8 A Polychromatic Portrait of Speech Rhythm

STEVEN GREENBERG

8.1 Introduction

This chapter's central thesis is that speech rhythm plays a critical role in the parsing and understanding of spoken language. Moreover, rhythm provides a sensory framework for low-frequency (3 Hz–25 Hz) acoustic modulations to interact with other modulatory signals such as visual speech cues. Such sensory modulation provides a potential gateway for accessing endogenous neural activity oscillating at comparable frequencies, hence facilitating cognitive access to linguistic information useful for interpretating inherently ambiguous acoustic signals. The rest of this chapter discusses a variety of experimental and theoretical evidence in support of this thesis.

8.2 A Brief History of Early Speech Modulation Research

The prosodic properties of spoken language have attracted a lot of attention in recent years (e.g., Loukina et al., 2011; Ravignani and Norton, 2017; Poeppel and Assaneo, 2020). Some of this newfound interest is due to an increasing awareness that the linguistic (especially phonemic) models of yesteryear are not entirely in accord with how speech and often intense is processed in the real world, where background noise and reverberation are commonplace and often intense (Assmann and Summerfield, 2004).

For many years, speech models focused on the signal's spectral structure as the basis for decoding spoken language (e.g., Jakobson et al., 1952/1963; Fletcher, 1953; Ladefoged, 1968/1995; Allen, 1995; Stevens, 1998). These frequency-based models assume that processing of the speech signal amounts to linking spectral patterns with phonemic elements, which in turn are used to recognize words and other (e.g., phrasal) linguistic elements. In this context, "spectral" refers to the frequency analysis performed by the auditory system, one that provides a tonotopically structured profile (with respect to auditory neural discharge patterns) as a function of time (i.e., a frequency versus time representation) that some region(s) of the brain

https://doi.org/10.1017/9781009295888.011 Published online by Cambridge University Press

(presumably, higher-level language areas) associate with specific phonetic properties, either articulatory-acoustic features (e.g., Chomsky and Halle, 1968; Jakobson, 1968), "landmarks" (Stevens, 1998), or individual "phonemes" (e.g., Ladefoged, 1968/1995), in a cognitive (and neurological) effort to infer the words spoken (in their appropriate sequence). The detailed mechanism(s) by which this neural time–frequency activity is linked to linguistic elements isn't described in detail, nor is it a means for connecting such linguistic entities to higher-level constructs such as syllables, words, or phrases (and other meaningful elements).

Although there were clues early on that nonphonemic elements might play an important role (e.g., Dudley, 1939, 1940; Kozhnekov and Chistovich, 1965; Liberman et al., 1967), the spectral-phonemic perspective formed the foundation of mainstream speech research until recently. This spectral focus played an especially prominent role in automatic speech recognition (ASR) technology in which lexical models were based on strings of phonemic elements (hereafter "phonemes" [Rabiner and Juang, 2008]). Considerable effort was put into developing "language models" and "pronunciation dictionaries" to compensate for sub-par recognition of the phonetic constituents of the words spoken. Such systems required advance knowledge of the likely words spoken ("language models") to have a reasonable chance of accurately inferring them (e.g., Jelinek, 1998).

8.2.1 The Modulation Spectrum: Origins and Motivation

The seeds for an alternative approach, based on slow modulations in the speech waveform, were planted in the 1970s when the first steps were taken towards dethroning the spectral-phonemic framework.

The original goal of Houtgast and Steekeken's (1973, 1985) research was a practical one. They wanted a metric to predict when speech would be intelligible in a variety of listening environments (e.g., concert auditoria, theaters, and classrooms). They found that the acoustic properties most predictive of intelligibility (i.e., the ability to accurately report the words spoken in the correct order) were not spectral (i.e., acoustic frequency) but rather ones based on extremely slow modulations in the speech signal. How slow? So slow that when expressed in units of frequency, they were well below the threshold of tonal pitch (ca. 30 Hz). In other words, the key acoustic features for intelligibility are orthogonal to the classic spectral elements that had long dominated phonetics and speech research.

To quantify these very low-frequency modulations, Houtgast and Steeneken developed a novel method for their computation and visualization called the modulation spectrum. By analogy with the frequency spectrum, they measured the amount of energy in a bank of modulation-frequency filters. Highly intelligible speech, presented in high signal-to-noise-ratio conditions, exhibited a peak in the modulation spectrum at ca. 4 Hz–5 Hz. However, there was considerable energy down to ca. 3 Hz and up to ca. 12 Hz. The 4 Hz–5 Hz peak implied there was something special about syllables, whose durational properties (125 ms–250 ms) correspond closely to that time frame (when expressed in frequency units). The modulation spectrum turned out to be an excellent predictor of intelligibility because of its sensitivity to acoustic interference from reverberation (the result of acoustic reflections off hard surfaces such as walls, floors, and ceilings) or background noise (e.g., multi-talker speech babble or street noise).

https://doi.org/10.1017/9781009295888.011 Published online by Cambridge University Press

In extremely noisy environments, where intelligibility is severely degraded, they found that the peak of the modulation spectrum is significantly attenuated or even flattened.

Despite Houtgast and Steeneken's compelling data, the modulation spectrum's specific connection to intelligibility was left unresolved.

8.2.2 Modulation's Role in an Early Form of Speech Synthesis

Deeper insight emerged in the 1990s when several studies investigated how low-frequency modulations in the speech waveform impact intelligibility. But before discussing this research, it's helpful to summarize the pioneering research of Homer Dudley, an engineer working at Bell Laboratories in the 1930s. Dudley wanted to improve the quality of speech transmitted over the telephone (his workplace was a subsidiary of AT&T). To do so, he developed two analog tools, the voder and the vocoder (the latter being a more sophisticated version of the former). These were analog filter banks designed to "quantize" speech into a more compact form than the original signal allowed. In two groundbreaking articles (Dudley 1939, 1940), Dudley made a key distinction between the "carrier" associated with spectral features distinguishing phonetic segments (aka "temporal fine structure" [Smith et al., 2002]) and the "modulator" (whose time course roughly follows syllabic elements). The latter was the precursor to what is now called the "speech envelope," but it is important to note that Dudley's original terminology applied to an engineering application designed to simulate human speech, not to a perceptual or neurological property.

