frontiers in
HUMAN NEUROSCIENCE

REVIEW ARTICLE
published: 28 May 2014
doi: 10.3389/fnhum.2014.00311

Cortical entrainment to continuous speech: functional roles and interpretations

Nai Ding [1] **** and Jonathan Z. Simon*** [2,3,4] *******

[1] *Department of Psychology, New York University, New York, NY, USA*
[2] *Department of Electrical and Computer Engineering, University of Maryland College Park, College Park, MD, USA*
[3] *Department of Biology, University of Maryland College Park, College Park, MD, USA*
[4] *Institute for Systems Research, University of Maryland College Park, College Park, MD, USA*

Edited by:
Sonja A. E. Kotz, Max Planck Institute for Human Cognitive and Brain Sciences, Germany

Reviewed by:
István Winkler, University of Szeged, Hungary
Jonas Obleser, Max Planck Institute for Human Cognitive and Brain Sciences, Germany

****Correspondence:***
Nai Ding, Department of Psychology, New York University, New York, NY 10012, USA
e-mail: gahding@gmail.com;
Jonathan Z. Simon, Department of Electrical and Computer Engineering, University of Maryland College Park, College Park, MD 20742, USA
e-mail: jzsimon@umd.edu

Auditory cortical activity is entrained to the temporal envelope of speech, which corresponds to the syllabic rhythm of speech. Such entrained cortical activity can be measured from subjects naturally listening to sentences or spoken passages, providing a reliable neural marker of online speech processing. A central question still remains to be answered about whether cortical entrained activity is more closely related to speech perception or non-speech-specific auditory encoding. Here, we review a few hypotheses about the functional roles of cortical entrainment to speech, e.g., encoding acoustic features, parsing syllabic boundaries, and selecting sensory information in complex listening environments. It is likely that speech entrainment is not a homogeneous response and these hypotheses apply separately for speech entrainment generated from different neural sources. The relationship between entrained activity and speech intelligibility is also discussed. A tentative conclusion is that theta-band entrainment (4–8 Hz) encodes speech features critical for intelligibility while delta-band entrainment (1–4 Hz) is related to the perceived, non-speech-specific acoustic rhythm. To further understand the functional properties of speech entrainment, a splitter's approach will be needed to investigate (1) not just the temporal envelope but what specific acoustic features are encoded and (2) not just speech intelligibility but what specific psycholinguistic processes are encoded by entrained cortical activity. Similarly, the anatomical and spectro-temporal details of entrained activity need to be taken into account when investigating its functional properties.

Keywords: auditory cortex, entrainment of rhythms, speech intelligibility, speech perception in noise, speech envelope, cocktail party problem

INTRODUCTION

Speech recognition is a process that maps an acoustic signal onto the underlying linguistic meaning. The acoustic properties of speech are complex and contain temporal dynamics on several time scales (Rosen, 1992; Chi et al., 2005). The time scale most critical for speech recognition is on the order of hundreds of milliseconds (1–10 Hz), and the temporal fluctuations on this time scale are usually called the *temporal envelope* (**Figure 1A**). Single neuron neurophysiology from animal models has shown that neurons in primary auditory cortex encode the analogous temporal envelope of other non-speech sounds by phase locked neural firing (Wang et al., 2003). In contrast, the finer scale acoustic properties that decide the pitch and timbre of speech at each time moment (acoustic fragments lasting a few 100 ms) are likely to be encoded using a spatial code, by either individual neurons (Bendor and Wang, 2005) or spatial patterns of cortical activity (Walker et al., 2011).

In the last decade or so, cortical entrainment to the temporal envelope of speech has been demonstrated in humans using magnetoencephalography (MEG; Ahissar et al., 2001; Luo and Poeppel, 2007), electroencephalography (EEG; Aiken and Picton, 2008), and electrocorticography (ECoG; Nourski et al., 2009). This envelope following response can be recorded from subjects listening to sentences or spoken passages and therefore provides an online marker of neural processing of continuous speech. Envelope entrainment has mainly been seen in the waveform of low-frequency neural activity (<8 Hz) and in the power envelope of high-gamma activity (Pasley et al., 2012; Zion Golumbic et al., 2013). Although the phenomenon of envelope entrainment has been well established, its underlying neural mechanisms, and functional roles remain controversial. It is still under debate whether entrained cortical activity is more closely tied to the physical properties of the acoustic stimulus or to higher level language related processing that is directly related to speech perception. A number of studies have shown that cortical entrainment to speech is strongly modulated by top–down cognitive functions such as attention (Kerlin et al., 2010; Ding and Simon, 2012a; Mesgarani and Chang, 2012; Zion Golumbic et al., 2013) and therefore is not purely a bottom-up response. On the other hand, cortical entrainment to the sound envelope is seen for non-speech sound (Lalor et al., 2009; Hämäläinen et al., 2012; Millman et al., 2012; Wang et al., 2012; Steinschneider et al., 2013) and therefore does not rely on speech-specific neural processing. In this article, we first summarize a number of hypotheses about the functional roles of envelope entrainment, and then review the literature about how envelope entrainment is affected by speech intelligibility.

