NeuroImage 85 (2014) 761–768

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing

Keith B. Doelling [a], Luc H. Arnal [a], Oded Ghitza [c], David Poeppel [a,b,*]

[a] Department of Psychology, New York University, USA
[b] NYUAD Institute, New York University Abu Dhabi, P.O. Box 129188, Abu Dhabi, UAE
[c] Department of Biomedical Engineering, Boston University, USA

ARTICLE INFO

Article history:
Accepted 7 June 2013
Available online 19 June 2013

Keywords:
Neural oscillation
Auditory cortex
Speech
Acoustic edge
MEG
Perceptual parsing

ABSTRACT

A growing body of research suggests that intrinsic neuronal slow (<10 Hz) oscillations in auditory cortex appear to track incoming speech and other spectro-temporally complex auditory signals. Within this framework, several recent studies have identified critical-band temporal envelopes as the specific acoustic feature being reflected by the phase of these oscillations. However, how this alignment between speech acoustics and neural oscillations might underpin intelligibility is unclear. Here we test the hypothesis that the 'sharpness' of temporal fluctuations in the critical band envelope acts as a temporal cue to speech syllabic rate, driving delta–theta rhythms to track the stimulus and facilitate intelligibility. We interpret our findings as evidence that sharp events in the stimulus cause cortical rhythms to re-align and parse the stimulus into syllable-sized chunks for further decoding. Using magnetoencephalographic recordings, we show that by removing temporal fluctuations that occur at the syllabic rate, envelope-tracking activity is reduced. By artificially reinstating these temporal fluctuations, envelope-tracking activity is regained. These changes in tracking correlate with intelligibility of the stimulus. Together, the results suggest that the sharpness of fluctuations in the stimulus, as reflected in the cochlear output, drive oscillatory activity to track and entrain to the stimulus, at its syllabic rate. This process likely facilitates parsing of the stimulus into meaningful chunks appropriate for subsequent decoding, enhancing perception and intelligibility.

Introduction

Because auditory signals unfold over time, at multiple scales, the process of decoding input sounds to link them to meaningful objects or concepts requires integrating sensory information over time. In speech perception, this temporal integration must occur in at least two (and arguably more) distinct timescales which relate to syllabic-level (~200 ms or ~5 Hz) and phonemic-level (~25 ms or ~40 Hz) information. Several models have suggested that this type of multi-time resolution analysis and integration could be performed in auditory cortex using neuronal oscillations – corresponding to these two temporal windows of integration (~5 Hz, theta; ~40 Hz, gamma) – to parse the sound input at these separate timescales (Ghitza, 2011; Poeppel, 2003). It is hypothesized, in particular, that the phase of the slow oscillation (nested with gamma) locks to the syllabic rhythm to optimally decode and integrate syllabic and phonemic speech features (Giraud and Poeppel, 2012).

In this magnetoencephalography (MEG) study, we focus on the role of the longer temporal window, most readily corresponding to delta–theta oscillations, to gain a better mechanistic understanding of how neuronal activity in this band might underpin auditory perception and speech comprehension. Recently, much research has focused on slow neural oscillations and their relationship to auditory stimuli (Cogan and Poeppel, 2011; Ding and Simon, 2009; Howard and Poeppel, 2010, 2012; Luo and Poeppel, 2007, 2012; Peelle et al., 2013). In addition, the relevance of low-modulation frequency oscillations to multi-sensory perception has been demonstrated, for example in naturalistic scenes or the well-studied cocktail party scenario (Kerlin et al., 2010; Luo et al., 2010; Zion Golumbic et al., 2013). There is an emerging consensus that the phase of slow oscillations precisely tracks the stimulus acoustics. However whether this stimulus–response alignment across time is necessary for speech comprehension remains debated (Howard and Poeppel, 2010; versus Luo and Poeppel, 2007; Peelle et al., 2013). One hypothesis is that cortical delta–theta oscillations track the critical band envelopes of the stimulus — a feature which carries crucial cues regarding segmental and syllabic information (Rosen, 1992)[1]. Despite the

Abbreviations: CALM, categorization and learning module; CACoh, cerebro-acoustic coherence; MEG, magnetoencephalography.

* Corresponding author at: Department of Psychology, New York University, 6 Washington Place, New York, NY 10003, USA.
E-mail address: david.poeppel@nyu.edu (D. Poeppel).

1053-8119/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.neuroimage.2013.06.035

[1] Note the distinction between the temporal amplitude envelope of the (full-band) stimulus, on the one hand, and the auditory critical band envelopes (i.e., at the cochlear output), on the other (Ghitza et al., 2013).

body of research showing this oscillation tracking the envelope, it remains unclear which aspects of the stimulus drive this response. One plausible hypothesis generated from the Giraud and Poeppel (2012) model suggests that it is the onsets of syllables that produce temporal fluctuations, which entrain slow neural oscillations at the syllabic rate. Here, we test this hypothesis by filtering these fluctuations in very particular ways and analyzing the effect on oscillatory entrainment. As such, the principal goal of this study is to understand more clearly the mechanisms of slow oscillation envelope tracking and, in particular, to uncover aspects in the temporal domain of the stimulus that drive this neuronal activity.

It has recently been demonstrated that theta envelope tracking of speech is enhanced by stimulus intelligibility (Peelle and Davis, 2012; Peelle et al., 2013), while earlier work showed similar neural phase-locking for sentences played backwards (no intelligibility) and forwards (Howard and Poeppel, 2010). Thus the question of whether the linguistic content of the stimuli induces a top-down 'amplification' of the oscillation-based envelope-tracking mechanism is debated. As a result, a secondary goal of this study is to investigate how envelope tracking relates to intelligibility and to understand its putative function in the broader context of speech perception.

This neurophysiological experiment builds on a recent behavioral study that manipulated the temporal acoustic features of speech to delineate the role of low frequency (syllabic) cues in speech intelligibility (Ghitza, 2012). Artificially removing exactly those temporal fluctuations in the critical band envelopes that relate to the syllabic rate (2–9 Hz) significantly reduces the intelligibility of the degraded speech. However, when brief noise bursts are added to the degraded stimulus precisely where the 'acoustic landmarks'[2] of the original *would* have been, the error rate drops by about 50%. The interpretation proposed to explain this psychophysical effect is that removing these cues disrupts the ability of cortical delta–theta oscillations to track the stimulus envelope. While removing slow fluctuations from the stimulus reduced intelligibility, reinstating temporal cues artificially by using transient edges at landmark positions enhanced intelligibility.

