Neural Oscillations and the Perception of Spoken Language

Ivan Iotzov

January 2019

1 Introduction

Great progress has been made in recent years on the goal of decoding the neural computations that undergird human speech processing. This review is intended to serve as a brief overview of this work, focusing especially on neural oscillations. I will be examining the role of both spontaneous and exogenously driven neural oscillations in speech processing. Although it has been well established that neural oscillations track the amplitude envelope of speech, whether this reflects a bottom-up processing of stimulus acoustics or a top-down speech tracking that is integral to speech comprehension is a matter of ongoing debate. In this introduction I will review some basic properties of auditory speech signals, as well as the putative origin and function of various endogenous oscillations that have been measured in the brain. The second chapter will go into more detail on the goals and methods of the auditory processing system in general. The third chapter will be dedicated to discussing the endogenous oscillations in the brain, their interaction with stimulus-driven exogenous oscillations, and the theorized role of these oscillations in speech processing. Lastly, I will attempt to integrate all of this information and draw broader conclusions about the functional role of these oscillations in speech processing and avenues for future work that would fill gaps in the literature. I will also address the debate in the field about whether this type of oscillatory activity represents bottom-up processing of stimulus acoustics or more top-down processing of linguistic stimulus features.

1.1 Speech Rhythms

Spoken language can be roughly categorized as a quasi-periodic, hierarchically organized, sequence of auditory events that encode some information. In *The Speech Chain* [6], Peter Denes elegantly laid out the encoding, decoding, and processing that must take place in order for some information to successfully be transferred from one person to another in the form of speech.

In this chain (illustrated in Fig.1), information is transformed in a number of ways when moving from a speaker's intention to a listener's understanding. But, for the purposes of this review, it is important to recognize the conceptual contribution of treating speech comprehension as a hierarchical process that extracts information from an acoustic signal and passes it up a chain of increasing complexity until the content of the message is decoded. In this context, we can then move to model the stages of the human speech comprehension system as transformations of various types on the original speech input signal.

One of the most fundamental questions in the field is whether language processing is accomplished by a general architecture that is also applied to all other sounds that are captured by the ear, or whether language has a special set of processing rules that apply only to it [68]. There is evidence to support both conclusions, with some arguing that language is not a special case at all, and all of the functions that we see when examining the language faculties in humans can be explained by a general, distributed network that is applied an all acoustic analysis [7]. Others see the structure of human language and its uniqueness in the animal kingdom [3, 19] as an indication that there is special architecture that supports its use and comprehension. While there

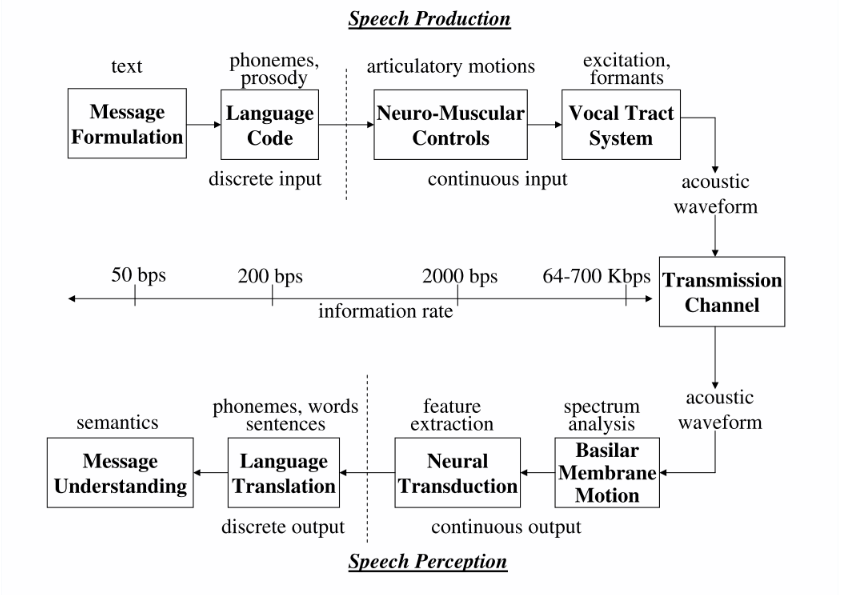


Figure 1: The chain of transformations that constitute the 'speech chain' as described by Denes [6, 57]

is evidence pointing in both directions, there seems to be more suggesting that the faculty of language is indeed special, and that special neural circuits are recruited to facilitate its comprehension [54].

The breadth and variety of human languages is huge, with ~7000 [64] languages currently spoken across the world. Among these 7000 languages there are a large, but consistent number of speech sounds that are produced by humans (see Fig.A.1). One of the primary challenges of both linguists and neuroscientists working in the area of speech processing is how the human brain is capable of dealing with information that comes in such a diverse array of languages, each with their own lexical, semantic, and stylistic rules.

Despite all this variation, every human language is intelligible to those that speak it. Indeed, the ability to learn languages outside of one's native language points to the universality of the mechanisms on which speech comprehension is built. In the case of speech, modulations less than 16 Hz, with a notable peak at 3-5 Hz, are the basis for the syllabic rhythm [25, 27].

One way to characterize these rhythms is through analysis of the modulation spectrum of speech. The modulation spectrum is defined as 'the spectrum of the temporal envelope of sound and reflects how fast sound intensity fluctuates over time' [13]. In a study of a large audio corpus of 9 languages, Ding et al. [13] found a remarkably consistent peak for the modulation spectrum of spoken language at 2-10 Hz, with a particular peak at ~5 Hz. This same 5Hz rate has also been found in empirical studies of syllabic rate in a number of languages [52] and seems to be a consistent, intrinsic property of speech, regardless of the language that is being spoken.

These temporal modulations are likely important to the processing of spoken language as they provide consistent acoustic landmarks for the chunking of incoming acoustic signals into discrete units that can then be decoded. Examination of the modulation spectrum has not been the only method of measuring the rhythmic aspects of speech. There is a rich literature in the field of linguistics dealing with how and on what basis different languages are arranged in time. In this system of classification that has been developed, languages can be stress-timed or syllable-timed [48, 5]. Stress-timed languages (such as English or Dutch) rely on a series of stressed or un-stressed syllables to generate the rhythm that is perceived in speech. Syllable-timed languages (such as French) rely only on the timing of the syllables themselves, and not on the stresses that are placed on them. Note that Japanese has been placed in a class of its own, said to have a unique mora based timing scheme.

This effort by linguists to classify languages into different timing schemes is primarily an effort to force the quasi-periodic rhythm of human speech into a more isochronous form. If there is an underlying isochronicity below the relatively messy rhythms that are present in natural language, then there is a relatively easy way to describe how humans can perceive speech that is independent of language and can also be easily compared to the perception of musical rhythms which are fundamentally based in their isochronicity.

This paradigm has been challenged in a number of ways [58, 26], with critics primarily pointing to the subjective and often contradictory classification of languages that can change based on the corpus of speech used and the judgments of the observer. The problem of classification in this paradigm has led some to focus solely on stress timing as a rhythmic framework in all languages [30, 43]. This framework also has some empirical justification as several ERP studies have found that the brain can detect violations of expected stress patterns [61, 39, 29].

