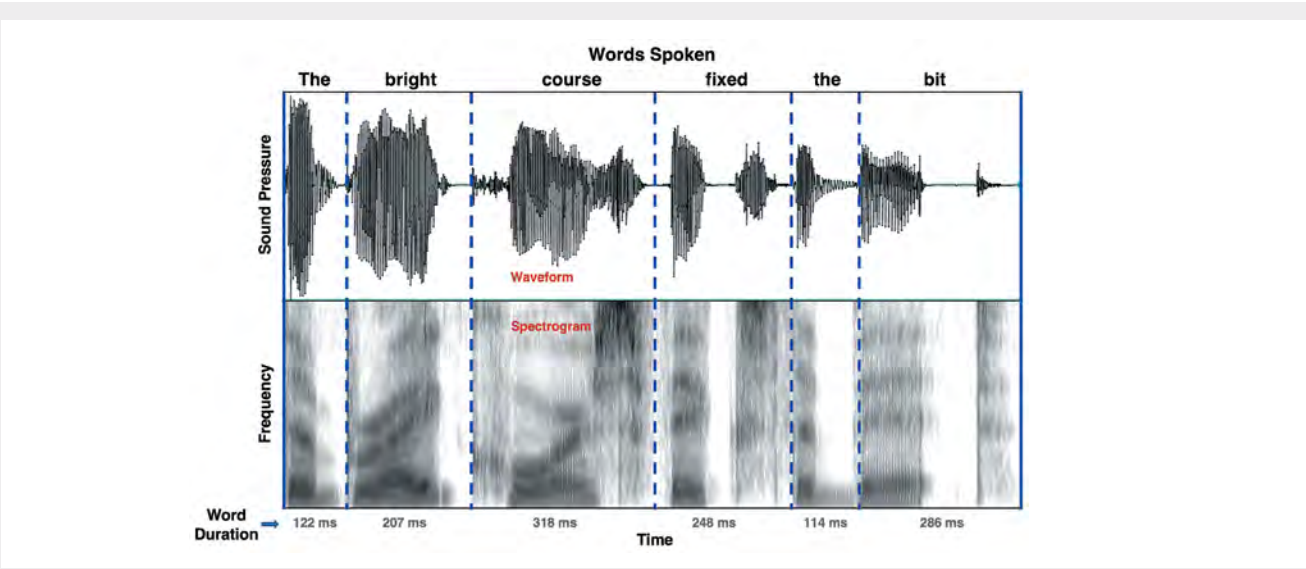


Figure 1. *The speech waveform (**top**) and associated spectrogram (**bottom**) for a sample speech signal. The words spoken are indicated above the waveform, which consists of both fast and slow modulations. The slower ones reflect syllabic and segmental rhythms.* ***Dotted vertical blue lines*** *separate adjoining words (which are also single syllables). Their durations are shown below the spectrogram.*

Here I examine how linguistic elements such as "phonetic segments," syllables, words, phrases, and sentences are impacted by the cognitive form of rhythm ("prosody") as well as by several physical attributes: modulation, phase, duration, frequency, and intensity.

Rhythm's physical form can be visualized via the acoustic signal's waveform (**Figure 1**). It contains both fast (i.e., higher frequency) and slow (i.e., very low frequency) sound pressure fluctuations. The fast modulations, the "temporal fine structure," are often associated with pitch and other tonal properties (Smith et al., 2002) but lie outside the scope of the present discussion.

Rhythm is reflected in the very slow modulations in the waveform, known as the "speech envelope," and in the motion of the speech articulators (Tucker and Wright, 2020), especially the opening and closing of the jaw, as well as the movement of the lips and tongue during speech production (Stevens, 1998). These parallel movements are the acoustic expression of speech rhythm. There is also a highly visible component, the so-called "speech-reading" cues associated with the articulatory movements that interact with certain elements of the acoustic signal to produce perceptual "objects" at the phonetic (van Wassenhove et al., 2007) and lexical (Winn, 2018) levels. The interaction between the audio and visual speech signals, especially under challenging listening conditions, shields the speaker's message from the deleterious impact of background noise and other forms of acoustic interference (Assmann and Summerfield, 2004), something especially important for the hearing impaired.

The Dynamics of Rhythm

The motion of the articulators, especially the jaw, establishes the upper and lower bounds of the speech envelope's energy swings. These slow articulatory movements largely coincide with the linguistic element known as the "syllable." Although a syllable may contain just a single phonetic segment (e.g., "a") or as many as seven (e.g., "strengths"), most syllables contain just two or three (Greenberg, 1999). Although the average duration of a syllable is about 200 ms in American English (Greenberg, 1999) and 165 ms in Japanese (Arai and Greenberg, 1997), their length can vary from about 100 ms to about 330 ms. Such durational properties are important for the next discussion because they can also be expressed in terms of "modulation frequency," a key quantitative metric for representing speech rhythms across a range of temporal scales and is also important for speech intelligibility (the ability to decode and understand the words spoken in a phrase, sentence, or longer utterance).