In Dudley's view, both the carrier and the modulator were essential for speech to sound natural and intelligible. Because the bandwidth of the telephone was limited in those days, Dudley sought to determine how much of the signal's bandwidth could be reduced (i.e., "compressed") without serious impact on intelligibility or quality. Dudley noted that a relatively small number (ca. 10) of frequency channels was required for the speech to be intelligible (and sound natural). Through this simple demonstration, the importance of auditory frequency analysis was demonstrated using a primitive form of what later came to be known as speech synthesis (or "text to speech").

8.2.3 The Importance of Low-Frequency Modulations in Speech Perception

The broader significance of Dudley's demonstration and Houtgast and Steeneken's intelligibility assessment tool emerged with new research in the 1990s. The next step came from a group headed by Reinier Plomp (see Plomp, 2003, for a review of his research). One of Plomp's students, Rob Drullman (Drullman et al., 1994), performed, as part of his PhD research, a set of perceptual experiments in which the low-frequency modulations in the acoustic speech signal were selectively low- or high-pass filtered (keeping other acoustic properties untouched), and the impact of such manipulations on intelligibility were assessed. "Smearing" (Drullman's term) of syllabic elements, resulting from low-pass filtering of the modulations in the speech signal, seriously impaired intelligibility. Listeners found it challenging to parse the speech stream into distinct, identifiable words. Drullman's results were roughly consistent with Houtgast and Steeneken's earlier studies. Modulation frequencies below 16 Hz were found to

https://doi.org/10.1017/9781009295888.011 Published online by Cambridge University Press

be essential for highly intelligible speech, and modulations in the vicinity of 3 Hz–8 Hz were shown to be especially important (in line with the intuition that syllabic elements are indeed important).

In the US, Shannon et al. (1995) at the House Ear Institute used an updated version of Dudley's vocoder to demonstrate the importance of low-frequency modulations using a novel method. In place of voiced speech, a broadband white noise signal served as the input to the vocoder. This noise signal was modulated by the original voiced version of the speech signal. In other words, the original speech signal's fine time structure was replaced by white noise (i.e., equal energy per Hz band). The modulatory properties of the original speech signal were varied, along with specific details based on the spectral bandwidth and number of frequency channels. A single-channel modulator comprising the full spectral bandwidth of the signal (analogous to the speech envelope) was not intelligible. Only when the noise signal was filtered into four, eight, or more frequency channels were the resulting signals intelligible, indicating that the modulatory properties of the speech signal needed to be manipulated in a specific way. Another way of stating their finding is that a diversity of slow modulation patterns, derived from a broad range of distinct frequency channels, is essential for producing artificial speech that is readily understood. This noise-excited vocoder provided a demonstration that the rhythmic properties of speech could not be satisfactorily represented by a modulatory envelope derived from a unitary source (i.e., the full-band speech signal or even an octave-band signal derived from just a single region of the spectrum).

This key insight, buttressed by Drullman's studies, suggested that the "speech envelope" is inadequate to fully characterize the rhythmic properties of speech. Instead, speech is most faithfully represented by a series of distinct low-frequency modulation patterns emanating from different parts of the tonotopically organized auditory array. How these tonotopically distinct modulation patterns give rise to a unitary sense of rhythm is not well understood.

At this juncture in the modulation story, my colleagues and I began a series of studies (summarized in Greenberg, 2006, 2022) to delve deeper into how these low-frequency modulation patterns impact intelligibility, as described later in the chapter. But before doing so, let's examine some early research on speech synthesis before returning to our consideration of slow waveform modulations.

8.3 Modulations and Rhythm in Speech Technology

A group of researchers in Japan sought to create more natural-sounding speech than was possible using Dudley's original vocoder or through a different method known as vocal-tract-based synthesis.

Beginning in the late 1980s, researchers at ATR (Advanced Telecommunications Research Institute in Japan) began to experiment with high-quality recordings of human talkers speaking a broad range of different words that provided a comprehensive repertoire of phonetic, syllabic, lexical, and phrasal contexts (i.e., read sentences and paragraphs). The recorded material sounded very natural, so the key step to deploy this recorded corpus to generate novel material was to define the appropriate acoustic "target" units and then choose a closely matched snippet

https://doi.org/10.1017/9781009295888.011 Published online by Cambridge University Press

of speech and connect it to another snippet with minimum distortion. Digital splicing algorithms were developed to figure out which snippet most closely approximated the desired target for each linguistic element being modeled. This method is called unit selection (Iwahashi et al., 1992; Black and Campbell, 1995), which soon became the foundation for most high-quality speech synthesis in which novel words and sentences could be generated from the limited amount (often just several hours) of pre-recorded material. The next challenge was to figure out how to splice the chosen snippets together to produce an output that minimized "glitches" and other distortions to optimize intelligibility. The approach came to be known as "concatenative synthesis" and quickly became the dominant form of speech synthesis. Only in recent years has concatenative synthesis been superseded by more sophisticated (and more natural-sounding) methods (summarized by Karagiannakos, 2021).

Concatenative synthesis's key insight is that prosodic properties, such as tempo and meter, matter greatly, both for naturalness and intelligibility. By using spoken material as the core of the synthesis engine, the ATR researchers (implicitly) recognized the importance of low-frequency modulation patterns for producing high-quality spoken material.

Efforts to incorporate modulation models in ASR began in the 1990s (Greenberg and Kingsbury, 1997; Kingsbury et al., 1998; Kanedera et al., 1999). However, these early efforts didn't produce a true breakthrough due to their failure to "translate" modulation patterns into meaningful linguistic elements (as well as their failure to incorporate modulation phase appropriately into sub-word and lexical representations). Later efforts, using more realistic modulation models, improved word recognition slightly (e.g., Hermansky, 2010; Avila et al., 2019), but optimal utilization of modulation patterns and rhythm would have to wait until the advent of "attention" and "transformers" (and other broad-context approaches) to ASR in the late 2010s and early 2020s (e.g., Vaswani et al., 2017).