1.1.1 Classes of Phonemes

Phonemes can either be classified by their acoustic properties, or by the mode of their generation in natural human speech. In the latter method, they are typically divided into stops, nasals, fricatives, laterals, rhotics, clicks, and vowels [41]. Though, no language makes use of every speech sound that can be produced by the human vocal tract, and some sounds are found much more rarely than others. These distinctions are far more useful for the linguistic analysis of speech than they are for the acoustic analysis, so we turn to the main acoustic differentiation of speech sounds: periodic, voiced speech versus aperiodic, unvoiced speech. The difference between these two can be clearly seen from both their waveforms, as well as their frequency domain spectra (See Fig.2).

The differences between these two classes of speech sounds is explained through differences in their modes of production. For instance, the unvoiced fricative [ʃ] (the first syllable in Fig.2) looks like random noise, while the voiced vowel [ʊ] appears to have a definite structure with periodic features. This is because [ʃ] is produced by the constriction of airflow out of the mouth through use of the tongue, while [ʊ] is produced by the modification of vibrations generated in the vocal cords. The fact that

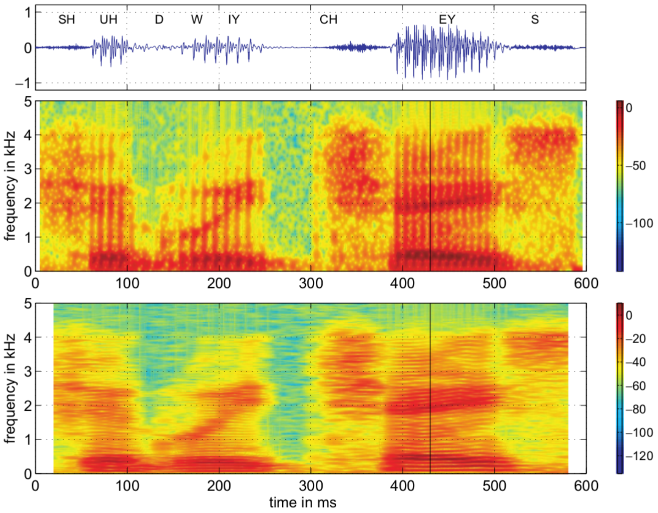


Figure 2: Waveform and 3D Spectrogram of the phrase 'Should we chase' [57]

voiced sounds are generated by vocal cord vibrations gives them a special structure that makes them much easier to identify in the context of feature extraction, and makes them important markers in computer speech recognition [28].

How the brain parses differences between different types of phonemes, particularly those that lack discernable spectral content (such as fricatives) is still an area of active study. The ability to discern between meaningless noise and a meaningful phoneme must rely on information outside of just the spectral content of the speech signal itself, as the two might be indistinguishable based on just this information. Instead, it is likely that timing and rhythm play an important role in the differentiation of speech sounds. The order and precise timing of the delivery of a speech sound has a strong effect on how it is perceived and these context effects have strong influences on the human perception of speech and language.

1.1.2 Speech Formants

An important feature of voiced speech sounds that seems to be critical in their recognition are the formants of the vowel sound. Formants are defined as being distinctive frequency components of a speech signal that are generated by the resonance of the human vocal tract [6]. They can be seen clearly in the speech spectrogram (See Fig. 4) as areas of particularly high energy within a certain frequency band. Formants come about due to the resonance of the vocal tract. By moving the articulators (lips, tongue, etc.), the size and shape of the cavities in the vocal tract are altered and therefore so are the resonant frequencies. All vowel sounds are produced through this manipulation of the resonant frequencies of the vocal tract which gives them a characteristic, periodic structure that can be readily identified.

By convention, vowel formants are assigned numbers in order of ascending fre-

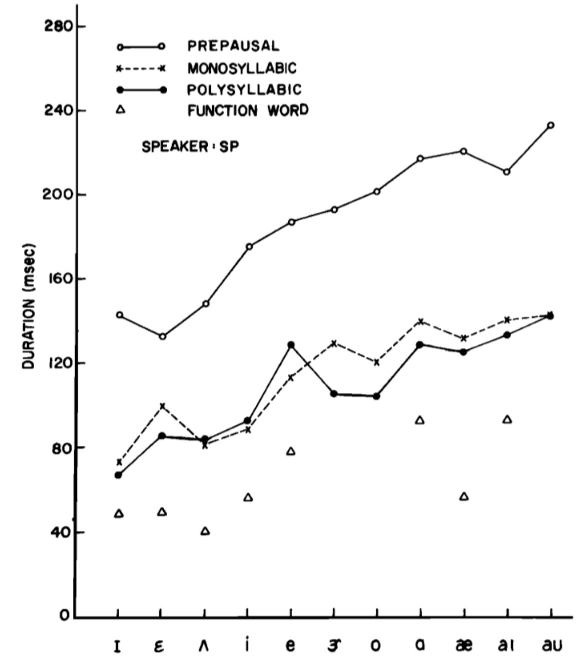


Figure 3: Average duration of a selection of vowels under three different speaking conditions from one speaker [69]

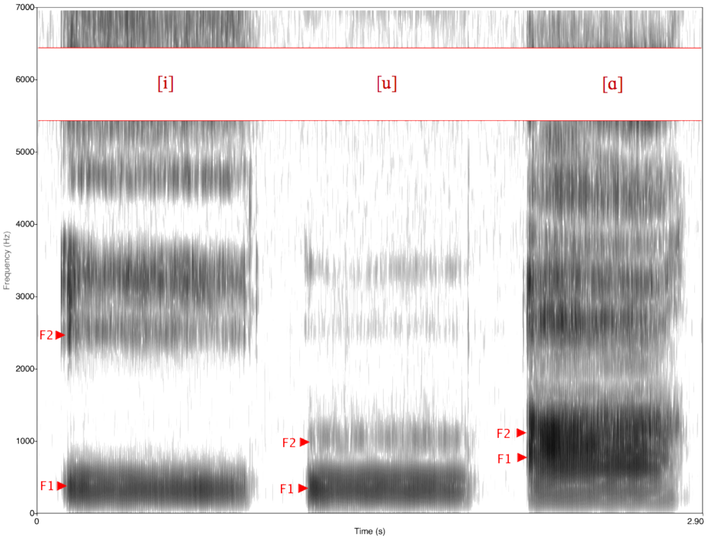


Figure 4: Spectrogram of a native speaker of American English pronouncing the vowels [i, u, ɑ] [71].

quency. So, the lowest frequency formant of a vowel is assigned F_1, the next lowest F_2, and so on. Many vowels can be distinguished from one another simply by the use of these first two formants [62], making them important markers in both human and computer speech recognition. For example, the vowels in Fig.4 can be differentiated from one another simply by reference to their F_1 and F_2 frequencies. The importance of these formants to biological auditory comprehension was demonstrated in part by Young and Sachs (1979) [72], who showed that auditory nerve discharges in response to auditory stimuli reflect the formants when vowels are presented.

