J Neurophysiol 116: 2497–2512, 2016.
First published September 7, 2016; doi:10.1152/jn.00074.2016.

High-frequency neural activity predicts word parsing in ambiguous speech streams

Anne Kösem,[1,2,3] Anahita Basirat,[1,4] Leila Azizi,[1] and Virginie van Wassenhove[1]

[1]*Cognitive Neuroimaging Unit, CEA DRF/I2BM, Institut National de la Santé et de la Recherche Médicale, Université Paris-Sud, Université Paris-Saclay, Gif/Yvette, France;* [2]*Radboud University, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, The Netherlands;* [3]*Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands; and* [4]*SCALab, Centre National de la Recherche Scientifique UMR 9193, Université Lille, Lille, France*

Submitted 26 January 2016; accepted in final form 3 September 2016

Kösem A, Basirat A, Azizi L, van Wassenhove V. High-frequency neural activity predicts word parsing in ambiguous speech streams. *J Neurophysiol* 116: 2497–2512, 2016. First published September 7, 2016; doi:10.1152/jn.00074.2016.—During speech listening, the brain parses a continuous acoustic stream of information into computational units (e.g., syllables or words) necessary for speech comprehension. Recent neuroscientific hypotheses have proposed that neural oscillations contribute to speech parsing, but whether they do so on the basis of acoustic cues (bottom-up acoustic parsing) or as a function of available linguistic representations (top-down linguistic parsing) is unknown. In this magnetoencephalography study, we contrasted acoustic and linguistic parsing using bistable speech sequences. While listening to the speech sequences, participants were asked to maintain one of the two possible speech percepts through volitional control. We predicted that the tracking of speech dynamics by neural oscillations would not only follow the acoustic properties but also shift in time according to the participant's conscious speech percept. Our results show that the latency of high-frequency activity (specifically, beta and gamma bands) varied as a function of the perceptual report. In contrast, the phase of low-frequency oscillations was not strongly affected by top-down control. Whereas changes in low-frequency neural oscillations were compatible with the encoding of prelexical segmentation cues, high-frequency activity specifically informed on an individual's conscious speech percept.

speech segmentation; neural entrainment; bistability; MEG; phase

NEW & NOTEWORTHY

A critical problem the brain faces when analyzing speech is how to parse a continuous stream of information into relevant linguistic units. With the use of bistable speech streams that could be perceived as two distinct word sequences repeated over time, our results show that high-frequency activity reflects the word sequence participants perceived. Our study suggests that high-frequency activity reflects the conscious representation of speech after segmentation.

LISTENING TO SPEECH REQUIRES that essential linguistic units (phonemes, syllables, words) are computed online while hearing a continuous stream of acoustic information (Poeppel et al. 2008). This segmentation problem has been discussed in recent theoretical and neurocomputational models of speech processing, which describe brain oscillations as active parsers of the auditory speech signals (Ding and Simon 2014; Ghitza 2011; Giraud and Poeppel 2012; Hyafil et al. 2015a; Peelle and Davis 2012; Poeppel 2003; Poeppel et al. 2008). In particular, two main oscillatory regimes are deemed fundamental for the encoding of speech. First, low-frequency neural oscillations in the delta to theta range (2–8 Hz) have been shown to follow natural speech rhythms, enabling the tracking of the temporal structure of acoustic speech features such as syllables and words (Ahissar et al. 2001; Ding and Simon 2013; Doelling et al. 2014; Gross et al. 2013; Luo and Poeppel 2007, 2012; Millman et al. 2013; Peelle and Davis 2012; Rimmele et al. 2015; Zion Golumbic et al. 2013). Second, high-frequency neural activities, including the beta (20–30 Hz) and gamma bands (>40 Hz), have been hypothesized to encode the fine-grained properties of the speech signal such as phonetic features (Ghitza 2011; Giraud and Poeppel 2012; Poeppel 2003; Poeppel et al. 2008).

In this context, an important question is whether the entrainment of low-frequency neural oscillations (LFO) by speech is sufficient to define the segmentation boundaries of perceived syllables and words. LFO could first impact speech parsing by tracking the salient acoustic cues in speech (Doelling et al. 2014; Ghitza 2011; Giraud and Poeppel 2012; Hyafil et al. 2015a) and thus primarily reflect stimulus-driven neural entrainment, which is known to modulate the perception of sounds in a periodical fashion (Henry and Obleser 2012; Ng et al. 2012). Under this hypothesis, the phase of LFO could be reset by the sharp temporal fluctuations in the speech envelope (Doelling et al. 2014; Giraud and Poeppel 2012). LFO could primarily be modulated by the acoustics of the speech signal so that a particular phase of the LFO would be associated with the acoustic edges demarcating the boundaries between speech units. We will refer to this mechanism as "acoustic parsing," a bottom-up mechanism driven by the analysis of the acoustic signal.

However, acoustic parsing is insufficient for the extraction of linguistic tokens, considering that in continuous speech, words and syllables are not always delimited by sharp acoustic edges (Maddieson 1984; Stevens 2002). In particular, if neural oscillations passively track the fluctuations of the speech envelope, phase reset would be predicted to occur at the onset of vowels, which are the features that carry the most important energy fluctuations (Stevens 2002). This would be problematic for speech segmentation, considering that a majority of words and syllables start with consonants (Maddieson 1984). Parsing mechanisms may thus require top-down processing informed

Address for reprint requests and other correspondence: A. Kösem, Donders Institute for Cognitive Neuroimaging, Kapittelweg 29 6525 EN Nijmegen, The Netherlands (e-mail: a.kosem@donders.ru.nl).

0022-3077/16

by the representational availability of syllables or words in a given language (Mattys et al. 2005); we will refer to this hypothesized parsing mechanism as “linguistic parsing.” LFO are known to be under top-down attentional control: both attention and stimulus expectation can modulate the phase of entrained neural oscillations bringing periods of high neural excitability in phase with stimulus presentation, thereby facilitating the detection of the attended sensory inputs (Besle et al. 2011; Cravo et al. 2013; Gomez-Ramirez et al. 2011; Lakatos et al. 2008; Schroeder and Lakatos 2009a; Stefanics et al. 2010). In complex auditory environments, the control of neural oscillations by attention has been shown to be beneficial for speech processing, as well (Rimmele et al. 2015; Zion Golumbic et al. 2013), suggesting that when speech perception is under attentional or volitional control, LFO may correlate with the outcome of word comprehension.

Additionally, recent evidence suggests that LFO play a role in the parsing of linguistic content (Ding et al. 2016). Delta oscillations were shown to delineate the perceived linguistic structure (phrases and sentences) within continuous speech, suggesting that LFO may be actively relevant for the parsing of smaller linguistic units such as words or syllables, which is the focus of this study. So far, the strength of LFO entrainment has been reported to systematically correlate with speech intelligibility (Ahissar et al. 2001; Ding and Simon 2013; Doelling et al. 2014; Gross et al. 2013; Peelle et al. 2013; Rimmele et al. 2015), implying that LFO may be relevant for word segmentation. However, and importantly, speech intelligibility was also confounded with changes in the acoustic properties of the speech signal, leaving open the possibility that the observed modulations of LFO were driven by acoustic cues. In fact, in a different series of experiments controlling for acoustic properties, no direct link between speech intelligibility and neural entrainment of LFO was found (Millman et al. 2015; Peña and Melloni 2012; Zoefel and VanRullen 2015). All in all, these results suggest that LFO may govern attention and temporal expectation mechanisms that regulate the gain of the acoustic information but may not reflect top-down syllable/word segmentation per se.