We hypothesize that temporal cues that reflect the syllabic rate are at the origin of the envelope-tracking phenomenon, which in turn constitutes a crucial condition for continuous speech to be intelligible. Specifically, we propose that acoustic landmarks entrain intrinsic cortical oscillations to permit the extraction of temporal primitives and subsequently finer grained speech features in a decoding stage. This quasi-periodicity generates the envelope tracking behavior, which could have the capacity to parse the stimulus into syllable-size representations.

Materials and methods

Participants

16 right-handed participants (9 females; mean age 23 years, range 18–31) took part in the experiment after providing informed consent and received compensation for their participation. Handedness was determined using the Edinburgh Handedness Inventory (Oldfield, 1971). All participants were self-reported as having normal hearing and no neurological deficits. One participant was removed because he did not input his behavioral ratings as instructed. Another was removed due to too much noise in the MEG data. Consequently, the data from a total of 14 participants were analyzed. The study was approved by the local Institutional Review Board (New York University's Committee on Activities Involving Human Subjects).

[2] By 'acoustic landmarks' we refer to vocalic landmarks, or glide landmarks, or acoustically abrupt landmarks (sometimes termed 'acoustic edges'). See Stevens (2002).

Stimuli

Twenty stimuli, spoken strings of seven digits, were chosen from a set initially used in a behavioral study (see Ghitza, 2012). These stimuli were filtered into sixteen critical bands logarithmically spaced between 230 and 3800 Hz; the Hilbert envelope of each was manipulated into one of five conditions (described below) and then combined with a noise carrier with bandwidth equal to that of the critical band before being linearly summed across critical bands (see Fig. 1A). Each stimulus was between 2 and 3 s in duration (sampling rate 11 kHz). The 100 stimuli were presented four times to each participant in pseudo-randomized order.

Envelope alterations

In the *Control* condition (Fig. 1B), each critical band envelope was low-pass filtered at 10 Hz. These stimuli are an adaptation of stimuli used by Drullman et al. (1994) and are known to be highly intelligible. The *Noθ* condition (Fig. 1C) consisted of critical band envelopes with a band-stop filter from 2 to 9 Hz. Effectively, this removes all temporal cues in the envelope that relate to the syllabic rate of the stimulus. In the *Chθ* condition (Fig. 1D), the peaks in each critical band envelope are replaced with peaks of uniform height and shape. This, essentially, distils the stimulus down to only the temporal cues relating to syllabic rate. It removes most acoustic–phonetic information and leaves only information pertaining to the peak amplitude of each syllable in each critical band. The *Noθ + Chθ* condition is the linear sum of the Noθ and Chθ conditions creating a stimulus in which the natural syllabic temporal fluctuations in the Control condition are replaced by the artificial Chθ. The *Noθ + Glbθ* is the same as the Noθ + Chθ conditions save for the Chθ peak picking operation which is done on the whole broadband envelope. This is an extreme version of the Noθ + Chθ removing all acoustic phonetic information and leaving only a noise burst at the peak amplitude of each syllable.

Task

The stimuli were delivered diotically via MEG-compatible tubephones (E-A-RTONE 3A 50 Ω, Etymotic Research) attached to E-A-RLINK foam plugs inserted into the ear canal and presented at normal conversational sound levels (~72 dB SPL).

For each trial, participants listened to one stimulus and were asked to rate the stimulus in terms of its intelligibility on a scale from 1 (poor) to 3 (good). In the original behavioral study, participants were asked to repeat the last four digits of each stimulus. This was not so viable in the MEG setup as the head movements associated with speech production can create noise as well as change the orientation of the participants' head during the experiment. Trials for each condition were randomly interleaved and each stimulus was presented four times. Mean Intertrial Interval (ITI) was 1 s with a standard deviation of .3 s. Scores for each stimulus are averaged across repetitions per subject. The aim of the psychophysics was to collect behavioral data during scanning that were compatible and comparable to the data published by Ghitza (2012).

Recording

MEG recording

Neuromagnetic signals were measured using a 157-channel whole-head axial gradiometer system (KIT, Kanazawa Institute of Technology, Japan). Five electromagnetic coils were attached to a participant's head to monitor head position during MEG recording. The locations of the coils were determined with respect to three anatomical landmarks (nasion, left and right preauricular points) on