1.2 Endogenous Brain Rhythms

There exist a number of neural rhythms that are innately present in the human brain. Roughly speaking, these are divided into α (8 - 12 Hz), β (12 - 30 Hz), θ (4 - 8 Hz), δ (1 - 4 Hz), γ (30 - 80 Hz), and high γ (80 - 150 Hz) [47, 59]. These are roughly divided by the frequency of the oscillations and their suspected behavioral correlates. There is clearly overlap between these categorizations and it is difficult to make very clear distinctions between the various classes of neural oscillations. Despite this, there are some clear behavioral correlates in each of these frequency bands, and there have been attempts to localize the neural circuits responsible for their generation [45].

Some of these rhythms are particularly relevant for the study of speech processing as they appear to have the same frequency as the typical speech rhythm and their power in EEG signals increases as a function of intelligibility. Particularly important to speech processing appear to be the δ, γ, and θ rhythms [24, 44]. In the case of the θ rhythm, Oded Ghitza has been a particularly strong advocate for the idea that the θ rhythm is crucial to parsing the speech signal into syllabic chunks which are the fundamental unit of speech processing. He refers to this fundamental unit as the 'theta-syllable' [23] and claims that it is central to the theory of speech comprehension as a faculty facilitated by cortical entrainment to speech signals. In this view, the theta-syllable is defined as 'a theta-cycle long speech segment located between two successive vocalic nuclei'. A vocalic nucleus is the critical section of a syllable (often a vowel) that can be preceded or followed by other marginal sounds (often consonants). For example, the word *window* can be divided into the syllables /ˈwɪn/ and /doʊ/. In the /ˈwɪn/ syllable, the /ɪ/ sound functions as the nucleus while the /w/ and /n/ sounds are on the margins of that nucleus.

In this interpretation, the theta rhythm is essentially the 'master' rhythm that is responsible for the main chunking of the incoming speech stimulus and leads to the efficient parsing of syllables. This is crucial to understanding speech comprehension because of one of the most important mysteries in this area is how the brain chunks information into syllables that can then be parsed and transformed into higher-order representations. Without the ability to transform incoming acoustic information into smaller 'chunks', the brain would have no ability to separate out syllabic information and parse what the speech information in a given chunk of time is. Instead, it would appear more as an undifferentiated mass.

2 Neural Auditory Processing

In the above sections, the emphasis has been on the properties of the acoustic stimulus itself, its important features, and the mechanisms by which it is detected by the human auditory system. This chapter is dedicated to the transformation and integration of that acoustic information after it has been detected by the auditory system. Once an acoustic stimulus has been transformed into neural impulses by the machinery of the inner and outer ear, it proceeds through the brainstem and into the cortex. Roughly, information travels from the cochlea to the cochlear nuclei of the brainstem, the superior olive, the inferior colliculus, the medial geniculate nucleus of the thalamus, and finally to the primary auditory cortex [32, 70]. This outline of the auditory pathway is a gross oversimplification but useful for this review as it shows that there is some processing of incoming speech signals being done before the neural impulse reaches the brain. Brainstem and subcortical innervation sites are more numerous than described above, and there are also significant efferent connections from the higher auditory processing areas all the way down to the cochlear neurons themselves [37, 70]. This overview is meant to provide a rough sketch to demonstrate the hierarchical and interconnected nature of the auditory processing pathways as well as to give a general view of how information proceeds from the ear to the cortex. Here, we will mainly concern ourselves with the processing of neural signals once they reach the primary auditory cortex and other related areas.

This pathway shares some features of the visual processing pathway. It is hierarchical, but not strictly so, and it contains significant back-projections and divergent pathways [70, 32]. Hickok and Poeppel, two prominent voices in the field, maintain that there are actually two parallel cortical pathways for acoustic information much like the parallel processing pathways found in the visual system [32, 31, 33]. The ventral auditory processing stream is thought to underlie the conversion from acoustic signals into lexical and semantic representations. The analogy could be drawn to the ventral visual stream, which is thought to mainly be responsible for visual object recognition. In a similar way, the ventral auditory stream underlies the ability to recognize acoustic 'objects' and connect those to phonological or semantic meanings [51, 60]. The dorsal stream of the auditory pathway has a less well-understood function, but it is thought to be involved in sensorimotor integration in much the same way as the dorsal visual stream is. In particular, Hickok and Poeppel theorize that this network is crucial in the development of speech as it integrates the sounds that are perceived and facilitates the motor learning task of learning to speak.

A competing theory put forward by Angela Friederici [20] is that there are actually two ventral pathways and two parallel dorsal pathways. In this theory, there is one ventral pathway from roughly Brodmann's area 45 to the temporal cortex and another from the frontal operculum (FOP) to the uncinate fascile (UF). These two ventral pathways are hypothesized to mainly be responsible for language processing and the processing of adjacent elements in an audio signal. The two dorsal pathways are thought to connect from the temporal cortex to the premotor cortex, as well as from the temporal cortex to Brodmann's area 44. The first pathway is thought to mainly support auditory motor functions similar to those proposed by Hickok and Poeppel, while the second is thought to be involved in more high-level language processing functions. Specifically, the second pathway is thought to provide a complement to the ventral streams in that it connects information that is not adjacent in the auditory stream and allows for grammatical analysis and connections.

2.1 Primary Auditory Cortex

The primary auditory cortex (A1) is the main target of auditory information from the sensory neurons in the cochlea and is the backbone of auditory perception. It is located on the superior temporal gyrus (see Fig. 5) and can be divided into a primary area and various belt or peripheral areas [56]. A1 is organized tonotopically, meaning that there is a separation by frequency of the incoming auditory signals, and that neighboring neurons respond to neighboring frequencies [42]. Organization of A1 is analogous to the topographical mapping that can be found in V1, and mirrors the tonotopic organization of the basilar membrane, discussed above. Therefore, A1 can be said to have a frequency 'axis', groupings of neurons that react selectively to the frequency content of incoming auditory signals. Additionally, A1 contains an orthogonal 'axis' that processes information related to the binaural aspects of incoming auditory signals and may serve localization or source identification purposes [56].

