Published in final edited form as:

Brain Lang. 2012 September ; 122(3): 151–161. doi:10.1016/j.bandl.2011.12.010.

Temporal Context in Speech Processing and Attentional Stream Selection: A Behavioral and Neural perspective

Elana M. Zion Golumbic^{a,b}, David Poeppel^c, and Charles E. Schroeder^{a,b}

^aDepartments of Psychiatry and Neurology, Columbia University Medical Center, 710 W 168th st, New York NY 10032, USA

^bThe Nathan Kline Institute for Psychiatric Research, 140 Old Orangeburg Road, Orangeburg, NY 10962, USA

^cDepartment of Psychology, New York University, 6 Washington Place, New York NY 10003, USA

Abstract

The human capacity for processing speech is remarkable, especially given that information in speech unfolds over multiple time scales concurrently. Similarly notable is our ability to filter out of extraneous sounds and focus our attention on one conversation, epitomized by the ‘Cocktail Party’ effect. Yet, the neural mechanisms underlying on-line speech decoding and attentional stream selection are not well understood. We review findings from behavioral and neurophysiological investigations that underscore the importance of the temporal structure of speech for achieving these perceptual feats. We discuss the hypothesis that entrainment of ambient neuronal oscillations to speech’s temporal structure, across multiple time-scales, serves to facilitate its decoding and underlies the selection of an attended speech stream over other competing input. In this regard, speech decoding and attentional stream selection are examples of ‘*active sensing*’, emphasizing an interaction between proactive and predictive top-down modulation of neuronal dynamics and bottom-up sensory input.

1. Introduction

In speech communication, a message is formed over time as small packets of information (sensory and linguistic) are transmitted and received, accumulating and interacting to form a larger semantic structure. We typically perform this task effortlessly, oblivious to the inherent complexity that on-line speech processing entails. It has proven extremely difficult to fully model this basic human capacity, and despite decades of research, the question of how a stream of continuous acoustic input is decoded in real-time to interpret the intended message remains a foundational problem.

The challenge of understanding mechanisms of on-line speech processing is aggravated by the fact that speech is rarely heard under ideal acoustic conditions. Typically, we must filter out extraneous sounds in the environment and focus only on a relevant portion of the

© 2012 Elsevier Inc. All rights reserved.

Corresponding author: Elana Zion Golumbic (ezg2101@columbia.edu), Department of Neurology, Columbia University Medical Center, 710 W 168th st, RM 721, New York NY 10032, USA, Tel: (914) 589-6770, Fax: (212) 305-5445.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

auditory scene, as exemplified in the ‘Cocktail Party Problem’, famously coined by Cherry (1953). While many models have dealt with the issue of segregating multiple streams of concurrent input (Bregman, 1990), the specific attentional process of selecting one particular stream over all others, which is the cognitive task we face daily, has proven extremely difficult to model and is not well understood.

This is a truly interdisciplinary question, of fundamental importance in numerous fields, including psycholinguistics, cognitive neuroscience, computer science and engineering. Converging insights from research across these different disciplines have led to growing recognition of the centrality of the temporal structure of speech for adequate speech decoding, stream segregation and attentional stream selection (Giraud & Poeppel, 2011; Shamma, Elhilali, & Micheyl, 2011). According to this perspective, since speech is first and foremost a temporal stimulus, with acoustic content unfolding and evolving over time, decoding the momentary *spectral content* (‘fine structure’) of speech benefits greatly from considering its position within the longer *temporal context* of an utterance. In other words, the timing of events within a speech stream is critical for its decoding, attribution to the correct source, and ultimately to whether it will be included in the attentional ‘spotlight’ or forced to the background of perception.

Here we review current literature on the role of the temporal structure of speech in achieving these cognitive tasks. We discuss its importance for speech intelligibility from a behavioral perspective as well as possible neuronal mechanisms that represent and utilize this temporal information to guide on-line speech decoding and selective attention at a ‘Cocktail Party’.

2. The temporal structure of speech

2.1 The centrality of temporal structure to speech processing/decoding

At the core of most models of speech processing is the notion that speech tokens are distinguishable based on their spectral makeup (Fant, 1960; Stevens, 1998). Inspired by the tonotopic organization in the auditory system (Eggermont, 2001; Pickles, 2008), theories for speech decoding, be it by brain or by machine, have emphasized on-line spectral analysis of the input for the identification of basic linguistic units (primarily phonemes and syllables; (Allen, 1994; French & Steinberg, 1947; Pavlovic, 1987). Similarly, many models of auditory stream segregation rely on frequency separation as the primary cue for stream segregation, attributing consecutive inputs with overlapping spectral structure to the same stream (Beauvois & Meddis, 1996; Hartmann & Johnson, 1991; McCabe & Denham, 1997). However, other models have recognized that relying purely on spectral cues is insufficient for robust speech decoding and stream segregation, especially under noisy conditions (Elhilali, Ma, Micheyl, Oxenham, & Shamma, 2009; McDermott, 2009; Moore & Gockel, 2002). These models emphasize the central role of temporal cues in providing a context within which the spectral content is processed.

Psychophysical studies clearly demonstrate the critical importance of speech’s temporal structure for its intelligibility. For example, speech becomes unintelligible if it is presented at an unnaturally fast or slow rate, even if the fine structure is preserved (Ahissar et al., 2001; Ghitza & Greenberg, 2009) or if its temporal envelope is smeared or otherwise disrupted (Arai & Greenberg, 1998; Drullman, 2006; Drullman, Festen, & Plomp, 1994a, 1994b; Greenberg, Arai, & Silipo, 1998; Stone, Fullgrabe, & Moore, 2010). In particular, intelligibility of speech crucially depends on the integrity of the modulation spectrum’s amplitude in the region between 3 and 16 Hz (Kingsbury, Morgan, & Greenberg, 1998). In the extreme case of complete elimination of temporal modulations (flat envelope) speech intelligibility is reduced to 5%, demonstrating that without temporal context, the fine structure alone carries very little information (Drullman et al., 1994a, 1994b). In contrast,

high degrees of intelligibility can be maintained even with significant distortion to the spectral structure, as long as the temporal envelope is preserved. In a seminal study, Shannon et al. (1995) amplitude modulated white noise using temporal envelopes of speech which preserved temporal cues and amplitude fluctuations but without any fine structure information. Listeners were able to correctly report the phonetic content from this amplitude-modulated noise, even when the envelope was extracted from as few as four frequency bands of speech. Speech intelligibility was also robust to other manipulations of fine structure, such as spectral rotation (Blessner, 1972) and time-reversal of short segments (Saberi & Perrott, 1999), further emphasizing the importance of temporal contextual cues as opposed to momentary spectral content for speech processing. Moreover, hearing-impaired patients using cochlear implant rely heavily on the temporal envelope for speech recognition, since the spectral content is significantly degraded by such devices (Chen, 2011; Friesen, Shannon, Baskent, & Wang, 2001).

These psychophysical findings raise an important question: What is the functional role of the temporal envelope of speech, and what contextual cues does it provide that renders it crucial for intelligibility? In the next sections we discuss the temporal structure of speech and how it may relate to speech processing.

2.2 A Hierarchy of Time Scales in Speech

The temporal envelope of speech reflects fluctuations in speech amplitude at rates between 2 and 50Hz (Drullman et al., 1994a, 1994b; Greenberg & Ainsworth, 2006; Rosen, 1992). It represents the timing of events in the speech stream, and provides the temporal framework in which the fine structure of linguistic content is delivered. The temporal envelope itself contains variations at multiple distinguishable rates which probably convey different speech features. These include phonetic segments, which typically occur at a rate of ~20–50Hz; syllables which occur at a rate of 3–10Hz; and phrases whose rate is usually between ~0.5–3Hz (Drullman, 2006; Greenberg, Carvey, Hitchcock, & Chang, 2003; Poeppel, 2003; see Figure 1).

Although the various speech features evolve over differing time scales, they are not processed in isolation; rather speech comprehension involves integrating information across different time scales in a hierarchical and interdependent manner (Poeppel, 2003; Poeppel, Idsardi, & van Wassenhove, 2008). Specifically, longer-scale units are made up of several units of a shorter time-scale: words and phrases contain multiple syllables, and each syllable is comprised of multiple phonetic segments. Not only is this hierarchy a description of the basic structure of speech, but behavioral studies demonstrate that this structure alters perception and assists intelligibility. For example, Warren and colleagues (1990) found that presenting sequences of vowels at different time scales creates markedly different percepts. Specifically, when vowel duration was > 100ms (i.e. within the syllabic rate of natural speech) they maintained their individual identity and could be named. However, when vowels were 30–100ms long they lost their perceptual identities and instead the entire sequence was perceived as a word, with its emergent percept bearing little resemblance to the actual phonemes it was comprised of. There are many other psychophysical examples that the interpretation of individual phonemes and syllables is affected by prior input (Mann, 1980; Norris, McQueen, Cutler, & Butterfield, 1997; Repp, 1982), supporting the view that these *smaller units are decoded as part of a longer temporal unit rather than in isolation*. Similar effects have been reported for longer time scales. Pickett and Pollack (1963) show that single words become more intelligible when they are presented within a longer phrase versus in isolation. Several linguistic factors also exert contextual influences over speech decoding and intelligibility, including semantic (Allen, 2005; Borsky, Tuller, & Shapiro, 1998; Connine, 1987; Miller, Heise, & Lichten, 1951), syntactic (Cooper & Paccia-Cooper, 1980; Cooper, Paccia, & Lapointe, 1978) and lexical (Connine & Clifton, 1987; Ganong,

1980) influences. Discussing the nature of these influences is beyond the scope of this paper; however they emphasize the underlying hierarchical principle in speech processing. The momentary acoustic input is insufficient for decoding; rather it is greatly affected by the temporal context they are embedded in, often relying on integrating information over several hundred milliseconds (Holt & Lotto, 2002; Mattys, 1997).