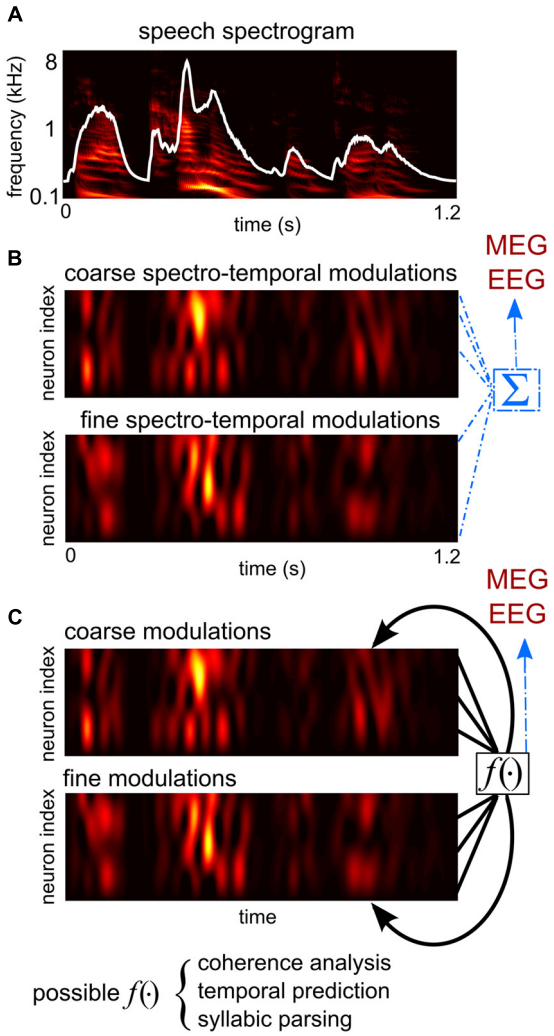

FIGURE 1 | A schematic illustration of hypotheses proposed to explain the generation of cortical entrainment to the speech envelope. (A) The spectro-temporal representation of speech, obtained from a cochlear model (Yang et al., 1992). The broad-band temporal envelope of speech, the sum of the spectro-temporal representation over frequency, is superimposed in white. **(B)** An illustration of the collective feature tracking hypothesis and the onset tracking hypothesis. The colored images show time courses of the dendritic activity of two example groups of neurons, hypothetically in primary and associative auditory areas. One group encodes the slow temporal modulations and coarse spectral modulations of sound intensity, i.e., the spectro-temporal envelope of speech, which contain major phonetic cues. The other group encodes the slow temporal changes of cues computed from the spectro-temporal fine structure, e.g., the pitch contour and the trajectory of the sound source location. According to the collective feature tracking hypothesis, magnetoencephalography (MEG)/electroencephalography (EEG) measurements are the direct sum of dendritic activity across all such neural populations in primary and associative auditory areas. The onset tracking hypothesis is similar, but instead neurons encoding the temporal edges of speech dominate cortical activity and thus drive MEG/EEG measurable responses. **(C)** An illustration of the syllabic parsing hypothesis and the sensory selection hypotheses. These hypotheses assume certain computations that integrate over distributively-represented auditory features. The syllable parsing hypothesis hypothesizes neural operations integrating features belonging to the same syllable. The sensory selection hypotheses propose either a temporal coherence analysis or a temporal predictive analysis.

FUNCTIONAL ROLES OF CORTICAL ENTRAINMENT

A number of hypotheses have been proposed about what aspects of speech, ranging from its acoustic features to its linguistic meaning, are encoded by entrained cortical activity. A few dominant hypotheses are summarized and compared (**Table 1**). Other unresolved questions about cortical neural entrainment, e.g., what the biophysical mechanisms generating cortical entrainment are, and whether entrained neural activity is related to spontaneous neural oscillations, are not covered here (see discussions in e.g., Schroeder and Lakatos, 2009; Howard and Poeppel, 2012; Ding and Simon, 2013b).

ONSET TRACKING HYPOTHESIS

Speech is dynamic and is full of acoustic "edges," e.g., onsets and offsets. These edges usually occur at syllable boundaries and are well characterized by the speech envelope. It is well known that a reliable macroscopic brain response can be evoked by an acoustic edge. Therefore, it has been proposed that neural entrainment to the speech envelope is a superposition of discrete, edge/onset related brain responses (Howard and Poeppel, 2010). Consistent with this hypothesis, it has been shown that the sharpness of acoustic edges, i.e., how quickly sound intensity increases, strongly influences cortical tracking of the sound envelope (Prendergast et al., 2010; Doelling et al., 2014). A challenge of this hypothesis, however, is that speech is continuously changing and it remains a problem as to which acoustic transients can be counted as edges.

If this hypothesis is true, a question naturally follows about whether envelope entrainment can provide insights that cannot be learned using the traditional event-related response approach. The answer is yes. Cortical responses, including edge/onset related auditory evoked responses, are stimulus-dependent, and quickly adapt to the spectro-temporal structure of the stimulus (Zacharias et al., 2012; Herrmann et al., 2014). Therefore, even if envelope entrainment is just a superposition of event-related responses, it can still provide insights about the properties of cortical activity when it is adapted to the acoustic properties of speech.

COLLECTIVE FEATURE TRACKING HYPOTHESIS

When sound enters the ear, it is decomposed into narrow frequency bands in the auditory periphery and is further decomposed into multi-scale acoustic features in the central auditory system, such as pitch, sound source location information, and coarse spectro-temporal modulations (Shamma, 2001; Ghitza et al., 2012). In speech, most acoustic features coherently fluctuate in time and these coherent fluctuations are captured by the speech envelope. If a neuron or a neural population encodes an acoustic feature, its activity is synchronized to the strength of that acoustic feature. As a result, neurons or neural networks that are tuned to coherently fluctuating speech features are activated coherently (Shamma et al., 2011).