Bilateral destruction of A1 results in cortical deafness, resulting in a total loss of hearing faculties in those affected. In order to manifest, this disorder requires total bilateral destruction of A1. Due to this, as well as the relatively high degree of redundancy in subcortical auditory structures, the disorder is quite rare [55]. Additionally, subcortical processing of auditory signals may preserve some auditory capacities even in patients with this condition [2]. More commonly, following damage to the cortical auditory areas are the conditions of auditory agnosia, pure word deafness, and

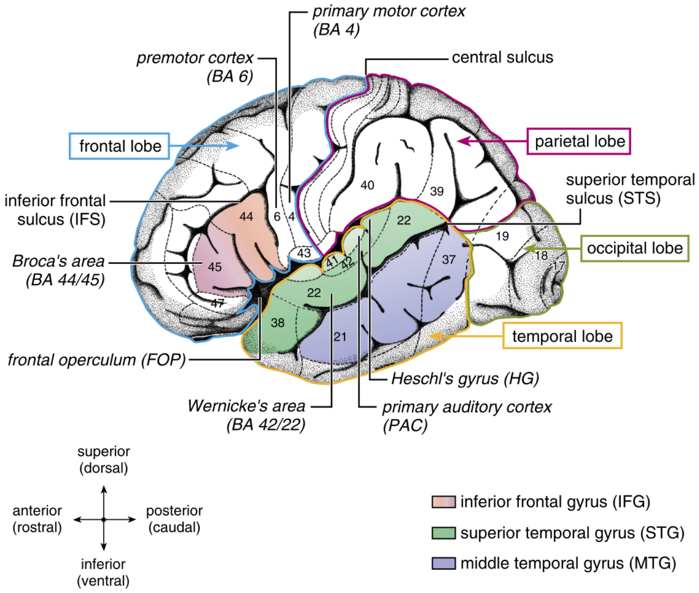


Figure 5: This figure shows the gross anatomy of the left hemisphere of the human brain. The lobes, Brodmann's areas, and other areas of interest have been highlighted. [20]

phonoagnosia.

Auditory agnosia is a condition in which patients are incapable of identifying sounds such as coughing or whistling, but show no evidence of impaired auditory speech comprehension. Pure word deafness is essentially the opposite of auditory agnosia, where patients are incapable of comprehending speech while maintaining their ability to speak, read, and identify sounds. Interestingly, some patients with pure word deafness retain their ability to extract information about a speaker (such as age, sex, etc.) based on their speech [55]. Phonoagnosia is a condition analogous to prosopagnosia (inability to recognize familiar faces), where patients lack the ability to recognize familiar voices.

These conditions provide insights into the different functions that are performed by the auditory processing system and demonstrate that these functions can be dissociated without disrupting the system completely. This observation points to the fact that the human auditory system is attempting to extract multiple types of information from an incoming auditory signal and that the methods for extracting this information involve separate cortical processing mechanisms.

3 Neural Entrainment & Stimulus-Driven Dynamics

Entrainment is defined as the synchronization of two or more oscillators that can generate their own rhythms in the absence of any rhythmic input. In the cases that will be discussed below, the oscillators are a neural population, and a rhythmic input stimulus (i.e. speech). This definition is intended to contrast the frequency and phase entrainment responses with others, such as the envelope-following response.

3.1 Brainstem Entrainment

The Auditory Brainstem Response (ABR) is a well-known neural response to auditory stimuli that was discovered more than 40 years ago [36, 35] and has seen widespread use in clinical settings for determining auditory thresholds or diagnosing neuropathologies [66] and is especially popular for hearing screening in infants. Measurement of this response is one of the most common clinical applications of auditory evoked potentials and the typical waveform is well characterized.

The test typically consists of presenting a subject with a series of click stimuli while recording neural activity through surface electrodes placed on the scalp. The averaged activity is then examined for the characteristic response that develops over a window of ~10ms and consists of a series of waveform peaks labeled *I-VII* [65, 1]. This activity is a stimulus-driven response that is generated in the brainstem and the auditory nerve and as such cannot be equated with the conscious perception of a stimulus, but does seem to be highly predictive of a subject's level of hearing loss [65].

More recently, there has been work involving the measurement of the ABR in response to complex stimuli, as opposed to the simple clicks or tones that were used previously. For example, Galbraith et al. (1995) [21] found that intelligible speech could be recovered from the ABR that is recorded while the subject is presented with speech stimuli. Further studies have been conducted looking at a diverse range of stimuli including words, phrases, and music, and this work is ongoing. ABR still remains a reliable indicator of healthy auditory processing precisely because it is a more low level response that is easily disrupted by problems at the level of the ear or afferent auditory fibers.

However, due to the low-level nature of the ABR, it is not a very useful response for studying the higher-level auditory responses, such as speech processing. The fact that it responds mainly to changes that occur at the level of the ear limits its usefulness in studying cortical processing of auditory input.

3.2 Cortical Entrainment and Envelope Following

In addition to entrainment occurring in the brainstem, it is also possible to measure cortical entrainment in response to acoustic stimuli with methods such as electroencephalography and magnetoencephalography. This type of entrainment typically has the best correspondence to the temporal modulation of the amplitude envelope of the speech signal [10, 8, 49, 34]. This cortical entrainment seems to be linked to a number of behavioral factors, such as attention [15, 73] as well as engagement with the stimulus [14]. This entrainment is clearly an exogenous, stimulus-driven process, which is illustrated by the consistent responses elicited across subjects by the same audio or visual stimuli [4, 53].

In discussing neural entrainment to any feature in natural speech, it is important to recall that speech is hierarchically structured and therefore the auditory stream is encoding information at different rates. For example, the syllabic rate of a certain speech segment may be 4Hz, but this is concurrent with a phrasal rate of, say, 2Hz. Even if the 4Hz rhythm is the only one that is 'actually' contained within the stimulus, there should still be a response in the brain to the 2Hz phrasal rhythm in order for us to be able to comprehend the speech. This phenomenon has been demonstrated empirically [12] and corresponds with the idea that the brain is tracking rhythmic properties of acoustic stimuli.

It has been demonstrated that cortical neurons entrain to the amplitude oscillations present in speech stimuli, but the causal mechanism of this entrainment is still being debated in the field. Some claim that it is caused by entrainment to low-level features of speech sounds (defined as the sound amplitude and spectral content), basically a version of an auditory steady-state potential (such as those found in the brainstem discussed in section 3.1). Others maintain that this phase entrainment is a product of more high-level speech sound features and reflects a processing of information carried in the speech signal by the brain.

There is a growing body of literature that supports the hypothesis that cortical entrainment to the temporal modulation of speech is the critical process that enables speech comprehension [44, 46, 73, 16]. This entrainment is found in both frequency (FM) and amplitude (AM) modulation, though the two are not independent [9]. Ding & Simon (2009) [9] find that for a signal that is both frequency ($f_{FM} = 40Hz$) and amplitude ($f_{AM} \leq 15Hz$) modulated, there is an auditory steady-state response at both f_{FM} as well as at f_{AM} but that the response at f_{FM} is amplitude and frequency modulated with fundamental frequency f_{AM}. This points to the fact that when we speak about entrainment of signals in the brain there are multiple different 'sites' of entrainment and that neural oscillations can be entrained to multiple features of a stimulus.

Zoefel and VanRullen [74] address this question by presenting subjects with mixed speech/noise stimuli that retain the patterns of higher-level features, but do not have the fluctuations in spectral content and sound amplitude that some claim is the basis for the phase entrainment response. They found that neural phase entrainment occurs despite the missing low-level content, indicating that it is the higher-level features that drive this response. Additionally, they find that reversing their speech/noise mixture stimuli does not eliminate the phase entrainment response. This finding seems to indicate that the entrainment is in response to higher-level features, but only those that are acoustic in nature and not linguistic features, as those are absent in time-reversed speech.