The robustness of speech processing to co-articulation, reverberation, and phonemic restorations provides additional evidence that a global percept of the speech structure does not simply rely on linear combination of all the smaller building blocks. Some have explicitly argued for a ‘reversed hierarchy’ with the auditory system (Harding, Cooke, & König, 2007; Nahum, Nelken, & Ahissar, 2008; Nelken & Ahissar, 2006), and specifically of time scales for speech perception (Cooper et al., 1978; Warren, 2004). This approach advocates for global pattern recognition akin to that proposed in the visual system (Bar, 2004; Hochstein & Ahissar, 2002). Along with a bottom-up process of combining low level features over successive stages, there is a dynamic and critical top-down influence with higher-order units constraining and assisting in the identification of lower-order units. Recent models for speech decoding have proposed this can be achieved by a series of hierarchical processors which sample the input at different time scales in parallel, with the processors sampling longer timescales imposing constraints on processors of a finer time-scale (Ghitza, 2011; Ghitza & Greenberg, 2009; Kiebel, von Kriegstein, Daunizeau, & Friston, 2009; Tank & Hopfield, 1987).

2.3 Parsing and integrating – utilizing temporal envelope information

A core issue in the hierarchical perspective for speech processing is the need to parse the sequence correctly into smaller units, organize and integrate them in a hierarchical fashion in order to perform higher-order pattern recognition (Poeppel, 2003). This poses an immense challenge to the brain (as well as to computers), to convert a stream of continuously fluctuating input into a multi-scale hierarchical structure, and to do so on-line as information accumulates. It is suggested that the temporal envelope of the speech contains cues that assist in parsing the speech stream into smaller units. In particular, two basic units of speech which are important for speech parsing, the phrasal and syllabic levels, can be discerned and segmented from the temporal envelope.

The syllable has long been recognized as a fundamental unit in speech which is vital for parsing continuous speech (Eimas, 1999; Ferrand, Segui, & Grainger, 1996; Mehler, Dommergues, Frauenfelder, & Segui, 1981). The boundaries between syllables are evident in the fluctuations of envelope energy over time, with different “bursts” of energy indicating distinct syllables (Greenberg & Ainsworth, 2004; Greenberg et al., 2003; Shastri, Chang, & Greenberg, 1999). Thus, parsing into syllable-size packages can be achieved relatively easily based on the temporal envelope. Although syllabic parsing has been emphasized in the literature, some advocate that at the core of speech processing is an even longer time scale – that of phrases –which constitutes the backbone for integrating the smaller units of language – phonetic segments, syllables and words – into comprehensible linguistic streams (Nygaard & Pisoni, 1995; Shukla, Nespor, & Mehler, 2007; Todd, Lee, & O’Boyle, 2006). The acoustic properties for identifying phrasal boundaries are much less well defined than those of syllabic boundaries, however it is generally agreed that prosodic cues are key for phrasal parsing (Cooper et al., 1978; Fodor & Ferreira, 1998; Friederici, 2002; Hickok, 1993). These may include metric characteristics of the language (Cutler & Norris, 1988; Otake, Hatano, Cutler, & Mehler, 1993) and more universal prosodic cues (Endress & Hauser, 2010) as well as cues such as stress, intonation and speeding up/slowing down of speech and the insertion of pauses (Arciuli & Slowiczek, 2007; Lehiste, 1970; Wightman, Shattuck-Hufnagel, Ostendorf, & Price, 1992). Many of these prosodic cues are represented in the temporal envelope, mostly in subtle variation from the mean amplitude or rate of

syllables (Rosen, 1992). For example, stress is reflected in higher amplitude and (sometimes) longer durations, and changes in tempo are reflected in shorter or longer pauses. Thus, the temporal envelope contains cues for speech parsing on both the syllabic and phrasal levels. It is suggested that the degraded intelligibility observed when the speech envelope is distorted is due to lack of cues by which to parse the stream and place the fine structure content into its appropriate context within the hierarchy of speech (Greenberg & Ainsworth, 2006; Greenberg & Ainsworth, 2004).

2.4 Making Predictions

One factor that might assist greatly in both phrasal and syllabic parsing - as well as hierarchical integration - is the ability to anticipate the onset and offsets of these units before they occur. It is widely demonstrated that temporal predictions assist perception in more simple rhythmic contexts. Perceptual judgments are enhanced for stimuli occurring at an anticipated moment, compared to stimuli occurring randomly or at unanticipated times (Barnes & Jones, 2000). These studies have led to the 'Attention in Time' hypothesis (Jones, Johnston, & Puente, 2006; Large & Jones, 1999; Nobre, Correa, & Coull, 2007; Nobre & Coull, 2010) which posits that attention can be directed to particular points in time when relevant stimuli are expected, similar to the allocation of spatial attention. A more recent construction of this idea, the "Entrainment Hypothesis" (Schroeder & Lakatos, 2009b), specifies temporal attention in rhythmic terms directly relevant to speech processing.

It follows that speech decoding would similarly benefit from temporal predictions regarding the precise timing of events in the stream as well as syllabic and phrasal boundaries. The concept of prediction in sentence processing, i.e. beyond the bottom-up analysis of the speech stream, has been developed in detail by Phillips and Wagner (2007) and others. In particular, the 'Entrainment Hypothesis' emphasizes utilizing rhythmic regularities to form predictions about upcoming events, which could be potentially beneficial for speech processing. Temporal prediction may also play a key role in the perception of speech in noise and in stream selection, discussed below (Elhilali et al., 2009; Elhilali & Shamma, 2008; Schroeder & Lakatos, 2009b).

Another cue that likely contributes to forming temporal predictions about upcoming speech events is congruent visual input. In many real-life situations auditory speech is accompanied by visual input of the talking face (an exception being a telephone conversation). It is known that congruent visual input substantially improves speech comprehension, particularly under difficult listening conditions (Campanella & Belin, 2007; Campbell, 2008; Sumby & Pollack, 1954; Summerfield, 1992). Importantly, these cues are *predictive* (Schroeder, Lakatos, Kajikawa, Partan, & Puce, 2008; van Wassenhove, Grant, & Poeppel, 2005). Visual cues of articulation typically precede auditory input by ~150–100ms (Chandrasekaran, Trubanova, Stillitano, Caplier, & Ghazanfar, 2009), though they remain perceptually linked to congruent auditory input at offsets of up to 200 ms (van Wassenhove, Grant, & Poeppel, 2007). These time frames allow ample opportunity for visual cues to influence the processing of upcoming auditory input. In particular, congruent visual input is critical for making temporal predictions, as it contains information about the temporal envelope of the acoustics (Chandrasekaran et al., 2009; Grant & Greenberg, 2001) as well as prosodic cues (Bernstein, Eberhardt, & Demorest, 1989; Campbell, 2008; Munhall, Jones, Callan, Kuratate, & Vatikiotis-Bateson, 2004; Scarborough, Keating, Mattys, Cho, & Alwan, 2009). A recent study by Arnal et al. (2011) emphasizes the contribution of visual input to making predictions about upcoming speech input on multiple time-scales. Supporting this premise, from a neural perspective there is evidence auditory cortex is activated by lip reading prior to arrival of auditory speech input (Besle et al., 2008), and visual input brings about a phase-reset in auditory cortex (Kayser, Petkov, & Logothetis, 2008; Thorne, De Vos, Viola, & Debener, 2011).

To summarize, the temporal structure of speech is critical for speech processing. It provides contextual cues for transforming continuous speech into units on which subsequent linguistic processes can be applied for construction and retrieval of the message. It further contains predictive cues - from both the auditory and visual components of speech – which can facilitate rapid on-line speech processing. Given the behavioral importance of the temporal envelope, we now turn to discuss neural mechanisms that could represent and utilize this temporal information for speech processing and for the selection of a behaviorally relevant speech stream among concurrent competing input.

3. Neuronal tracking of the temporal envelope

3.1 Representation of temporal modulations in the auditory pathway

Given the importance of parsing and hierarchical integration for speech processing, how are these processes implemented from a neurobiologically mechanistic perspective? Some evidence supporting a link between the temporal windows in speech and inherent preferences of the auditory system can be found in studies of responses to amplitude modulated (AM) tones or noise and other complex natural sounds such as animal vocalizations. Physiological studies in monkeys demonstrate that envelope encoding occurs throughout the auditory system, from the cochlea to the auditory cortex, with slower modulation rates represented at higher stations along the auditory pathway (reviewed by Bendor & Wang, 2007; Frisina, 2001; Joris, Schreiner, & Rees, 2004). Importantly, neurons in primary auditory cortex demonstrate a preference to slow modulation rates, in ranges prevalent in natural communication calls (Creutzfeldt, Hellweg, & Schreiner, 1980; de Ribaupierre, Goldstein, & Yeni-Komshian, 1972; Funkenstein & Winter, 1973; Gaese & Ostwald, 1995; Schreiner & Urbas, 1988; Sovijärvi, 1975; Wang, Merzenich, Beitel, & Schreiner, 1995; Whitfield & Evans, 1965). In fact, these ‘naturalistic’ modulation rates are obvious in spontaneous as well as stimulus-evoked activity (Lakatos et al., 2005). Similar patterns of hierarchical sensitivity to AM rate along the human auditory system were reported by Giraud et al. (2000) using fMRI, showing that the highest-order auditory areas are most sensitive to AM modulations between 4–8Hz, corresponding to the syllabic rate in speech. Behavioral studies also show that human listeners are most sensitive to AM rates comparable to those observed in speech (Chi, Gao, Guyton, Ru, & Shamma, 1999; Viemeister, 1979). These converging findings demonstrate that the auditory system contains inherent tuning to the natural temporal statistical structure of communication calls, and the apparent temporal hierarchy along the auditory pathway could be potentially useful for speech recognition, in which information from longer time-scales inform and provides predictions for processing at shorter time scales.

It is also well established that presentation of AM tones or other periodic stimuli, evokes a periodic steady-state response at the modulating frequency that is evident in local field potentials (LFP) in nonhuman subjects, as well as intracranial EEG recordings (ECoG) and scalp EEG in humans, as an increase in power at that frequency. While the steady state response has most often been studied for modulation rates higher than those prevalent in speech (~40Hz, Galambos, Makeig, & Talmachoff, 1981; Picton, Skinner, Champagne, Kellett, & Maiste, 1987), it has also been observed for low-frequency modulations, in ranges prevalent in the speech envelope and crucial for intelligibility (1–16Hz, Liégeois-Chauvel, Lorenzi, Trébuchon, Jean Régis, & Chauvel, 2004; Picton et al., 1987; Rodenburg, Verweij, & van der Brink, 1972; Will & Berg, 2007). This auditory steady state response (ASSR) indicates that a large ensemble of neurons synchronize their excitability fluctuations to the amplitude modulation rate of the stimulus. A recent model which emphasizes tuning properties to temporal modulations along the entire auditory pathway, successfully reproduced this steady state response as recorded directly from human auditory cortex (Dugué, Le Bouquin-Jeannès, Edeline, & Faucon, 2010).