Analogously to the speech envelope being the summation of the power of all speech features at each time moment, the large-scale neural entrainment to speech measured by MEG/EEG can be the summation of neural activity tracking different acoustic features of speech (**Figure 1B**). It is therefore plausible to hypothesize that macroscopic speech entrainment is a passive summation of microscopic neural tracking of acoustic features

Table 1 | A summary of major hypotheses about the functional roles of cortical entrainment to speech.

Hypothesis	Underlying neural computations	Reference
Onset tracking	Temporal edge detection	Howard and Poeppel (2010)
Collective feature tracking	Spectro-temporal feature coding	Ding and Simon (2012b),Ghitza et al. (2012)
Syllabic parsing	Binding features of the same syllable; discretization	Giraud and Poeppel (2012),Ghitza (2013)
Sensory selection I	Temporal coherence-based binding of auditory features	Shamma et al. (2011),Ding and Simon (2012a)
Sensory selection II	Modulation of neuronal excitability; temporal prediction	Schroeder and Lakatos (2009)

across neurons/networks (Ding and Simon, 2012b). Based on this hypothesis, the MEG/EEG speech entrainment is a marker of a collective cortical representation of speech but does not play any additional roles in regulating neuronal activity.

The onset tracking hypothesis can be viewed as a special case of the collective feature tracking hypothesis, when the acoustic features driving cortical responses are restricted to a set of discrete edges. The collective feature tracking hypothesis, however, is more general since it allows features to be continuously changing and also incorporates features that are not associated with sharp intensity changes, such as changes in the pitch contour (Obleser et al., 2012), and sound source location. Under the onset tracking hypothesis, entrained neural activity is a superposition of onset/edge-related auditory evoked responses. Under the more general collective feature tracking hypothesis, at a first-order approximation, entrained activity is a convolution between speech features, e.g., the temporal envelopes in different narrow frequency bands, and the corresponding response functions, e.g., the response evoked by a very brief tone pip in the corresponding frequency band (Lalor et al., 2009; Ding and Simon, 2012b).

SYLLABIC PARSING HYPOTHESIS

During speech recognition, the listener must segment a continuous acoustic signal into a sequence of discrete linguistic symbols, into the units of, e.g., phonemes, syllables or words. The boundaries between phonemes, and especially syllables, are relatively well encoded by the speech envelope (Stevens, 2002; Ghitza, 2013, see also Cummins, 2012). Furthermore, the average syllabic rate ranges between 5 and 8 Hz across languages (Pellegrino et al., 2011) and the rate for stressed syllables is below 4 Hz for English (Greenberg et al., 2003). Therefore it has been hypothesized that neural entrainment to the speech envelope plays a role in creating a syllabic level, discrete, representation of speech (Giraud and Poeppel, 2012). In particular, it has been hypothesized that each cycle of the cortical theta oscillation (4–8 Hz) is aligned to the portion of speech signal in between of two vowels, corresponding to two adjacent peaks in the speech envelope. Auditory features within a cycle of theta oscillation are then used to decode the phonetic information of speech (Ghitza, 2011, 2013). Therefore, according to this hypothesis, speech entrainment does not only passively track acoustic features but also reflects the language-based packaging of speech into syllable size chunks. Since syllables play different roles in segmenting syllable-timed language and stress-timed language (Cutler et al., 1986), further cross-language research may further elucidate which of these neural processes are represented in envelope tracking activity.

SENSORY SELECTION HYPOTHESIS

In everyday listening environments, speech is often embedded in a complex acoustic background. Therefore, to understand speech, a listener must segregate speech from the listening background and process it selectively. A useful strategy for the brain would be to find and selectively process moments in time (or spectro-temporal instances in a more general framework) that are dominated by speech and ignore the moments dominated by the background (Wang, 2005; Cooke, 2006). In other words, the brain might robustly encode speech by taking glimpses at the temporal (or spectro-temporal) features that contain critical speech information. The rhythmicity of speech (Schroeder and Lakatos, 2009; Giraud and Poeppel, 2012), and the temporal coherence between acoustic features (Shamma et al., 2011), are both reflected by the speech envelope and so become critical cues for the brain to decide where the useful speech information lies. Therefore, envelope entrainment may play a critical role in the neural segregation of speech and the listening background.

In a complex listening environment, cortical entrainment to speech has been found to be largely invariant to the listening background (Ding and Simon, 2012a; Ding and Simon, 2013a). Two possible functional roles have been hypothesized for the observed background-invariant envelope entrainment. One is that the brain uses temporal coherence to bind together acoustic features belonging to the same speech stream and envelope entrainment may reflect computations related to this coherence analysis (Shamma et al., 2011; Ding and Simon, 2012a). The other is that envelope entrainment is used by the brain to predict which moments contain more information about speech than the acoustic background and then guide the brain to selectively process those moments (Schroeder et al., 2008; Schroeder and Lakatos, 2009; Zion Golumbic et al., 2012).