What, exactly, the functional role of this entrainment to the speech envelope is remains a topic of debate in the field, but there has recently been strong evidence to support the hypothesis that cortical entrainment to the speech envelope facilitates comprehension of the speech signal [10, 11, 50]. These observations do show that there is a link between speech intelligibility and stronger neural entrainment, but it is not possible to tell what the precise cause is. It could be due to a stronger evoked response to a stimulus that is easier to hear and understand. Or, it could be due to stronger recruitment of the oscillatory mechanisms that facilitate the intelligibility of the speech.

Some studies have looked to address this problem by dissociating these two possible mechanisms of neural entrainment. One by Kösem et al. [40] attempted to do this by quickly changing the speech rate at the end of a sentence. If the observed neural entrainment to the stimulus is indeed caused by internal oscillatory mechanics, then these oscillators would entrain to the initial part of the sentence and when the speech rate is changed at the end of the sentence, it will be mis-perceived. This is indeed what they found. Participants were asked to indicate whether they heard a short [ɑ] or a long [aː] at the end of the sentence, and their answer was dependent on whether the speech preceding this vowel was presented at a fast or a slow rate. This study points to the fact that neural entrainment does not simply reflect an evoked response to low level features of the speech input, but an endogenous, ongoing neural process of speech parsing that influences our perception.

The fact that cortical neurons track the amplitude envelope of speech signals is now well-known and is considered an important part of the speech processing pathway. But, the interaction between natural cortical rhythms (discussed in Section 1.2) and the envelope following response is still not completely understood, and its functional role remains a subject of debate. There is abundant evidence that broad-spectrum EEG signals can be used to decode which speaker is being attended to by a subject in a cocktail party scenario [34, 67], but these studies only offer a partial picture of the role of cortical envelope-following responses in speech comprehension.

For instance, Zion Golumbic et al. [73] were able to localize the envelope tracking response both in terms of cortical location as well as which frequency band was recruited for entrainment using electrocorticography (ECoG) in a cocktail party scenario with two speakers. They found found that there is a significant ability to predict which speaker is being attended in two different frequency bands that they term 'low-frequency' (1-7 Hz) and 'high gamma' (70-150 Hz). Further, they found that cortical locations in which these two frequency bands possessed significant predictive power differed. Both showed a very strong response in the superior temporal gyrus, which is to be expected given that is the location of the early cortical auditory areas, but the high gamma band showed a more widespread distribution around the brain, while the low-frequency band was more concentrated in frontal and temporal areas.

These findings point to the conclusion that the envelope following response is not monolithic, and that there are multiple functional roles filled by the envelope following response.

One of the most popular theories currently is that speech signals are essentially a quasi-rhythmic input which certain innate cortical rhythms are recruited to track. This tracking then enables both the exclusion of irrelevant speech signals [34, 50] as well as the syllabic parsing that is thought to be accomplished by the innate theta oscillations in the cortex [16, 22].

Interestingly, speech remains intelligible despite a large amount of information in the speech signal being destroyed. Speech information is resistant to both degradation of temporal as well as spectral information [63, 17]. Barring total elimination of temporal and spectral information, it is only when both are sufficiently degraded that speech becomes unintelligible [18]. Related to this, the phenomenon of speaker gender identification is also reliant on similar information and can be modulated by changing the information present in the temporal and the spectral domain. Elliott and Theunissen [18] found that speaker gender identification rate can be significantly degraded by removing spectral modulations between 3 and 7 cycles/Hz and that this specifically reduced the subjects' ability to correctly identify female speakers. They postulated that this is because the female vocal spectrum has more power in this specific range and this is what contributes to 'sounding like' a female speaker. Clearly, there is more information than just that necessary for comprehension that is embedded in the temporal and spectral modulations of speech and it is important to parse out what is necessary for the comprehension of speech and what information is supplemental, but not necessary to the bare comprehension of speech.

This brings up a large controversy in this field of study, which is the question of whether neural entrainment occurs because of the recruitment of natural neural rhythms that are simply used for alignment and parsing of the incoming speech signal, or whether these oscillations are actually generated by the speech signal itself and do not rely on oscillatory activity that is always ongoing in the cortex. This distinction is important because it colors how the increase in stimulus-aligned activity is interpreted. In the case of innate, ongoing oscillations being recruited to speech comprehension purposes, the activity that is elicited by presentation of a stimulus is used in 'chunking' the incoming speech signal into segments that can be processed for lexical information and then combined to give some type of semantic meaning to a sentence. On the other hand, if the increase in stimulus-aligned activity is due to the emergence of a signal in the brain that tracks the amplitude envelope of the incoming speech signal, then it would seem that the analysis being performed in the brain is not of the 'chunking' type, but involves synchronizing to the rate of the incoming speech signal and using information other than the relation of the incoming speech signal to innate oscillatory activity to process and comprehend speech.

4 Conclusion

We know that language is a way of encoding information in a systematic, temporally organized manner and transmitting it in a way that is intelligible to other speakers of that language. We also know that the information is hierarchically organized, and that there is meaning embedded in each level of this hierarchy. Various attempts have been made to look at features within the speech signal and use these to decode what is going on inside the brain. These attempts, though, have mostly just focused on one speech feature at a time (e.g. the amplitude envelope, spectral content, modulation spectrum, etc.). It is becoming more and more clear that an approach that is informed by the hierarchical nature of speech will be critical in developing models that accurately reflect all of the levels of processing that are ongoing in the brain when speech is being decoded. Simply looking to one feature as a broad reflection of the processing that is going on is not sufficient to describe the richness that is embedded in human communication and the models that are based on such an approach are bound to come up short in many ways. In order to make more informed and accurate temporal predictions, it is my opinion, expressed in this review, that a more holistic and hierarchical modeling approach must be taken. It is only by looking at slower, sentence and phrasal rhythms in conjunction with faster temporal fine-structure, and everything in between, that we will be able to have a full and complete view of the information that the brain is attempting to decode and then make use of.

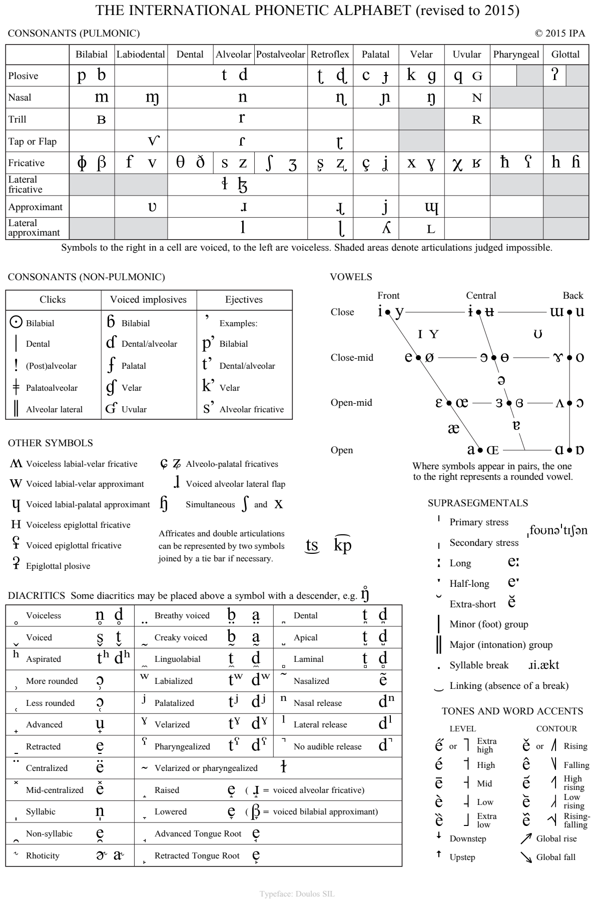
In this review, I have laid out the background and basis for various theories on the role of oscillatory mechanics in speech comprehension. I have gone over the various properties of spoken language that are used by the brain in order to decode the information present in speech, as well as some of the basic and fundamental transformations that speech information undergoes as part of the speech comprehension chain, before it is processed in the cortex. Most importantly, I have looked to highlight the ongoing debate in the field about the existence and role of neural entrainment to a stimulus in speech comprehension.