The auditory system's inclination to synchronize neuronal activity according to the temporal structure of the input has raised the hypothesis that this response is related to entrainment of cortical neuronal oscillations to the external driving rhythms. Moreover, it is suggested that such neuronal entrainment is ideal for temporal coding of more complex auditory stimuli, such as speech (Giraud & Poeppel, 2011; Schroeder & Lakatos, 2009b; Schroeder et al., 2008). This hypothesis is motivated, in part, by the strong resemblance between the characteristic time scales in speech and other communication sounds and those of the dominant frequency bands of neuronal oscillations - namely the delta (1–3Hz), theta (3–7Hz) and beta/gamma (20–80Hz) bands. Given these time-scale similarities, neuronal oscillations are well positioned for parsing continuous input in behaviorally relevant time scale and hierarchical integration. This hypothesis is further developed in the next sections, based on evidence for LFP and spike 'tracking' of natural communication calls and speech.

3.2 Neuronal encoding of the temporal envelope of complex auditory stimuli

Several physiological studies in primary auditory cortex of cat and monkey have found that the temporal envelope of complex vocalizations is significantly correlated with the temporal patterns of spike activity in spatially distributed populations (Gehr, Komiya, & Eggermont, 2000; Nagarajan et al., 2002; Rotman, Bar-Yosef, & Nelken, 2001; Wang et al., 1995). These studies show bursts of spiking activity time-locked to major peaks in the vocalization envelope. Recently, it has been shown that the temporal pattern of LFPs is also phase-locked to the temporal structure of complex sounds, particularly in frequency ranges of 2–9Hz and 16–30Hz (Chandrasekaran, Turesson, Brown, & Ghazanfar, 2010; Kayser, Montemurro, Logothetis, & Panzeri, 2009). Moreover, in those studies the timing of spiking activity was also locked to the phase of the LFP, indicating that the two neuronal responses are inherently linked (see also (Eggermont & Smith, 1995; Montemurro, Rasch, Murayama, Logothetis, & Panzeri, 2008). Critically, Kayser et al. (2009) show that spikes and LFP phase (<30Hz) carry *complementary* information, and their combination proved to be the most informative about complex auditory stimuli. It has been suggested that the phase of the LFP, and in particular that of underlying neuronal oscillations which generate the LFP, control the timing of neuronal excitability and gate spiking activity so that they occur only at relevant times (Elhilali, Fritz, Klein, Simon, & Shamma, 2004; Fries, 2009; 2007; 2008; Lakatos et al., 2005; Mazzoni, Whittingstall, Brunel, Logothetis, & Panzeri, 2010; Schroeder & Lakatos, 2009a; Womelsdorf et al., 2007). According to this view, and as explicitly suggested by Buzsáki and Chrobak (1995), neuronal oscillations supply the temporal *context* for processing complex stimuli and serve to direct more intricate processing of the acoustic *content*, represented by spiking activity, to particular points in time, i.e. peaks in the temporal envelope. In the next section we discuss how this framework may relate to the representation of natural speech in human auditory cortex.

3.3 Low-frequency tracking of speech envelope

Few studies have attempted to directly quantify neuronal dynamics to speech in humans, perhaps since traditional EEG/MEG methods often are geared toward strictly time-locked responses to brief stimuli or well defined events within a stream, rather than long continuous and fluctuating stimuli. Nonetheless, the advancement of single-trial analysis tools has allowed for some examination of the relationship between the temporal structure of speech and the temporal dynamics of neuronal activity while listening to it. Ahissar et al. (2005; 2001) compared the spectrum of the MEG response when subjects listened to sentences, to the spectrum of the sentences' temporal envelope and found that they were highly correlated, with a peak in the theta band (3–7Hz). Critically, when speech was artificially compressed yielding a faster syllabic rate, the theta-band peak in the neural spectrum shifted accordingly, as long as the speech remained intelligible. The match between the neural and acoustic signals broke down as intelligibility decreased with additional compression (see

also Ghitza & Greenberg, 2009). This finding is important for two reasons: First, it shows that not only does the auditory system respond at similar frequencies to those prevalent in speech, but there is dynamic control and adjustment of the theta-band response to accommodate different speech rates (within a circumscribed range). Thus it is well positioned for encoding the temporal envelope of speech, given the great variability across speakers and tokens. Second, and critically, this dynamic is limited to within the frequency bands where speech intelligibility is preserved, demonstrating its specificity for speech processing. In contrast, using ECoG recordings, Nourski et al. (2009) found that gamma-band responses in human auditory cortex continue to track an auditory stimulus even past the point of non-intelligibility, suggesting that the higher frequency response is more stimulus-driven and is not critical for intelligibility.

More direct support for the notion that the phase of low-frequency neuronal activity tracks the temporal envelope of speech is found in recent work from several groups using scalp recorded EEG and MEG (Abrams, Nicol, Zecker, & Kraus, 2008; Aiken & Picton, 2008; Bonte, Valente, & Formisano, 2009; Lalor & Foxe, 2010; Luo, Liu, & Poeppel, 2010; Luo & Poeppel, 2007). Luo and colleagues (2007, 2010) show that low-frequency phase pattern (1–7Hz; recorded by MEG) is highly replicable across repetitions of the same speech token, and reliably distinguishes between different tokens (illustrated in Figure 2 top). Importantly, as in the studies by Ahissar et al. (2001), this effect was related to speech intelligibility and was not found for distorted speech. Similarly, distinct alpha phase patterns (8–12Hz) were found to characterize different vowels (Bonte et al., 2009). Further studies have directly shown that the temporal pattern of low-frequency auditory activity tracks the temporal envelope of speech (Abrams et al., 2008; Aiken & Picton, 2008; Lalor & Foxe, 2010).

These scalp-recorded speech-tracking responses are likely dominated by sequential evoked responses to large fluctuations in the temporal envelope (e.g. word beginnings, stressed syllables), akin to the auditory N1 response which is primarily a theta-band phenomenon (see Aiken & Picton 2008). Thus, a major question with regard to the above cited findings is whether speech-tracking is entirely sensory-driven, reflecting continuous responses to the auditory input, or whether it is influenced by neuronal entrainment facilitating accurate speech tracking in a top-down fashion. While this question has not been fully resolved empirically, the latter view is strongly supported by the findings that speech tracking breaks down when speech is no longer intelligible (due to compression or distortion), coupled with known top-down influences of attention on the N1 (Hillyard, Hink, Schwent, & Picton, 1973; particularly relevant in this context is the N1 sensitivity to temporal attention: Lange, 2009; Lange & Röder, 2006; Sanders & Astheimer, 2008). Moreover, an entirely sensory-driven approach would greatly hinder speech processing particularly in noisy settings (Atal & Schroeder, 1978), whereas an entrainment-based mechanism utilizing temporal predictions is more robust to filtering out irrelevant background noise (see section 4).

Distinguishing between sensory-driven evoked responses and modulatory influences of intrinsic entrainment is not simple. However, one way to determine the relative contribution of entrainment is by comparing the tracking responses to rhythmic and non rhythmic stimuli, or comparing speech-tracking responses at the beginning versus the end of an utterance, i.e. before and after temporal regularities have supposedly been established. It should be noted that the neuronal patterns hypothesized to underlie such entrainment, (e.g. theta and gamma band oscillations), exist in absence of any stimulation and therefore intrinsic (e.g. Giraud et al., 2007; Lakatos et al., 2005).

3.4 Hierarchical representation of the temporal structure of speech

A point that has not yet been explored with regard to neuronal tracking of the speech envelope, but is of pressing interest, is how it relates to the hierarchical temporal structure of

speech. As discussed above, in processing speech the brain is challenged to convert a single stream of complex fluctuating input into a hierarchical structure across multiple time scales. As mentioned above, the dominant rhythms of cortical neuronal oscillations - delta, theta and beta/gamma bands - strongly resemble the main time scales in speech. Moreover, ambient neuronal activity exhibits hierarchical cross-frequency coupling relationships, such that higher frequency activity tends to be nested within the envelope of a lower-frequency oscillation (Canolty et al., 2006; Canolty & Knight, 2010; Lakatos et al., 2005; Monto, Palva, Voipio, & Palva, 2008), just as spiking activity is nested within the phase of neuronal oscillations (Eggermont & Smith, 1995; Mazzoni et al., 2010; Montemurro et al., 2008) mirroring the hierarchical temporal structure of speech. Some have suggested that the similarity between the temporal properties of speech and neural activity are not coincidental but rather, that spoken communication evolved to fit constraints posed by the neuronal mechanisms of the auditory system and the motor articulation system (Gentilucci & Corballis, 2006; Liberman & Whalen, 2000). From an 'Active Sensing' perspective (Schroeder, Wilson, Radman, Scharfman, & Lakatos, 2010), it is likely that the dynamics of motor system, articulatory apparatus, and related cognitive operations constrain the evolution and development of both spoken language and auditory processing dynamics. These observations have led to suggestions that neuronal oscillations are utilized on-line as a means for sampling the unfolding speech content at multiple temporal scales simultaneously and form the basis for adequately parsing and integrating the speech signal in the appropriate time windows (Poeppel, 2003; Poeppel et al., 2008; Schroeder et al., 2008). There is some physiological evidence for parallel processing in multiple time scales in audition, emphasizing the theta and gamma bands in particular (Ding & Simon, 2009; Theunissen & Miller, 1995), with a noted asymmetry of the sensitivity to these two time scales in the right and left auditory cortices (Boemio, Fromm, Braun, & Poeppel, 2005; Giraud et al., 2007; Poeppel, 2003). Further, the perspective of an array of oscillators parsing and processing information at multiple time scales has been mirrored in computational models attempting to describe speech processing (Ahissar & Ahissar, 2005; Ghitza & Greenberg, 2009) as well as beat perception in music (Large, 2000). The findings that low-frequency neuronal activity tracks the temporal envelope of complex sounds and speech provide an initial proof-of-concept that this occurs. However, a systematic investigation of the relationship between different rhythms in the brain and different linguistic features is required, as well as investigation of the hierarchical relationship across frequencies.