WHICH HYPOTHESIS IS TRUE? AN ANALYSIS-BY-SYNTHESIS ACCOUNT OF SPEECH PROCESSING

Speech processing is a complicated process that can be roughly divided into an analysis stage and a synthesis stage. In the analysis stage, speech sounds are decomposed into primitive auditory features, a process that starts from the cochlea and applies mostly equally to the auditory encoding of both speech and non-speech sounds. A later synthesis stage, in contrast, combines multiple

auditory features to create speech perception, including, e.g., binding spectro-temporal cues to determine phonemic categories, or integrating multiple acoustic cues to segregate a target speech stream from an acoustic background. The onset tracking hypothesis and the collective feature tracking hypothesis both view speech entrainment as a passive auditory encoding mechanism belonging to the analysis stage. Note, however, that the analysis stage does include some integration over separately represented features also. For example, neural processing of pitch and spectral modulations requires integrating information across frequency. Functionally, however, the purpose of integrating features in the analysis stage is to extract higher level auditory features rather than to construct linguistic/perceptual entities.

The syllabic parsing hypothesis and the sensory selection hypothesis propose functional roles of cortical entrainment in the synthesis stage. They hypothesize that cortical entrainment is involved in combining features into linguistic units, e.g., syllables, or perceptual units, e.g., speech streams (**Figure 1C**). These additional functional roles may be implemented in two ways: an active mechanism would be one that entrained cortical activity, as a large-scale voltage fluctuation, directly regulating syllabic parsing or sensory selection (Schroeder et al., 2008; Schroeder and Lakatos, 2009). A passive mechanism would be one where neural computations related to syllabic parsing or sensory selection would generate spatially coherent neural signals that are measurable by macroscopic recording tools.

Although clearly distinctive from each other, the four hypotheses may all be true for different functional areas of the brain and describe different neural generators for speech entrainment. Onset detection, feature tracking, syllabic parsing, and sensory selection are all neural computations necessary for speech recognition and all of them are likely to be synchronized to the speech rhythm carried by the envelope. Therefore, these neural computations may all be reflected by cortical entrainment to speech, and may only differ in their fine-scale neural generators. It remains unclear, however, whether these fine-scale neural generators can be resolved by macroscopic recording tools such as MEG and EEG.

Future studies are needed to explicitly test these hypotheses, or explicitly modify them, to determine which specific acoustic features and which specific psycholinguistic processes are relevant to cortical entrainment. For example, to dissociate the onset tracking hypothesis and the collective feature tracking hypothesis, one approach is to create explicit computational models for them and test which model would fit the data better. To test the syllabic parsing hypothesis, it will be important to calculate the correlation between cortical entrainment and relevant behavioral measures, e.g., misallocation of syllable boundaries (Woodfield and Akeroyd, 2010). To test the sensory selection hypothesis, stimuli that vary in their temporal probability or coherence among spectro-temporal features are likely to be revealing.

ENVELOPE ENTRAINMENT AND SPEECH INTELLIGIBILITY

ENTRAINMENT AND ACOUSTIC MANIPULATION OF SPEECH

As indicated by its name, envelope entrainment is correlated with the speech envelope, an acoustic property of speech. Nevertheless, neural encoding of speech must underlie the ultimate goal of decoding its meaning. Therefore, it is critical to identify if cortical entrainment to speech is related to any behavioral measure during speech recognition, such as speech intelligibility.

A number of studies have compared cortical activity entrained to intelligible speech and unintelligible speech. One approach is to vary the acoustic stimulus and analyze how cortical entrainment changes within individual subjects. Some studies have found that cortical entrainment to normal sentences is similar to cortical entrainment to sentences that are played backward in time (Howard and Poeppel, 2010; Peña and Melloni, 2012; though see Gross et al., 2013).

A second way to reduce intelligibility is to introduce different types of acoustic interference. When speech is presented together with stationary noise, delta-band (1–4 Hz) cortical entrainment to the speech is found to be robust to noise until the listeners can barely hear speech, while theta-band (4–8 Hz) entrainment decreases gradually as the noise level increases (Ding and Simon, 2013a). In this way, theta-band entrainment is correlated with noise level and also speech intelligibility, but delta-band entrainment is not. When speech is presented together with a competing speech stream, cortical entrainment is found to be robust against the level of the competing speech stream even though intelligibility drops (Ding and Simon, 2012a; theta- and delta-band activity was not analyzed separately there).

A third way to reduce speech intelligibility is to degrade the spectral resolution through noise-vocoding, which destroys spectro-temporal fine structure but preserves the temporal envelope (Shannon et al., 1995). When the spectral resolution of speech decreases, it has been shown that theta-band cortical entrainment reduces (Peelle et al., 2013; Ding et al., 2014) but delta-band entrainment enhances (Ding et al., 2014). In contrast, when background noise is added to speech and the speech-noise mixture is noise vocoded, it is found that both delta- and theta-band entrainment is reduced by vocoding (Ding et al., 2014).

A fourth way to vary speech intelligibility is to directly manipulate the temporal envelope (Doelling et al., 2014). When the temporal envelope in the delta-theta frequency range is corrupted, cortical entrainment in the corresponding frequency bands degrades and so does speech intelligibility. When a processed speech envelope is used to modulate a broadband noise carrier, the stimulus is not intelligible but reliable cortical entrainment is nevertheless seen.

In many of these studies investigating the correlation between cortical entrainment and intelligibility, a common issue is that stimuli which differ in intelligibly also differ in acoustic properties. This makes it is difficult to determine if changes in cortical entrainment arise from changes in speech intelligibility or from changes in acoustic properties. For example, speech syllables generally have a sharper onset than offset, so reversing speech in time changes those temporal characteristics. Similarly, when the spectral resolution is reduced, neurons tuned to fine spectral features are likely to be deactivated. Therefore, based on the studies reviewed here, it can only be tentatively concluded that, when critical speech features are manipulated, speech intelligibility, and theta-band entrainment are affected in similar ways while delta-band entrainment is not. It remains unclear about

whether speech intelligibility causally modulates cortical entrainment or that auditory encoding, reflected by cortical entrainment, influences downstream language processing and therefore become indirectly related to intelligibility.