8.4 Modulation as an Integrative Framework

The power of modulation models lies mostly in their ability to integrate sensory (and cognitive) information across a range of scenarios that would otherwise be challenging for purely spectral models. This limitation of the spectral-phoneme approach is of prime concern, as listeners rarely converse in pristine (i.e., little or no noise) environments. Honking autos, infants crying, barking dogs, cocktail-party babble, and so forth can pose a challenge for decoding speech. In many listening environments, speech's acoustic signature changes into something quite distinct from the waveform emanating from the speaker's vocal tract. And yet most listeners have little difficulty engaging in social discourse under such conditions (much of the time).

The basis of such resilience is unclear, but low-frequency modulation patterns are probably key. The importance of these slow fluctuations may have to do with two other sources of low-frequency modulation – one sensory, the other neural.

Visual speech cues (aka "speech reading") can facilitate comprehension, especially in challenging environments (Grant and Walden, 2000) and for the hearing-impaired (Grant et al., 1998). The opening and closing of the lips, the up and down motion of the jaw, the to-and-fro movement of the tongue, provide a coarse visual analog of certain features of the waveform's

https://doi.org/10.1017/9781009295888.011 Published online by Cambridge University Press

modulation (see Chapter 2). Such visual cues are associated primarily with two linguistic features: "place of articulation" (Skipper et al., 2007), which serves to distinguish among certain consonants (e.g., [p] versus [t] versus [k]), and syllable dynamics germane to lexical and post-lexical parsing (Grant and Walden, 1996). So important are visual cues that their presence can transform largely unintelligible signals into readily understandable speech in challenging listening conditions (e.g., background noise, multi-speech babble, extreme reverberation [Grant and Walden, 2000]). Why is this sensory boost so important for modulation models?

It is likely because of the interaction of distinct but complimentary cues capable of providing the sort of linguistic specificity required to successfully decode the speech signal across a broad range of listening conditions. For example, place-of-articulation information is most closely associated with acoustic frequencies between 1,500 Hz and 3,500 Hz (Stevens, 1998), while syllable parsing is most facilitated by that region of the spectrum ca. 3 kHz (Grant and Walden, 1996). It is the blending of classic phonetic cues with prosodic information that likely provides the foundation for speech intelligibility and comprehension.

8.5 The Speech Envelope as a Model of Auditory Modulation

The "speech envelope" is often used as a proxy for auditory neural activity in studies of linguistic rhythm (e.g., Aiken and Picton, 2008; Ding et al., 2017; Oganian and Chang, 2019; Poeppel and Assaneo, 2020). The envelope's popularity stems from its ability to portray modulation in simple terms (e.g., Figure 8.1). It aids the visualization of neural activity that may be associated with such rhythmic qualities as speaking rate, metrical beat, and prosodic meter.

However, the term "speech envelope" is potentially misleading, as there is no physiological evidence that the excitation of auditory neural elements (e.g., clusters of neurons) modulates to the sort of coarse envelope signal portrayed. Rather, the "envelope" is best thought of as an "artistic" (i.e., graphical) rendering of what the excitatory signal is believed to be at various level(s) of the brain.

Acoustic signals are spectrally filtered, initially in the auditory periphery (i.e., cochlea, auditory nerve), and are subject to additional frequency analysis upstream. Because of this filtering, the auditory signal driving the activity of single units and neural clusters likely differs from the full-spectrum speech envelope portrayed so frequently in the literature.

Auditory filtering decomposes the speech signal into neural waveforms that vary across the tonotopic axis. The speech envelope pattern most closely matching the composite waveform is weighted most heavily to frequencies below 1 kHz (a result of the low-frequency bias of the acoustic speech signal; see Fant, 1960). But such low-frequency elements are not necessarily the primary sources of speech rhythm (either sensed or perceived). Nor is this spectral region necessarily the most important for intelligibility (Miller, 1951; Fletcher, 1953).

Rather, spoken language's resilience and richness are likely a consequence of its "polychromatic" qualities. For example, visual speech cues reinforce (and enhance) linguistic information derived from the acoustic signal (e.g., Grant et al., 1998; Skipper et al., 2007). The situational

https://doi.org/10.1017/9781009295888.011 Published online by Cambridge University Press

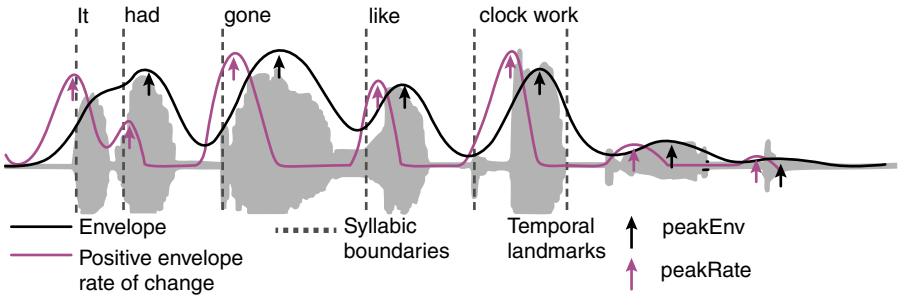


Figure 8.1 The speech envelope illustrated.
The "speech envelope" for a single sentence (shown in black). The temporal fine structure is portrayed in very light gray. Syllable boundaries are indicated by dotted vertical lines. The contour shows the "rate of change" in envelope energy, analogous to "delta features" used in certain ASR applications. Note how coarse the speech envelope is relative to the speech signal's finer details. The envelope contour pertains to the broadband, unfiltered signal.
Adapted from Oganian and Chang (2019).

setting (aka "semantic context") also contributes (by constraining the number of likely options) (e.g., Pollack and Picket, 1964; Greenberg and Christiansen, 2007). It is these multifaceted qualities of spoken language that are fundamental to its semantic and emotional impact. Only a glimmer of this polychromatic character is evident in the acoustic signal. By confining the analysis of speech rhythm to the acoustic envelope alone, one risks overlooking other contributing elements (such as those discussed below).