In modulation-frequency units, syllables range between 3 Hz (for long-duration examples) and 10 Hz (for short-duration examples) (**Figure 2**). Syllables form the backbone of speech's modulation spectrum, a reflection of the articulatory dynamics associated with the opening and closing of the jaw during speaking, which modulates the amplitude of the acoustic signal. The intensity of a speech sound is closely related to the aperture of the oral cavity. More energy is released during the vocalic portion of the syllable when the opening is wide, whereas much less energy is released when the aperture is reduced during the production of (most) consonants. Hence, one can liken a syllable's waveform to an "energy arc" (Greenberg, 2006) where there are rises and falls in energy that closely follow the amplitude characteristics of the individual phonological constituents within a syllable. This is illustrated for the two-syllable word "seven" in a three-dimensional representation of the speech signal called a "spectro-temporal profile" (STeP; **Figure 3**). The STeP shows the energy dynamics of the speech signal using hundreds of instances of the same word that have been averaged to derive a composite representation (Greenberg et al., 2003).

Waveform modulations are also associated with a syllable's constituent phonetic segments (or "phones"), ranging in duration between ca. 50 ms (20 Hz) and ca. 150 ms (7 Hz). These faster undulations are nested within

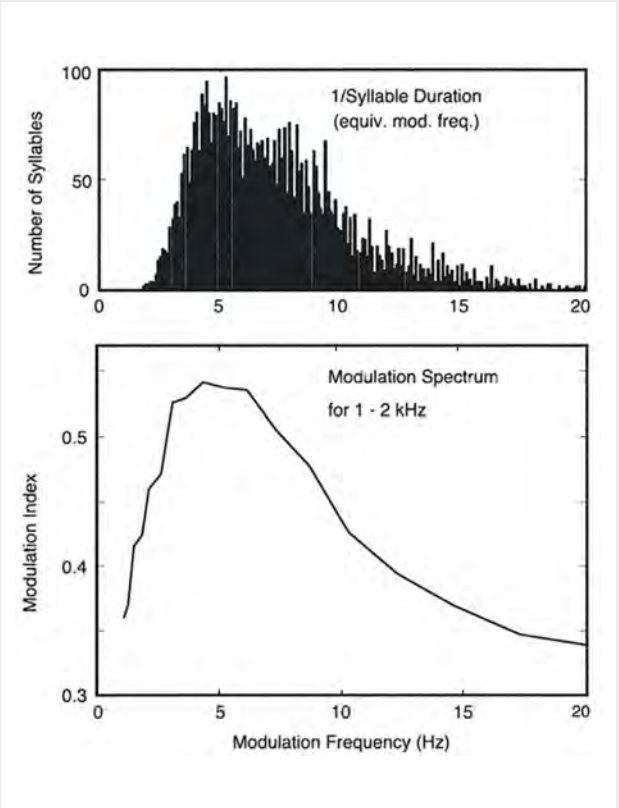


Figure 2. *The relationship between the distribution of syllable duration (transformed into equivalent modulation frequency [equiv. mod. freq.] units) (**top**) and the modulation spectrum of the same material (Japanese spontaneous speech) (**bottom**) as calculated for the octave region between 1 and 2 kHz.*

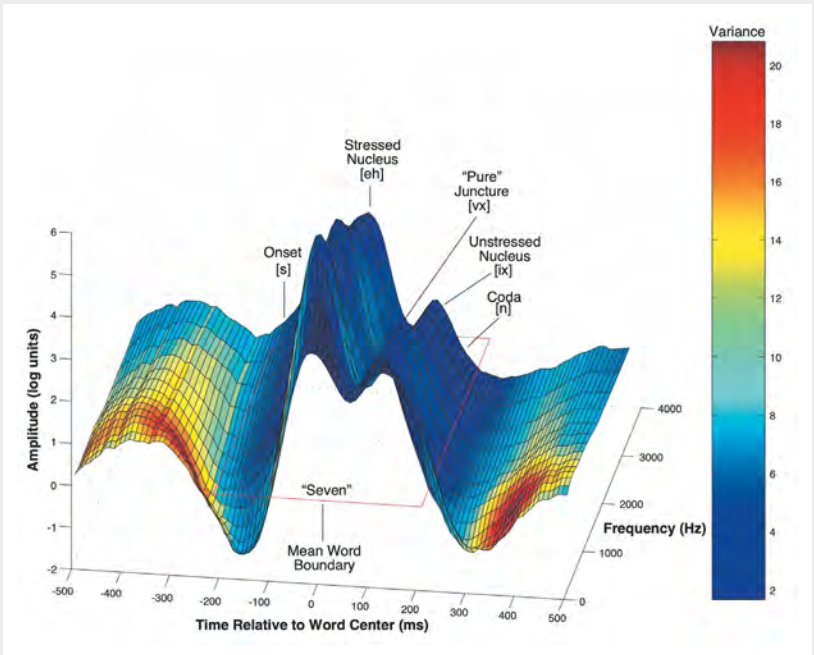


Figure 3. *A spectro-temporal profile (STeP) of the word "seven," a normalized averaging of hundreds of instances from the OGI Numbers corpus. The STeP shows the signal modulation patterns associated with the onset (s), nucleus, and coda (n) constituents of two syllables, the first stressed (eh) and the second unstressed (ix), to highlight the waveform dynamics of the spoken material. The pure juncture lies in the trough between the stressed and unstressed vocalic nuclei.*

syllabic modulations, imparting a phonetic detail required to achieve lexical clarity and semantic precision.