Crucially, speech models posit that the entrainment of LFO by syllabic and phrasal speech rates are associated with a modulation of high-frequency activity (HFA) by LFO (Ding and Simon 2014; Giraud and Poeppel 2012; Hyafil et al. 2015a). LFO are known to orchestrate periods of inhibition and excitation for HFA, notably in the beta and gamma bands. This is achieved via cross-frequency coupling, meaning that an increase in HFA power occurs at particular phases of LFO (Akam and Kullmann 2014; Canolty et al. 2006; Canolty and Knight 2010; Hyafil et al. 2015b; Lakatos et al. 2005). During speech listening, HFA has been predicted to be enhanced as syllables and words unfold over time but inhibited at their boundaries (Ding and Simon 2014; Giraud and Poeppel 2012; Hyafil et al. 2015a). Speech models thus predict that the inhibition of HFA would also mark the onsets and offsets of the parsing windows used to segment the acoustic signals into speech units.

In this experiment, we were interested in understanding whether neural oscillatory dynamics predicted linguistic parsing when the acoustic properties of speech were maintained identically over time yet yielded different conscious percepts, a phenomenon known as “verbal transformations” (Basirat et al. 2012; Billig et al. 2013; Sato et al. 2006, 2007; Warren 1968) (Fig. 1*A*). For instance, hearing the word fly steadily repeated over time “. . . . flyflyflyflyflyflyflyflyfly . . .” will typically result in perceiving alternatively “life” or “fly.” Four different speech sequences were used and could be perceived as two distinct French words: “lampe” ([lãp]) or “plan” ([plã]), and “képi” ([kepi]) or “piquer” ([pike]); or pseudo-words: “pse” ([psə]) or “sep” ([səp]), and “tapa” ([tapa]) or “pata” ([pata]). Participants were asked to maintain one or the other percept during the presentation of a given speech sequence (Fig. 1*A*). Because the acoustics of the speech signal were constant over time, the changes in word percept could only be attributed to linguistic parsing. We predicted that if LFO actively participated in linguistic parsing in a manner consistent with participants’ conscious perception, LFO should track the speech signal at distinct latencies for each competing percept when under volitional control (Fig. 1*B*). Alternatively, if LFO tracked the acoustic cues irrespective of conscious speech perception, no changes should be seen when contrasting two perceptual reports given the same acoustic presentation. Additionally, in the context of the introduced speech models (Ding and Simon 2014; Giraud and Poeppel 2012; Hyafil et al. 2015a), we expected that HFA power should also follow speech dynamics and that the tracking should similarly shift in time according to the boundaries of the perceived syllables and words (Fig. 1*B*). Our results show small latency modulations of the LFO phase response but strong latency modulations of the power of HFA that are consistent with linguistic parsing. The functional dissociation between these two neural markers is discussed in detail.

MATERIALS AND METHODS

Participants

Twenty participants (8 women, mean age 23 yr) took part in the study. All were right-handed native French speakers with normal hearing. All participants were naive as to the purpose of the study. Before taking part in the study, each participant provided a written informed consent in accordance with the Declaration of Helsinki (2008) and the Ethics Committee on Human Research at NeuroSpin (Gif-sur-Yvette, France). Two participants were rejected because of noisy magnetoencephalography (MEG) recordings (rejection after visual inspection of MEG raw data, before MEG analysis), one participant did not finish the task, and two participants did not correctly perform the “volitional” verbal transformation task because they could not voluntarily hear the required percept (<10% report) in at least one of the sequences. Hence, 15 participants (5 women, mean age 23 yr) were considered for the reported analysis.

Experimental Paradigm

Stimuli. Four auditory sequences (adapted from Basirat et al. 2012 and Sato et al. 2007) were presented binaurally to participants via Etymotic earphones (Etymotic Research, Elk Grove Village, IL) at a comfortable hearing level. One sequence consisted of the repetition of the monosyllabic French word “lampe” ([lãp], French equivalent of “lamp”). The sequence was bistable and could also be perceived as the repetition of the word “plan” ([plã], French equivalent of “map”). The second sequence consisted of the repetition of the bisyllabic word “képi” ([kepi], French equivalent of “kepi”), which could also be perceived as the repetition of the word “piquer” ([pike], French equivalent of “to sting”). Two other sequences consisted of the repetition of pseudo-words that were either monosyllabic “sep”