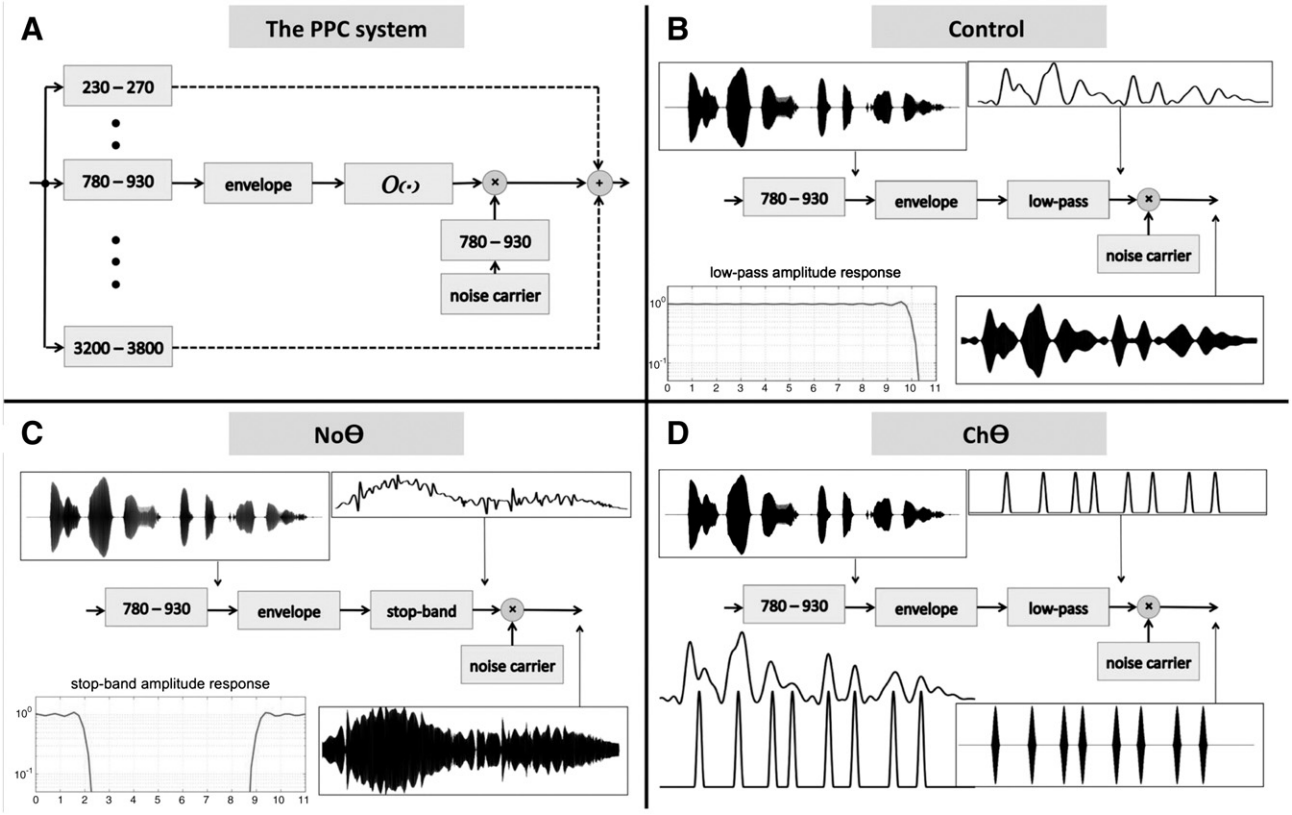

Fig. 1. Schematic of stimulus creation (figure from Ghitza, 2012). The figure shows the processing steps for each initial waveform. A. Each stimulus is filtered into 16 logarithmically-spaced critical bands from 230 to 3800 Hz, the Hilbert envelope is derived and an operator O for each condition (identified in B, C and D) is executed. Finally, the processed bands are linearly summed. B; *Control*. Operator O is a low-pass filter of the envelope at 10 Hz. C; *Noθ*. The operator is a stop-band filter from 2 to 9 Hz. D; *Chθ*. The operator is a peak picking code (PPC) in which each peak in the envelope is replaced by a peak of uniform height and shape.

the scalp using 3D digitizer software (Source Signal Imaging, Inc.) and digitizing hardware (Polhemus, Inc.). The coils were localized to the MEG sensors, at both the beginning and the end of the experiment. The MEG data were acquired with a sampling rate of 1000 Hz, filtered online between 1 Hz and 200 Hz, with a notch filter at 60 Hz.

Analysis

MEG analysis

All recorded responses were noise-reduced off-line using the CALM algorithm (Adachi et al., 2001). All further preprocessing and analysis were performed using the FieldTrip toolbox (http://fieldtrip.fcdonders.nl; Oostenveld et al., 2011) in MATLAB (version 7.10.0; MathWorks, Inc.). Trials were visually inspected and those with obvious artifacts such as channel jumps or resets were removed. An independent component analysis as implemented in FieldTrip was used to correct for eyeblink-, eye movement-, and heartbeat-related artifacts. Time–frequency information from 1 to 40 Hz (.5 Hz resolution from 1 to 10 Hz; 1 Hz resolution from 10 to 40 Hz) was extracted using a wavelet analysis in 10 ms steps. For the time–frequency analyses, the stimulus envelope was processed in the same manner. The phase difference between the stimulus envelope and the recorded data at each frequency was extracted by calculating the phase angle of the cross-spectral density between the stimulus envelope and each individual channel.

Cerebro-acoustic coherence (CACoh)

To measure the extent to which the recorded neural data tracked the stimulus, we used a measure introduced by Peelle et al., 2013. The measure finds the Phase-Locking Value between recorded data from each neural channel and the envelope of the stimulus. Specifically, we used the following equation for each frequency.

$$\mathrm{Coh}_{CA,f} = \frac{\left|\sum_t \left(e^{i\theta_{CA,t}} \sqrt{P_{C,t} \cdot P_{A,t}}\right)\right|}{\sqrt{\sum_t \left(P_{C,t} \cdot P_{A,t}\right)}}$$

where t refers to each time point in each trial, P_C and P_A refer to power at a specific frequency from recorded "cerebral" data and from the "acoustic" envelope respectively and θ_{CA} refers to the phase difference between the neural recording and the stimulus envelope.

Fig. 2 demonstrates an exemplar of the kind of data we record from participants. We use CACoh to compare the 10-Hz low-pass filtered envelope of the stimulus to the recorded neural data at each frequency. A time–frequency wavelet analysis is performed on each channel and on the stimulus envelope to extract the power spectrum and the cross-spectral density and is inputted to the CACoh equation. The correlational analyses between our CACoh values and two descriptive measures of our speech stimuli are shown below.

Channel selection

As we are interested in auditory cortical responses, channels were selected for analysis on the basis of the magnitude and signal-to-noise ratio of their recorded M100 response to a 400 ms 1000 Hz sinusoidal tone recorded in a pre-test and averaged over 200 trials. ISI between trials was randomly interleaved between 0.9, 1.1 and 1.3 s. In each quadrant of channel space (anterior left (AL); posterior left (PL); anterior right (AR); posterior right (PR)), the five channels with the largest auditory response were selectively averaged together and analyzed.

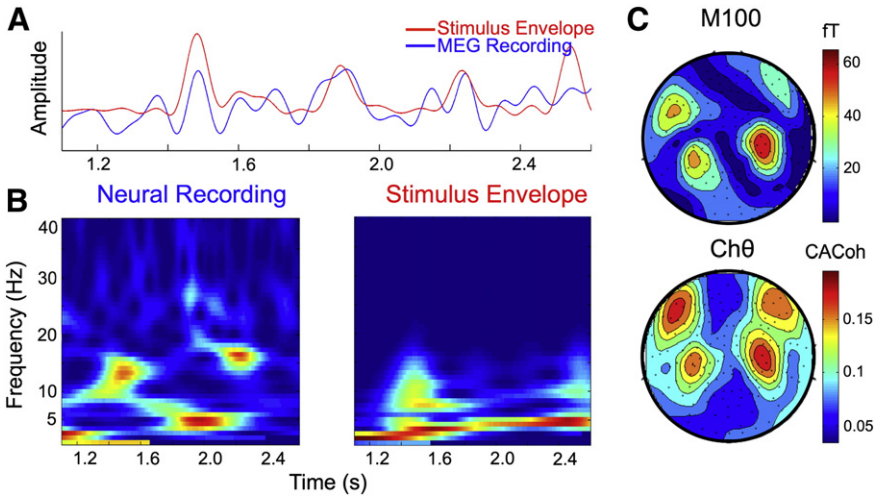

Fig. 2. A. Example neural data (*blue*) recorded of the last 1.5 s of one trial, averaged over 20 channels, independently selected as having strong auditory responses. In *red*, the envelope of the stimulus presented during the same trial. There is a noticeable overlap between the two waveforms. B. Time–frequency analysis of the example recording and stimulus envelope. The time–frequency analysis is used as input for calculation of cerebro-acoustic coherence for individual frequencies. Actual analysis performed over entire trial. C. (*Top*) Topography of averaged M100 to single sinusoidal 1 kHz tone across subjects. (*Bottom*) Topography of averaged CACoh values for condition Chθ across subjects. The similarity in topographies suggests that the channels we select based on the M100 response reflect activity from auditory areas.