VARIABILITY BETWEEN LISTENERS

A second approach to address the correlation between neural entrainment and speech intelligibility is to investigate the variability across listeners. Peña and Melloni (2012) compared neural responses in listeners who speak the tested language and listeners who do not speak the tested language. It was found that language understanding does not significantly change the low-frequency neural responses, but it does change high-gamma band neural activity. Within the group of native speakers, the intelligibility score still varied broadly in the challenging listening conditions. Delta-band, but not theta-band, cortical entrainment has been shown to correlate with intelligibility scores for individual listeners in a number of studies (Ding and Simon, 2013a; Ding et al., 2014; Doelling et al., 2014). The advantage of investigating inter-subject variability is that it avoids modifications of the sound stimuli. Nevertheless, it still cannot identify whether the individual differences in speech recognition arise from the individual differences in auditory processing (Ruggles et al., 2011), language related processing, or cognitive control.

The speech intelligibility approach in general, suffers from a drawback that it is the end point of the entire speech recognition chain, and is not targeted at specific linguistic computations, e.g., allocating the boundaries between syllables. Furthermore, when the acoustic properties of speech are degraded, speech recognition requires additional cognitive control and the involved neural processing networks adapt (Du et al., 2011; Wild et al., 2012; Erb et al., 2013; Lee et al., 2014). Therefore, just from a change in speech intelligibility, it is difficult to trace what kinds of neural processing are affected.

DISTINCTIONS BETWEEN DELTA- AND THETA-BAND ENTRAINMENT

In summary of these different approaches, when the acoustic properties of speech are manipulated, theta-band entrainment often shows changes that correlate with speech intelligibility. For the same stimulus, however, the speech intelligibility measured from individual listeners is often correlated with delta-band entrainment. To explain this dichotomy, here we hypothesize that theta-band entrainment encodes syllabic-level acoustic features critical for speech recognition, while delta-band entrainment is more closely related to the perceived acoustic rhythm rather than the phonemic information of speech. This hypothesis is also consistent with the fact that speech modulations between 4 and 8 Hz are critical for intelligibility (Drullman et al., 1994a,b; Elliott and Theunissen, 2009) while temporal modulations below 4 Hz include prosodic information of speech (Goswami and Leong, 2013) and it is the frequency range important for music rhythm perception (Patel, 2008; Farbood et al., 2013).

ENVELOPE ENTRAINMENT TO NON-SPEECH SOUNDS

Although speech envelope entrainment may show correlated changes with speech intelligibility when the acoustic properties of speech are manipulated, speech intelligibility is probably not a major driving force for envelope entrainment. A critical evidence is that envelope entrainment can be observed for non-speech sounds in humans and both speech and non-speech sounds in animals. Here, we briefly review human studies on envelope entrainment for non-speech sounds (see e.g., Steinschneider et al., 2013 for a comparison between envelope entrainment in human and animal models).

Traditionally, envelope entrainment has been studied using the auditory steady-state response (aSSR), a periodic neural response tracking the stimulus repetition rate or modulation rate. An aSSR at a given frequency can be elicited by, e.g., a click or tone-pip train repeating at the same frequency (Nourski et al., 2009; Xiang et al., 2010), and by amplitude or frequency modulation at that frequency (Picton et al., 1987; Ross et al., 2000; Wang et al., 2012). Although the cortical aSSR can be elicited in a broad frequency range (up to ∼100 Hz), speech envelope entrainment is likely to be related to the slow aSSR in the corresponding frequency range, i.e., below 10 Hz (see Picton, 2007 for a review of the robust aSSR of 40 Hz and above). More recently, cortical entrainment has also been demonstrated for sounds modulated by an irregular envelope (Lalor et al., 2009). Low-frequency (<10 Hz) cortical entrainment to non-speech sound shares many properties with cortical entrainment to speech. For example, when envelope entrainment is modeled using a linear system-theoretic model, the neural response is qualitatively similar for speech (Power et al., 2012) and amplitude-modulated tones (Lalor et al., 2009). Furthermore, low-frequency (<10 Hz) cortical entrainment to non-speech sound is also strongly modulated by attention (Elhilali et al., 2009; Power et al., 2010; Xiang et al., 2010), and the phase of entrained activity is predictive of listeners' performance in some sound-feature detection tasks (Henry and Obleser, 2012; Ng et al., 2012).

SUMMARY

Cortical entrainment to the speech envelope provides a powerful tool to investigate online neural processing of continuous speech. It greatly extends the traditional event-related approach that can only be applied to analyze the response to isolated syllables or words. Although envelope entrainment has attracted researchers' attention in the last decade, it is still a less well-characterized cortical response than event-related responses. The basic phenomenon of envelope entrainment has been reliably seen in EEG, MEG, and ECoG, even at the single-trial level (Ding and Simon, 2012a; O'Sullivan et al., 2014). Hypotheses have been proposed about the neural mechanisms generating cortical entrainment and its functional roles, but these hypotheses remain to be explicitly tested. To test these hypotheses, a computational modeling approach is likely to be effective. For example, rather than just calculating the correlation between neural activity and the speech envelope, more explicit computational models can be proposed and used to fit the data (e.g., Ding and Simon, 2013a). Furthermore, to understand what linguistic computations are achieved by entrained cortical activity, more fine-scaled behavioral measures are likely to be required, e.g., measures related to syllable boundary allocation rather than the general measure of intelligibility. Finally, the

anatomical, temporal, and spectral specifics of cortical entrainment should be taken into account when discussing its functional roles (Peña and Melloni, 2012; Zion Golumbic et al., 2013; Ding et al., 2014).