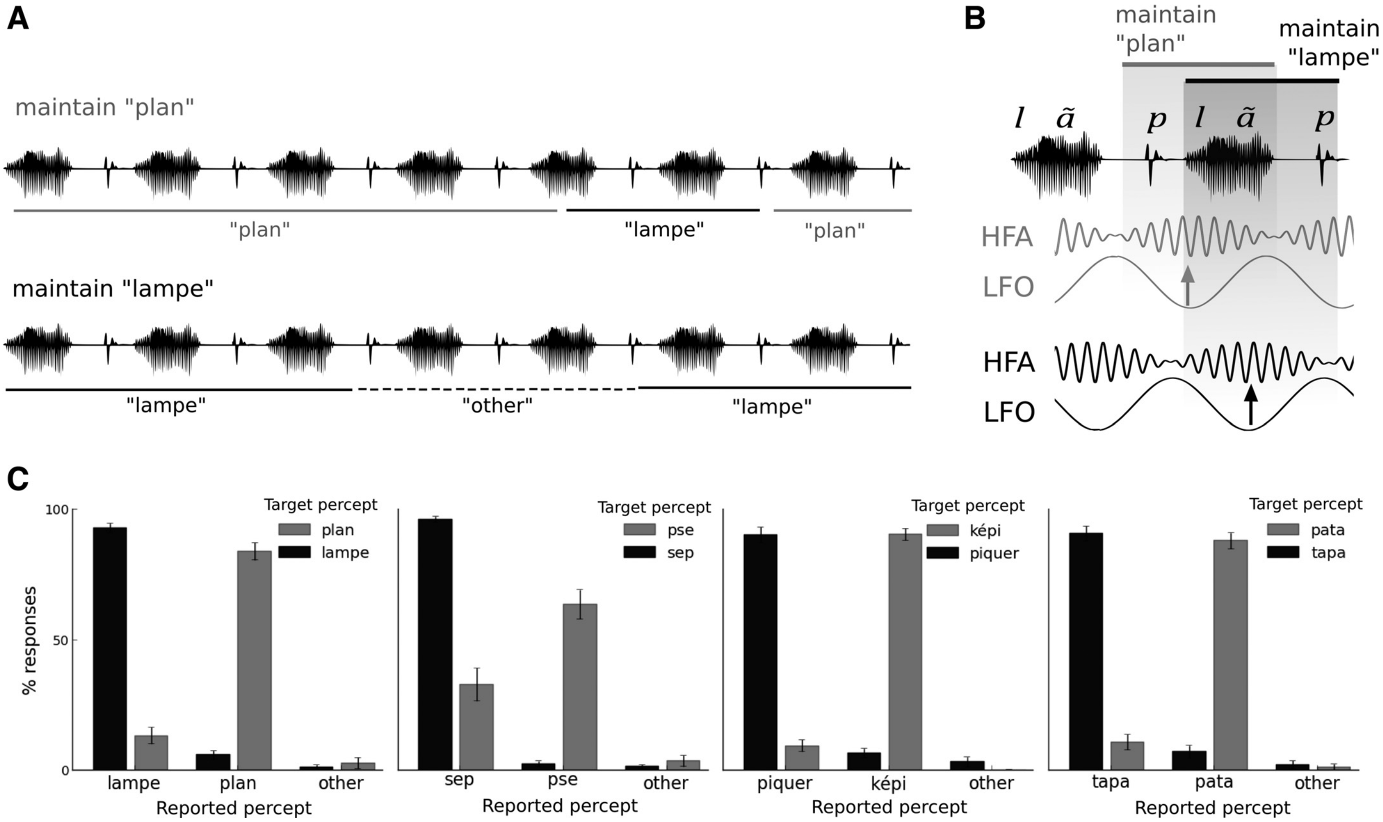

Fig. 1. Bistable speech segmentation: design (*A*), hypothesis (*B*), and behavioral reports (*C*). *A*: participants were asked to maintain a given percept a long as possible while listening to a bistable speech sequence. Four sequences of interest were presented: repetition of the word "lampe" ([lãp]), repetition of "sep" ([səp]), repetition of "képi" ([kepi]), and repetition of "pata" ([pata]). The sequences were bistable and could also be perceived as repetitions of the word "plan" ([plã]), "pse" ([psə]), "piquer" ([pike]), and "tapa" ([tapa]), respectively. Participants listened twice to each sequence and were asked to maintain either one or the other of the possible bistable speech percepts (e.g., maintain "plan" or maintain "lampe"). Subjects reported online their current percept by keeping a button pressed. Three buttons were given: two buttons for the bistable percepts (e.g., "plan" and "lampe") and one button for "other" if subjects perceived another utterance. *B*: if low-frequency oscillations (LFO) reflect linguistic parsing mechanisms, we predicted that changes in the latency of speech tracking by LFO would define the boundaries of speech segmentation. In this example, for a given acoustic signal, the latency (indexed by the phase) of the low-frequency neural response was predicted to vary depending on whether participants perceive "plan" or "lampe". Because the phase of low-frequency neural oscillations may modulate the excitability of high-frequency activity (HFA), changes in the latency of HFA were expected so that lowest HFA would be aligned with the word boundaries. *C*: on average, for any given speech sequence, participants succeeded in maintaining the instructed percept. Bars represent the proportion of responses when participants were asked to maintain the perception of one word during one sequence presentation (black) or the other word in another sequence presentation (gray). Errors bars denote SE.

([səp]), leading to the alternative percept "pse" ([psə]), or bisyllabic "pata" ([pata]), which could also be heard as the repetition of the pseudo-word "tapa" ([tapa]). All syllables in the auditory sequences were recorded (16-bit resolution, 22.05-kHz sampling rate) in a soundproof room by a native French speaker (A. Kösem). The speaker pronounced each syllable naturally and maintained an even intonation and vocal intensity while producing the sequences. Stimuli sequences were constructed using the Praat freeware (Boersma 2002). For the bisyllabic sequences, one syllable of each token [pa], [ta], [ke], and [pi] was selected; the criterion consisted of selecting the syllable that matched as closely as possible the sequence rate of 3 Hz (1 syllable per 333 ms). All syllables had equalized sound levels based on root mean square (RMS). The selected syllables were assembled to form the word "képi" and the pseudo-word "pata," and each word was repeated 100 times to form the sequences. For the monosyllabic word and the pseudo-word sequences, one clearly articulated token [psə], [səp], [plã], and [lãp] of 333-ms duration was selected from the recordings and repeated 150 times. In all recordings, the syllabic length was 333 ms, leading to a repetition rate of 1.5 Hz in bisyllabic sequences and 3 Hz in monosyllabic sequences.

Procedure. First, before beginning the main experiment, participants were familiarized with the stimuli via the spontaneous verbal transformation task (Basirat et al. 2012), in which participants hear a sequence of repeated acoustic utterances yielding bistable auditory percepts. Participants were asked to spontaneously report their perception while listening to these auditory sequences. After completing this familiarization phase, participants performed a variation of the verbal transformation paradigm in which they were asked to voluntarily maintain hearing one of the possible speech percepts for as long as possible while listening to the sequence. Participants were instructed to perceive the sequence as the repetition of a target word, without vocalizing the word or imposing a rhythm during the presentation of the sequences. Specifically, we asked participants to hear the external speaker repeating the word and not to covertly produce the sequences. The four auditory sequences were presented twice, and the instructed target word was counterbalanced for each presentation. Hence, in one presentation, participants were asked to maintain the target "képi," "pata," "sep," or "lampe," and in the second presentation of the same sequence, they were asked to maintain the alternative target percept "piquer," "tapa," "pse," or "plan," respectively. The two successive presentations of a sequence constituted one block, and blocks were presented in random order across individuals. During a given speech sequence, participants were asked to continuously depress the button corresponding to the currently perceived utterance and to switch buttons as soon as, and every time, their perception changed. One button was assigned for each of the two expected percepts in a sequence ("képi" and "piquer"; "pata" and "tapa"; "pse" and "sep"; "plan" and "lampe"), and a third button (labeled "other") was used to report any other percepts participants might have experienced during the sequence.

MEG Analysis

Data acquisition. Neuromagnetic brain recordings were collected in a magnetically shielded room using the whole head Elekta Neuromag Vector View 306 MEG system (Elekta Neuromag, Helsinki, Finland) equipped with 102 triple-sensor elements (2 orthogonal planar gradiometers and 1 magnetometer per sensor location). Shielding against environmental noise was provided by MaxShield (Elekta Neuromag). Participants were seated in an upright position. Each participant's head position was measured before each block with four head position coils (HPI) placed over frontal and mastoid areas. MEG recordings were sampled at 1 kHz and online bandpass filtered between 0.03 and 330 Hz. The electro-occulograms (EOG; horizontal and vertical eye movements) and electrocardiogram (ECG) were recorded simultaneously with the MEG.

Data preprocessing. The signal space separation (SSS) method was applied to decrease the impact of external noise (Taulu et al. 2003). SSS correction, head movement compensation, and bad channel rejection was done using MaxFilter software (Elekta Neuromag). Signal-space projections were computed by principal component analysis (PCA) using Graph software (Elekta Neuromag) to correct for eye blinks and cardiac artifacts (Uusitalo and Ilmoniemi 1997).

Data analysis. MEG analyses were performed using MNE-Python (Gramfort et al. 2013, 2014). The analyses were performed on gradiometer data, known to be less sensitive to environmental noise (Hämäläinen et al. 1993; Vrba 2002). The trials for evoked responses, phase quantification, and cross-correlation analyses were computed by segmenting continuous data into 2-s epochs centered on the burst of the [p] plosive in the speech signal. The trials for spectral analyses were computed by segmenting data in 8.2-s epochs to ensure a high spectral resolution of low-frequency dynamics. A rejection criterion for gradiometers with peak-to-peak amplitude exceeding 4.000 e^{-10} T/m was applied to select the epoch data. Trials in which participants failed to maintain the target percept were excluded from further analysis, as assessed by participants' explicit button presses while listening to speech; trials which were preceded or followed by a change in button press by 500 ms were also excluded. Hence, the analysis included trials in which the perceptual reports matched the target percepts: for instance, we will call "lampe" trials those trials in which participants were instructed to maintain the percept "lampe" and actually reported having heard "lampe." In total, ~23% ± 9.6% (mean ± SD) of epochs were rejected.