8.6 The Contribution of Higher-Frequency Acoustic Information to Intelligibility

The minimum bandwidth required to reliably transmit intelligible speech is 300 Hz–3,400 Hz (Miller, 1951; Fletcher, 1953). Because spectral bandwidth was both precious and costly in the early days of telephony, AT&T needed to know the minimum bandwidth required to ensure the signal was reliably intelligible (defined as 96% of the words being correct or better). Early "articulation" studies highlighted the importance of frequencies between 1,500 Hz and 3,400 Hz (Fletcher, 1953; Allen, 1995; Stevens, 1998).

What is this observation's relevance for speech rhythm?

It is relevant because the lower frequencies (<1,500 Hz) contribute far more to the speech envelope than the mid- and higher frequencies. Although the speech modulation envelope appears reasonable (by eye), it excludes certain portions of the spectrum essential for intelligibility. The higher speech frequencies contribute less to the acoustic signal's waveform on account of their lower amplitude (see Fant, 1960). But this does not mean these higher frequencies contribute less to speech communication – on the contrary. Because of the auditory system's acute sensitivity to the mid-range portion of the spectrum (1,500–4,000 Hz) along with the presence of spectral filtering, auditory waveforms associated with this higher-frequency

https://doi.org/10.1017/9781009295888.011 Published online by Cambridge University Press

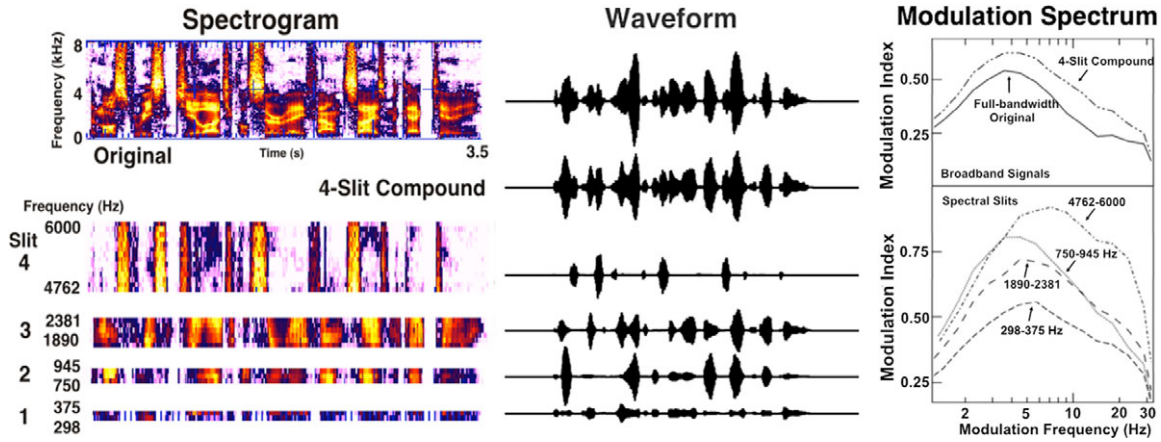


Figure 8.2 Speech waveforms, spectrograms, and modulation spectra across acoustic frequencies. Spectrographic and time domain representations of the single sentence "The most recent geological survey found seismic activity" (Greenberg et al., 1998). The waveforms are plotted on the same amplitude scale, while the scale of the original, unfiltered signal is compressed by a factor of five for illustrative clarity. The frequency axis of the spectrographic display of the channels has been nonlinearly compressed for illustrative purposes. Note the quasi-orthogonal temporal registration of the waveform modulation pattern across frequency channels. On the right are modulation spectra (magnitude component) associated with each of four, 1/3-octave channels. The peak of the spectrum (in all but the highest channel) lies between 4 Hz and 6 Hz. Note the large amount of energy in the higher-modulation frequencies associated with the highest-frequency channel. The modulation spectra of the four-channel compound and the original, unfiltered signal are illustrated for comparison (top panel).
From Greenberg et al. (1998).

region are modulated as much as, if not more than, waveforms associated with the lower part of the speech spectrum (Figure 8.2). It is these all-important higher-frequency elements that are missing from the prototypical "speech envelope."

The portion of the acoustic frequency spectrum whose waveform most closely approximates this envelope model is not highly intelligible; indeed, it sounds muffled and indistinct (see Miller, 1951), due to two factors. The portion of the spectrum essential for accurate consonant decoding is mostly above 1,500 Hz, and these higher frequencies facilitate the micro-parsing (at the syllable level) of the speech stream (see Kösem et al., 2023, for data consistent with this assertion).

A study by Smoorenburg (1992) is instructive for illustrating the importance of the higher frequencies. Using conventional audiometric measures, he found that the nature of the acoustic background was critical for determining which portion of the spectrum contributed most to intelligibility. In quiet conditions (i.e., no background noise), the most accurate predictor is a listener's pure-tone threshold below 2 kHz. However, in high-noise environments, the best predictor is tonal thresholds above 2 kHz. Such results suggest it is the higher frequencies in the signal (and auditory tonotopic array) that underlie the resilience of speech comprehension in adverse listening conditions, an important finding for ameliorating hearing impairment among the hard of hearing.

The harmful impact of background noise (and other forms of acoustic interference) on intelligibility may be due, in part, to the challenge of parsing (into syllabic and lexical elements)

https://doi.org/10.1017/9781009295888.011 Published online by Cambridge University Press

the incoming speech signal, thereby making it difficult to distinguish one syllable from another. This is where speech rhythm likely comes into play; it provides a linguistic foundation with which to segregate, analyze, and recognize the signal's syllabic sequences (and their articulatory constituents) and recode the signal within a semantic framework that can be readily transformed into a meaningful interpretation for action.