AUTHOR CONTRIBUTIONS

Nai Ding and Jonathan Z. Simon wrote and approved the paper.

ACKNOWLEDGMENT

The work is supported by NIH grant R01 DC 008342.

REFERENCES

Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., and Merzenich, M. M. (2001). Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proc. Natl. Acad. Sci. U.S.A.* 98, 13367–13372. doi: 10.1073/pnas.201400998

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

Bendor, D., and Wang, X. (2005). The neuronal representation of pitch in primate auditory cortex. *Nature* 436, 1161–1165. doi: 10.1038/nature03867

Chi, T., Ru, P., and Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *J. Acoust. Soc. Am.* 118, 887–906. doi: 10.1121/1.1945807

Cooke, M. (2006). A glimpsing model of speech perception in noise. *J. Acoust. Soc. Am.* 119, 1562–1573. doi: 10.1121/1.2166600

Cummins, F. (2012). Oscillators and syllables: a cautionary note. *Front. Psychol.* 3:364. doi: 10.3389/fpsyg.2012.00364

Cutler, A., Mehler, J., Norris, D., and Segui, J. (1986). The syllable's differing role in the segmentation of French and English. *J. Mem. Lang.* 25, 385–400. doi: 10.1016/0749-596X(86)90033-1

Ding, N., Chatterjee, M., and Simon, J. Z. (2014). Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. *Neuroimage* 88C, 41–46. doi: 10.1016/j.neuroimage.2013.10.054 [Epub ahead of print].

Ding, N., and Simon, J. Z. (2012a). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proc. Natl. Acad. Sci. U.S.A.* 109, 11854–11859. doi: 10.1073/pnas.1205381109

Ding, N., and Simon, J. Z. (2012b). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *J. Neurophysiol.* 107, 78–89. doi: 10.1152/jn.00297.2011

Ding, N., and Simon, J. Z. (2013a). Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. *J. Neurosci.* 33, 5728–5735. doi: 10.1523/JNEUROSCI.5297-12.2013

Ding, N., and Simon, J. Z. (2013b). Power and phase properties of oscillatory neural responses in the presence of background activity. *J. Comput. Neurosci.* 34, 337–343. doi: 10.1007/s10827-012-0424-6

Doelling, K., Arnal, L., Ghitza, O., and Poeppel, D. (2014). Acoustic landmarks drive delta-theta oscillations to enable speech comprehension by facilitating perceptual parsing. *Neuroimage* 85, 761–768. doi: 10.1016/j.neuroimage.2013.06.035

Drullman, R., Festen, J. M., and Plomp, R. (1994a). Effect of reducing slow temporal modulations on speech reception. *J. Acoust. Soc. Am.* 95, 2670–2680. doi: 10.1121/1.409836

Drullman, R., Festen, J. M., and Plomp, R. (1994b). Effect of temporal envelope smearing on speech reception. *J. Acoust. Soc. Am.* 95, 1053–1064. doi: 10.1121/1.408467

Du, Y., He, Y., Ross, B., Bardouille, T., Wu, X., Li, L., et al. (2011). Human auditory cortex activity shows additive effects of spectral and spatial cues during speech segregation. *Cereb. Cortex* 21, 698–707. doi: 10.1093/cercor/bhq136

Elhilali, M., Xiang, J., Shamma, S. A., and Simon, J. Z. (2009). Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene. *PLoS Biol.* 7:e1000129. doi: 10.1371/journal.pbio.1000129

Elliott, T., and Theunissen, F. (2009). The modulation transfer function for speech intelligibility. *PLoS Comput. Biol.* 5:e1000302. doi: 10.1371/journal.pcbi.1000302

Erb, J., Henry, M. J., Eisner, F., and Obleser, J. (2013). The brain dynamics of rapid perceptual adaptation to adverse listening conditions. *J. Neurosci.* 33, 10688–10697. doi: 10.1523/JNEUROSCI.4596-12.2013

Farbood, M. M., Marcus, G., and Poeppel, D. (2013). Temporal dynamics and the identification of musical key. *J. Exp. Psychol. Hum. Percept. Perform.* 39, 911–918. doi: 10.1037/a0031087

Ghitza, O. (2011). Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input rhythm. *Front. Psychol.* 2:130. doi: 10.3389/fpsyg.2011.00130

Ghitza, O. (2013). The theta-syllable: a unit of speech information defined by cortical function. *Front. Psychol.* 4:138. doi: 10.3389/fpsyg.2013.00138

Ghitza, O., Giraud, A.-L., and Poeppel, D. (2012). Neuronal oscillations and speech perception: critical-band temporal envelopes are the essence. *Front. Hum. Neurosci.* 6:340. doi: 10.3389/fnhum.2012.00340

Giraud, A.-L., and Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. *Nat. Neurosci.* 15, 511–517. doi: 10.1038/nn.3063

Goswami, U., and Leong, V. (2013). Speech rhythm and temporal structure: converging perspectives? *Lab. Phonol.* 4, 67–92. doi: 10.1515/lp-2013-0004