While there is still much work to be done, it has been fairly well characterized that correlations can be detected between the amplitude envelope of speech, and the neural oscillations that can be detected in the brain via methods such as EEG and MEG. The main question now is whether these oscillations represent a low-level bottom-up response that is purely a reflection of the acoustic properties of the stimulus itself, or whether this is a more complex, more top-down response that reflects a processing of higher level linguistic elements beyond pure acoustics. In my opinion, there is much more going on here than mere bottom-up representations of stimulus acoustics. The studies that show oscillations in the brain at phrasal rates that are not actually represented in the stimulus acoustics go some way to demonstrate this. Additionally, 'cocktail party' type studies that show that there is not a broad representation of all stimuli that are heard, but rather only those that are attended to, also point in this direction.

That all being said, more work still needs to be done to demonstrate the causal relationship between this phenomenon of speech tracking and speech comprehension. There have already been some studies that demonstrate the effect of top-down factors such as attention on the strength of speech tracking, which is supportive of the idea that it is more than just a bottom-up stimulus driven response, but other methods of showing this are necessary if it is to be accepted more widely. For instance, some groups have designed experiments that manipulate linguistic info while leaving the stimulus acoustics alone [12, 38]. These types of clever experimental designs help to demonstrate that there are indeed aspects of linguistic tracking that are present when this speech tracking phenomenon occurs.

There remains much work to be done, but there is room in the field for improved experiments and experimental designs. Future experiments could benefit from the types of acoustic and linguistic dissociations that have been utilized by some researchers. Additionally, there is also more research that could be done in the domain of showing the top-down modulation of speech tracking by higher level factors such as attention. In sum, there is still disagreement and a need for more experimental data, but there is an emerging consensus of evidence and researchers that are showing that the brain is tracking linguistic structures hierarchically in much the same way that speech is structured from a linguistic perspective. This theory answers some old questions about the rate of speech and how speech processing is carried from acoustics through to comprehension, but opens a new world of questions about how these oscillations created by the stimulus interact with endogenous oscillations and contribute to overall perception of time as well as speech.

A Appendix



THE INTERNATIONAL PHONETIC ALPHABET (revised to 2015)

CONSONANTS (PULMONIC)

© 2015 IPA

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k ɡ	q ɢ	▨	ʔ ▨
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ	▨	▨
Trill	ʙ			r				▨	ʀ		▨
Tap or Flap		ⱱ		ɾ		ɽ		▨			▨
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative	▨	▨		ɬ ɮ						▨	▨
Approximant		ʋ		ɹ		ɻ	j	ɰ			▨
Lateral approximant	▨	▨		l		ɭ	ʎ	ʟ		▨	▨

Symbols to the right in a cell are voiced, to the left are voiceless. Shaded areas denote articulations judged impossible.

CONSONANTS (NON-PULMONIC)

Clicks	Voiced implosives	Ejectives
ʘ Bilabial	ɓ Bilabial	ʼ Examples:
ǀ Dental	ɗ Dental/alveolar	pʼ Bilabial
ǃ (Post)alveolar	ʄ Palatal	tʼ Dental/alveolar
ǂ Palatoalveolar	ɠ Velar	kʼ Velar
ǁ Alveolar lateral	ʛ Uvular	sʼ Alveolar fricative

VOWELS

	Front		Central		Back
Close	i • y		ɨ • ʉ		ɯ • u
		ɪ ʏ		ʊ	
Close-mid	e • ø		ɘ • ɵ		ɤ • o
			ə		
Open-mid	ɛ • œ		ɜ • ɞ		ʌ • ɔ
		æ	ɐ		
Open	a • ɶ				ɑ • ɒ

Where symbols appear in pairs, the one to the right represents a rounded vowel.

OTHER SYMBOLS

ʍ Voiceless labial-velar fricative

w Voiced labial-velar approximant

ɥ Voiced labial-palatal approximant

ʜ Voiceless epiglottal fricative

ʢ Voiced epiglottal fricative

ʡ Epiglottal plosive

ɕ ʑ Alveolo-palatal fricatives

ɺ Voiced alveolar lateral flap

ɧ Simultaneous ʃ and x

Affricates and double articulations can be represented by two symbols joined by a tie bar if necessary. t͜s k͡p

SUPRASEGMENTALS

ˈ	Primary stress	ˌfoʊnəˈtɪʃən
ˌ	Secondary stress	
ː	Long	eː
ˑ	Half-long	eˑ
˘	Extra-short	ĕ
|	Minor (foot) group	
‖	Major (intonation) group	
.	Syllable break	ɹi.ækt
‿	Linking (absence of a break)	

DIACRITICS Some diacritics may be placed above a symbol with a descender, e.g. ŋ̊

◌̥	Voiceless	n̥ d̥	◌̤	Breathy voiced	b̤ a̤	◌̪	Dental	t̪ d̪
◌̬	Voiced	s̬ t̬	◌̰	Creaky voiced	b̰ a̰	◌̺	Apical	t̺ d̺
ʰ	Aspirated	tʰ dʰ	◌̼	Linguolabial	t̼ d̼	◌̻	Laminal	t̻ d̻
◌̹	More rounded	ɔ̹	ʷ	Labialized	tʷ dʷ	◌̃	Nasalized	ẽ
◌̜	Less rounded	ɔ̜	ʲ	Palatalized	tʲ dʲ	ⁿ	Nasal release	dⁿ
◌̟	Advanced	u̟	ˠ	Velarized	tˠ dˠ	ˡ	Lateral release	dˡ
◌̠	Retracted	e̠	ˤ	Pharyngealized	tˤ dˤ	̚	No audible release	d̚
◌̈	Centralized	ë	~	Velarized or pharyngealized	ɫ			
◌̽	Mid-centralized	e̽	◌̝	Raised	e̝ (ɹ̝ = voiced alveolar fricative)			
◌̩	Syllabic	n̩	◌̞	Lowered	e̞ (β̞ = voiced bilabial approximant)			
◌̯	Non-syllabic	e̯	◌̘	Advanced Tongue Root	e̘			
˞	Rhoticity	ɚ a˞	◌̙	Retracted Tongue Root	e̙			

TONES AND WORD ACCENTS

LEVEL			CONTOUR		
e̋ or	˥	Extra high	ě or	˩˥	Rising
é	˦	High	ê	˥˩	Falling
ē	˧	Mid	e᷄	˦˥	High rising
è	˨	Low	e᷅	˩˨	Low rising
ȅ	˩	Extra low	e᷈	˧˦˧	Rising-falling
ꜜ	Downstep		↗	Global rise	
ꜛ	Upstep		↘	Global fall	

Typeface: Doulos SIL

Figure A.1: The International Phonetic Alphabet as defined by the International Phonetic Association

References

[1] Neil Bhattacharyya. *Auditory Brainstem Response Audiometry*. 2017.