What are the properties associated with the higher end of the speech spectrum so critical for intelligibility? As we've seen, the parsing of spoken material at the syllable level appears to be especially focused on the spectral region around 3 kHz. Knowing the number of and stress patterns of syllables within and across words is helpful for comprehension (McQueen and Dilley, 2020), especially under adverse conditions (Grant and Walden, 1996). Such knowledge may reflect how lexical elements are stored in the brain, as can be seen in the "tip of the tongue" phenomenon (Brown and McNeil, 1966), where syllable number, stress pattern, and initial consonant appear to be key for lexical recall (and imply that words are more than mere strings of phonemic elements [Greenberg, 1999, 2006]). Another way of stating such findings is that prosodic knowledge (e.g., syllable number and stress patterning) can reduce the listener's uncertainty regarding the identity of words spoken under a broad range of listening conditions. Hence, low-frequency modulation is a critical component of rhythm, and high-frequency modulatory patterns appear to be important for syllable parsing.

8.7 Modulation Phase and Speech Intelligibility

Houtgast and Steeneken's intelligibility metric measured only the modulation spectrum's magnitude component (a limitation of their analog instrumentation). Although their metric did an excellent job of estimating intelligibility across a variety of acoustic environments, their original formulation of the modulation spectrum was not designed to provide a more granular measure for use in speech perception models.

The phase (i.e., relative timing) component of the modulation spectrum provides a set of acoustic features useful for distinguishing phonetic segments, especially consonants, from one another. In effect, modulation phase provides a means of integrating critical phonetic information such as "place" and "manner" of articulation into a unified modulation representation. This is because phase, as used in the complex modulation spectrum (CMS), is a proxy for temporal alignment across the acoustic frequency axis. When articulatory-acoustic cues signaling place and manner of articulation are out of alignment with other properties of the speech signal, intelligibility deteriorates. This is what happens in highly reverberant environments such as noisy restaurants or transportation hubs (e.g., bus terminals, train stations).

Greenberg and Arai (2001, 2004) developed a novel method for combining the phase and magnitude components of the modulation spectrum into a unified representation that closely corresponds to the intelligibility of spoken English sentences (Figure 8.3). It is likely that the processing of speech in the auditory system involves certain neural operations akin to the computations involved in the CMS (see Elhilali, 2019). And as cited elsewhere in this chapter, modulation phase has been shown to (modestly) improve ASR performance (Kanedera et al., 1999; Hermansky, 2010; Avila et al., 2019).

https://doi.org/10.1017/9781009295888.011 Published online by Cambridge University Press

https://doi.org/10.1017/9781009295888.011 Published online by Cambridge University Press

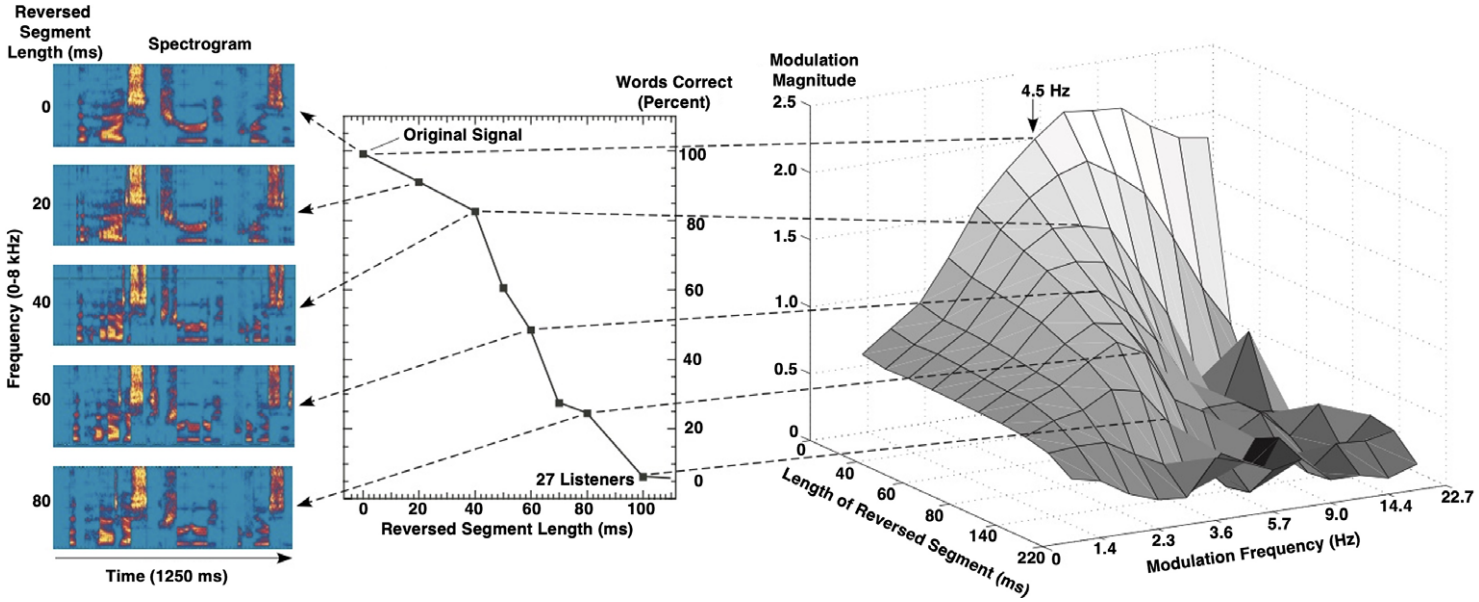


Figure 8.3 Speech intelligibility and the CMS.

The CMS integrates the magnitude and phase components into a single value. The sentence material's intelligibility for a listening experiment 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 CMS. The spectro-temporal properties (and articulatory-acoustic features) also deteriorate appreciably under such conditions.

Reprinted from Greenberg (2022). The figure is an adaptation of one originally published in Greenberg and Arai (2001, 2004).