Fluctuations on a longer timescale than the syllable are often referred to as "prosodic," although there may be modulatory patterns within a syllable that are also of prosodic significance. These prosodic patterns are reflected in the modulation spectrum's lower limb (<3 Hz). Perceptually, these very low frequency modulations are instantiated in a syllable's *prominence* relative to neighboring syllables in a word, phrase, or sentence. These emphasized syllables are "accented" or "stressed" (Beckman, 1992). The intensity and duration of a syllable's vocalic core (known as the "nucleus") relative to nearby nuclei are the most important physical attributes of prominence (Silipo and Greenberg, 1999), although other physical properties play a role and have been incorporated into an automatic prosodic prominence labeling system, AutoSAL (Greenberg, 2005, Fig. 11).

It is not just the energy within an utterance that varies, but also its fundamental frequency (f_0; "pitch") contour. Such pitch variation may mark the transition from one grammatical phrase to another (tone and break indices [ToBI]; Silverman et al., 1992), helping the listener parse the speech signal for better comprehension. A computational version (AuToBI) uses pitch contour patterns as well as syllable duration and intensity to parse utterances (Rosenberg, 2010).

How important are speech rhythm and slow waveform modulations for intelligibility? As early as 1939, a Bell Labs engineer, Homer Dudley, recognized the importance of slow modulations for creating intelligible speech with his invention of the "vocoder" (Dudley, 1939). He distinguished between the fast-moving "carrier" (i.e., the spectro-temporal "fine structure") and the more slowly moving "modulator," making it clear that both are essential for creating intelligible speech (Bunnell, 2022). A vocoder consists of a series of band-pass filters, simulating the frequency analysis of the auditory system used to create a perceptual model of the speech signal that is more compact than the original. Modern-day applications of the vocoder are found in a variety of text-to-speech applications (Kawahara, 2015) and have been fine tuned to create much more natural sounding speech than Dudley's (1939) original version.

Dudley's (1939) insight received renewed interest in the 1970s when an automated system was developed for predicting intelligibility in acoustic environments like concert halls, theaters, and worship spaces (Houtgast and Steeneken, 1973). Key to the system's success was a method for quantifying the amount of energy in each frequency channel of the very slowly moving modulations in the speech signal. Houtgast and Steeneken dubbed their metric the "modulation spectrum" because it quantified the amount of energy in each frequency channel of modulation. In this context, "frequency channel" refers to a unit of time considerably longer (50 ms to 2 s) than the temporal units associated with tonal spectral audibility (50 μs to 20 ms) in human listeners. Houtgast and Steeneken noted that the contour of the modulation spectrum could be used to distinguish intelligible from unintelligible speech, especially in noisy and reverberant environments.

Why does the modulation spectrum's profile predict speech intelligibility so well? An intuitive explanation is that speech energy (i.e., acoustic "reflections") added back to the speech signal with a certain delay smooths the contours of the slow modulations in ways that degrade critical linguistic information within the syllable. The waveform modulations containing critical phonetic cues are no longer crisply defined, thereby compromising a listener's ability to extract sufficient phonetic detail to decode and interpret the speech signal. This intuition is consistent with a study by Drullman et al. (1994), who low-pass filtered the slow modulations using a procedure that "smeared" (i.e., "blurred") the boundaries between adjacent syllables, thereby distorting speech-relevant information in the modulation "packets." **Figure 4** shows how intelligibility declines as the *complex* modulation spectrum diminishes in amplitude (Greenberg and Arai, 2004).

The modulation spectra of intelligible speech material exhibit a peak between 4 and 8 Hz, the key range for syllabic information. There is also complex modulation energy between 8 and 16 Hz, the region most closely associated with phonetic segmental cues. The lower branch of the modulation spectrum (<4 Hz) is associated with highly prominent (i.e., accented) syllables. When the intelligibility decreases, so does the amount of energy in the modulation spectrum, especially in the critical 4- to 8-Hz region.

The importance of slow modulations for intelligibility was demonstrated in a different way by Shannon et al. (1995). In place of a conventional speech waveform with