4. Selective speech tracking – implications for the 'Cocktail Party' problem

4.1 The challenge of solving the 'Cocktail Party' problem

Studying the neural correlates of speech perception is inherently tied to the problem of selective attention since, more often than not, we listen to speech amid additional background sounds which we must filter out. Though auditory selective attention, as exemplified by the 'Cocktail Party' problem (Cherry, 1953), has been recognized for over 50 years as an essential cognitive capacity, its precise neuronal mechanisms are unclear.

Tuning in to a particular speaker in a noisy environment involves at least two processes: identifying and separating different sources in the environment (stream segregation) and directing attention to one task relevant one (stream selection). It is widely accepted that attention is necessary for stream selection, although there is an ongoing debate whether attention is required for stream segregation as well (Alain & Arnott, 2000; Cusack, Deeks, Aikman, & Carlyon, 2004; Sussman & Orr, 2007). Besides activating dedicated fronto-parietal 'attentional networks' (Corbetta & Shulman, 2002; Posner & Rothbart, 2007), attention influences neural activity in auditory cortex (Bidet-Caulet et al., 2007; Hubel, Henson, Rupert, & Galambos, 1959; Ray, Niebur, Hsiao, Sinai, & Crone, 2008; Woldorff et

al., 1993). Moreover, attending to different attributes of a sound can modulate the pattern of auditory neural response, demonstrating feature-specific top-down influence (Brechmann & Scheich, 2005). Thus, with regard to the ‘Cocktail Party’ problem, directing attention to the features of an attended speaker may serve to enhance their cortical representation and thus facilitate processing at the expense of other competing input.

What are those features and how does attention operate, from a neural perspective, to enhance their representation? In a ‘Cocktail Party’ environment, speakers can be differentiated based on acoustic features (such fundamental frequency (pitch), timbre, harmonicity, intensity), spatial features (location, trajectory) and temporal features (mean rhythm, temporal envelope pattern). In order to focus auditory attention on a specific speaker, we most likely make use of a combination of these features (Fritz, Elhilali, David, & Shamma, 2007; Hafer, Sarampalis, & Loui, 2007; McDermott, 2009; Moore & Gockel, 2002). One way in which attention influences sensory processing is by sharpening neuronal receptive fields to enhance the selectivity for attended features, as found in the visual system (Reynolds & Heeger, 2009). For example, utilizing the tonotopic organization of the auditory system, attention could selectively enhance responses to an attended frequency while responsiveness at adjacent spectral frequencies is suppressed (Fritz, Shamma, Elhilali, & Klein, 2003; Fritz, Elhilali, & Shamma, 2005). Similarly, in the spatial domain, attention to a particular spatial location enhances responses in auditory neurons spatially-tuned to the attended location, as well as cortical responses (Winkowski & Knudsen, 2006, 2008; Wu, Weissman, Roberts, & Woldorff, 2007). Given these tuning properties, if two speakers differ in their fundamental frequency and spatial location, biasing the auditory system towards these features could assist in its selective processing. Indeed, many current models emphasize frequency separation and spatial location as the main features underlying stream segregation (Beauvois & Meddis, 1996; Darwin & Hukin, 1999; Hartmann & Johnson, 1991; McCabe & Denham, 1997)

However, recent studies challenge the prevalent view that tonotopic and spatial separation are sufficient for perceptual stream segregation, since they fail to account for observed influence of the relative timing of sounds on streaming percepts and neural responses (Elhilali et al., 2009; Elhilali & Shamma, 2008; Shamma et al., 2011). Since different acoustic streams evolve differently over time (i.e. have unique temporal envelopes), neuronal populations encoding the spectral and spatial features of a particular stream share common temporal patterns. In their model, Elhilali and colleagues (2008, 2009) propose that the temporal coherence among these neural populations serves to group them together and form a unified, and distinct percept of an acoustic source. Thus, temporal structure has a key role in stream segregation. Furthermore, Elhilali and colleagues suggest that an effective system would utilize the temporal regularities within the sound pattern of a particular source to form predictions regarding upcoming input from that source, which are then compared to the actual input. This predictive capacity is particularly valuable for stream selection, as it allows directing attention not only to particular spectral and spatial feature of an attended speaker, but also to points in time when attended events are projected to occur, in line with the ‘Attention in Time’ hypothesis discussed above (Barnes & Jones, 2000; Jones et al., 2006; Large & Jones, 1999; Nobre et al., 2007; Nobre & Coull, 2010). Given the fundamental role of oscillations in controlling the timing of neuronal excitability, it is particularly suited for implementing such temporal selective attention.

4.2 The Selective Entrainment Hypothesis

In attempt to link cognitive behavioral to its underlying neurophysiology, Schroeder and Lakatos (2009b) recently proposed a Selective Entrainment Hypothesis suggesting that the auditory system selectively tracks the temporal structure of only one behaviorally-relevant stream (Figure 2 bottom). According to this view, the temporal regularities within a speech

stream enable the system to synchronize the timing of high neuronal excitability with predicted timing of events in the attended stream, thus selectively enhancing its processing. This hypothesis also implies that events from the unattended stream will often occur at low-excitability phase, producing reduced neuronal responses, and rendering them easier to ignore. The principle of Selective Entrainment has thus far been demonstrated using brief rhythmic stimuli in LFP recordings in monkeys as well as ECoG and EEG in humans (Besle et al., 2011; Lakatos et al., 2008; Lakatos et al., 2005; Stefanics et al., 2010). In these studies, the phase of entrained oscillations at the driving frequency was set either in-phase or out-of-phase with the stimulus, depending on whether it was to be attended or ignored. These findings demonstrate the capacity of the brain to control the timing of neuronal excitability (i.e. the phase of the oscillation) based on task demands.

Extending the Selective Entrainment Hypothesis to naturalistic and continuous stimuli is not straightforward, due to their temporal complexity. Yet, although speech tokens share common modulation rates, they differ substantially in their *precise* time course and thus can be differentiated in their relative phases (see Figure 2A). Selective Entrainment to speech would, thus, manifest in precise phase locking (tracking) to the temporal envelope of an attended speaker, with minimal phase locking to the unattended speaker. Initial evidence for selective entrainment to continuous speech was provided by Kerlin et al. (2010) who showed that attending to different speakers in a scene manifests in unique temporal pattern of delta and theta activity (combination of phase and power). We have recently expanded on these findings (Zion Golumbic et al., 2010, in preparation) and demonstrated using ECoG recordings in humans that, indeed, both delta and theta activity selectively track the temporal envelope of an attended, but not an unattended speaker nor the global acoustics in a simulated Cocktail Party.

5. Conclusion: Active Sensing

The hypothesis put forth in this paper is that dedicated neuronal mechanisms encode the temporal structure of naturalistic stimuli, and speech in particular. The data reviewed here suggest that this temporal representation is critical for speech parsing, decoding, and intelligibility, and may also serve as a basis for attentional selection of a relevant speech stream in a noisy environment. It is proposed that neuronal oscillations, across multiple frequency bands, provide the mechanistic infrastructure for creating an on-line representation of the multiple time scales in speech, and facilitate the parsing and hierarchical integration of smaller speech units for speech comprehension. Since speech unfolds over time, this representation needs to be created and constantly updated ‘on the fly’, as information accumulates. To achieve this task optimally, the system most likely does not rely solely on bottom-up input to shape the momentary representation, but rather it also creates internal predictions regarding forthcoming stimuli to facilitate its processing. Computational models specifically emphasize the centrality of predictions for successful speech decoding and stream segregation (Denham & Winkler, 2006; Elhilali & Shamma, 2008). As noted above, multiple cues carry predictive information, including the intrinsic rhythmicity of speech, prosodic information and visual input, all of which contribute to forming expectations regarding upcoming events in the speech stream. In that regard, speech processing can be viewed as an example for ‘**active sensing**’, in which, rather than being passive and reactive the brain is proactive and predictive in that it shapes its own internal activity in anticipation of probable upcoming input (Jones et al., 2006; Nobre & Coull, 2010; Schroeder et al., 2010).

Using prediction to facilitate neural coding of current and future events has often been suggested as a key strategy in multiple cognitive and sensory functions (Eskandar & Assad, 1999; Hosoya, Baccus, & Meister, 2005; Kilner, Friston, & Frith, 2007). ‘Active sensing’

promotes a mechanistic approach to understanding how prediction influences speech processing in that it connects the allocation of *attention* to a particular speech stream, with predictive “tuning” of auditory neuron ensemble dynamics to the anticipated temporal qualities of that stream. In speech as in other rhythmic stimuli, entrainment of neuronal oscillations is well suited to serve active sensing, given their rhythmic nature and their role in controlling the timing of excitability fluctuations in the neuronal ensembles comprising the auditory system. The findings reviewed here support the notion that attention controls and directs entrainment to a behaviorally relevant stream, thus facilitating its processing over all other input in the scene. The rhythmic nature of oscillatory activity generates predictions as to when future events might occur and thus allows for ‘sampling’ the acoustic environment with an appropriate rate and pattern. Of course, sensory processing cannot rely only on predictions, and when the real input arrives it is used to update the predictive model and create new predictions. One such mechanism used often by engineers to model entrainment to a quasi-rhythmic source and proposed as a useful mechanism for speech decoding is a phase-lock loop (PLL, Ahissar & Ahissar, 2005; Gardner, 2005; Ghitza, 2011; Ghitza & Greenberg, 2009). A PLL consists of a phase-detector which compares the phase of an external input with that of an ongoing oscillator, and adjusts the phase of the oscillator using feedback to optimize its synchronization with the external signal.

As discussed above, another source of predictive information for speech processing is visual input. It has been demonstrated that visual input brings about phase-resetting in auditory cortex (Kayser et al., 2008; Lakatos et al., 2009; Thorne et al., 2011), suggesting that viewing the facial movements of speech provides additional potent cues for adjusting the phase of entrained oscillations and optimizing temporal predictions (Schroeder et al., 2008). Luo et al. (2010) show that both auditory and visual cortices align phases with great precision in naturalistic audiovisual scenes. However, determining the degree of influence that visual input has on entrainment to speech and its role in active sensing invites additional investigation.