Channel selection was performed for each individual participant. Fig. 2C shows the topography of the averaged M100 response across subjects compared to the average CACoh response to the Chθ condition (the condition with the largest CACoh response) from 2 to 4 Hz. The similarity of the topographies suggests the neural data we present is coming from similar brain regions as the M100 response — known to be generated in auditory cortex (Lutkenhoner and Steinstrater, 1998; Reite et al., 1994).

Sharpness and mean amplitude

The stimulus is characterized by using two variables defined at the cochlear output; both are derived from a signal, which is the linear sum (with equal weight) of all smoothed critical band envelopes. The two variables are: (1) the *sharpness* of temporal fluctuations, defined as the mean positive first derivative values of the summed envelope, and (2) the *mean amplitude* of the summed envelope. The mean amplitude also showed influence on intelligibility of the stimulus. We discuss the comparison between amplitude and sharpness in the results.

Results

Intelligibility and sharpness

Intelligibility ratings (Fig. 3) closely mirror Ghitza's (2012) psychophysical findings. We tested differences between conditions in a one-way repeated measures ANOVA and found a main effect of condition ($F = 11.6$, $p < .0001$). Using a post-hoc Tukey–Kramer multiple comparisons test, we determined that both the Noθ and Chθ conditions were significantly less intelligible than the Control condition (Noθ, $p < .0001$; Chθ, $p < .0001$). Furthermore, summing these two inputs into the Noθ + Chθ condition resulted in a significant increase in intelligibility as compared to the Noθ condition ($p < .0001$). The Noθ + Glbθ condition, however, did not show a significant increase in intelligibility as compared to the Noθ ($p = .13$). Thus, while the participant rating scale used to gauge intelligibility in this study replicated most effects found in the more detailed psychophysical study, the lack of effect between the Noθ + Glbθ condition and the Noθ condition likely reflects the reduction in power of this metric as compared to an intelligibility assessment by measuring digit recognition rate.

No correlation is observed between intelligibility and sharpness with all conditions included ($R^2 = .016$, $p = .21$). As Chθ is the only condition in which acoustic–phonetic information is actively removed, it may be the only condition in which the amount of information in the stimulus to be "decoded" is the limiting factor for the intelligibility rating (see section on The removal of Chθ). Indeed, the Chθ has the most sharpness (Fig. 3) of any condition while also being the least intelligible. When the Chθ condition is removed, the correlation between

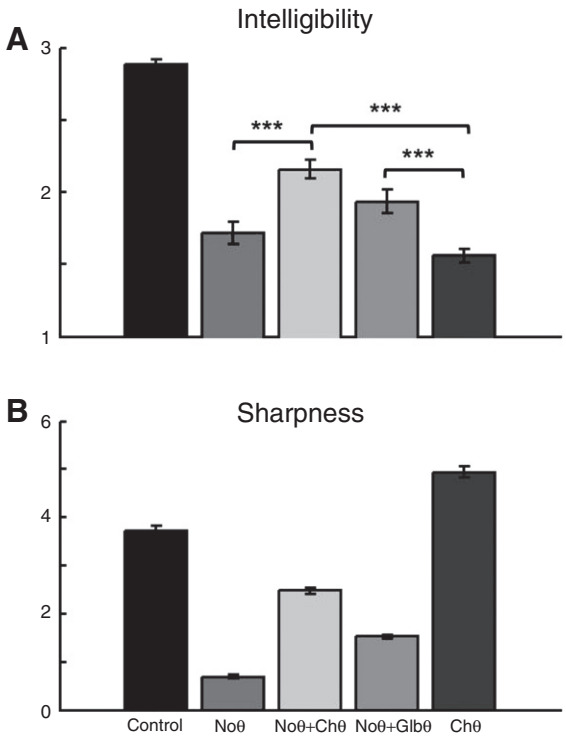

Fig. 3. Behavioral data: intelligibility and sharpness. A. Increased intelligibility ratings from Noθ and Chθ to Noθ + Chθ. No significant difference between Noθ and Noθ + Glbθ. B. Sharpness metric. All conditions are significantly different from one another.

sharpness and intelligibility is very strong ($R^2 = .7$, $p < .0001$). Similarly strong is the correlation between intelligibility and mean amplitude ($R^2 = .58$, $p < .0001$). A partial correlation between sharpness and intelligibility, controlling for the mean amplitude, still remains strong ($R^2 = .44$, $p < .0001$). Conversely, a partial correlation between mean amplitude and intelligibility, controlling for sharpness, while still significant, is dramatically reduced ($R^2 = .07$, $p < .05$). Stimulus sharpness and mean amplitude were not correlated ($R^2 = .03$, $p = .08$).

Cerebro-acoustic coherence

We hypothesized that the ability of the auditory cortices to track the critical band envelopes of the stimulus should (i) be largely driven by the sharpness of fluctuations in the stimulus, and (ii) constitute an important factor towards comprehension. To measure envelope tracking, we use cerebro-acoustic coherence (CACoh) as an index, which measures phase locking between the neural signal and the summed critical band envelopes. We predict that this measure should correlate both with sharpness and with intelligibility.

We grouped all conditions into one set and looked for correlations between CACoh and either the sharpness or the intelligibility rating. Fig. 4 shows the results. We found a significant positive correlation of CACoh and sharpness in the range of 2–4 Hz ($p < .01$, Bonferroni corrected) and a negative correlation between CACoh and both sharpness and intelligibility at 9.5–12 Hz ($p < .01$, Bonferroni corrected). No positive correlation existed between CACoh and intelligibility.

We then removed the Chθ (as we did in correlating sharpness with intelligibility) from the stimulus conditions and found that a positive correlation with intelligibility emerged in the same 2–4 Hz range as with sharpness ($p < .01$, Bonferroni corrected; Fig. 4, bottom row). Interestingly, there was no correlation of response power – rather than CACoh – at any frequency with sharpness or intelligibility.

To investigate this effect of envelope tracking at the syllabic rate further, we then separated the conditions and compared differences between conditions in the 2–4 Hz range for z-scored CACoh. Fig. 5A shows a significant increase in CACoh from Noθ to both Noθ + Chθ ($p < .05$) and Noθ + Glbθ ($p < .05$) with all selected channels included. Splitting these channels by region (Fig. 5B) shows that the anterior channels (both left and right) show a significant increase in envelope tracking from Noθ to Noθ + Chθ (AL, $p < .05$; AR, $p < .001$). Differences in posterior regions were not significant.