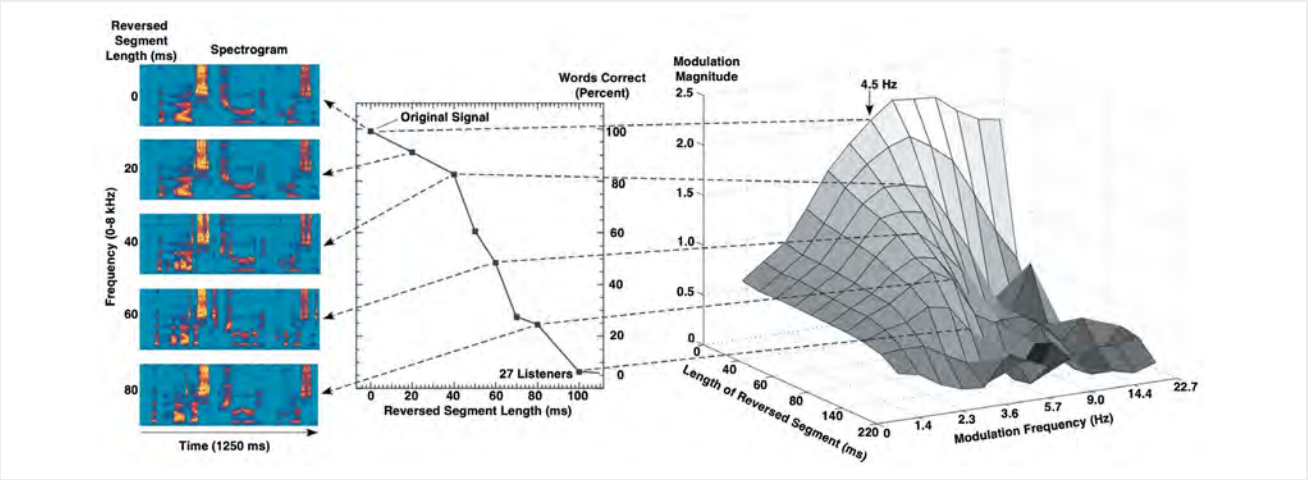


Figure 4. *The relationship between the complex magnitude of the modulation spectrum and speech intelligibility. The complex modulation spectrum integrates the magnitude and phase components into a single value. The sentence material's intelligibility was manipulated by locally time- reversing the speech signal over different segment lengths. As the reversed-segment duration increases beyond 40 ms, intelligibility declines precipitously, as does the magnitude of the complex modulation spectrum. The spectro-temporal properties also deteriorate appreciably under such conditions. Reprinted from Greenberg and Arai (2004).*

its harmonic (i.e., "voiced") structure, the carrier signal used was white noise. But the modulator of the original speech signal was retained, used to modulate the white-noise carrier in ways reminiscent of a coarse-grained spectrum analyzer. These slow modulations vary depending on whether they are derived from the low-, mid-, or high-frequency region of the speech signal's acoustic spectrum. Shannon et al. (1995) discovered that intelligible speech was only possible if the slow modulations were combined across different regions of the acoustic frequency spectrum. In other words, *a diverse set of slow modulators was required to preserve the linguistic information* contained in the original signal when transformed into a vocoded, noise-excited version.

To summarize, these pioneering studies demonstrated that low-frequency modulations in the speech waveform convey information critical to intelligibility. But these early demonstrations left unaddressed a variety of questions regarding how such information unlocks neurological pathways involved in speech comprehension and understanding.

It was at this point that my colleagues and I performed several studies to shed more light on *why* these slow modulations figure so importantly in speech perception. We asked five basic questions.

Q1: How much can the slow modulations be perturbed without impacting intelligibility?

Q2: Does the modulation spectrum vary across acoustic (and hence auditory) frequency?

Q3: How do the modulations across the acoustic frequency spectrum interact with each other?

Q4: Can the visual modality interact with the auditory speech signal to provide a measure of redundancy?

Q5: Can such perceptual data be linked to such linguistic "objects" as the phonetic segment, syllable, and word?

Our studies showed that

A1: The slow modulations can be perturbed only to a limited degree without seriously compromising intelligibility (Greenberg and Arai, 2004).

A2: The modulation spectrum does indeed vary appreciably across the acoustic frequency spectrum (Greenberg et al., 2003).

A3: Intelligibility is moderately sensitive to the *phase* (i.e., timing) of the slow modulations across acoustic frequency (Greenberg and Arai, 2004).

A4: Visual speech rhythmic patterns do interact with the acoustic signal but in an asymmetrical way. Intelligibility is far more tolerant of audiovisual asynchrony when the visual component leads the audio rather than vice versa (Grant and Greenberg, 2001).

A5: Different parts of the modulation spectrum are associated with distinct linguistic elements (Greenberg et al., 2003). Highly prominent (i.e., accented, stressed) syllables are associated with the lower limb (3-5 Hz) of the modulation spectrum, whereas less prominent syllables are associated with its upper limb (6-8 Hz). A syllable's prominence influences the phonetic realization of both consonant and vocalic segments (Greenberg, 2005).

Speech Rhythm in Broader Perspective

Prosody's power to connect with listeners is well-known to those engaged in entertainment, politics, or preaching. Much of this emotive force is an embodiment of specific properties of the speech signal, especially the emphasis placed on specific syllables (and words) and their timing relative to their less prominent counterparts. Such rhythms can help the listener navigate the speech stream to separate the semantic "wheat" from the "chaff."

Parsing and chunking speech are important for other reasons too. Speech is inherently ambiguous. Listen to a short snippet of, say, a single syllable or word and try to guess what the speaker is saying. This experiment was performed nearly 60 years ago by Pickett and Pollack (1963) and Pollack and Pickett (1964) who found that several words in succession (ca. 1 s) were required to reliably recognize the words in both read and conversational speech (**Figure 5**). This ambiguity of individual linguistic elements places a premium on predicting which sounds and words are likely to follow. Human listeners routinely do this, and recent brain-imaging data show that parts of the frontal cortex, especially the prefrontal region, are heavily involved (Park et al., 2015). This is where speech rhythm may play an especially important role because it serves as an organizing, active framework for critical cerebral centers to "latch on" to relevant neural activity in various parts of the brain (Schroeder et al., 2010).

However, speech rhythm entrails more than syllabic prominence and pitch contours. The emotional valence of what is said can be equally (if not more) important. The prosody of emotion has long been the subject of study and speculation (Scherer, 2003). Darwin (1873) was perhaps the first to suggest that emotion has deep roots in our phylogenetic history and that primitive elements of prosody can be found in vocalizations of certain nonhuman species (Ravignani et al., 2019).