To summarize, the findings and concepts reviewed here support the idea of active sensing as a key principle governing perception in a complex natural environment. They support a view that the brain’s representation of stimuli, and particularly those of natural and continuous stimuli, emerge from an interaction between proactive, top-down modulation of neuronal activity and bottom-up sensory dynamics. In particular, they stress an important functional role for ambient neuronal oscillations, directed by attention, in facilitating selective speech processing and encoding its hierarchical temporal context, within which the content of the semantic message can be adequately appreciated.

Acknowledgments

The compilation and writing of this review paper was supported by NIH grant 1 F32 MH093061-01 to EZG and MH060358.

References

- Abrams DA, Nicol T, Zecker S, Kraus N. Right-Hemisphere Auditory Cortex Is Dominant for Coding Syllable Patterns in Speech. *J Neurosci*. 2008; 28(15):3958–3965. [PubMed: 18400895]
- Ahissar, E.; Ahissar, M. Processing of the Temporal Envelope of Speech. In: Konig, R.; Heil, P.; Buder, E.; Scheich, H., editors. *The Auditory Cortex: A Synthesis of Human and Animal Research*. London: Lawrence Erlbaum Associates, Inc; 2005. p. 295-313.
- Ahissar E, Nagarajan S, Ahissar M, Protopapas A, Mahncke H, Merzenich MM. Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proc Natl Acad Sci U S A*. 2001; 98(23):13367–13372. [PubMed: 11698688]

- Aiken SJ, Picton TW. Human Cortical Responses to the Speech Envelope. *Ear Hear.* 2008; 29(2):139–157. [PubMed: 18595182]
- Alain C, Arnott SR. Selectively Attending to Auditory Objects. *Front Biosci.* 2000; 5:d202–212. [PubMed: 10702369]
- Allen JB. How do humans process and recognize speech? *Speech and Audio Processing, IEEE Transactions on.* 1994; 2(4):567–577.
- Allen, JB. *Articulation and Intelligibility.* San Rafael: Morgan and Claypool publishers; 2005.
- Arai, T.; Greenberg, S. Speech intelligibility in the presence of cross-channel spectral asynchrony. 12–15 May 1998; Paper presented at the IEEE International Conference on Acoustics, Speech and Signal Processing; 1998.
- Arciuli J, Slowiczak LM. The where and when of linguistic word-level prosody. *Neuropsychologia.* 2007; 45(11):2638–2642. [PubMed: 17475291]
- Arnal LH, Wyart V, Giraud AL. Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. *Nat Neurosci.* 2011; 14(6):797–801. [PubMed: 21552273]
- Atal, B.; Schroeder, M. Predictive coding of speech signals and subjective error criteria. Apr 1978; Paper presented at the Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '78; 1978.
- Bar M. Visual objects in context. *Nat Rev Neurosci.* 2004; 5(8):617–629. [PubMed: 15263892]
- Barnes R, Jones MR. Expectancy, Attention, and Time. *Cognit Psychol.* 2000; 41(3):254–311. [PubMed: 11032658]
- Beauvois MW, Meddis R. Computer simulation of auditory stream segregation in alternating-tone sequences. *J Acoust Soc Am.* 1996; 99(4):2270–2280. [PubMed: 8730073]
- Bendor D, Wang X. Differential neural coding of acoustic flutter within primate auditory cortex. *Nat Neurosci.* 2007; 10(6):763–771. [PubMed: 17468752]
- Bernstein LE, Eberhardt SP, Demorest ME. Single-channel vibrotactile supplements to visual perception of intonation and stress. *J Acoust Soc Am.* 1989; 85(1):397–405. [PubMed: 2522107]
- Besle J, Fischer C, Bidet-Caulet AI, Lecaigard F, Bertrand O, Giard M. Visual Activation and Audiovisual Interactions in the Auditory Cortex during Speech Perception: Intracranial Recordings in Humans. *J Neurosci.* 2008; 28(52):14301–14310. [PubMed: 19109511]
- Besle J, Schevon CA, Mehta AD, Lakatos P, Goodman RR, McKhann GM, et al. Tuning of the Human Neocortex to the Temporal Dynamics of Attended Events. *J Neurosci.* 2011; 31(9):3176–3185. [PubMed: 21368029]
- Bidet-Caulet A, Fischer C, Besle J, Aguera P, Giard M, Bertrand O. Effects of selective attention on the electrophysiological representation of concurrent sounds in the human auditory cortex. *J Neurosci.* 2007; 27(35):9252–9261. [PubMed: 17728439]
- Blessner B. Speech Perception Under Conditions of Spectral Transformation: I. Phonetic Characteristics. *J Speech Hear Res.* 1972; 15(1):5–41. [PubMed: 5012812]
- Boemio A, Fromm S, Braun A, Poeppel D. Hierarchical and asymmetric temporal sensitivity in human auditory cortices. *Nat Neurosci.* 2005; 8(3):389–395. [PubMed: 15723061]
- Bonte M, Valente G, Formisano E. Dynamic and Task-Dependent Encoding of Speech and Voice by Phase Reorganization of Cortical Oscillations. *J Neurosci.* 2009; 29(6):1699–1706. [PubMed: 19211877]
- Borsky S, Tuller B, Shapiro LP. “How to milk a coat:” The effects of semantic and acoustic information on phoneme categorization. *J Acoust Soc Am.* 1998; 103(5):2670–2676. [PubMed: 9604360]
- Brechmann A, Scheich H. Hemispheric shifts of sound representation in auditory cortex with conceptual listening. *Cereb Cortex.* 2005; 15(5):578–587. [PubMed: 15319313]
- Bregman, AS. *Auditory Scene Analysis: The Perceptual Organization of Sound.* MIT Press; 1990.
- Buzsáki G, Chrobak JJ. Temporal structure in spatially organized neuronal ensembles: a role for interneuronal networks. *Curr Opin Neurobiol.* 1995; 5(4):504–510. [PubMed: 7488853]
- Campanella S, Belin P. Integrating face and voice in person perception. *Trends Cogn Sci.* 2007; 11(12):535–543. [PubMed: 17997124]

- Campbell R. The processing of audio-visual speech: empirical and neural bases. *Philos Trans R Soc Lond, B, Biol Sci.* 2008; 363(1493):1001–1010. [PubMed: 17827105]
- Canolty R, Edwards E, Dalal S, Soltani M, Nagarajan S, Kirsch H, et al. High gamma power is phase-locked to theta oscillations in human neocortex. *Science.* 2006; 313(5793):1626–1628. [PubMed: 16973878]
- Canolty RT, Knight RT. The functional role of cross-frequency coupling. *Trends Cogn Sci.* 2010; 14(11):506–515. [PubMed: 20932795]
- Chandrasekaran C, Trubanova A, Stillitano Sb, Caplier A, Ghazanfar AA. The Natural Statistics of Audiovisual Speech. *PLoS Comput Biol.* 2009; 5(7):e1000436. [PubMed: 19609344]
- Chandrasekaran C, Tureson HK, Brown CH, Ghazanfar AA. The Influence of Natural Scene Dynamics on Auditory Cortical Activity. *J Neurosci.* 2010; 30(42):13919–13931. [PubMed: 20962214]
- Chen F. The relative importance of temporal envelope information for intelligibility prediction: A study on cochlear-implant vocoded speech. *Med Eng Phys.* 2011 In Press, Corrected Proof.
- Cherry EC. Some experiments on the recognition of speech, with one and two ears. *J Acoust Soc Am.* 1953; 25:975–979.
- Chi T, Gao Y, Guyton MC, Ru P, Shamma S. Spectro-temporal modulation transfer functions and speech intelligibility. *J Acoust Soc Am.* 1999; 106(5):2719–2732. [PubMed: 10573888]
- Connine CM. Constraints on interactive processes in auditory word recognition: The role of sentence context. *J Mem Lang.* 1987; 26(5):527–538.
- Connine CM, Clifton CJ. Interactive use of lexical information in speech perception. *J Exp Psychol Hum Percept Perform.* 1987; 13(2):291–299. [PubMed: 2953858]
- Cooper, WE.; Paccia-Cooper, J. *Syntax and speech.* Cambridge, Mass: Harvard University Press; 1980.
- Cooper WE, Paccia JM, Lapointe SG. Hierarchical coding in speech timing. *Cognit Psychol.* 1978; 10(2):154–177.
- Corbetta M, Shulman GL. Control of goal-directed and stimulus-driven attention in the brain. *Nat Rev Neurosci.* 2002; 3(3):201–215. [PubMed: 11994752]
- Creutzfeldt O, Hellweg F, Schreiner C. Thalamocortical transformation of responses to complex auditory stimuli. *Exp Brain Res.* 1980; 39(1):87–104. [PubMed: 6247179]
- Cusack R, Decks J, Aikman G, Carlyon RP. Effects of Location, Frequency Region, and Time Course of Selective Attention on Auditory Scene Analysis. *J Exp Psychol Hum Percept Perform.* 2004; 30(4):643–656. [PubMed: 15301615]
- Cutler A, Norris D. The role of strong syllables in segmentation for lexical access. *J Exp Psychol Hum Percept Perform.* 1988; 14(1):113–121.
- Darwin CJ, Hukin RW. Auditory objects of attention: The role of interaural time differences. *J Exp Psychol Hum Percept Perform.* 1999; 25(3):617–629. [PubMed: 10385981]
- de Ribaupierre F, Goldstein MH, Yeni-Komshian G. Cortical coding of repetitive acoustic pulses. *Brain Res.* 1972; 48:205–225. [PubMed: 4645205]
- Denham SL, Winkler I. The role of predictive models in the formation of auditory streams. *Journal of Physiology-Paris.* 2006; 100(1–3):154–170.
- Ding N, Simon JZ. Neural Representations of Complex Temporal Modulations in the Human Auditory Cortex. *J Neurophysiol.* 2009; 102(5):2731–2743. [PubMed: 19692508]
- Drullman, R. The significance of temporal modulation frequencies for speech intelligibility. In: Greenberg, S.; Ainsworth, W., editors. *Listening to Speech: an Auditory Perspective.* Mahwah NJ: Erlbaum; 2006. p. 39-48.
- Drullman R, Festen JM, Plomp R. Effect of reducing slow temporal modulations on speech reception. *J Acoust Soc Am.* 1994a; 95(5):2670–2680. [PubMed: 8207140]
- Drullman R, Festen JM, Plomp R. Effect of temporal envelope smearing on speech reception. *J Acoust Soc Am.* 1994b; 95(2):1053–1064. [PubMed: 8132899]
- Dugué P, Le Bouquin-Jeannès R, Edeline JM, Faucon G. A physiologically based model for temporal envelope encoding in human primary auditory cortex. *Hearing Res.* 2010; 268(1–2):133–144.