Data were analyzed in two regions of interest by selecting gradiometers covering the left and right temporal areas (the selected sensors are depicted in Fig. 2 as black dots). A spatial filter was used for channel averaging in each region of interest. The spatial filter was estimated on the basis of the signal-space projection of the covariance of the evoked responses (Tesche et al. 1995) to each sequence; it consisted of signed weights on each sensor, based on their contribution to the evoked response and their polarity. This was done to enhance the contribution of sensors that were strongly modulated by the evoked component of the signal and to alleviate sensor cancellation due to opposite signal polarities.

ERF analysis. For ERF analysis, epochs were filtered between 1 and 40 Hz. The comparisons of evoked responses between conditions were computed using a nonparametric permutation test in the time dimension. Correction for multiple comparisons was performed with cluster-level statistics (cluster $\alpha = 0.05$), using as base statistic a one-way *F*-test computed at each time sample (Maris and Oostenveld 2007). For illustration, we show in Fig. 2 the topography of the 3-Hz component of the evoked response. To do so, 3-Hz evoked amplitude was estimated by using Morlet wavelet transform (4 cycles) on the evoked data, which was then averaged across sequences.

Frequency analysis. MEG signals were divided in epochs of 8.2 s to compute the power spectrum density (PSD) by using a Welch's average periodogram method for each experimental condition (perceived speech), for each hemisphere, and on a per-speech sequence and per-individual basis. Repeated-measures ANOVA were performed at observed frequency peaks of the power spectra: the entrainment frequency (3 Hz), 1.5 Hz and its harmonics (4.5, 6 Hz), and alpha oscillations (8–12 Hz). This was done to assess the contribution of each frequency to the overall brain response. The other factors included were sequence type (4 levels: "kepi," "pata," "lampe," and "sep"), reported percept (2 levels: "*percept 1*" and "*percept 2*", e.g., "lampe" and "plan"), and hemisphere (2 levels: right and left).

Phase analysis. The phase of the 3- and 1.5-Hz entrained oscillatory responses and the phase-locking value (PLV; also called phase-locking factor or intertrial coherence) were computed at the onset of the plosive burst. The PLV is a measure of phase consistency across trials (Tallon-Baudry et al. 1996) and is defined as

$$\mathrm{PLV(t)} = \frac{1}{n}\left|\sum_{k=1}^{n} e^{j\theta(t,k)}\right|,$$

where n is the number of trials and $\theta(t,k)$ is the instantaneous phase at time t and trial k. Single-trial MEG data (2-s epochs centered on the plosive) were convolved with a 3-cycle Morlet wavelet at 1.5-Hz frequency for the computation of the 1.5-Hz preferential phase and with a 4-cycle Morlet wavelet at 3-Hz frequency to compute the 3-Hz preferential phase. To assess statistical significance of phase shifts between the two percepts during the presentation of one speech sequence (e.g., "plan" and "lampe"), we computed the difference in the preferential phase of entrainment on a per-individual basis. The 95% confidence intervals (CI) of the distribution of phase differences across participants was estimated with a bootstrapping method based on 10,000 resamples of the distribution with replacement (Fisher 1995). Phase distributions were considered statistically different between the percepts if zero lay outside the measured confidence interval ($P \leq 0.\ 05$, uncorrected for multiple comparisons), i.e., if zero was lower than the 2.5% percentile or higher than the 97.5% percentile of the bootstrap distribution.

Cross-correlation measures of speech envelope with MEG signals. To estimate the modulations of HFA amplitude that followed the speech envelope, normalized cross-correlations between the amplitude of neural oscillations and the dynamics of the speech envelope were computed. First, the speech envelope was estimated by using a filter bank that models the passage of the signal through the cochlea (Ghitza 2011; Glasberg and Moore 1990). The filter bank was designed as a set of parallel bandpass finite impulse response (FIR) filters, each tuned to a different frequency. The center frequencies of interest were chosen from 250 and 3,000 Hz. The center frequency f and the critical bandwidth of each filter were computed following the method of Glasberg and Moore (1990). Second, a Hilbert transform was applied to each filtered signal, and the absolute value of each Hilbert transform was averaged to obtain the final envelope. The amplitude of neural oscillations ranging from 6 to 140 Hz (in 2-Hz steps) was computed on the 2-s-long epochs on a per-trial basis by filtering the MEG signal for each frequency band (FIR filter, bandwidth ±2 Hz for frequencies <20 Hz, bandwidth ±5 Hz for frequencies ≥20 Hz) and by taking the absolute value of the Hilbert transform applied to each filtered signal.

Normalized cross-correlations were computed between the amplitude of the MEG signal for each frequency band and the cosine of the phase of the speech envelope at 3 Hz [FIR filter, bandwidth (2.5, 3.5 Hz), in mono- and bisyllabic sequences] and 1.5 Hz [FIR filter, bandwidth (1, 2 Hz), in bisyllabic sequences only]. The resulting cross-correlograms thus indicate how the amplitude of high-frequency oscillations consistently tracks the dynamics of 3- and 1.5-Hz speech envelopes over trials. For each individual, we tested if the cross-correlograms significantly differed between the percept conditions using spectrotemporal cluster permutation statistics (Maris and Oostenveld 2007). A one-way *F*-test was first computed at each time and frequency sample. Samples were selected if the *P* value associated to the *F*-test was <0.05 and were clustered on the basis of spectrotemporal adjacency. The sum of the *F* values within a cluster was used as the cluster-level statistic. The reference distribution for cluster-level statistics was computed by performing 1,000 permutations of the data between the two conditions. Clusters were considered significant if the probability of observing a cluster test statistic of that size in the reference distribution was <0.05. The significant clusters indicated, on a per-individual basis, in which frequency bands the difference between maintained percepts was most pronounced.

The resulting differences between brain responses to successfully maintained percepts could originate either from a difference in the strength of the speech-brain coupling as a function of the perceived speech or from a difference in the latency of the cross-correlation (which could be interpreted as temporal shifts in neural speech tracking). To test these two hypotheses, we measured both the maximal value and the latency of the cross-correlation for each frequency band of each significant cluster, comprising beta (12–30 Hz), gamma (40–80 Hz), and high-gamma (90–130 Hz) frequency ranges. The latency was estimated by computing the phase of the 3-Hz component of the cross-correlation at the onset of the plosive.

RESULTS

Volitional Maintenance of Conscious Speech Percepts

During MEG recording, participants listened twice to the same ambiguous speech sequence and were asked to maintain one or the other possible speech percepts. Participants continuously reported their percept by keeping one of three possible response buttons pressed: one button was used for each of the two expected perceptual outcomes, and a third when a different percept was heard. Overall, participants reported that the task was easy to perform and that they could easily hear the speaker pronouncing either one of the two word sequences. Consistent with their introspection on the task, participants successfully maintained the required speech percept in all conditions: the percept to be maintained was heard significantly more than the alternative percepts [$F(2,28) = 17.6$, $P < 0.001$; Fig. 1*C*]. Although post hoc analysis showed that the percept "pse" was significantly harder to maintain compared with other percepts (64% maintenance, compared with 84–96% in the other conditions), it remained significantly dominant compared with the alternative percept "sep" (33%) in this condition. These results confirmed that with this task, the perception of the repeated word in the sequence could be manipulated while keeping the acoustic signal constant. Volitional control also limited the biases observed during spontaneous bistable perception in which participants typically report hearing mainly one word repeated in the sequence and not the two bistable percepts in balance (see Basirat et al. 2012; Sato et al. 2007).