8.8 Rhythm as a Linguistic Parser

There are other reasons to believe phonemic elements are not the only, or even the optimal, way to linguistically model the speech stream. From both an acoustic and a spectral perspective, the speech signal has been likened to "scrambled eggs" (Hockett, 1960) with respect to how phonemic elements unfold acoustically when displayed in 2-D or 2.5-D representations (e.g., a spectrogram). There is considerable overlap between contiguous segments with respect to their phonetic properties. It is indeed a challenge to pinpoint where a phonetic element ends and the following one begins, especially for syllable-medial (typically vocalic) constituents (Greenberg, 2006). Rhythmic properties may facilitate the delineation of phonetic elements within a syllable by highlighting onsets, codas, and the interstitial nuclei.

The phonetic quality of a consonantal segment is heavily influenced by a variety of linguistic factors, the most important being its position within the syllable (e.g., onset, nucleus, coda), the syllable's degree of prominence/stress (high, mid, low), and its approximate "place of articulation" (front, center, back of the oral cavity). The interaction of such features is what (mostly) determines how a consonant is phonetically realized within a syllable (Greenberg, 2006). Vocalic segments are sensitive to such factors as well, but their phonetic expression differs from consonants in that intensity, duration, and vowel height are the most relevant features. Vowels are more intense, longer, and lower in tongue height in highly prominent syllables than their less prominent counterparts (Greenberg et al., 2003; Cho, 2005).

8.9 Speech, Rhythm, and the Brain

How this parsing, analysis, and recoding are performed within the brain is not well understood. This is where speech rhythm is likely to play a prominent role, as it provides a sensory framework for low-frequency acoustic modulation to interact with other modulatory signals such as visual speech cues. Such sensory modulation provides a gateway for accessing endogenous neural activity that modulates at comparable (low) frequencies and incorporates a variety of information useful for interpretating ambiguous acoustic signals (see Chapter 3).

Low-frequency neural rhythms (aka "oscillations") have been long thought to play a role in the processing of spoken language (Lenneberg, 1967) and certain other behaviors (Buzsaki, 2006, 2021; Chapter 3). Such cortical oscillations have been recorded in the language areas of the brain (e.g., Meyer, 2018; Poeppel and Assaneo, 2020; Greenberg, 2022). Although the specific function(s) of "delta" (0.5 Hz–3.5 Hz), "theta" (3.5 Hz–7 Hz), "beta" (10 Hz–25 Hz), and "gamma" (25 Hz–60 Hz) oscillations remains controversial (Buzsaki, 2006, 2021), it has been proposed that such endogenous rhythms are involved with certain aspects of linguistic processing (e.g., Greenberg, 2011; Ding et al., 2016).

It is tempting to link various brain rhythms to the processing and interpretation of specific linguistic constituents based largely on temporal properties. The periodicity of theta rhythm (3.5–7 Hz) coincides with that of syllabic elements (140–300 ms) across languages (Ding et al., 2016), while "beta" oscillations (10–25 Hz) are similar in duration to the typical phonetic segment (40–100 ms) (Ding et al., 2016). And the ultra-low-frequency "delta" waves (0.5 Hz–3.5 Hz) are

https://doi.org/10.1017/9781009295888.011 Published online by Cambridge University Press

comparable in length to spoken phrases and sentences (300 ms–2,000 ms). However, it's unlikely the brain operates in such a rigid, lockstep fashion as some have proposed (e.g., Ghitza, 2011).

Speech rhythm is often realized via certain modulatory properties of the signal. Prominent syllables are generally more intense than their less prominent counterparts (Greenberg et al., 2003), which typically precede or follow them. This variation in amplitude (or "voice intensity") helps pinpoint the most informative parts of the utterance to aid the listener's parsing and interpretation of the communication (Greenberg, 2022), especially helpful in adverse listening conditions (Assman and Summerfield, 2004). The most prominent (i.e., highly stressed) syllables are longer in duration than their more lightly or unstressed counterparts, and such patterning is reflected in the modulation spectrum by a pronounced peak between 3 Hz and 6 Hz (Greenberg et al., 2003, Figure 4). Shorter, less prominent syllables are represented in the modulation spectrum with energy in the 7 Hz to 12 Hz range.

8.10 Rhythmic-Centric Representations

It is widely acknowledged that such prosodic features as syllable prominence, stress accent, and phrasal intonation are as (if not more) important for linguistic communication than the classic notion of the phoneme. This is because prosody can influence the phonetic realization of much of what's spoken (e.g., Greenberg, 2006). It also plays an outsized role in heavily emotional speech where certain syllables are often emphasized (aka hyper-articulation) to ensure the listener truly gets the speaker's point (Lindblom, 1990).

It has been hypothesized that the root cause of certain communication disorders may lie in neural timing issues as evidenced in deviant patterns of brain oscillations (Giraud and Poeppel, 2012; Leong and Goswami, 2014; Ding et al., 2016, 2017).

Models of spoken (and possibly written) language would likely benefit from integrating articulatory, acoustic, visual, and motor features into a unified theoretical framework that focuses on the interaction among different sources of linguistic information (e.g., McQueen and Dilley, 2020). Such a unified perspective could aid in the development of new (and hopefully more efficacious) methods for teaching children to read and speak effectively (Leong and Goswami, 2014), as well as provide more effective ways of treating and ameliorating communication disorders impacting so many across their lifespan.

8.11 Acknowledgements

I thank the following colleagues for collaborating on research mentioned in this article: Takayuki Arai, Hannah Carvey, Shuangyu Chang, Oded Ghitza, Jeff Good, Ken Grant, Leah Hitchcock, Joy Hollenback, Brian Kingsbury, Nelson Morgan, and Rosaria Silipo. I would also like to thank reviewers Bettina Braun and Ting Huang for their helpful suggestions for improving this chapter. Thanks, as well, to the organizers of this special volume, Lars Meyer, Antje Strauss, and Caroline Duchow. Some of the material in this chapter was

https://doi.org/10.1017/9781009295888.011 Published online by Cambridge University Press

presented on March 12, 2021, at a virtual workshop in memory of John Ohala, organized by John Kingston and Didier Demolin.