[2] Marianna Cavinato et al. “Preservation of Auditory P300-Like Potentials in Cortical Deafness”. In: *PLoS ONE* 7.1 (2012), pp. 1–6.

[3] Noam Chomsky. *Knowledge of Language*. Vol. 8. 2. 1986, pp. 139–159.

[4] Samantha S. Cohen, Simon Henin, and Lucas C. Parra. “Engaging narratives evoke similar neural activity and lead to similar time perception”. In: *Scientific Reports* 7.1 (2017), pp. 1–10.

[5] R.M. Dauer. “Stress-timing and Syllable-Timing reanalysed”. In: *Journal of Phonetics* 11 (1983), pp. 51–62.

[6] P.B. Denes and E.N Pinson. *The Speech Chain*. 1993, pp. 1–9.

[7] Frederic Dick et al. “Language deficits, localization, and grammar: Evidence for a distributive model of language breakdown in aphasic patients and neurologically intact individuals”. In: *Psychological Review* 108.4 (2001), pp. 759–788.

[8] Nai Ding, Monita Chatterjee, and Jonathan Z. Simon. “Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure”. In: *NeuroImage* 88 (2014), pp. 41–46.

[9] Nai Ding and Jonathan Z Simon. “Neural Representations of Complex Temporal Modulations in the Human Auditory Cortex”. In: *Journal of Neurophysiology* 102.5 (2009), pp. 2731–2743.

[10] Nai Ding and Jonathan Z. Simon. “Cortical entrainment to continuous speech: functional roles and interpretations”. In: *Frontiers in Human Neuroscience* 8 (2014).

[11] Nai Ding and Jonathan Z. Simon. “Neural coding of continuous speech in auditory cortex during monaural and dichotic listening”. In: *Journal of Neurophysiology* 107.1 (2012), pp. 78–89.

[12] Nai Ding et al. “Cortical tracking of hierarchical linguistic structures in connected speech”. In: *Nature Neuroscience* 19.1 (2015), pp. 158–164.

[13] Nai Ding et al. “Temporal modulations in speech and music”. In: *Neuroscience & Biobehavioral Reviews* 81 (2017), pp. 181–187.

[14] Jacek P. Dmochowski et al. “Extracting multidimensional stimulus-response correlations using hybrid encoding-decoding of neural activity”. In: *NeuroImage* May (2017), pp. 1–13.

[15] Jacek P Dmochowski et al. “Multidimensional stimulus-response correlation reveals supramodal neural responses to naturalistic stimuli”. In: (2016), pp. 1–18.

[16] Keith B. Doelling et al. “Acoustic landmarks drive delta-theta oscillations to enable speech comprehension by facilitating perceptual parsing”. In: *NeuroImage* 85 (2014), pp. 761–768.

[17] Rob Drullman, Joost M. Festen, and Reinier Plomp. “Effect of temporal envelope smearing on speech reception”. In: *The Journal of the Acoustical Society of America* 95.2 (1994), pp. 1053–1064.

[18] Taffeta M. Elliott and Frédéric E. Theunissen. “The modulation transfer function for speech intelligibility”. In: *PLoS Computational Biology* 5.3 (2009).

[19] J.A. Fodor. *The modularity of mind: an essay on faculty psychology.* 1983. 1983, pp. 878–914.

[20] Angela D. Friederici. “The Brain Basis of Language Processing: From Structure to Function”. In: *Physiological Reviews* 91.4 (2011), pp. 1357–1392.

[21] Gary C. Galbraith et al. *Intelligible speech encoded in the human brain stem frequency-following response.* 1995.

[22] Oded Ghitza. “The theta-syllable: A unit of speech information defined by cortical function”. In: *Frontiers in Psychology* 4.MAR (2013), pp. 1–5.

[23] Oded Ghitza, Anne-Lise Giraud, and David Poeppel. “Neuronal oscillations and speech perception: critical-band temporal envelopes are the essence”. In: *Frontiers in Human Neuroscience* 6 (2013).

[24] Oded Ghitza and Steven Greenberg. “On the possible role of brain rhythms in speech perception: Intelligibility of time-compressed speech with periodic and aperiodic insertions of silence”. In: *Phonetica* 66.1-2 (2009), pp. 113–126.

[25] Anne-Lise Giraud and David Poeppel. “Cortical oscillations and speech processing: emerging computational principles and operations.” In: *Nature neuroscience* 15.4 (2012), pp. 511–7.

[26] Esther Grabe and Ee Ling Low. “Durational variability in speech and the rhythm class hypothesis”. In: *Papers in laboratory phonology* 7.4 (Apr. 2002), pp. 515–546.

[27] Steven Greenberg et al. “Temporal properties of spontaneous speech - A syllable-centric perspective”. In: *Journal of Phonetics* 31.3-4 (2003), pp. 465–485.

[28] Ricardo Gutierrez-Osuna. *Organization of Speech Sounds.* 2017.

[29] Saskia Haegens and Elana Zion Golumbic. “Rhythmic facilitation of sensory processing: A critical review”. In: *Neuroscience and Biobehavioral Reviews* 86.December 2017 (2018), pp. 150–165.

[30] B Hayes. “The phonology of rhythm in English”. In: *Linguistic Inquiry* 15.1 (1984), pp. 33–74.

[31] Gregory Hickok and David Poeppel. “Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language”. In: *Cognition* 92.1-2 (2004), pp. 67–99.

[32] Gregory Hickok and David Poeppel. “The cortical organization of speech processing”. In: *Nature Reviews Neuroscience* 8.5 (2007), pp. 393–402.

[33] Gregory Hickok and David Poeppel. *Towards a functional neuroanatomy of speech perception.* 2000.

[34] Cort Horton, Ramesh Srinivasan, and Michael D’Zmura. “Envelope responses in single-trial EEG indicate attended speaker in a ‘cocktail party’”. In: *Journal of Neural Engineering* 11.4 (2014), p. 046015.

[35] D. L. Jewett, M. N. Romano, and J. S. Williston. “Human Auditory Evoked Potentials: Possible Brain Stem Components Detected on the Scalp”. In: *Science* 167.3924 (1970), pp. 1517–1518.

[36] Don L. Jewett and John S. Williston. “Auditory-evoked far fields averaged from the scalp of humans”. In: *Brain* 94.4 (1971), pp. 681–696.