Low-Frequency Phase Response is Not Indicative of Perceived Word Segmentation Boundary

Changes in the phase of the 3-Hz oscillatory component as a function of perceived word during monosyllabic sequences. The syllabic rate of all speech sequences was set to 3 Hz to keep within the natural range of syllabicity described across all languages (Greenberg et al. 2003; Poeppel 2003). Bistable speech percepts were thus effectively repeated at 3 Hz in monosyllabic sequences ("lampe" and "sep") but at 1.5 Hz in bisyllabic sequences ("képi" and "pata"). As predicted, auditory cortices showed a strong phase locking at the syllabic rate (i.e., 3 Hz) in all conditions and over temporal sensors bilaterally (Fig. 2*A*).

We compared the neural responses between trials in which participants successfully maintained the target percepts: for instance, the trials of the percept condition "lampe" correspond to the trials in which participants were instructed to maintain the target "lampe" and reported having heard "lampe"; the "plan" trials correspond to the trials for which participants were instructed to maintain "plan" and reported having heard "plan." If the entrainment of LFO by speech rhythms implements acoustic parsing, no changes in LFO should be seen, because the acoustic signal was identical for both percept conditions. Alternatively, if linguistic parsing modulates LFO, substantial temporal shifts in the LFO should be observed. In this study, the LFO would realign at different time points of the acoustic signal depending on participant speech report. Each duty cycle of the LFO would define landmarks that are relevant for linguistic parsing (e.g., word's onset and offset), and the extent of the parsing window (or the duty cycle of the LFO) should contain sufficient acoustic information to result in the syllabic or word representation (Fig. 1*B*). The temporal shift in tracking could be measured as a phase shift of the entrained oscillation, meaning that a certain speech acoustic feature should occur at a different phase of the LFO depending on the perceived word. We first tested this hypothesis by computing the 3-Hz phase-locking value (PLV) and preferential phases of brain responses elicited by the presentation of the monosyllabic sequences ("sep" and "lampe"). PLVs and preferential phases were computed at the onset of the plosive burst ([p]) for all possible perceptual outcomes to directly assess whether a given acoustic landmark was associated with the same phase characteristics of LFO irrespective of perception. First, the PLVs did not significantly differ between the left and right hemispheres [$F(1,14) < 1$] or between the perceptual outcomes within each sequence [main effect of percept: $F(1,14) = 2.6$, $P = 0.13$; interaction between percept and sequence type: $F(3,56) < 1$; Table 1], suggesting that the strength of low-frequency neural entrainment was comparable irrespective of participants' perceptual report. The 3-Hz PLVs were nevertheless of different strengths for each sequence [main effect of sequence: $F(3,56) = 9.5$, $P < 0.001$; Table 1]. The monosyllabic sequences elicited a stronger PLV than bisyllabic sequences (Table 1, post hoc Tukey-Kramer test: "képi" vs. "pata," $P = 0.3$; "képi" vs. "lampe," $P < 0.001$; "képi" vs. "sep," $P < 0.001$; "pata" vs. "lampe," $P < 0.001$; "pata" vs. "sep," $P = 0.005$; "lampe" vs. "sep," $P = 0.04$).

Second, in line with a possible top-down modulation of LFO, the preferential phase response varied as a function of the perceived utterance within the same sequence: perceiving the word "lampe" was associated with a phase advance of −8° (95% CI = [−15.5°, −0.3°]) compared with "plan" (Fig. 2*B*). Similarly, a phase advance of −9° (95% CI = [−17.7°, −1.5°]) in the 3-Hz oscillatory response distinguished the perceived syllable "sep" from "pse." No other changes in the phase or in the evoked responses at 3 Hz were observed between the different percepts of the bisyllabic sequences (Fig. 2, *B* and *C*). Hence, in both monosyllabic sequences, the