- Eggermont JJ. Between sound and perception: reviewing the search for a neural code. *Hearing Res.* 2001; 157(1–2):1–42.
- Eggermont JJ, Smith GM. Synchrony between single-unit activity and local field potentials in relation to periodicity coding in primary auditory cortex. *J Neurophysiol.* 1995; 73(1):227–245. [PubMed: 7714568]
- Eimas PD. Segmental and syllabic representations in the perception of speech by young infants. *J Acoust Soc Am.* 1999; 105(3):1901–1911. [PubMed: 10089612]
- Elhilali M, Fritz JB, Klein DJ, Simon JZ, Shamma SA. Dynamics of Precise Spike Timing in Primary Auditory Cortex. *J Neurosci.* 2004; 24(5):1159–1172. [PubMed: 14762134]
- Elhilali M, Ma L, Micheyl C, Oxenham AJ, Shamma SA. Temporal Coherence in the Perceptual Organization and Cortical Representation of Auditory Scenes. *Neuron.* 2009; 61(2):317–329. [PubMed: 19186172]
- Elhilali M, Shamma SA. A cocktail party with a cortical twist: How cortical mechanisms contribute to sound segregation. *J Acoust Soc Am.* 2008; 124(6):3751–3771. [PubMed: 19206802]
- Endress AD, Hauser MD. Word segmentation with universal prosodic cues. *Cognit Psychol.* 2010; 61(2):177–199. [PubMed: 20573342]
- Eskandar EN, Assad JA. Dissociation of visual, motor and predictive signals in parietal cortex during visual guidance. *Nat Neurosci.* 1999; 2(1):88–93. [PubMed: 10195185]
- Fant, G. *Acoustic Theory of Speech Production.* The Hague: Mouton; 1960.
- Ferrand L, Segui J, Grainger J. Masked Priming of Word and Picture Naming: The Role of Syllabic Units. *J Mem Lang.* 1996; 35:708–723.
- Fodor, JD.; Ferreira, F. *Reanalysis in sentence processing.* Dordrecht: Kluwer; 1998.
- French NR, Steinberg JC. Factors governing the intelligibility of speech sounds. *J Acoust Soc Am.* 1947; 19:90–119.
- Friederici AD. Towards a neural basis of auditory sentence processing. *Trends Cogn Sci.* 2002; 6(2): 78–84. [PubMed: 15866191]
- Fries P. Neuronal Gamma-Band Synchronization as a Fundamental Process in Cortical Computation. *Annu Rev Neurosci.* 2009; 32(1):209–224. [PubMed: 19400723]
- Friesen LM, Shannon RV, Baskent D, Wang X. Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants. *J Acoust Soc Am.* 2001; 110(2):1150–1163. [PubMed: 11519582]
- Frisina RD. Subcortical neural coding mechanisms for auditory temporal processing. *Hearing Res.* 2001; 158(1–2):1–27.
- Fritz J, Shamma S, Elhilali M, Klein DJ. Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nat Neurosci.* 2003; 6:1216–1223. [PubMed: 14583754]
- Fritz JB, Elhilali M, David SV, Shamma SA. Auditory attention -- focusing the searchlight on sound. *Curr Opin Neurobiol.* 2007; 17(4):437–455. [PubMed: 17714933]
- Fritz JB, Elhilali M, Shamma SA. Differential Dynamic Plasticity of A1 Receptive Fields during Multiple Spectral Tasks. *J Neurosci.* 2005; 25(33):7623–7635. [PubMed: 16107649]
- Funkenstein HH, Winter P. Responses to acoustic stimuli of units in the auditory cortex of awake squirrel monkeys. *Exp Brain Res.* 1973; 18(5):464–488. [PubMed: 4212320]
- Gaese BH, Ostwald J. Temporal coding of amplitude and frequency modulation in the rat auditory cortex. *Eur J Neurosci.* 1995; 7(3):438–450. [PubMed: 7773441]
- Galambos R, Makeig S, Talmachoff PJ. A 40-Hz auditory potential recorded from the human scalp. *Proc Natl Acad Sci U S A.* 1981; 78(4):2643–2647. [PubMed: 6941317]
- Ganong, Wr. Phonetic categorization in auditory word perception. *J Exp Psychol Hum Percept Perform.* 1980; 6(1):110–125. [PubMed: 6444985]
- Gardner, FM. *Phaselock Techniques.* 3. New York: Wiley; 2005.
- Gehr DD, Komiya H, Eggermont JJ. Neuronal responses in cat primary auditory cortex to natural and altered species-specific calls. *Hearing Res.* 2000; 150(1–2):27–42.
- Gentilucci M, Corballis MC. From manual gesture to speech: A gradual transition. *Neurosci Biobehav Rev.* 2006; 30(7):949–960. [PubMed: 16620983]

- Ghitza O. Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input rhythm. *Frontiers in Psychology: Auditory Cognitive Neuroscience*. 2011; 2:130.10.3389/fpsyg.2011.00130
- Ghitza O, Greenberg S. On the Possible Role of Brain Rhythms in Speech Perception: Intelligibility of Time-Compressed Speech with Periodic and Aperiodic Insertions of Silence. *Phonetica*. 2009; 66(1–2):113–126. [PubMed: 19390234]
- Giraud AL, Kleinschmidt A, Poeppel D, Lund TE, Frackowiak RSJ, Laufs H. Endogenous Cortical Rhythms Determine Cerebral Specialization for Speech Perception and Production. *Neuron*. 2007; 56(6):1127–1134. [PubMed: 18093532]
- Giraud AL, Lorenzi C, Ashburner J, Wable J, Johnsrude I, Frackowiak R, et al. Representation of the Temporal Envelope of Sounds in the Human Brain. *J Neurophysiol*. 2000; 84(3):1588–1598. [PubMed: 10980029]
- Giraud, A-L.; Poeppel, D. Speech Perception from a Neurophysiological Perspective. In: Poeppel, OTD.; Popper, AN.; Fay, RR., editors. *The Human Auditory Cortex*. Berlin: Springer; 2011.
- Grant, KW.; Greenberg, S. Speech intelligibility derived from asynchronous processing of auditory-visual information. Paper presented at the Proceedings of the Workshop on Audio-Visual Speech Processing; Scheelsminde, Denmark. 2001.
- Greenberg, S.; Ainsworth, W. *Listening to Speech: an Auditory Perspective*. Mahwah, NJ: Erlbaum; 2006.
- Greenberg, S.; Ainsworth, WA. *Speech Processing in the Auditory System*. New York: Springer-Verlag; 2004.
- Greenberg S, Arai T, Silipo R. Speech intelligibility derived from exceedingly sparse spectral information. *Int Conf Spoken Lang Proc*. 1998:2803–2806.
- Greenberg S, Carvey H, Hitchcock L, Chang S. Temporal properties of spontaneous speech: a syllable-centric perspective. *J Phonetics*. 2003; 31:465–485.
- Hafta, E.; Sarampalis, A.; Loui, P. Auditory attention and filters. In: Yost, W., editor. *Auditory Perception of Sound Sources*. Springer-Verlag; 2007.
- Harding, S.; Cooke, M.; König, P. Auditory Gist Perception: An Alternative to Attentional Selection of Auditory Streams?. Paper presented at the 4th International Workshop on Attention in Cognitive Systems; Hyderabad, India: WAPCV; 2007.
- Hartmann WM, Johnson D. Stream Segregation and Peripheral Channeling. *Music Perception: An Interdisciplinary Journal*. 1991; 9(2):155–183.
- Hickok G. Parallel Parsing: Evidence from Reactivation in Garden-Path Sentences. *J Psycholinguist Res*. 1993; 22(2):239–250.
- Hillyard S, Hink R, Schwent V, Picton T. Electrical signs of selective attention in the human brain. *Science*. 1973; 182:177–180. [PubMed: 4730062]
- Hochstein S, Ahissar M. View from the Top: Hierarchies and Reverse Hierarchies in the Visual System. *Neuron*. 2002; 36(5):791–804. [PubMed: 12467584]
- Holt LL, Lotto AJ. Behavioral examinations of the level of auditory processing of speech context effects. *Hearing Res*. 2002; 167(1–2):156–169.
- Hosoya T, Baccus SA, Meister M. Dynamic predictive coding by the retina. *Nature*. 2005; 436(7047):71–77. [PubMed: 16001064]
- Hubel DH, Henson CO, Rupert A, Galambos R. “Attention” Units in the Auditory Cortex. *Science*. 1959; 129(3358):1279–1280. [PubMed: 13658956]
- Jones MR, Johnston HM, Puente J. Effects of auditory pattern structure on anticipatory and reactive attending. *Cognit Psychol*. 2006; 53(1):59–96. [PubMed: 16563367]
- Joris PX, Schreiner CE, Rees A. Neural Processing of Amplitude-Modulated Sounds. *Physiol Rev*. 2004; 84(2):541–577. [PubMed: 15044682]
- Kayser C, Montemurro MA, Logothetis NK, Panzeri S. Spike-Phase Coding Boosts and Stabilizes Information Carried by Spatial and Temporal Spike Patterns. *Neuron*. 2009; 61(4):597–608. [PubMed: 19249279]
- Kayser C, Petkov CI, Logothetis NK. Visual Modulation of Neurons in Auditory Cortex. *Cereb Cortex*. 2008; 18(7):1560–1574. [PubMed: 18180245]