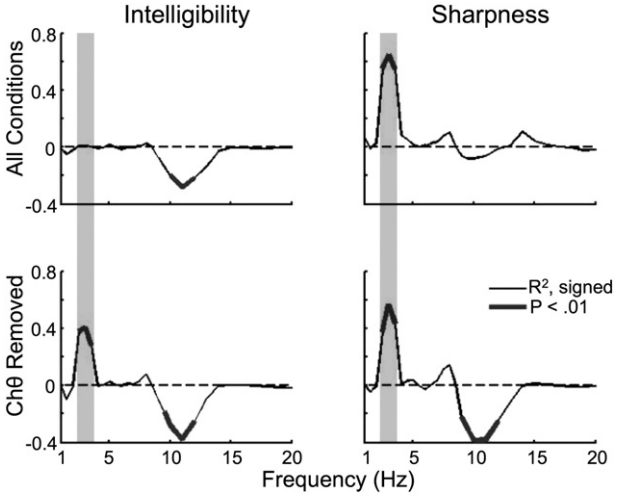

Fig. 4. Cerebro-acoustic coherence. CACoh at the syllabic rate of the materials used here (2.5–4 Hz) correlates with both sharpness (right panels) and intelligibility (left). *Right panels*. Signed R^2 values for correlation between CACoh and sharpness. Thick line shows Bonferroni corrected at $p < .01$. Shaded region shows frequency range of interest. Correlation is robust with inclusion of all stimuli. *Left panels*. Correlation between CACoh and intelligibility ratings. Legend same as right. Correlation at syllabic rate only with inclusion of stimuli containing acoustic–phonetic information.

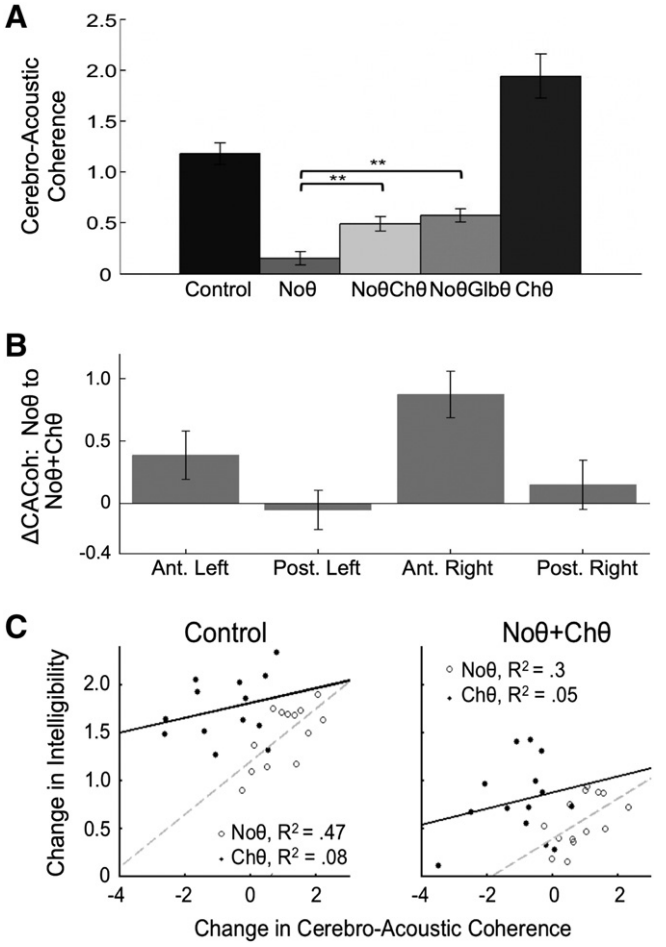

Fig. 5. Relation between behavioral scores and CACoh. A. CACoh averaged across 4 auditory regions. Significant increase from Noθ to both Noθ + Chθ and Noθ + Glbθ. B. Difference in CACoh between Noθ and Noθ + Chθ. Anterior regions show significance. C. Change in CACoh in anterior right channels between Noθ (gray, dotted line) and most intelligible conditions (Control, left, and Noθ + Chθ, right) correlates with change in intelligibility. No correlation with Chθ (dark, solid line). One outlier was removed in the Control panel as ΔCACoh was more than 2 standard deviations away from the mean of the group.

To evaluate whether this increase in envelope tracking is related to the increase in intelligibility between Noθ and Noθ + Chθ, we look at the set of channels with the greatest difference between the two conditions (anterior right) and correlated the change in intelligibility with the change in envelope tracking between conditions using each subject as a point in the correlation. The results show (in Fig. 5C) that when comparing the Noθ condition to either Control ($R^2 = .47$, $p < .01$; one outlier removed) or to Noθ + Chθ ($R^2 = .30$, $p < .05$) the correlation is positive and significant. However, when comparing either of those conditions to the Chθ condition there is no significant correlation (Control: $R^2 = .08$, $p = .36$; Noθ + Chθ: $R^2 = .05$, $p = .43$). In the other regions, there is similarly no significant correlation between the Chθ condition and the more intelligible conditions. The only other significant correlation is between Noθ and Control in AL ($R^2 = .44$, $p < .05$). Comparing these low intelligibility conditions to Noθ + Glbθ does not show this same effect. This is likely to be due to the lack of significant difference in intelligibility ratings between Noθ and Noθ + Glbθ.

Discussion

This study demonstrates a clear relationship between envelope tracking in the auditory cortex and intelligibility of a speech signal.

While this relationship has been suggested previously (Luo and Poeppel, 2007; Peelle et al., 2013), this particular method enables us to shed light on the nature of this relationship. Specifically, we suggest that reliable envelope tracking requires a sufficient degree of temporal envelope fluctuations at the cochlear output. These fluctuations are driven by acoustic landmarks in the broadband stimulus. Such a mechanism could enable comprehension of continuous speech by parsing the input into interpretable – roughly syllable-sized – chunks or temporal primitives, thereby organizing the decoding process.

Temporal fluctuations in speech

Our stimuli are designed to control the amount of temporal envelope fluctuation *at the syllabic rate* while maintaining as much of the acoustic–phonetic information as possible. In order to quantify the degree of fluctuation we used the measure of envelope sharpness, which represents rapid changes in the summed critical band envelopes. We chose this metric in particular as it follows from a prediction – generated by both the Giraud and Poeppel (2012) and Tempo (Ghitza, 2011) models, lately summarized in Ghitza et al. (2013) – that temporal fluctuations in speech entrain delta–theta oscillations. The results show that sharpness correlates with intelligibility for stimuli that maintain their acoustic–phonetic information.