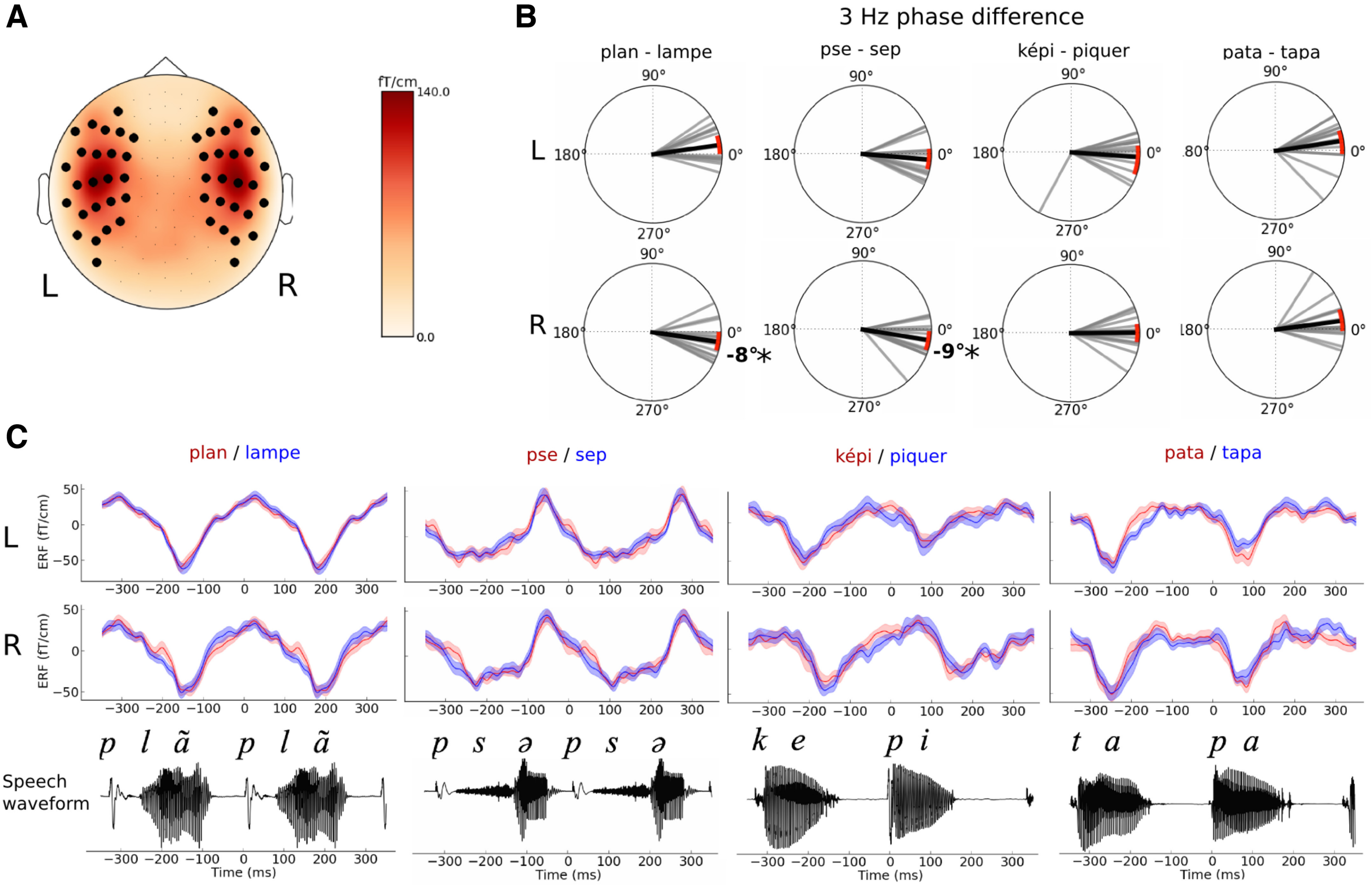

Fig. 2. Characteristics of the 3-Hz neural response to speech. *A*: scalp topography of the 3-Hz auditory evoked response. Black dots illustrate the position of the selected gradiometers over the left (L) and right (R) hemispheres. *B*: phase differences of the 3-Hz response contrasting perceived speech utterances. Polar plots report the 3-Hz phase difference between the 2 perceptual outcomes (*top*, left hemispheric sensors; *bottom*, right hemispheric sensors). Each gray bar is an individual's phase difference between the 2 perceptual outcomes in a given condition. The black bar corresponds to the mean average phase difference across all participants. Red arcs are 95% confidence intervals (CI). *C*: grand average auditory evoked response fields (ERFs) that correspond to left sensors (*top*) or right sensors (*middle*). Speech waveforms are shown for each sequence (*bottom*). No significant changes in the ERF were observed between percepts. Shaded areas denote SE.

perceptual outcomes were associated with phase shifts of the 3-Hz response (Fig. 2*B*). Although these phase shifts were consistent across monosyllabic conditions, the confidence intervals of the differences were nevertheless close to zero and included zero when Bonferroni correction was used for multiple comparisons. To ensure that these results were not specific to the temporal reference or the acoustic landmark used in the computations of the phase responses, the same analysis was carried out when phase responses were locked 50 and 100 ms before or after plosive onset and the main observations were replicated.

Table 1. *PLVs at 3 Hz observed in left and right temporal sensors as a function of perceived speech for each speech sequence*

Percept	PLV_L	Percept	PLV_L
lampe	0.62	plan	0.65
sep	0.56	pse	0.60
képi	0.48	piquer	0.48
tapa	0.50	pata	0.54
Percept	**PLV_R**	**Percept**	**PLV_R**
lampe	0.61	plan	0.62
sep	0.60	pse	0.56
képi	0.49	piquer	0.48
tapa	0.48	pata	0.55

Phase-locking values (PLVs) did not significantly change between percept conditions [no main effect of percept: $F(1,14) = 2.6$, $P = 0.13$; interaction between percept and sequence type: $F < 1$]. L, left; temporal sensor; R, right temporal sensor.

These results suggest that if neural tracking of speech is subject to top-down modulations, then the effect may be weak. Phase shifts of 8° or 9° translate into temporal shifts of ~9 or 10 ms, respectively, suggesting that the neural response to the plosive [p] when participants perceived the word "plan" was delayed by 9–10 ms compared with when participants perceived "lampe." Notably, the latency shifts observed in the monosyllabic sequences conditions were smaller than would have been expected on the basis of LFO as parsers, considering that the temporal distance between the percepts' onset and offset features was of higher magnitude. In the "lampe" sequences, the plosive and the onset of the consonant [l] were separated by 80 ms, and the silent gap prior to the plosive was 90 ms. For the "sep" sequences, the plosive and the onset of the consonant [s] were 50 ms apart with a silent gap of 70 ms. Thus the minimal temporal shift of the parsing window necessary to distinguish the two monosyllabic percepts should have been

50–80 ms, corresponding to phase shifts in neural entrainment of at least 55° to 85°. Because the reported phase shifts were much smaller (8° to 9°), our results for monosyllabic sequences suggest that the duty cycles of the entrained LFO are insufficient to account for linguistic parsing.

Phase characteristics of the 1.5-Hz oscillatory response as a function of the perceived word during bisyllabic sequences. In addition to the 3-Hz auditory peak response, significant peak responses were found in the power spectral density (PSD) of the MEG brain responses (Fig. 3*A*). Specifically, what appeared as the subharmonic and harmonic components of the acoustic signals were observed in the PSD consistent with the repetition rates of the mono- and bisyllabic words, namely, at 1.5, 3, 4.5, and 6 Hz. The canonical alpha rhythm (8–12 Hz) was also readily seen. The contribution of each observed frequency differed according to the sequence. In fact, we observed that the power of the 1.5-Hz oscillatory response was significantly enhanced when participants were listening to bisyllabic speech utterances compared with monosyllabic ones (Fig. 3, *A* and *B*). ANOVA revealed a significant main effect of frequency peak response [$F(4,56) = 29.9$, $P < 0.001$] and of the sequence [$F(3,42) = 3.4$, $P = 0.027$], as well as a significant interaction between frequency peak response and sequence [$F(12,168) = 11.4$, $P < 0.001$]. Tukey-Kramer post hoc analysis showed a significant difference in the power of the 1.5-Hz response between bisyllabic and monosyllabic sequences ("képi" vs. "pata," $P = 0.9$; "képi" vs. "lampe," $P < 0.001$; "képi" vs. "sep," $P < 0.001$; "pata" vs. "lampe", $P < 0.001$, "pata" vs. "sep," $P = 0.002$; "lampe" vs. "sep," $P = 1$), suggesting that 1.5-Hz dynamics were more prominent for bi- than monosyllabic processing. The 1.5-Hz power was not indicative of the perceived word within a sequence [$F(1,14) < 1$; Fig. 3*B*] and did not show significant differences across hemispheres [$F(1,14) < 1$]. The results were qualitatively similar after correction for 1/f noise distribution as reported by Kösem et al. (2014) and Nozaradan et al. (2011).

There are two possible origins for the observation that bisyllabic words induce significant 1.5-Hz auditory responses (Fig. 3, *A* and *B*). First, the 1.5-Hz response could be elicited by a subharmonic component already present in the speech signals, considering that bisyllabic sequences consist of the repetition of acoustic patterns at 1.5 Hz. The 1.5-Hz response observed in the PSD of MEG activity could thus reflect a passive bottom-up frequency tagging of the auditory response. Second, the 1.5-Hz response could also be under the influence of top-down mechanisms. The two hypotheses cannot be fully