- Kerlin JR, Shahin AJ, Miller LM. Attentional Gain Control of Ongoing Cortical Speech Representations in a “Cocktail Party”. *J Neurosci*. 2010; 30(2):620–628. [PubMed: 20071526]
- Kiebel SJ, von Kriegstein K, Daunizeau J, Friston KJ. Recognizing Sequences of Sequences. *PLoS Comput Biol*. 2009; 5(8):e1000464. [PubMed: 19680429]
- Kilner J, Friston K, Frith C. Predictive coding: an account of the mirror neuron system. *Cognitive Processing*. 2007; 8(3):159–166. [PubMed: 17429704]
- Kingsbury BED, Morgan N, Greenberg S. Robust speech recognition using the modulation spectrogram. *Speech Communication*. 1998; 25(1–3):117–132.
- Lakatos P, Chen CM, O’Connell MN, Mills A, Schroeder CE. Neuronal Oscillations and Multisensory Interaction in Primary Auditory Cortex. *Neuron*. 2007; 53(2):279–292. [PubMed: 17224408]
- Lakatos P, Karmos G, Mehta AD, Ulbert I, Schroeder CE. Entrainment of Neuronal Oscillations as a Mechanism of Attentional Selection. *Science*. 2008; 320(5872):110–113. [PubMed: 18388295]
- Lakatos P, O’Connell MN, Barczak A, Mills A, Javitt DC, Schroeder CE. The Leading Sense: Supramodal Control of Neurophysiological Context by Attention. *Neuron*. 2009; 64(3):419–430. [PubMed: 19914189]
- Lakatos P, Shah AS, Knuth KH, Ulbert I, Karmos G, Schroeder CE. An Oscillatory Hierarchy Controlling Neuronal Excitability and Stimulus Processing in the Auditory Cortex. *J Neurophysiol*. 2005; 94(3):1904–1911. [PubMed: 15901760]
- Lalor EC, Foxe JJ. Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. *Eur J Neurosci*. 2010; 31(1):189–193. [PubMed: 20092565]
- Lange K. Brain correlates of early auditory processing are attenuated by expectations for time and pitch. *Brain Cogn*. 2009; 69(1):127–137. [PubMed: 18644669]
- Lange K, Röder B. Orienting Attention to Points in Time Improves Stimulus Processing Both within and across Modalities. *J Cognit Neurosci*. 2006; 18(5):715–729. [PubMed: 16768372]
- Large EW. On synchronizing movements to music. *Human Movement Science*. 2000; 19(4):527–566.
- Large EW, Jones MR. The dynamics of attending: How people track time-varying events. *Psychol Rev*. 1999; 106(1):119–159.
- Lehiste, I. *Suprasegmentals*. Cambridge, Mass: M.I.T. Press; 1970.
- Lieberman AM, Whalen DH. On the relation of speech to language. *Trends Cogn Sci*. 2000; 4(5):187–196. [PubMed: 10782105]
- Liégeois-Chauvel C, Lorenzi C, Trébuchon A, Jean Régis, Chauvel P. Temporal Envelope Processing in the Human Left and Right Auditory Cortices. *Cereb Cortex*. 2004; 14(7):731–740. [PubMed: 15054052]
- Luo H, Liu Z, Poeppel D. Auditory Cortex Tracks Both Auditory and Visual Stimulus Dynamics Using Low-Frequency Neuronal Phase Modulation. *PLoS Biol*. 2010; 8(8):e1000445. [PubMed: 20711473]
- Luo H, Poeppel D. Phase Patterns of Neuronal Responses Reliably Discriminate Speech in Human Auditory Cortex. *Neuron*. 2007; 54(6):1001–1010. [PubMed: 17582338]
- Mann VA. Influence of preceding liquid on stop-consonant perception. *Percept Psychophys*. 1980; 28(5):407–412. [PubMed: 7208250]
- Mattys S. The use of time during lexical processing and segmentation: A review. *Psychonomic Bulletin & Review*. 1997; 4(3):310–329.
- Mazzoni A, Whittingstall K, Brunel N, Logothetis NK, Panzeri S. Understanding the relationships between spike rate and delta/gamma frequency bands of LFPs and EEGs using a local cortical network model. *NeuroImage*. 2010; 52(3):956–972. [PubMed: 20026218]
- McCabe SL, Denham MJ. A model of auditory streaming. *J Acoust Soc Am*. 1997; 101(3):1611–1621.
- McDermott JH. The cocktail party problem. *Curr Biol*. 2009; 19(22):R1024–R1027. [PubMed: 19948136]
- Mehler J, Dommergues JY, Frauenfelder U, Segui J. The syllable’s role in speech segmentation. *Journal of Verbal Learning and Verbal Behavior*. 1981; 20(3):298–305.
- Miller GA, Heise GA, Lichten W. The intelligibility of speech as a function of the context of the test materials. *J Exp Psychol*. 1951; 41(5):329–335. [PubMed: 14861384]

- Montemurro MA, Rasch MJ, Murayama Y, Logothetis NK, Panzeri S. Phase-of-Firing Coding of Natural Visual Stimuli in Primary Visual Cortex. *Curr Biol*. 2008; 18(5):375–380. [PubMed: 18328702]
- Monto S, Palva S, Voipio J, Palva JM. Very Slow EEG Fluctuations Predict the Dynamics of Stimulus Detection and Oscillation Amplitudes in Humans. *J Neurosci*. 2008; 28(33):8268–8272. [PubMed: 18701689]
- Moore BCJ, Gockel H. Factors Influencing Sequential Stream Segregation. *Acta Acustica United with Acustica*. 2002; 88:320–332.
- Munhall KG, Jones JA, Callan DE, Kuratate T, Vatikiotis-Bateson E. Visual Prosody and Speech Intelligibility. *Psychol Sci*. 2004; 15(2):133–137. [PubMed: 14738521]
- Nagarajan SS, Cheung SW, Bedenbaugh P, Beitel RE, Schreiner CE, Merzenich MM. Representation of Spectral and Temporal Envelope of Twitter Vocalizations in Common Marmoset Primary Auditory Cortex. *J Neurophysiol*. 2002; 87(4):1723–1737. [PubMed: 11929894]
- Nahum M, Nelken I, Ahissar M. Low-Level Information and High-Level Perception: The Case of Speech in Noise. *PLoS Biol*. 2008; 6(5):e126. [PubMed: 18494561]
- Nelken, I.; Ahissar, M. High-Level and Low-Level Processing in the Auditory System: The Role of Primary Auditory Cortex. In: Divenyi, PL.; Greenberg, S.; Meyer, G., editors. *Dynamics of Speech Production and Perception*. Amsterdam, NLD: IOS Press; 2006. p. 343–354.
- Nobre AC, Correa A, Coull JT. The hazards of time. *Curr Opin Neurobiol*. 2007; 17(4):465–470. [PubMed: 17709239]
- Nobre, AC.; Coull, JT. *Attention and Time*. Oxford, UK: Oxford University Press; 2010.
- Norris D, McQueen JM, Cutler A, Butterfield S. The Possible-Word Constraint in the Segmentation of Continuous Speech. *Cognit Psychol*. 1997; 34(3):191–243. [PubMed: 9466831]
- Nourski KV, Reale RA, Oya H, Kawasaki H, Kovach CK, Chen H, et al. Temporal Envelope of Time-Compressed Speech Represented in the Human Auditory Cortex. *J Neurosci*. 2009; 29(49):15564–15574. [PubMed: 20007480]
- Nygaard, L.; Pisoni, D. Speech Perception: New directions in research and theory. In: Miller, J.; Eimas, P., editors. *Handbook of Perception and Cognition: Speech, Language and Communication*. San Diego, CA: Academic; 1995. p. 63–96.
- Otake T, Hatano G, Cutler A, Mehler J. Mora or Syllable? Speech Segmentation in Japanese. *J Mem Lang*. 1993; 32(2):258–278.
- Pavlovic CV. Derivation of primary parameters and procedures for use in speech intelligibility predictions. *J Acoust Soc Am*. 1987; 82(2):413–422. [PubMed: 3624646]
- Phillips, C.; Wagers, M. Relating Structure and Time in Linguistics and Psycholinguistics. In: Gaskell, G., editor. *Oxford Handbook of Psycholinguistics*. Oxford, UK: Oxford University Press; 2007. p. 739–756.
- Pickett JM, Pollack I. Intelligibility of excerpts from fluent speech: effects of rate of utterance and duration of excerpt. *Lang Speech*. 1963; 6(3):151–164.
- Pickles, JO. *An introduction to the physiology of hearing*. 3. London: Emerald; 2008.
- Picton TW, Skinner CR, Champagne SC, Kellett AJC, Maiste AC. Potentials evoked by the sinusoidal modulation of the amplitude or frequency of a tone. *J Acoust Soc Am*. 1987; 82(1):165–178. [PubMed: 3624637]
- Poeppl D. The analysis of speech in different temporal integration windows: cerebral lateralization as ‘asymmetric sampling in time’. *Speech Commun*. 2003; 41:245–255.
- Poeppl D, Idsardi WJ, van Wassenhove V. Speech perception at the interface of neurobiology and linguistics. *Philos Trans R Soc Lond, B, Biol Sci*. 2008; 363(1493):1071–1086. [PubMed: 17890189]
- Posner MI, Rothbart MK. Research on Attention Networks as a Model for the Integration of Psychological Science. *Annu Rev Psychol*. 2007; 58(1):1–23. [PubMed: 17029565]
- Ray S, Niebur E, Hsiao SS, Sinai A, Crone NE. High-frequency gamma activity (80–150Hz) is increased in human cortex during selective attention. *Clinical neurophysiology: official journal of the International Federation of Clinical Neurophysiology*. 2008; 119(1):116–133. [PubMed: 18037343]