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

Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., et al. (2013). Speech rhythms and multiplexed oscillatory sensory coding in the human brain. *PLoS Biol.* 11:e1001752. doi: 10.1371/journal.pbio.1001752

Hämäläinen, J. A., Rupp, A., Soltész, F., Szücs, D., and Goswami, U. (2012). Reduced phase locking to slow amplitude modulation in adults with dyslexia: an MEG study. *Neuroimage* 59, 2952–2961. doi: 10.1016/j.neuroimage.2011.09.075

Henry, M. J., and Obleser, J. (2012). Frequency modulation entrains slow neural oscillations and optimizes human listening behavior. *Proc. Natl. Acad. Sci. U.S.A.* 109, 20095–20100. doi: 10.1073/pnas.1213390109

Herrmann, B., Schlichting, N., and Obleser, J. (2014). Dynamic range adaptation to spectral stimulus statistics in human auditory cortex. *J. Neurosci.* 34, 327–331. doi: 10.1523/JNEUROSCI.3974-13.2014

Howard, M. F., and Poeppel, D. (2010). Discrimination of speech stimuli based on neuronal response phase patterns depends on acoustics but not comprehension. *J. Neurophysiol.* 104, 2500–2511. doi: 10.1152/jn.00251.2010

Howard, M. F., and Poeppel, D. (2012). The neuromagnetic response to spoken sentences: co-modulation of theta band amplitude and phase. *Neuroimage* 60, 2118–2127. doi: 10.1016/j.neuroimage.2012.02.028

Kerlin, J. R., Shahin, A. J., and Miller, L. M. (2010). Attentional gain control of ongoing cortical speech representations in a "cocktail party". *J. Neurosci.* 30, 620–628. doi: 10.1523/JNEUROSCI.3631-09.2010

Lalor, E. C., Power, A. J., Reilly, R. B., and Foxe, J. J. (2009). Resolving precise temporal processing properties of the auditory system using continuous stimuli. *J. Neurophysiol.* 102, 349–359. doi: 10.1152/jn.90896.2008

Lee, A. K., Larson, E., Maddox, R. K., and Shinn-Cunningham, B. G. (2014). Using neuroimaging to understand the cortical mechanisms of auditory selective attention. *Hear. Res.* 307, 111–120. doi: 10.1016/j.heares.2013.06.010

Luo, H., and Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron* 54, 1001–1010. doi: 10.1016/j.neuron.2007.06.004

Mesgarani, N., and Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature* 485, 233–236. doi: 10.1038/nature11020

Millman, R. E., Prendergast, G., Hymers, M., and Green, G. G. (2012). Representations of the temporal envelope of sounds in human auditory cortex: can the results from invasive intracortical "depth" electrode recordings be replicated using non-invasive MEG "virtual electrodes"? *Neuroimage* 64, 185–196. doi: 10.1016/j.neuroimage.2012.09.017

Ng, B. S. W., Schroeder, T., and Kayser, C. (2012). A precluding but not ensuring role of entrained low-frequency oscillations for auditory perception. *J. Neurosci.* 32, 12268–12276. doi: 10.1523/JNEUROSCI.1877-12.2012

Nourski, K. V., Reale, R. A., Oya, H., Kawasaki, H., Kovach, C. K., Chen, H., et al. (2009). Temporal envelope of time-compressed speech represented in the human auditory cortex. *J. Neurosci.* 29, 15564–15574. doi: 10.1523/JNEUROSCI.3065-09.2009

Obleser, J., Herrmann, B., and Henry, M. J. (2012). Neural oscillations in speech: don't be enslaved by the envelope. *Front. Hum. Neurosci.* 6:250. doi: 10.3389/fnhum.2012.00250

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., et al. (2014). Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cereb. Cortex.* doi: 10.1093/cercor/bht355

Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., Shamma, S. A., Crone, N. E., et al. (2012). Reconstructing speech from human auditory cortex. *PLoS Biol.* 10:e1001251. doi: 10.1371/journal.pbio.1001251

Patel, A. D. (2008). *Music, Language, and the Brain.* New York, NY: Oxford University Press.

Peelle, J. E., Gross, J., and Davis, M. H. (2013). Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cereb. Cortex* 23, 1378–1387. doi: 10.1093/cercor/bhs118

Pellegrino, F., Coupé, C., and Marsico, E. (2011). Across-language perspective on speech information rate. *Language* 87, 539–558. doi: 10.1353/lan.2011.0057

Peña, M., and Melloni, L. (2012). Brain oscillations during spoken sentence processing. *J. Cogn. Neurosci.* 24, 1149–1164. doi: 10.1162/jocn_a_00144

Picton, T. W. (2007). "Audiometry using auditory steady-state responses," in *Auditory Evoked Potentials: Basic Principles and Clinical Application,* eds. R. F. Burkard, J. J. Eggermont, and M. Don (Baltimore: Lippincott Williams & Wilkins), 441–462.