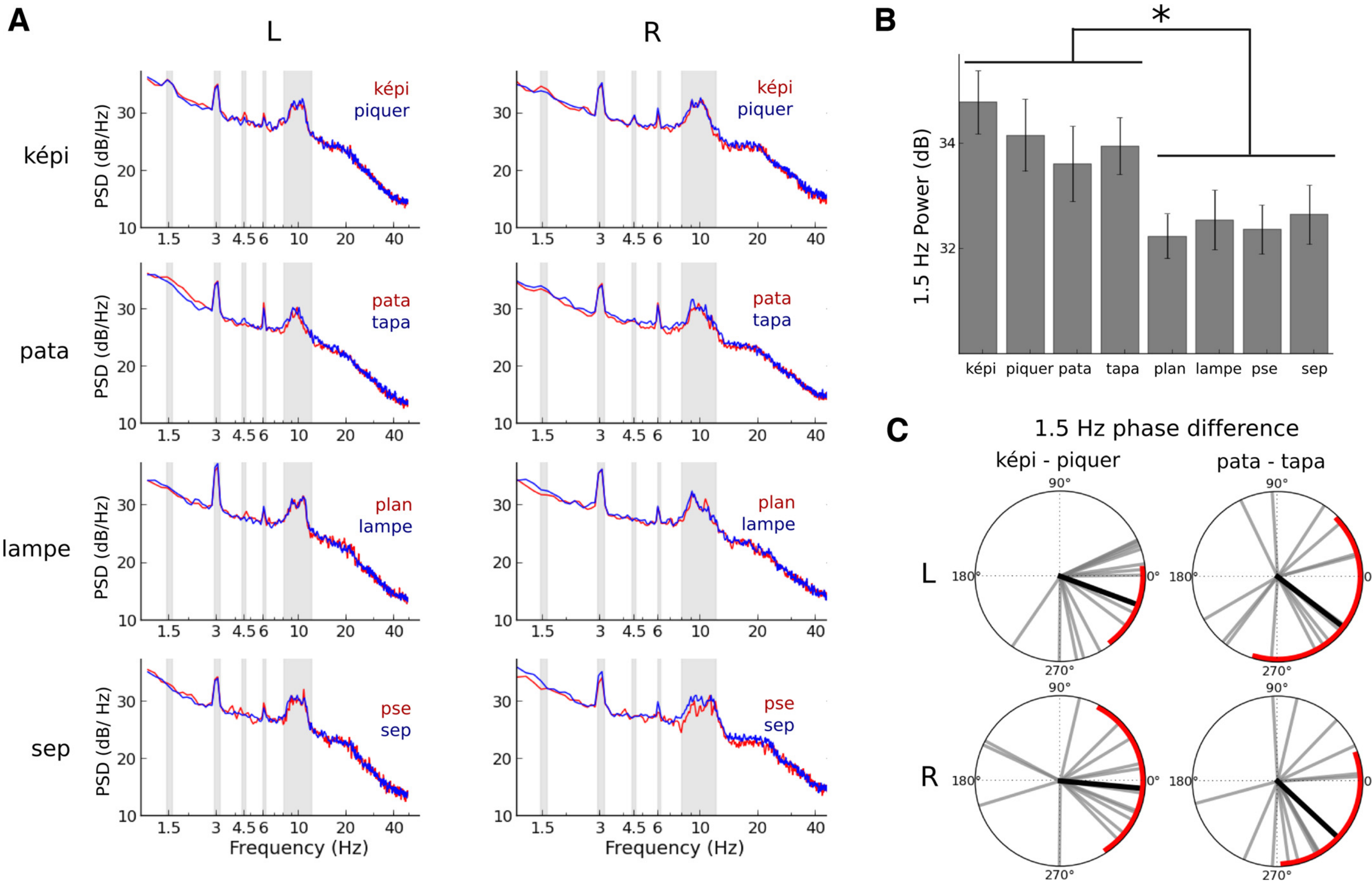

Fig. 3. Modulations of low-frequency neural entrainment and characteristics of (sub)harmonic peak responses. *A*: power spectral density (PSD). The red and blue traces correspond to the PSD of brain activity in response to the presentation of the speech sequences "képi," "pata," "lampe," and "sep." Color codes for the percept that was maintained. Gray areas highlight the frequencies of interest in the spectra, e.g., 1.5, 3, 4.5, and 6 Hz and alpha band (8–12 Hz). *B*: increased 1.5-Hz power for bisyllabic speech sequences. 1.5 Hz power was significantly higher when participants listened to bisyllabic sequences compared with monosyllabic sequences. The power did not significantly vary between the 2 bisyllabic sequences or between the 2 monosyllabic sequences. Errors bars denote SE. *C*: no significant 1.5-Hz phase differences of the LFO are observed when contrasting perceived speech utterances. Each gray bar is an individual's 1.5-Hz phase difference observed when contrasting the 2 alternative percepts of a sequence. The black bar corresponds to the average phase shift across all participants. Red arcs correspond to 95% CI.

disentangled in our study, given that 1.5-Hz peaks were observable in the PSDs of the envelope of the bisyllabic speech sequences (Fig. 4). Nevertheless, previous reports have shown that delta oscillations are not purely stimulus driven and also may be involved in the encoding of abstract linguistic structures (Buiatti et al. 2009; Ding et al. 2016).

Similar to the 3-Hz oscillatory component in the monosyllabic sequences, the perceptual changes in the “képi” and “pata” sequences were expected to be accompanied by modulations of the 1.5-Hz oscillatory phase. No significant changes of the 1.5-Hz PLV were observed when the two perceptual outcomes of the same bisyllabic sequences were compared [main effect of percept: $F(1,14) < 1$; interaction between percept and sequence: $F(1,14) = 2.6$, $P = 0.12$; Table 2]. Nonsignificant phase shifts were observed between the two perceptual outcomes (Fig. 3*C*). As previously mentioned, if the phase of 1.5-Hz LFO marked bisyllabic boundaries for acoustic or linguistic parsing, patterns of out-of-phase shifts would have been observed because syllables composing the word were 333 ms apart (Fig. 1*B*), but this is not what we observed.

Overall, our results do not provide strong evidence that the neural tracking of speech by LFO is subject to top-down modulations. First, perceptual changes in the monosyllabic word sequences could be associated with phase shifts in speech tracking at the syllabic rate (3 Hz), but they were too small to account for a shift of the linguistic parsing window. Second, the subharmonic of the syllabic rate (1.5 Hz) was also observed in the auditory response when participants listened to sequences of bisyllabic words, but it is not entirely clear whether it only originates from bottom-up processing or whether top-down modulations intervene. Hence, these findings do not show direct evidence that LFO are indicative of segmentation boundaries that are directly relevant for conscious speech perception. Additional mechanisms likely come into play to account for the restructuring of information that would be consistent with the speech percept.

Changes in the Latency of HFA Reflect Conscious Speech Percepts on a Per-Individual Basis

Under the linguistic parsing hypothesis and the speech models being tested (Ding and Simon 2014; Giraud and Poeppel 2012; Hyafil et al. 2015a), HFA is coupled to low-frequency dynamics, and its periodical inhibition by LFO may mark segmentation boundaries. In the context of such cross-frequency coupling, we hypothesized that HFA may display latency shifts of the same magnitude as the phase shifts observed in LFO entrainment. To reliably quantify speech tracking of the HFA, we computed the cross-correlation between the phase of the speech envelope filtered at 3 Hz and the amplitude of the neural oscillations of frequencies spanning 6 to 140 Hz. Speech-neural response cross-correlograms have been used previously (Fontolan et al. 2014; Gross et al. 2013) to estimate the frequency bands in the neural signals that preferentially track the dynamics of speech, as well as to compute the latency between speech and the amplitude of neural oscillations. In this study, we specifically targeted the dynamics of the 3-Hz speech envelopes to capture the amplitude modulations that followed the syllabic rate. The resulting signal was an oscillation at 3 Hz whose phase corresponded to the latency between the speech sequence and the neural response. The latency of the cross-correlation was expected to be consistent within percept but different across percepts.