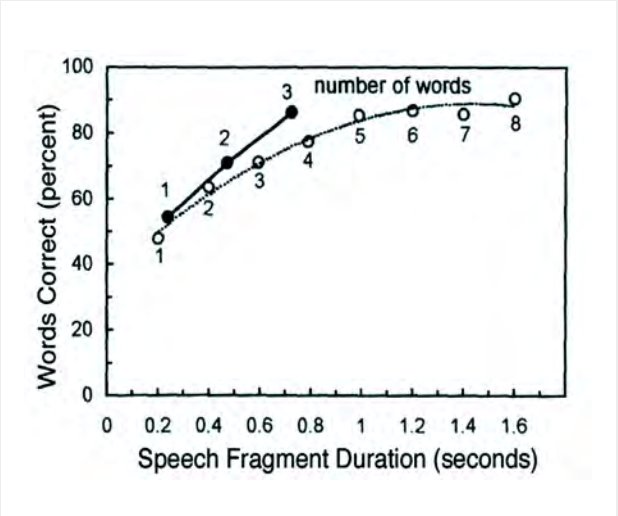


Figure 5. *Average identification score of words in fragments excised from read text (**closed circles**) and conversational speech (**open circles**) as a function of fragment duration. Based on data from Pickett and Pollack (1963) and Pollack and Pickett (1963) and adapted from Plomp (2002, p. 107).*

From this global perspective, rhythm can serve as both a mediating and unifying force. It acts as a mediator between the lower level, physical and sensory tiers and the higher, cognitive levels associated with semantic and situational analysis and interpretation (Hawkins, 2014). Rhythm is a unifier in that it combines what might otherwise be just an assortment of unrelated acoustic elements (e.g., harmonics and other frequency components) and groups them together to create sensory objects capable of signaling words, phrases, and concepts (Elhilali, 2019).

Speech Rhythms in the Brain

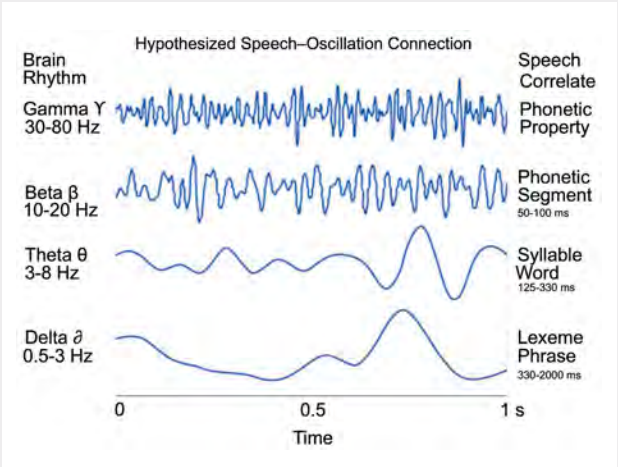
How is speech processed by the brain? Are exogenous signal properties, such as syllabic modulations, linked to endogenous neural activity associated with linguistic functions like phonetic analysis, word recognition, and semantic interpretation (Zhang and Ding, 2017)? Can neurological investigations elucidate not only the pertinent brain mechanisms (Friederici, 2011) but also shed light on acoustic biocommunication in nonhuman

species (Ravignani et al., 2019)? Possibly, as there are acoustic properties shared across many species, and the way different parts of the brain communicate with each other also appears to be similar across much of the animal kingdom.

It has long been known that low-frequency rhythms can be recorded from the scalp of human subjects (Berger, 1929), although their significance remained unclear for many decades. Over the past 30 years, many different brain rhythms have been studied (Buzsaki, 2006). They range in frequency from the relatively fast, gamma-ϒ (ca. 30-80 Hz) to the very slow, delta-δ (0.5-3 Hz) and to points in between: theta-θ (3-8 Hz), alpha-α (8-10 Hz), and beta-β (10-20 Hz) (**Figure 6**).

These brain rhythms are not perfectly periodic but fluctuate around an average frequency, with energy spanning a range of spectral components. But the rhythms of speech also deviate from lockstep periodicity. When rhythm is studied, it is the central tendency rather than a metronomic pattern that forms the focus of analysis. Thus, it is not surprising that rhythms internal to the brain don't follow a strict timetable but rather largely reflect synchronous, endogenous communication among cerebral regions pertinent to the behavior at hand.

Figure 6. *Hypothesized relationship between brain rhythms and speech processing over a range of timescales and neural oscillation frequencies. These neural oscillations reflect the synchronous activity of thousands (or millions) of neurons in the cerebral cortex and hippocampus responding to sensory stimulation (in this example, an acoustic speech signal). The different timescales of the oscillations are hypothesized to match the timescales of linguistic elements thought to be important for decoding and understanding the speech signal. Waveforms shown are solely for illustrative purposes.*



One of these endogenous rhythms, theta–θ, closely emulates the timing of syllables in spoken material (Poeppel and Assaneo, 2020) and may be "entrained" (i.e., extremely synchronized) to the signal's syllabic modulations. There are brain rhythms whose temporal properties are comparable to other linguistic elements, both shorter (phonetic segments; beta-β [50-100 ms, 10-20 Hz]) and longer (phrases; delta-δ [300-2,000 ms, 3-5 Hz]) than syllables (Etard and Reichenbach, 2019).