As a control, we tested the mean amplitude of the summed critical band envelopes. This metric also shows a strong correlation with intelligibility. We deem this to be trivially true; that the amount of information in the stimulus relates to intelligibility is not surprising. Critically, when mean amplitude is controlled, the relationship between sharpness and intelligibility remains strong. This suggests that the sharpness of the temporal fluctuations provides its own information to the listener.

It is important to note that the stimuli used in this study and in the behavioral study (Ghitza, 2012) are of low perplexity. They are not sentences with rich content but rather are strings of seven digits — spoken as phone numbers. Because the set of possible numbers is very limited (11 possible items: numbers 0 to 9 plus the letter O), these stimuli are highly predictive as compared to regular sentences. This qualitative difference from other similar studies may affect the behavioral responses, particularly in the more distorted conditions. However, we find it unlikely that the degree of perplexity of our material would have an effect on the physiological response we show here. The use of strings of digits rather than sentences, however, is likely to be the reason behind the syllabic rate (3 Hz) – and thus the CACoh effect – being at a lower frequency than considered typical (~5 Hz).

Slow oscillations track sharp fluctuations, enhancing intelligibility

Our data provide evidence that oscillation-based envelope tracking in early auditory cortex, at the syllabic rate, mediates a relationship between sharpness and intelligibility. We show a strong correlation between the sharpness of the temporal fluctuations and envelope tracking with all conditions included. This effect could be a general mechanism used by the auditory system for all stimuli. This would explain how Chθ – the least intelligible condition – evokes the strongest envelope tracking. In essence, we propose that large, rapid changes in the amplitude envelope – which are driven by acoustic landmarks or 'edges' in the full-band signal – entrain the phase of intrinsic neural oscillations and allow for oscillatory envelope tracking of regularities in the stimulus.

We have further shown that this tracking behavior correlates with enhanced intelligibility. A robust correlation exists between the intelligibility ratings given by participants and the amount of envelope tracking detected in their neural signal, at the syllabic rate of the stimuli. The question remains whether this type of neural activity is a necessary operation mediating the relationship between sharpness and intelligibility or if it is merely a byproduct of the effect with sharpness. The data in Fig. 5C partly address this issue. They show that individual differences in the increase in envelope tracking from Noθ to Noθ + Chθ correlate with the increase of intelligibility ratings across the same conditions (i.e., when the difference in sharpness is held constant across participants). This suggests that envelope tracking relates to intelligibility directly and is not merely a byproduct of fluctuations in the critical band envelope. Thus, the correlation between sharpness and intelligibility is mediated by envelope tracking (and modulated by individual differences in envelope-tracking ability). The models of Giraud and Poeppel (2012) and Ghitza (2011) explain this effect in terms of parsing. With sharper, better-defined fluctuations, the stimulus is easier to track and thus the auditory cortex is better able to parse the stimulus into relevant chunks for decoding.

It is important to note that parsing via envelope tracking is presumably mostly critical in the case of continuous (or everyday) speech. We do not propose that envelope tracking is a necessary prerequisite to initiate decoding in the case of single words or syllables, for example. In those cases, the stimulus is usually short enough that segmentation at the delta–theta band timescale is provided by the nature of the stimulus (see Ghitza, 2013).

Related recent work in developmental psychology is consistent with the view outlined here. A series of studies by Goswami and colleagues (e.g. Goswami et al., 2002; Richardson et al., 2004; Thomson and Goswami, 2008; Thomson et al., 2009) on developmental dyslexia show that deficits in the perception of the 'rise-time' (or edges) in the full-band amplitude envelope could provide an important contribution to the etiology of this disorder. They have shown that sensitivity to rise-time is a predictor of phonological awareness and reading acquisition across a number of languages (Goswami et al., 2011). In a theoretical framework (the 'temporal sampling framework', TSF; Goswami, 2011), they propose that impaired phase locking of slow neural oscillations in the auditory cortex to the amplitude envelope of the stimulus in part causes the deficit by preventing the establishment of robust phonological representations. Abnormal processing of the speech envelope in children has been associated with poor reading (Abrams et al., 2009) and reduced phase locking to low-frequency amplitude modulated noise has been shown in dyslexic adults as compared to controls (Hamalainen et al., 2012). The data we show here support the connection between envelope rise-times and low frequency phase locking, with an important caveat. While the TSF suggests that impaired phase locking results in poor sensitivity to auditory edges, our data implicate the reverse relationship. Specifically, poorly defined edges, (or in the case of dyslexia, an individual's insensitivity to well-defined edges) result in impaired phase locking and thus lead to impaired perception.

The removal of Chθ

One important requirement for a meaningful correlation between sharpness and intelligibility is that the Chθ stimulus condition must be excluded. We suggest that for intelligibility scores to be relevant to the parsability of the stimuli, every stimulus must contain acoustic–phonetic information to be decoded. In the Chθ condition, in each critical band, the phonetic content is replaced with information about the timing of syllables. While some acoustic–phonetic information is regained when critical bands are integrated, the vast majority of the content is destroyed. Thus, the stimulus is rated the least intelligible not because of how easily it can be parsed; rather it is because after it is parsed, there is not enough information to be decoded. As such, the Chθ condition must be separated from the others, as its intelligibility rating is largely the result of a different limiting factor. This condition provides an important control measure of envelope tracking behavior in the absence of phonetic content, supporting the concept that slow syllabic cues provide *a temporal framework* for

phonetic decoding. We suggest that these cues – and the envelope tracking mechanism – are necessary for a reliable extraction of phonetic content in the case of continuous speech but are not sufficient for comprehension.

Chθ and Glbθ

One point that complicates our interpretation of these results is the Noθ + Glbθ condition. Specifically, though both CACoh levels and intelligibility in this condition are not significantly different from those of Noθ + Chθ, no significant increase exists in intelligibility from Noθ to Noθ + Glbθ. Furthermore, there is no correlation of differences between Noθ and Noθ + Glbθ as there is for Noθ + Chθ (see Fig. 5C).

If our results regarding Noθ + Glbθ are on the right track, it is possible that they reflect the important role critical bands play in the neuronal mechanism of delta–theta tracking. Ghitza (2012) hypothesized that the difference between the time–frequency representations of the input rhythm at the cochlear output (Chθ) and the input rhythm at the waveform level (Glbθ) – see his Fig. 3 – could potentially be exploited by the envelope-tracking mechanism. While it is impossible to detect using MEG, it may well be that the theta-envelope tracking effect operates on the critical band level. Consistent with this hypothesis, a neurophysiological study shows – using much finer-grained methods – that in the auditory cortex, delta–theta band activity is most phase coherent between regions of core and belt areas with similar best frequencies (Farley and Norena, 2013). This finding supports the notion of delta–theta tracking occurring at each critical band rather than over the whole stimulus.