[37] E R Kandel, J H Schwartz, and T M Jessell. *Principles of Neural Science.* Vol. 3. 2000, p. 1414.

[38] Stephanie J Kayser et al. “Irregular Speech Rate Dissociates Auditory Cortical Entrainment, Evoked Responses, and Frontal Alpha.” In: *The Journal of neuroscience : the official journal of the Society for Neuroscience* 35.44 (2015), pp. 14691–701.

[39] Johannes Knaus, Richard Wiese, and Ulrike Janßen. “The processing of word stress: EEG studies on task-related components”. In: ... *of the International Congress of* ... August (2007), pp. 709–712.

[40] Anne Kösem and Virginie van Wassenhove. “Distinct contributions of low- and high-frequency neural oscillations to speech comprehension”. In: *Language, Cognition and Neuroscience* 32.5 (2017), pp. 536–544.

[41] Peter Ladefoged and Ian Maddieson. *The Sounds of the World's Languages.* 1996.

[42] Judith L Lauter et al. “Tonotopic organization in human auditory cortex revealed by positron emission tomography”. In: *Hearing Research* 20 (1985), pp. 199–205.

[43] James G. Martin. “Rhythmic (hierarchical) versus serial structure in speech and other behavior”. In: *Psychological Review* 79.6 (1972), pp. 487–509.

[44] Lars Meyer and Matthias Gumbert. “Synchronization of Electrophysiological Responses with Speech Benefits Syntactic Information Processing”. In: *Journal of Cognitive Neuroscience* 26.3 (2018), pp. 1–10.

[45] C. M. Michel et al. “Localization of the sources of EEG delta, theta, alpha and beta frequency bands using the FFT dipole approximation”. In: *Electroencephalography and Clinical Neurophysiology* 82.1 (1992), pp. 38–44.

[46] Benjamin Morillon and Charles E. Schroeder. “Neuronal oscillations as a mechanistic substrate of auditory temporal prediction”. In: *Annals of the New York Academy of Sciences* 1337.1 (2015), pp. 26–31.

[47] R. C. Muresan et al. “The Oscillation Score: An Efficient Method for Estimating Oscillation Strength in Neuronal Activity”. In: *Journal of Neurophysiology* 99.3 (2008), pp. 1333–1353.

[48] Lloyd H. Nakatani, Kathleen D. O’Connor, and Carletta H. Aston. “Prosodic aspects of American English speech rhythm”. In: *The Journal of the Acoustical Society of America* 69.S1 (1981), S82–S82.

[49] K. V. Nourski et al. “Temporal Envelope of Time-Compressed Speech Represented in the Human Auditory Cortex”. In: *Journal of Neuroscience* 29.49 (2009), pp. 15564–15574.

[50] James A. O’Sullivan et al. “Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG”. In: *Cerebral Cortex* 25.7 (2015), pp. 1697–1706.

[51] Geoffrey J M Parker et al. “Lateralization of ventral and dorsal auditory-language pathways in the human brain”. In: *NeuroImage* 24.3 (2005), pp. 656–666.

[52] François Pellegrino, Christophe Coupé, and Egidio Marsico. “Across-Language Perspective on Speech Information Rate”. In: *Language* 87.3 (2011), pp. 539–558.

[53] Agustin Petroni et al. “Age and sex modulate the variability of neural responses to naturalistic videos”. In: *bioRxiv* (2017).

[54] Steven Pinker and Ray Jackendoff. “The faculty of language: What’s special about it?” In: *Cognition* 95.2 (2005), pp. 201–236.

[55] Michael R. Polster and Sally B. Rose. “Disorders of auditory processing: Evidence for modularity in audition”. In: *Cortex* 34.1 (1998), pp. 47–65.

[56] Dale Purves et al. *Neuroscience. 2nd edition.* 2001.

[57] Lawrence R. Rabiner and Ronald W. Schafer. “Introduction to Digital Speech Processing”. In: *Foundations and Trends® in Signal Processing* 1.1–2 (2007), pp. 1–194.

[58] Franck Ramus, Marina Nespor, and Jacques Mehler. “Correlates of linguistic rhythm in the speech signal”. In: *Cognition* 73.3 (1999), pp. 265–292.

[59] Madhavi Rangaswamy et al. “Beta power in the EEG of alcoholics”. In: *Biological Psychiatry* 52.8 (2002), pp. 831–842.

[60] Josef P. Rauschecker and Sophie K. Scott. “Maps and streams in the auditory cortex: Nonhuman primates illuminate human speech processing”. In: *Nature Neuroscience* 12.6 (2009), pp. 718–724.

[61] Kathrin Rothermich, Maren Schmidt-Kassow, and Sonja A. Kotz. “Rhythm’s gonna get you: Regular meter facilitates semantic sentence processing”. In: *Neuropsychologia* 50.2 (Jan. 2012), pp. 232–244.

[62] Jan Schnupp, Israel Nelken, and Andrew King. *Auditory Neuroscience*. 2011, pp. 1–50.

[63] Rosaria Silipo, Steven Greenberg, and Takayuki Arai. “Temporal constraints on speech intelligibility as deduced from exceedingly sparse spectral representations”. In: *Proceedings of Eurospeech* (1999), pp. 2687–2690.

[64] Gary F Simons and Charles D Fennig. *How many languages are there in the world?* 2017.

[65] Yvonne S. Sininger. “Auditory Brain Stem Response for Objective Measures of Hearing”. In: *Ear and Hearing* 14.1 (1993), pp. 23–30.

[66] Erika Skoe and Nina Kraus. “Auditory Brain Stem Response to Complex Sounds: A Tutorial”. In: *Ear and Hearing* 31.3 (2010), pp. 302–324.

[67] Tobias de Taillez, Birger Kollmeier, and Bernd T. Meyer. “Machine learning for decoding listeners’ attention from electroencephalography evoked by continuous speech”. In: *European Journal of Neuroscience* June (2018), pp. 1–8.

[68] Sophia Uddin et al. “Hearing sounds as words: Neural responses to environmental sounds in the context of fluent speech”. In: *Brain and Language* 179.February (2018), pp. 51–61.

[69] Noriko Umeda. “Vowel duration in american english”. In: *Journal of the Acoustical Society of America* 58.2 (1975), pp. 434–445.

[70] Douglas B. Webster, Arthur N. Popper, and Richard R. Fay, eds. *The Mammalian Auditory Pathway: Neuroanatomy*. Vol. 1. Springer Handbook of Auditory Research. New York, NY: Springer New York, 1992.

[71] Wikimedia Commons. *Spectrogram iua*. 2005.

[72] Eric D. Young and Murray B. Sachs. “Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers”. In: *The Journal of the Acoustical Society of America* 66.5 (1979), pp. 1381–1403.

[73] Elana M. Zion Golumbic et al. “Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party"”. In: *Neuron* 77.5 (2013), pp. 980–991.

[74] Benedikt Zoefel and Rufin VanRullen. “EEG oscillations entrain their phase to high-level features of speech sound”. In: *NeuroImage* 124 (2016), pp. 16–23.