What is the significance of these neural oscillations? Are they merely tracking the signal's physical properties or do they reflect deeper processing germane to the analysis and interpretation of the linguistic message? Some studies indicate a marked entrainment to the waveform's syllabic rhythm on a cycle-by-cycle basis based on a variety of recording methods including electroencephalography (Fujii and Wan, 2014), magnetoencephalography (Poeppel and Assaneo, 2020), and electrocorticographic (Oganian and Chang, 2019).

However, there is evidence that at least some of these oscillations are linked to more profound processing, persisting well after cessation of the signal (van Bree et al., 2021). Several brain recording studies suggest that speech comprehension is most closely reflected in beta–β oscillations (Pefkou et al., 2017) and that a talker's speaking rate is faithfully reflected in theta–θ oscillations, which have also been linked to attentional processes (Fiebelkorn and Kastner, 2019) and is associated with parsing the speech into meaningful "chunks." Endogenous rhythms may also enhance the forecasting of phonetic segments, syllables, words, and phrasal structure in conversational speech. But how this linguistic "magic" is achieved is not currently well understood.

Rhythm in Developmental Perspective

Language, both spoken and written, requires time to acquire and master even by native speakers. It may not be until the age of 11 years that the child's linguistic prosody is fully formed (Polyanskaya and Ordin, 2015). Is it possible that the acquisition and mastery of a language depends on learning its rhythm? And if there is

some flaw in this skill might this deficit impair linguistic competence, at least for native speakers? There is some evidence that this is indeed the case, both in speech production (Fujii and Wan, 2014) and in reading (Leong and Goswami, 2014). Perhaps rhythm is a foundational property, one that holds the key to understanding language's neural bases.

The "Why" of Rhythm

Virtually all animals move, and such locomotion involves rhythmic motor activity, posing a challenge for sensory systems tasked with maintaining the illusion of stability for constantly changing stimuli. One way in which the brain can navigate such sensorimotor dynamics is through rhythmic patterns of neural activity (Lubinas et al., 2022) that submerge the intrinsic variability of sensory signals within nested hierarchies of cortical oscillations (Ghitza, 2011) that "translate" lower level features into more global, complex features of variable duration and cognitive complexity (Greenberg, 2011). Consistent with this perspective is a study that artificially distorted the rhythm of spoken sentences to disrupt intelligibility over the temporal range in which theta–θ oscillations are thought to operate (Ghitza and Greenberg, 2009). Perhaps the temporal patterning of spoken (and other forms of) communication evolved to "piggyback" on intrinsic rhythms of the brain (Kotz et al., 2018).

Rhythm's Future

Rhythm played a supporting role in the study of spoken language for most of the twentieth century, its importance only coming to the fore in the 1990s as perceptual and statistical studies highlighted rhythm's centrality for speech intelligibility and understanding. In recent years, this recognition has played a key role in integrating rhythm into speech synthesis technology to create more natural-sounding material (Bunnell, 2022) as well as incorporating rhythm into automatic speech-recognition models. Speech rhythm has also begun to be used in speech rehabilitation (Fujii and Wan, 2014), in foreign language instruction (Greenberg, 2018), and as an adjunct for teaching kids to read. And rhythm is now the center of attention for evolutionary studies of animal communication and its importance for the evolution of human language (Ravignani et al., 2019). The science of rhythm is in its infancy and is likely to provide further insights into language and other aspects of human behavior for years to come.

Acknowledgments

I thank the following colleagues for collaborating on the research mentioned in this article: Takayuki Arai, Hannah Carvey, Shuangyu Chang, Oded Ghitza, Jeff Good, Ken Grant, Leah Hitchcock, Joy Hollenback, Rosaria Silipo, and Thomas Christiansen. The research was funded in part by the National Science Foundation and the US Department of Defense. Special thanks to Arthur Popper for offering sage editorial advice on the preliminary versions of this article.

References

Arai, T., and Greenberg, S. (1997). The temporal properties of spoken Japanese are similar to those of English. *Proceedings of the Fifth European Conference on Speech Communication and Technology (Eurospeech 1997)*, Rhodes, Greece, September 22-25, 1997, pp. 1011-1014.

Assmann, P., and Summerfield, Q. (2004). The perception of speech under adverse conditions. In Greenberg, S., Ainsworth, W. A., Popper, A. N., and Fay, R. R. (Eds.), *Speech Processing in the Auditory System*. Springer, New York, NY, pp. 231-308.

Beckman, M. E. (1992*). Stress and Non-Stress Accent*. Fortis, Dordrecht, The Netherlands.

Berger, H. (1929). Über das Elektrenkephalogramm des Menschen. *Archiv für Psychiatrie und Nervenkrankheiten* 87, 527-570. https://doi.org/10.1007/BF01797193.

Bunnell, H. T. (2022). Speech synthesis: Toward a "voice" for all. *Acoustics Today* 18(1), 14-22. https://doi.org/10.1121/AT.2022.18.1.14.

Buzsaki, G. (2006). *Rhythms of the Brain*. Oxford University Press, New York, NY.

Cooper, G., and Meyer, L. B. (1960). *The Rhythmic Structure of Music*. University of Chicago Press, Chicago, IL.

Crossley-Holland, P. (1998/2002/2020). Rhythm. *Encyclopedia Britannica*. Available at https://www.britannica.com/art/rhythm-music.