If this is the case, it may explain why Noθ and Noθ + Glbθ exhibit differences in CACoh and sharpness levels, yet intelligibility differences are reduced. The addition of Glbθ enhances CACoh by providing sharp temporal fluctuations. However, because the input rhythm is not represented with the spectro-temporal richness of critical bands, the hypothesized envelope tracking mechanism is unable to parse as accurately as in the Noθ + Chθ condition — hence lower intelligibility. More data will be required to speak to the validity of this argument.

Frequency matching and temporal predictions

The frequency range of the neural effects we observe is between 2 and 4 Hz. We note that this range matches the average syllabic rate and modal frequency of the temporal envelope of the stimuli (~3 Hz). Such correspondence has been observed with other speech corpora, e.g., sentences in Ahissar et al. (2001). Thus, it seems that this neural oscillation tracks the stimulus at the syllabic rate because sharp fluctuations of critical band envelopes tend to occur at a rate corresponding to the intrinsic range of these oscillations.

These landmarks, then, are fortuitously placed, as they have the capacity not only to alert auditory cortices to the placement of the current syllable but also to align the oscillation cycle such that relevant neuronal populations will be at a high excitability phase at the onset of the *next* syllable. This concept makes contact with the more basic perspective of active sensing in purely rhythmic stimuli (e.g., Lakatos et al., 2008, 2013; Schroeder and Lakatos, 2009). In that context, sub-threshold oscillations in the delta range model rhythmic stimuli during attention. This stimulus-tracking modulates neural processing but is unlikely to be a source of stimulus content information. It is possible that the effect that we find is a reflection of this basic mechanism, integrated with the processing of more complex stimuli. If so, then there is potential for top-down control of this mechanism, based purely on when the relevant information might occur as there is in the case of perfectly rhythmic stimuli.

Some previous work has implicated a potential role of linguistic content as modulating top-down control of phase entrainment to speech and low-level speech processing (e.g., Hannemann et al., 2007; Obleser et al., 2007; Peelle, 2013; Peelle et al., 2013). These data suggest an interesting and puzzling interaction between stimulus content and the timing of occurrence. Our findings cannot support or refute these claims. However, it may well be that knowledge of a stimuli's content includes with it knowledge of its temporal nature and thus modulates the mechanism that we describe here.

We show increased correlation between individual differences in envelope tracking and intelligibility in anterior channels: an observation that is compatible with a possible top-down amplification of slow oscillation entrainment in the auditory cortices (cf. Arnal and Giraud, 2012; Besle et al., 2011; Schroeder et al., 2010). However, while entrainment was sustained, it did not increase across time, which may temper this interpretation (data not shown).

The effect in alpha range

Unexpectedly, we found a negative correlation between CACoh and intelligibility in the alpha range (9.5–12 Hz). It is likely that this is an effect of the critical band envelope filters. After investigating the modulation spectra of the conditions we found that the Control condition (low pass filtered at 10 Hz) may be driving this effect. A negative correlation exists between the modulation spectrum and intelligibility from 10 to 12.5 Hz only when Control condition is included (all conditions: signed $R^2 = -.37$, $p < .05$ Bonf. corrected; Control condition removed: $R^2 = -.0623$, $p = .51$, Bonf. corrected). Thus, it is likely that frequency power of the stimulus (used to normalize CACoh) is creating this effect.

An alternative hypothesis is generated by a similar result which has been found relating alpha power suppression to intelligibility (Obleser and Weisz, 2012). As the calculation of CACoh as defined here is normalized by power values, the result we find could be a reflection of this previously shown effect relating power (rather than phase-locking) to comprehension. However, a direct power analysis was unable to resolve this effect in our data. A brief power analysis has shown no correlation between power in the 2–4 Hz range and the intelligibility ratings showing that this cannot explain our effect in slow oscillations.

Conclusion

Our data paint an interesting picture of the role of neural envelope tracking in perceptual analysis of auditory signals, and ultimately in speech comprehension. Our interpretation of the data speaks first and foremost to the mechanism by which envelope-tracking activity is generated in auditory cortices. Namely, sharp fluctuations in critical band envelopes, driven by acoustic landmarks (e.g., edges), entrain the slow oscillations of auditory cortex, forcing the oscillation to track stimulus features (e.g., syllabic onsets) that occur at about its intrinsic rate. While envelope-tracking activity on its own is not sufficient for comprehension of continuous speech, it clearly seems to be necessary. The interaction between the sharpness of the stimulus and the intrinsic oscillations at this particular frequency promotes envelope tracking as a strong candidate to subserve the function of syllabic parsing, making it a crucial step towards reliably decoding and understanding naturally spoken language.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.neuroimage.2013.06.035.

Acknowledgments

This work is supported by NIH R01 DC05660 to D.P. We thank Jeff Walker for his expert technical support, and Benjamin Morillon for help with the analysis. Oded Ghitza is funded by a research grant from the United States Air Force Office of Scientific Research.

Conflict of Interest

The authors have no conflicts of interest.

References

Abrams, D.A., Nicol, T., Zecker, S., Kraus, N., 2009. Abnormal cortical processing of the syllable rate of speech in poor readers. J. Neurosci. 29, 7686–7693.

Adachi, Y., Shimogawara, M., Higuchi, M., Haruta, Y., Ochiai, M., 2001. Reduction of nonperiodic environmental magnetic noise in MEG measurement by continuously adjusted least square method. IEEE Trans. Appl. Supercond. 11, 669–672.

Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., 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.

Arnal, L.H., Giraud, A.L., 2012. Cortical oscillations and sensory predictions. Trends Cogn. Sci. 16, 390–398.

Besle, J., Schevon, C.A., Mehta, A.D., Lakatos, P., Goodman, R.R., McKhann, G.M., Emerson, R.G., Schroeder, C.E., 2011. Tuning of the human neocortex to the temporal dynamics of attended events. J. Neurosci. 31, 3176–3185.

Cogan, G.B., Poeppel, D., 2011. A mutual information analysis of neural coding of speech by low-frequency MEG phase information. J. Neurophysiol. 106, 554–563.

Ding, N., Simon, J.Z., 2009. Neural representations of complex temporal modulations in the human auditory cortex. J. Neurophysiol. 102, 2731–2743.