Box 8.1 Chapter Overview

Summary

Slow modulation of the acoustic speech signal plays a critical role in a listener's ability to decode and comprehend spoken language. This low-frequency (3 Hz–16 Hz) rhythmic patterning derives from an interaction of articulatory processes with information-encoding imperatives shielding linguistic elements from adverse acoustic conditions common in the real world.

Implications

Models of spoken language often focus on phonemic elements as key building blocks of lexical meaning. A prosodic perspective, incorporating rhythm and syllabic prominence as central features, provides a more comprehensive foundation for understanding how listeners decode the speech signal, especially in adverse listening environments.

Gains

A prosodic, rhythmic model of spoken language understanding provides a theoretical foundation for investigating the neural bases of spoken and written language. Low-frequency (3 Hz–25 Hz) cerebral oscillations offer potential insight into the neurological bases of linguistic behavior, in particular how listeners go from sound to meaning.

REFERENCES

Aiken, S. J., and Picton, T. W. (2008). Human cortical responses to the speech envelope. *Ear and Hearing*, 29(2), 139–157. https://doi.org/10.1097/AUD.0b013e31816453dc

Allen, J. B. (1995). How do humans process and recognize speech? In R. P. Ramachandran and R. J. Mammone (eds.), *Modern Methods of Speech Processing*, Springer International Series in Engineering and Computer Science, Volume 327, pp. 251–275. Boston, MA: Springer. https://doi.org/10.1007/978-1-4615-2281-2_11

Assmann, P., and Summerfield, Q. (2004). The perception of speech under adverse conditions. In S. Greenberg, W. A. Ainsworth, R. R. Fay, and A. N. Popper (eds.), *Speech Processing in the Auditory System*, pp. 231–308. New York: Springer-Verlag. https://doi.org/10.1007/0-387-21575

Avila, A. R., Kashirsagar, S. R., Tiwari, A., et al. (2019) Speech-based stress classification based on modulation spectral features and convolutional neural networks. 27th European Signal Processing Conference (EUSIPCO), pp. 1–5. https://ieeeplore.ieee.org/document/8903014

https://doi.org/10.1017/9781009295888.011 Published online by Cambridge University Press

Black, A., and Campbell, N. (1995). Optimising selection of units from speech databases for concatenative synthesis. Proceedings of EUROSPEECH 1995, pp. 581–584.

Brown, R., and McNeill, D. (1966). The "tip of the tongue" phenomenon. *Journal of Verbal Learning and Verbal Behavior*, 5(4), 325–337. https://doi.org/10.1016/S0022-5371(66)80040-3

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

Buzsaki, G. (2021). *The Brain from Inside Out*. New York: Oxford University Press.

Cho, T. (2005) Prosodic strengthening and featural enhancement: Evidence from acoustic and articulatory realizations of /ɑ,i/ in English. *Journal of the Acoustical Society of America*, 117, 3867–3878. https://doi.org/10.1121/1.1861893

Chomsky, N., and Halle, M. (1968). *The Sound Pattern of English*. New York: Harper and Row.

Ding, N., Melloni, L., Zhang, H., Tian, X., and Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, 19, 158–164. https://doi.org/10.1038/nn.4186

Ding, N., Patel, A. D., Chen, L., et al. (2017). Temporal modulations in speech and music. *Neuroscience and Biobehavioral Reviews (B)*, 81, 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. *Journal of the Acoustical Society of America*, 95, 1053–1064. https://doi.org/10.1121/1.408467

Dudley, H. (1939). Remaking speech. *Journal of the Acoustical Society of America*, 11, 169–177. https://doi.org/10.1121/1.1916020

Dudley, H. (1940). The carrier nature of speech. *Bell System Technical Journal*, 19, 495–515. https://doi.org/10.1002/j.1538-7305.1940.tb00843

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

Fant, G. (1960). *The Acoustic Theory of Speech Production*. Berlin: De Gruyter. https://doi.org/10.1515/9783110873429/html

Fletcher, H. (1953) *Speech and Hearing in Communication*, 2nd ed. New York: Van Nostrand.

Ghitza, O. (2011). Linking speech perception and neurophysiology: Speech decoding guided by cascaded oscillators locked to the input rhythm. *Frontiers in Psychology*, 27(2), 130. https://doi.org/10.3389/fpsyg.2011.00130

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

Grant, K. W., and Walden, B. E. (1996). Spectral distribution of prosodic information. *Journal of Speech and Hearing Research*, 39, 228–238. https://doi.org/10.1044/jshr.3902.228

Grant, K. W. and Walden, B. E. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *Journal of the Acoustical Society of America*, 108, 1197–1208. https://doi.org/10.1121/1.1288668

Grant, K. W., Walden, B. E., and Seitz, P. F. (1998) Auditory-visual speech recognition by hearing impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. *Journal of the Acoustical Society of America*, 103, 2677–2690. https://doi.org/10.1121/1.422788

Greenberg, S. (1999). Speaking in shorthand: A syllable-centric perspective for understanding pronunciation variation. *Speech Communication*, 29, 159–176. https://doi.org/10.1016/S0167-6393(99)00050-3

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

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

Greenberg, S. (2022). Speech, rhythm, and the brain. *Acoustics Today*, 18(3), 22–31. https://doi.org/10.1121/AT.2022.18.3.22

https://doi.org/10.1017/9781009295888.011 Published online by Cambridge University Press

Greenberg, S., and Arai, T. (2001) The relation between speech intelligibility and the complex modulation spectrum. Proceedings of the Seventh European Conference on Speech Communication and Technology (Eurospeech 2001), pp. 473–476. https://doi.org/10.21437/Eurospeech.2001-125

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., and Christiansen, T. U. (2007). Linguistic scene analysis and the importance of synergy. In T. Dau, J. Buchholz, J. Harte, and T. U. Christiansen (eds.), *Proceedings of the International Symposium on Auditory and Audiological Research: Auditory Signal Processing in Hearing-Impaired Listeners*, 1, pp. 351–363. https://proceedings.isaar.eu/index.php/isaarproc/article/view/2007-33

Greenberg, S., and Kingsbury, B. (1997). The modulation spectrogram: In pursuit of an invariant representation of speech. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 3, pp. 1647–1650.