Darwin, C. (1871). *The Expression of Emotions in Man and Animals*, 4th ed. Oxford University Press, Oxford, UK.

Ding N., Patel, A. D., Chen, L., Butler, H., Luo, C., and Poeppel, D. (2017). Temporal modulations in speech and music. *Neuroscience and Biobehavioral Reviews* 18, 181-187. https://doi.org/10.1016/j.neubiorev.2017.02.011.

Drullman, R., Festen, J. M., and Plomp, R. (1994). Effect of temporal envelope smearing on speech reception. *The Journal of the Acoustical Society of America* 95, 1053-1064. https://doi.org/10.1121/1.408467.

Dudley, H. (1939). Remaking speech. *The Journal of the Acoustical Society of America* 11, 169-177.

Elhilali, M. (2019). Modulation representations for speech and music. In Siedenburg, K., Saitis, C., McAdams, S., Popper, A. N., and Fay, R. R. (Eds.), *Timbre: Acoustics, Perception, and Cognition*, Springer, Cham, Switzerland, pp. 335-359. https://doi.org/10.1007/978-3-030-14832-4_12.

Etard, O., and Reichenbach, T. (2019). Neural speech tracking in the theta and in the delta frequency band differentially encode clarity and comprehension of speech in noise. *The Journal of Neuroscience* 39, 5750-5759. https://doi.org/10.1523/JNEUROSCI.1828-18.2019.

Fiebelkorn, I. C., and Kastner, S. (2019). A rhythmic theory of attention. *Trends in Cognitive Sciences* 23(2), 87-101. https://doi.org/10.1016/j.tics.2018.11.009.

Friederici, A. D. (2011). The brain basis of language processing: From structure to function. *Physiological Reviews* 91(4), 1357-1392. https://doi.org/10.1152/physrev.00006.2011.

Fujii, S., and Wan, C. Y. (2014). The role of rhythm in speech and language rehabilitation: The SEP hypothesis. *Frontiers of Human Neuroscience* 8, 777. https://doi.org/10.3389/fnhum.2014.00777.

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

Ghitza, O., and Greenberg, S. (2009). On the possible role of brain rhythms in speech perception: Intelligibility of time-compressed speech with periodic and aperiodic insertions of silence. *Phonetica* 66, 113-126. https://doi.org/10.1159/000208934.

Grant, K.W., and Greenberg, S. (2001). Speech intelligibility derived from asynchronous processing of auditory-visual information. *Proceedings of the International Conference on Auditory-Visual Speech Processing,* Aalborg, Denmark, September 7-9, 2001, pp. 132-137.

Greenberg, S. (1999). Speaking in shorthand — A syllable-centric perspective for understanding pronunciation variation. *Speech Communication* 29, 159-176.

Greenberg, S. (2005). From here to utility — Melding phonetic insight with speech technology. In Barry, W., and Domelen, W. (Eds.), *Integrating Phonetic Knowledge with Speech Technology*. Kluwer, Dordrecht, The Netherlands.

Greenberg, S. (2006) A multi-tier framework for understanding spoken language. In Greenberg, S., and Ainsworth, W. A. (Eds.), *Listening to Speech: An Auditory Perspective.* Lawrence Erlbaum, Mahwah, NJ, pp. 411-434.

Greenberg, S. (2011). Speak, memory — Wherefore art thou, invariance? *The Journal of the Acoustical Society of America* 130, 2374. https://doi.org/10.1121/1.3654514.

Greenberg, S. (2018) Deep language learning. *Acoustics Today* 14(1), 19-27.

Greenberg, S., and Arai, T. (2004). What are the essential cues for understanding spoken language? *IEICE Transactions on Information Systems* 87-D, 1059-1070.

Greenberg, S., Carvey, H., Hitchcock, L., and Chang, S. (2003). Temporal properties of spontaneous speech–a syllable-centric perspective. *Journal of Phonetics* 31, 465-485. https://doi.org/10.1016/j.wocn.2003.09.005.

Hawkins, S. (2014). Situational influences on rhythmicity in speech, music, and their interaction. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences* 369(1658), 20130398. https://doi.org/10.1098/rstb.2013.0398.

Houtgast, T., and Steeneken, H. (1973). The modulation transfer function in room acoustics as a predictor of speech intelligibility. *Acta Acustica* 28, 66-73.

Kawahara, H. (2015). Temporally variable multi-attribute morphing of arbitrarily many voices for exploratory research of speech prosody. In Hirose, K., and Tao, J. (Eds.), *Speech Prosody in Speech Synthesis: Modeling and Generation of Prosody for High Quality and Flexible Speech Synthesis.* Springer-Verlag, Berlin. Heidelberg, Germany. https://doi.org/10.1007/978-3-662-45258-5_8.

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

Leong, V., and Goswami, U. (2014). Impaired extraction of speech rhythm from temporal modulation patterns in speech in developmental dyslexia. *Frontiers in Human Neuroscience* 8, 96. https://doi.org/10.3389/fnhum.2014.00096.

Lubinas, C., Keitel, Oblesser, J., Poeppel, D., and Rimmele, J. (2022). Explaining flexible continuous speech comprehension from individual motor rhythms. *bioRxiv*. Available at https://www.biorxiv.org/content/10.1101/2022.04.01.486685v1.