Drullman, R., Festen, J.M., Plomp, R., 1994. Effect of reducing slow temporal modulations on speech reception. J. Acoust. Soc. Am. 95, 2670–2680.

Farley, B.J., Norena, A.J., 2013. Spatiotemporal coordination of slow-wave ongoing activity across auditory cortical areas. J. Neurosci. 33, 3299–3310.

Ghitza, O., 2011. Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input rhythm. Front. Psychol. 2, 130.

Ghitza, O., 2012. On the role of theta-driven syllabic parsing in decoding speech: intelligibility of speech with a manipulated modulation spectrum. Front. Psychol. 3, 238.

Ghitza, O., 2013. The theta-syllable: a unit of speech information defined by cortical function. Front. Psychol. 4, 138.

Ghitza, O., Giraud, A.L., Poeppel, D., 2013. Neuronal oscillations and speech perception: critical-band temporal envelopes are the essence. Front. Hum. Neurosci. 6, 340.

Giraud, A.L., Poeppel, D., 2012. Cortical oscillations and speech processing: emerging computational principles and operations. Nat. Neurosci. 15, 511–517.

Goswami, U., 2011. A temporal sampling framework for developmental dyslexia. Trends Cogn. Sci. 15, 3–10.

Goswami, U., Thomson, J., Richardson, U., Stainthorp, R., Hughes, D., Rosen, S., Scott, S.K., 2002. Amplitude envelope onsets and developmental dyslexia: a new hypothesis. Proc. Natl. Acad. Sci. U. S. A. 99, 10911–10916.

Goswami, U., Wang, H.L., Cruz, A., Fosker, T., Mead, N., Huss, M., 2011. Language-universal sensory deficits in developmental dyslexia: English, Spanish, and Chinese. J. Cogn. Neurosci. 23, 325–337.

Hamalainen, J.A., Rupp, A., Soltesz, F., Szucs, D., Goswami, U., 2012. Reduced phase locking to slow amplitude modulation in adults with dyslexia: an MEG study. NeuroImage 59, 2952–2961.

Hannemann, R., Obleser, J., Eulitz, C., 2007. Top-down knowledge supports the retrieval of lexical information from degraded speech. Brain Res. 1153, 134–143.

Howard, M.F., Poeppel, D., 2010. Discrimination of speech stimuli based on neuronal response phase patterns depends on acoustics but not comprehension. J. Neurophysiol. 104, 2500–2511.

Howard, M.F., Poeppel, D., 2012. The neuromagnetic response to spoken sentences: co-modulation of theta band amplitude and phase. NeuroImage 60, 2118–2127.

Kerlin, J.R., Shahin, A.J., Miller, L.M., 2010. Attentional gain control of ongoing cortical speech representations in a "cocktail party". J. Neurosci. 30, 620–628.

Lakatos, P., Karmos, G., Mehta, A.D., Ulbert, I., Schroeder, C.E., 2008. Entrainment of neuronal oscillations as a mechanism of attentional selection. Science 320, 110–113.

Lakatos, P., Musacchia, G., O'Connel, M.N., Falchier, A.Y., Javitt, D.C., Schroeder, C.E., 2013. The spectrotemporal filter mechanism of auditory selective attention. Neuron 77, 750–761.

Luo, H., Poeppel, D., 2007. Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. Neuron 54, 1001–1010.

Luo, H., Poeppel, D., 2012. Cortical oscillations in auditory perception and speech: evidence for two temporal windows in human auditory cortex. Front. Psychol. 3, 170.

Luo, H., Liu, Z., Poeppel, D., 2010. Auditory cortex tracks both auditory and visual stimulus dynamics using low-frequency neuronal phase modulation. PLoS Biol. 8, e1000445.

Lutkenhoner, B., Steinstrater, O., 1998. High-precision neuromagnetic study of the functional organization of the human auditory cortex. Audiol. Neurootol. 3, 191–213.

Obleser, J., Weisz, N., 2012. Suppressed alpha oscillations predict intelligibility of speech and its acoustic details. Cereb. Cortex 22, 2466–2477.

Obleser, J., Wise, R.J., Alex Dresner, M., Scott, S.K., 2007. Functional integration across brain regions improves speech perception under adverse listening conditions. J. Neurosci. 27, 2283–2289.

Oldfield, R.C., 1971. The assessment and analysis of handedness: the Edinburgh inventory. Neuropsychologia 9, 97–113.

Oostenveld, R., Fries, P., Maris, E., Schoffelen, J.M., 2011. FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Comput. Intell. Neurosci. 2011, 156869.

Peelle, J.E., 2013. Cortical responses to degraded speech are modulated by linguistic predictions. J. Acoust. Soc. Am. 133, 3387.

Peelle, J.E., Davis, M.H., 2012. Neural oscillations carry speech rhythm through to comprehension. Front. Psychol. 3, 320.

Peelle, J.E., Gross, J., Davis, M.H., 2013. Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. Cereb. Cortex 23, 1378–1387.

Poeppel, D., 2003. The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time'. Speech Commun. 41, 245–255.

Reite, M., Adams, M., Simon, J., Teale, P., Sheeder, J., Richardson, D., Grabbe, R., 1994. Auditory M100 component 1: relationship to Heschl's gyri. Brain Res. Cogn. Brain Res. 2, 13–20.

Richardson, U., Thomson, J.M., Scott, S.K., Goswami, U., 2004. Auditory processing skills and phonological representation in dyslexic children. Dyslexia 10, 215–233.

Rosen, S., 1992. Temporal information in speech: acoustic, auditory and linguistic aspects. Philos. Trans. R. Soc. Lond. B Biol. Sci. 336, 367–373.

Schroeder, C.E., Lakatos, P., 2009. Low-frequency neuronal oscillations as instruments of sensory selection. Trends Neurosci. 32, 9–18.

Schroeder, C.E., Wilson, D.A., Radman, T., Scharfman, H., Lakatos, P., 2010. Dynamics of active sensing and perceptual selection. Curr. Opin. Neurobiol. 20, 172–176.

Stevens, K.N., 2002. Toward a model for lexical access based on acoustic landmarks and distinctive features. J Acoust Soc Am. 111 (4), 1872–1891.

Thomson, J.M., Goswami, U., 2008. Rhythmic processing in children with developmental dyslexia: auditory and motor rhythms link to reading and spelling. J. Physiol. Paris 102, 120–129.

Thomson, J.M., Goswami, U., Baldeweg, T., 2009. The ERP signature of sound rise time changes. Brain Res. 1254, 74–83.

Zion Golumbic, E., Cogan, G.B., Schroeder, C.E., Poeppel, D., 2013. Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party". J. Neurosci. 33, 1417–1426.