Greenberg, S., Arai, T., and Silipo, R. (1998). Speech intelligibility derived from exceedingly sparse spectral information. Fifth International Conference on Spoken Language Processing (ICSLP 1998), Paper 0074. https://doi.org/10.21437/ICSLP.1998-473

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

Hermansky, H. (2010). History of modulation spectrum in ASR. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Dallas, TX, pp. 5458–5461. https://doi.org/10.1109/ICASSP.2010.5494907

Hockett, C. (1960). The origin of speech. *Scientific American*, 203, 88–96. https://doi.org/10.1038/scientificamerican0960-88

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

Houtgast, T., and Steeneken, H. (1985). A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. *Journal of the Acoustical Society of America*, 77, 1069–1077.

Iwahashi, N., Kaiki, N., and Sagisaka, S. (1992). Concatenative speech synthesis by minimum distortion criteria. Proceedings of 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing, San Francisco, CA, pp. 65–68. https://doi.org/10.1109/ICASSP.1992.226119

Jakobson, R. (1968). *Child Language, Aphasia and Phonological Universals*. The Hague: Mouton. http://dx.doi.org/10.1515/9783111353562

Jakobson, R., Fant, G., and Halle, M. (1952/1963) *Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates*. Cambridge, MA: MIT Press.

Jelinek, F. (1998) *Statistical Methods for Speech Recognition*. Cambridge, MA: MIT Press.

Kanedera, N., Arai, T., Hermansky, H., and Pavel, M. (1999). On the relative importance of various components of the modulation spectrum for automatic speech recognition. *Speech Communication*, 28, 43–55. https://doi.org/10.1016/S0167-6393(99)00002-3

Karagiannakos, S. (2021). Speech synthesis: A review of the best text to speech architectures with deep learning. AI Summer. https://theaisummer.com/text-to-speech/

Kingsbury, B., Morgan, N., and Greenberg, S. (1998). Robust speech recognition using the modulation spectrogram. *Speech Communication*, 25, 117–132. https://doi.org/10.1016/S0167-6393(98)00032-6

Kösem, A., Dai, B., McQueen, J. M., and Hagoort, P. (2023). Neural tracking of speech envelope does not unequivocally reflect intelligibility. *NeuroImage*, 272, 120040. https://doi.org/10.1016/j.neuroimage.2023.120040

Kozhnekov, V. A., and Chistovich, L. A. (1965). *Speech, Articulation, and Perception*. Moscow: Nauka (English translation published by US Department of Commerce).

Ladefoged, P. (1968/1995). *Preliminaries to Linguistic Phonetics*. Chicago: University of Chicago Press.

https://doi.org/10.1017/9781009295888.011 Published online by Cambridge University Press

Lenneberg, E. (1967). *Biological Foundations of Language*. New York: Wiley.

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

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, 74(6), 431–461. https://doi.org/10.1037/h0020279

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In W. J. Hardcastle and A. Marchal (eds.), *Speech Production and Speech Modelling*. NATO ASI Series, Volume 55, pp. 403–439. Dordrecht: Springer. https://doi.org/10.1007/978-94-009-2037-8_16

Loukina, A., Kochanski, G., Rosner, B., Shih, C., and Keane, E. (2011). Rhythm measures and dimensions of durational variation in speech. *Journal of the Acoustical Society of America*, 129, 3258–3270. https://doi.org/10.1121/1.3559709

McQueen, J. M., and Dilley, L. (2020). Prosody and spoken-word recognition. In C. Gussenhoven and A. Chen (eds.), *The Oxford Handbook of Language Prosody*, pp. 509–521. Oxford: Oxford university Press. https://pure.mpg.de/rest/items/item_3181263_3/component/file_3291232/content

Meyer, L. (2018). The neural oscillations of speech processing and language comprehension: State of the art and emerging mechanisms. *European Journal of Neuroscience*, 48, 2609–2621. https://doi.org/10.1111/ejn.13748

Miller, G. A. (1951) *Language and Communication*. New York: McGraw-Hill.

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

Plomp, R. (2003). *The Intelligent Ear*. London: Academic Press.

Poeppel, D., and Assaneo, M. F. (2020). Speech rhythms and their neural foundations. *Nature Reviews of 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 & Verbal Behavior*, 3(1), 79–84. https://doi.org/10.1016/S0022-5371(64)80062-1

Rabiner, L., and Juang, B. H. (2008). Historical perspective of the field of ASR/NLU. In J. Benesty, M. M. Sondhi, and Y. A. Huang (eds.), *Springer Handbook of Speech Processing*, pp. 521–538. Berlin: Springer. https://doi.org/10.1007/978-3-540-49127-9_26

Ravignani, A., and Norton, P. (2017) Measuring rhythmic complexity: A primer to quantify and compare temporal structure in speech, movement, and animal vocalizations. *Journal of Language Evolution*, 2(1), 4–19. https://doi.org/10.1093/jole/lz002

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

Skipper, J. I., van Wassenhove, V., Nusbaum, H. C., and Small, S. L. (2007). Hearing lips and seeing voices: How cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral Cortex*, 17, 2387–2399. https://doi.org/10.1093/cercor/bhl147

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

Smoorenburg, G. F. (1992). Speech reception in quiet and in noisy conditions by individuals with noise-induced hearing loss in relation to their tone audiogram. *Journal of the Acoustical Society of America*, 91, 421–437. https://doi.org/10.1121/1.402729

Stevens, K. (1998). *Acoustic Phonetics*. Cambridge, MA: MIT Press.

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017), Attention is all you need: Advances in neural information processing systems. Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS '17), Curran Associates Inc., Red Hook, NY, pp. 6000–6010.

https://doi.org/10.1017/9781009295888.011 Published online by Cambridge University Press