- Repp BH. Phonetic trading relations and context effects: new experimental evidence for a speech mode of perception. *Psychol Bull.* 1982; 92(1):81–110. [PubMed: 7134330]
- Reynolds JH, Heeger DJ. The Normalization Model of Attention. *Neuron.* 2009; 61(2):168–185. [PubMed: 19186161]
- Rodenburg M, Verweij C, van der Brink G. Analysis of evoked responses in man elicited by sinusoidally modulated noise. *Int J Audiol.* 1972; 11(5–6):283–293.
- Rosen S. Temporal Information in Speech: Acoustic, Auditory and Linguistic Aspects. *Phil Trans Biol Sci.* 1992; 336(1278):367–373.
- Rotman Y, Bar-Yosef O, Nelken I. Relating cluster and population responses to natural sounds and tonal stimuli in cat primary auditory cortex. *Hearing Res.* 2001; 152(1–2):110–127.
- Saberi K, Perrott DR. Cognitive restoration of reversed speech. *Nature.* 1999; 398(6730):760–760. [PubMed: 10235257]
- Sanders LD, Astheimer LB. Temporally selective attention modulates early perceptual processing: Event-related potential evidence. *Perceptual Psychophysiology.* 2008; 70(4):732–742.
- Scarborough R, Keating P, Mattys SL, Cho T, Alwan A. Optical Phonetics and Visual Perception of Lexical and Phrasal Stress in English. *Lang Speech.* 2009; 52:135–175. [PubMed: 19624028]
- Schreiner CE, Urbas JV. Representation of amplitude modulation in the auditory cortex of the cat. II Comparison between cortical fields. *Hearing Res.* 1988; 32(1):49–63.
- Schroeder CE, Lakatos P. The Gamma Oscillation: Master or Slave? *Brain Topogr.* 2009a; 22(1):24–26. [PubMed: 19205863]
- Schroeder CE, Lakatos P. Low-frequency neuronal oscillations as instruments of sensory selection. *Trends Neurosci.* 2009b; 32(1):9–18. [PubMed: 19012975]
- Schroeder CE, Lakatos P, Kajikawa Y, Partan S, Puce A. Neuronal oscillations and visual amplification of speech. *Trends Cogn Sci.* 2008; 12(3):106–113. [PubMed: 18280772]
- Schroeder CE, Wilson DA, Radman T, Scharfman H, Lakatos P. Dynamics of Active Sensing and perceptual selection. *Curr Opin Neurobiol.* 2010; 20(2):172–176. [PubMed: 20307966]
- Shamma SA, Elhilali M, Micheyl C. Temporal coherence and attention in auditory scene analysis. *Trends Neurosci.* 2011; 34(3):114–123. [PubMed: 21196054]
- Shannon RV, Zeng FG, Kamath V, Wyganski J, Ekelid M. Speech Recognition with Primarily Temporal Cues. *Science.* 1995; 270(5234):303–304. [PubMed: 7569981]
- Shastri, L.; Chang, S.; Greenberg, S. Syllable detection and segmentation using temporal flow neural networks. Paper presented at the 14th International Congress of Phonetic Sciences; San Francisco, CA. 1999.
- Shukla M, Nespore M, Mehler J. An interaction between prosody and statistics in the segmentation of fluent speech. *Cognit Psychol.* 2007; 54(1):1–32. [PubMed: 16782083]
- Sovijärvi A. Detection of natural complex sounds by cells in the primary auditory cortex of the cat. *Acta Physiol Scand.* 1975; 93(3):318–335. [PubMed: 1146577]
- Stefanics G, Hangya B, Hernádi I, Winkler I, Lakatos P, Ulbert I. Phase Entrainment of Human Delta Oscillations Can Mediate the Effects of Expectation on Reaction Speed. *J Neurosci.* 2010; 30(41):13578–13585. [PubMed: 20943899]
- Stevens, KN. *Acoustic Phonetics.* Cambridge, MA: MIT press; 1998.
- Stone MA, Fullgrabe C, Moore BCJ. Relative contribution to speech intelligibility of different envelope modulation rates within the speech dynamic range. *J Acoust Soc Am.* 2010; 128(4):2127–2137. [PubMed: 20968383]
- Sumby WH, Pollack I. Visual contribution to speech intelligibility in noise. *J Acoust Soc Am.* 1954; 26:212–215.
- Summerfield Q. Lipreading and Audio-Visual Speech Perception. *Philos Trans R Soc Lond B Biol Sci.* 1992; 335(1273):71–78. [PubMed: 1348140]
- Sussman ES, Orr M. The role of attention in the formation of auditory streams. *Attention, Perception and Psychophysics.* 2007; 69(1):136–152.
- Tank DW, Hopfield JJ. Neural computation by concentrating information in time. *Proc Natl Acad Sci USA.* 1987; 84(7):1896–1900. [PubMed: 3470765]

- Theunissen F, Miller JP. Temporal encoding in nervous systems: A rigorous definition. *J Comput Neurosci.* 1995; 2(2):149–162. [PubMed: 8521284]
- Thorne JD, De Vos M, Viola FC, Debener S. Cross-Modal Phase Reset Predicts Auditory Task Performance in Humans. *J Neurosci.* 2011; 31(10):3853–3861. [PubMed: 21389240]
- Todd, NPM.; Lee, CS.; O'Boyle, DJ. A sensorimotor Theory of Speech Perception: Implications for Learning, Organization and Recognition. In: Greenberg, S.; Ainsworth, WA., editors. *Listening to Speech: An Auditory Perspective.* Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc; 2006. p. 351-373.
- van Wassenhove V, Grant KW, Poeppel D. Visual speech speeds up the neural processing of auditory speech. *Proc Natl Acad Sci U S A.* 2005; 102(4):1181–1186. [PubMed: 15647358]
- van Wassenhove V, Grant KW, Poeppel D. Temporal window of integration in auditory-visual speech perception. *Neuropsychologia.* 2007; 45(3):598–607. [PubMed: 16530232]
- Viemeister NF. Temporal modulation transfer functions based upon modulation thresholds. *J Acoust Soc Am.* 1979; 66(5):1364–1380. [PubMed: 500975]
- Wang X, Merzenich MM, Beitel R, Schreiner CE. Representation of a species-specific vocalization in the primary auditory cortex of the common marmoset: temporal and spectral characteristics. *J Neurophysiol.* 1995; 74(6):2685–2706. [PubMed: 8747224]
- Warren, R. The Relation of Speech Preception to the Perception of Nonverbal Auditory Patterns. In: Greenberg, S.; Ainsworth, WA., editors. *Listening to Speech: An Auditory Perspective.* Mahwah, NJ: Laurence Erlbaum Associates, Inc; 2004. p. 333-350.
- Warren R, Bashford J, Gardner D. Tweaking the lexicon: Organization of vowel sequences into words. *Atten Percept Psychophys.* 1990; 47(5):423–432.
- Whitfield IC, Evans EF. Responses of auditory cortical neurons to stimuli of changing frequency. *J Neurophysiol.* 1965; 28(4):655–672. [PubMed: 14347626]
- Wightman CW, Shattuck-Hufnagel S, Ostendorf M, Price PJ. Segmental durations in the vicinity of prosodic phrase boundaries. *J Acoust Soc Am.* 1992; 91(3):1707–1717. [PubMed: 1564206]
- Will U, Berg E. Brain wave synchronization and entrainment to periodic acoustic stimuli. *Neurosci Lett.* 2007; 424(1):55–60. [PubMed: 17709189]
- Winkowski DE, Knudsen EI. Top-down gain control of the auditory space map by gaze control circuitry in the barn owl. *Nature.* 2006; 439(7074):336–339. [PubMed: 16421572]
- Winkowski DE, Knudsen EI. Distinct Mechanisms for Top-Down Control of Neural Gain and Sensitivity in the Owl Optic Tectum. *Neuron.* 2008; 60(4):698–708. [PubMed: 19038225]
- Woldorff MG, Gallen CC, Hampson SA, Hillyard SA, Pantev C, Sobel D, et al. Modulation of early sensory processing in human auditory cortex during auditory selective attention. *Proceedings of the National Academy of Sciences.* 1993; 90(18):8722–8726.
- Womelsdorf T, Schoffelen JM, Oostenveld R, Singer W, Desimone R, Engel AK, et al. Modulation of Neuronal Interactions Through Neuronal Synchronization. *Science.* 2007; 316(5831):1609–1612. [PubMed: 17569862]
- Wu CT, Weissman DH, Roberts KC, Woldorff MG. The neural circuitry underlying the executive control of auditory spatial attention. *Brain Res.* 2007; 1134(0):187–198. [PubMed: 17204249]
- Zion Golumbic, EM.; Bickel, S.; Cogan, GB.; Mehta, AD.; Schevon, CA.; Emerson, RG., et al. Low-frequency phase modulations as a tool for attentional selection at a cocktail party: insights from intracranial EEG and MEG in humans. Paper presented at the Neurobiology of Language Conference; San Diego, CA. 2010.

- The temporal structure of speech is critical for speech intelligibility and stream segregation
- Neuronal oscillations entrain to rhythms in speech, across multiple time scales
- Entrainment may serve to enhance the representation of an attended speaker at a ‘Cocktail Party’
- Speech decoding and attentional stream selection are prime examples of ‘*Active Sensing*’.

The temporal envelope of speech

Spectrum of speech envelopes

Figure 1.

Top: A 9-second long excerpt of natural, conversational speech, with its broad-band temporal envelope overlaid in red. The temporal envelope was extracted by calculating the RMS power of the broad-band sound wave, following the procedure described by (Lalor & Foxe, 2010). Bottom: A typical spectrum of the temporal envelope of speech, averaged over 8 tokens of natural conversational speech, each one ~10 seconds long.

Figure 2.

A schematic illustration of selective neuronal ‘entrainment’ to the temporal envelope of speech. On the left are the temporal envelopes derived from two different speech tokens, as well as their combination in a ‘Cocktail Party’ simulation. Although the envelopes share similar spectral makeup, with dominant delta and theta band modulations, they differ in their particular temporal/phase pattern (Pearson correlation $r = 0.06$). On the right is a schematic of the anticipated neural dynamics under the Selective Entrainment Hypothesis, filtered between 1–20Hz (zero-phase FIR filter). When different speech tokens are presented in isolation (top), the neuronal response tracks the temporal envelope of the stimulus, as suggested by several studies. When the two speakers are presented simultaneously, (bottom) it is suggested that attention enforces selective entrainment to the **attended** token. As a result, the neuronal response should be different when attending to different speakers in the same environment, despite the identical acoustic input. Moreover, the dynamics of the response when attending to each speaker should be similar to those observed when listening to the same speaker alone (with reduced/suppressed contribution of the sensory response to the unattended speaker).