Obermeier, C., Menninghaus, W., von Koppenfels, M., Raettig, T., Schmidt-Kassow, M., Otterbein, S., and Kotz, S. A. (2013). Aesthetic and emotional effects of meter and rhyme in poetry. *Frontiers in Psychology* 4, 10. https://doi.org/10.3389/fpsyg.2013.00010.

Oganian, Y., and Chang, E. F. (2019). A speech envelope landmark for syllable encoding in human superior temporal gyrus. *Science Advances* 5(11), 6279. https://doi.org/10.1126/sciadv.aay6279.

Park, H., Ince, R.A., Schyns, P. G., Thut, G., and Gross, J., (2015). Frontal top-down signals increase coupling of auditory low-frequency oscillations to continuous speech in human listeners. *Current Biology* 25, 1649-1653. https://doi.org/10.1016/j.cub.2015.04.049.

Pefkou, M., Arnal, L. H., Fontolan, L., and Giraud, A. L. (2017). θ-Band and β-band neural activity reflects independent syllable tracking and comprehension of time-compressed speech. *The Journal of Neuroscience* 37, 7930-7938. https://doi.org/10.1523/JNEUROSCI.2882-16.2017.

Pickett, J. M., and Pollack, I. (1963). Intelligibility of excerpts from fluent speech: Effects of rate of utterance and duration of excerpt. *Language and Speech* 6, 151-165.

Plomp, R. (2002). *The Intelligent Ear: On the Nature of Sound Perception*. Lawrence Erlbaum, Mahwah, NJ.

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

Pollack, I., and Pickett, J. M. (1964). Intelligibility of excerpts from fluent speech: auditory vs. structural context. *Journal of Verbal Learning and Verbal Behavior* 3(1), 79-84. https://doi.org/10.1016/S0022-5371(64)80062-1.

Polyanskaya, L., and Ordin, M. (2015). Acquisition of speech rhythm in first language. *The Journal of the Acoustical Society of America* 138(3), EL199-EL204. https://doi.org/10.1121/1.4929616.

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

Rosenberg, A. (2010). AutoToBI: A tool for automatic ToBI annotation. *Proceedings of Interspeech, 2010, 11th Annual Conference of the International Speech Communication Association*, Makuhari, Chiba, Japan, September 26-30, 2010, pp. 146-149.

Scherer, K. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication* 40, 227-256.

Schroeder, C. E., Wilson, D. A., Radman, T., Scharfman, H., and Lakatos, P. (2010). Dynamics of active sensing and perceptual selection. *Current Opinion in Neurobiology* 20, 172-176. https://doi.org/10.1016/j.conb.2010.02.010.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science* 270, 303-304. https://doi.org/10.1126/science.270.5234.303.

Silipo, R., and Greenberg, S. (1999). Automatic transcription of prosodic stress for spontaneous English. *Proceedings of the 14th International Congress of Phonetic Sciences,* San Francisco, CA, August 1-7, 1999, pp. 2351-2354.
Silverman, K. E., Beckman, M. E., Pitrelli, J. F., Ostendorf, M., Wightman, C. W., Price, P., Pierrehumbert, J. B., and Hirschberg. J. (1992). ToBI: A standard for labeling English prosody. *International Conference on Spoken Language Processing. ICSLP 1992,* Banff, AB, Canada, October 13-16, 1992.
Smith, Z. M., Delgutte, B., and Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature* 416, 87-90. https://doi.org/10.1038/416087a.
Stevens, K. (1998). *Acoustic Phonetics*. MIT Press, Cambridge, MA.
Tucker, B. V., and Wright, R. A. (2020). Speech acoustics of the world's languages. *Acoustics Today* 16(2), 56-64. https://doi.org/10.1121/AT.2020.16.2.56.
van Bree, S., Sohoglu, E., Davis, M. H, and Zoefel, B. (2021). Sustained neural rhythms reveal endogenous oscillations supporting speech perception. *PLoS Biology* 19(2), e3001142. https://doi.org/10.1371/journal.pbio.3001142.
van Wassenhove, V., Grant, K. W., and Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia* 45, 598-607. https://doi.org/10.1016/j.neuropsychologia.2006.01.001.
Winn, M. B. (2018). Speech: It's not as acoustic as you think. *Acoustics Today* 14(2), 43-49.
Zhang, W., and Ding, N. (2017). Time-domain analysis of neural tracking of hierarchical linguistic structures. *NeuroImage* 146, 333-340. https://doi.org/10.1016/j.neuroimage.2016.11.016.

About the Author

Steven Greenberg
steven@siliconspeech.com

Silicon Speech
El Paso, Texas 79912, USA

Steven Greenberg received his AB in linguistics from the University of Pennsylvania (Philadelphia) and his PhD in linguistics (with a strong minor in neuroscience) from the University of California, Los Angeles (Los Angeles). His thesis was *Neural Temporal Coding of Pitch and Vowel Quality*. He served as a research professor in the Department of Neurophysiology, University of Wisconsin-Madison (Madison) and held faculty appointments in Linguistics and the International Computer Science Institute at the University of California, Berkeley (Berkeley), and the Technical University of Denmark (Kongens Lyngby). He has edited several books on the auditory bases of spoken language and has organized conferences and symposia on speech and hearing science and technology.