Picton, T. W., Skinner, C. R., Champagne, S. C., Kellett, A. J. C., and Maiste, A. C. (1987). Potentials evoked by the sinusoidal modulation of the amplitude or frequency of a tone. *J. Acoust. Soc. Am.* 82, 165–178. doi: 10.1121/1.395560

Power, A. J., Foxe, J. J., Forde, E. J., Reilly, R. B., and Lalor, E. C. (2012). At what time is the cocktail party? A late locus of selective attention to natural speech. *Eur. J. Neurosci.* 35, 1497–1503. doi: 10.1111/j.1460-9568.2012.08060.x

Power, A. J., Lalor, E. C., and Reilly, R. B. (2010). Endogenous auditory spatial attention modulates obligatory sensory activity in auditory cortex. *Cereb. Cortex* 21, 1223–1230. doi: 10.1093/cercor/bhq233

Prendergast, G., Johnson, S. R., and Green, G. G. (2010). Temporal dynamics of sinusoidal and non-sinusoidal amplitude modulation. *Eur. J. Neurosci.* 32, 1599–1607. doi: 10.1111/j.1460-9568.2010.07423.x

Rosen, S. (1992). Temporal information in speech: acoustic, auditory and linguistic aspects. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 336, 367–373. doi: 10.1098/rstb.1992.0070

Ross, B., Borgmann, C., Draganova, R., Roberts, L. E., and Pantev, C. (2000). A high-precision magnetoencephalographic study of human auditory steady-state responses to amplitude-modulated tones. *J. Acoust. Soc. Am.* 108, 679–691. doi: 10.1121/1.429600

Ruggles, D., Bharadwaj, H., and Shinn-Cunningham, B. G. (2011). Normal hearing is not enough to guarantee robust encoding of suprathreshold features important in everyday communication. *Proc. Natl. Acad. Sci. U.S.A.* 108, 15516–15521. doi: 10.1073/pnas.1108912108

Schroeder, C. E., and Lakatos, P. (2009). Low-frequency neuronal oscillations as instruments of sensory selection. *Trends Neurosci.* 32, 9–18. doi: 10.1016/j.tins.2008.09.012

Schroeder, C. E., Lakatos, P., Kajikawa, Y., Partan, S., and Puce, A. (2008). Neuronal oscillations and visual amplification of speech. *Trends Cogn. Sci.* 12, 106–113. doi: 10.1016/j.tics.2008.01.002

Shamma, S. (2001). On the role of space and time in auditory processing. *Trends Cogn. Sci.* 5, 340–348. doi: 10.1016/S1364-6613(00)01704-6

Shamma, S. A., Elhilali, M., and Micheyl, C. (2011). Temporal coherence and attention in auditory scene analysis. *Trends Neurosci.* 34, 114–123. doi: 10.1016/j.tins.2010.11.002

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

Steinschneider, M., Nourski, K. V., and Fishman, Y. I. (2013). Representation of speech in human auditory cortex: is it special? *Hear. Res.* 305, 57–73. doi: 10.1016/j.heares.2013.05.013

Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *J. Acoust. Soc. Am.* 111, 1872–1891. doi: 10.1121/1.1458026

Walker, K. M., Bizley, J. K., King, A. J., and Schnupp, J. W. (2011). Multiplexed and robust representations of sound features in auditory cortex. *J. Neurosci.* 31, 14565–14576. doi: 10.1523/JNEUROSCI.2074-11.2011

Wang, D. (2005). "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation By Humans and Machines*, ed. P. Divenyi (New York: Springer), 181–197.

Wang, X., Lu, T., and Liang, L. (2003). Cortical processing of temporal modulations. *Speech Commun.* 41, 107–121. doi: 10.1016/S0167-6393(02)00097-3

Wang, Y., Ding, N., Ahmar, N., Xiang, J., Poeppel, D., and Simon, J. Z. (2012). Sensitivity to temporal modulation rate and spectral bandwidth in the human auditory system: MEG evidence. *J. Neurophysiol.* 107, 2033–2041. doi: 10.1152/jn.00310.2011

Wild, C. J., Yusuf, A., Wilson, D. E., Peelle, J. E., Davis, M. H., and Johnsrude, I. S. (2012). Effortful listening: the processing of degraded speech depends critically on attention. *J. Neurosci.* 32, 14010–14021. doi: 10.1523/JNEUROSCI.1528-12.2012

Woodfield, A., and Akeroyd, M. A. (2010). The role of segmentation difficulties in speech-in-speech understanding in older and hearing-impaired adults. *J. Acoust. Soc. Am.* 128, EL26–EL31. doi: 10.1121/1.3443570

Xiang, J., Simon, J., and Elhilali, M. (2010). Competing streams at the cocktail Party: exploring the mechanisms of attention and temporal integration. *J. Neurosci.* 30, 12084–12093. doi: 10.1523/JNEUROSCI.0827-10.2010

Yang, X., Wang, K., and Shamma, S. A. (1992). Auditory representations of acoustic signals. *IEEE Trans. Inf. Theory* 38, 824–839. doi: 10.1109/18.119739

Zacharias, N., König, R., and Heil, P. (2012). Stimulation – history effects on the M100 revealed by its differential dependence on the stimulus onset interval. *Psychophysiology* 49, 909–919. doi: 10.1111/j.1469-8986.2012.01370.x

Zion Golumbic, E. M., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., Mckhann, G. M., et al. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". *Neuron* 77, 980–991. doi: 10.1016/j.neuron.2012.12.037

Zion Golumbic, E. M., Poeppel, D., and Schroeder, C. E. (2012). Temporal context in speech processing and attentional stream selection: a behavioral and neural perspective. *Brain Lang.* 122, 151–161. doi: 10.1016/j.bandl.2011.12.010

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 10 March 2014; accepted: 27 April 2014; published online: 28 May 2014.
Citation: Ding N and Simon JZ (2014) Cortical entrainment to continuous speech: functional roles and interpretations. Front. Hum. Neurosci. ***8****:311. doi: 10.3389/fnhum.2014.00311*
This article was submitted to the journal Frontiers in Human Neuroscience.