We contrasted the cross-correlograms between the two percepts of a given sequence for each participant and performed between-trials spectrotemporal cluster analysis of the contrast. Significant changes in the cross-correlograms were found depending on the individual’s perceptual outcome for mono- and bisyllabic sequences. All participants presented significant differences between the monosyllabic percept conditions in both hemispheres, and relatively few participants had significant changes for bisyllabic sequences. For illustration, we report the outcome of this analysis for two participants (Fig. 5), contrasting perceiving “lampe” with perceiving “plan.” For *participant p04*, changes in percept were associated with differences in speech-brain cross-correlation that were more prominent in the gamma and high-gamma ranges (Fig. 5*A*), and for *participant p05*, differences in speech-brain cross-correlations were strongest for distinct gamma and high-gamma bands and for lower frequency responses (Fig. 5*A*). Crucially, the significant changes in cross-correlation were related to strong latency differences as reflected by the phase shifts of the 3-Hz cross-correlograms between perceiving “plan” and “lampe” (Fig. 5, *B* and *C*). Hence, the latency (quantified as the phase of the modulated HFA) shifted according to the speech envelope, and these shifts were associated with changes in conscious percepts (Fig. 5*C*). Additionally, although strong phase oppositions in

Table 2. *PLVs at 1.5 Hz observed in left and right temporal sensors as a function of perceived speech in bisyllabic speech sequences*

Percept	PLV_L	Percept	PLV_L
képi	0.30	piquer	0.25
tapa	0.22	pata	0.19
Percept	**PLV_R**	**Percept**	**PLV_R**
képi	0.27	piquer	0.22
tapa	0.19	pata	0.19

No significant changes of the 1.5-Hz PLV were observed when contrasting the two perceptual outcomes of the same bisyllabic sequences [main effect of percept: $F(1,14) < 1$; interaction between percept and sequence: $F(1,14) = 2.6$, $P = 0.12$].

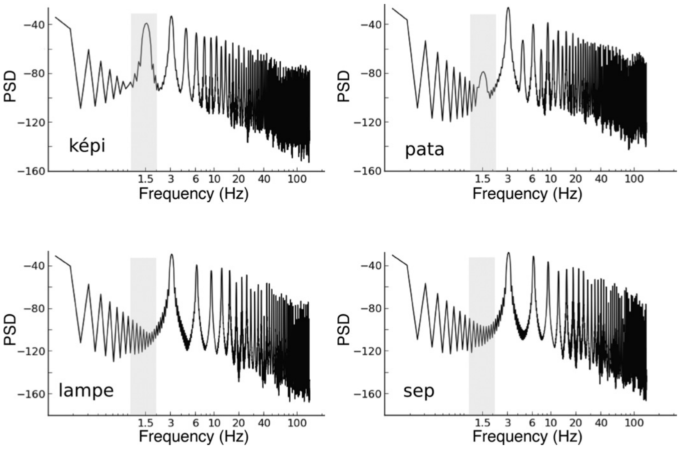

Fig. 4. Frequency power spectra of the envelopes of the acoustic stimuli. The 1.5-Hz subharmonic in the bisyllabic stimuli sequences could readily be seen in contrast to the monosyllabic stimuli. The 1.5-Hz component was also stronger in ”képi“ sequence compared with the ”pata“ sequence.

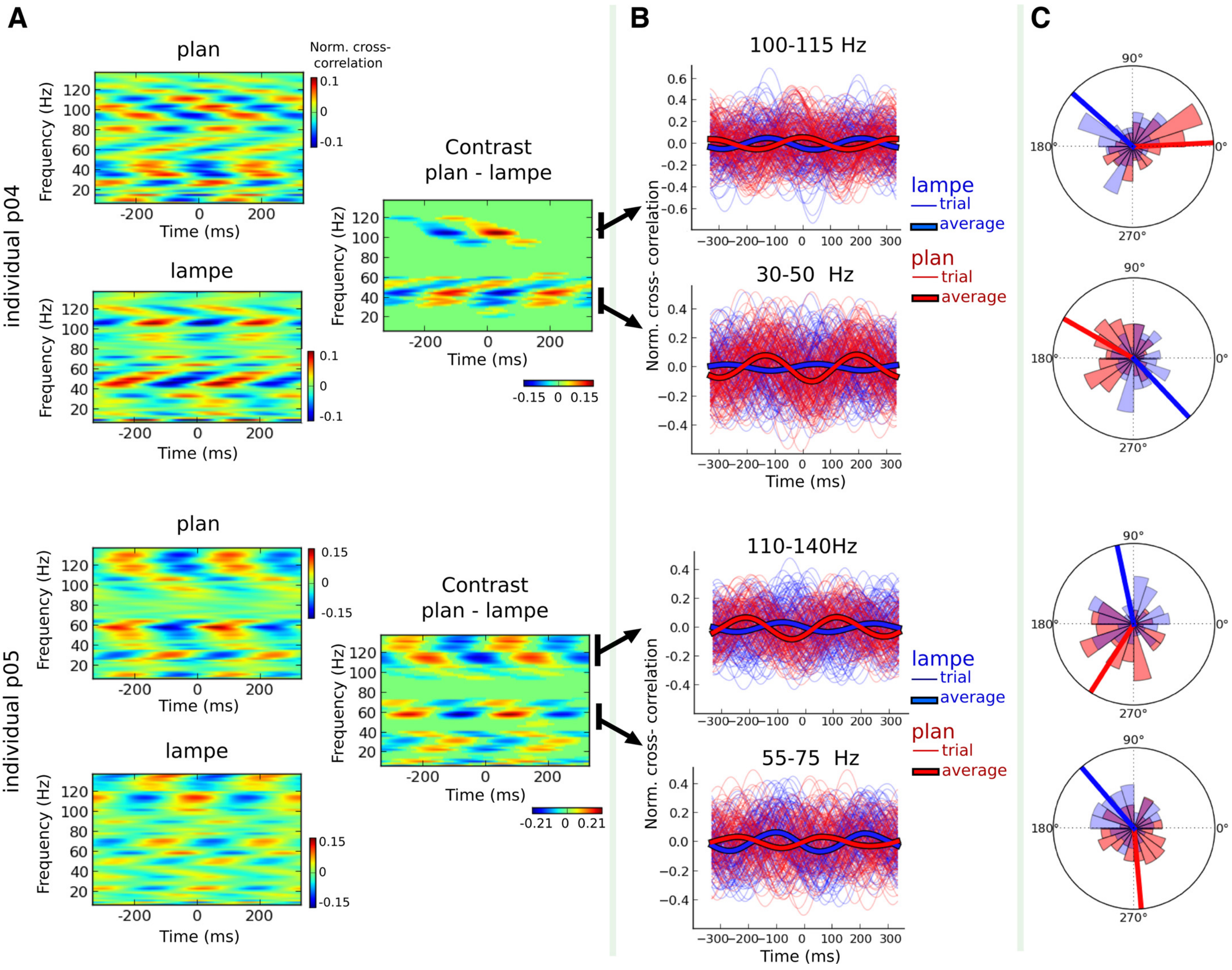

Fig. 5. High-frequency activity predicted an individual's conscious word percept. *A*: cross-correlations between the speech envelope and brain activity during the "lampe“ sequences for 2 participants (*p04* and *p05*). Cross-correlations significantly changed over time according to the perceived word. *Left*, the outcome of the cross-correlograms for each percept; *right*, the difference between the two percepts. Significant differences are reflected by any patch not colored green. *B*: time series of individual speech-HFA cross-correlations within significant clusters. The peak of the cross-correlations systematically occurred at distinct latencies as a function of the individual's perceived word despite the observed variability across participants with respect to the frequency specificity of the HFA and the latency of the maximal correlation. *C*: phase distributions depicting the peak latency of the speech-neural response cross-correlations. Bars denote the mean preferential phase for each percept condition. Strong differences in phase (i.e., in cross-correlations latencies) were observed between the percept conditions despite interindividual variability of absolute phase (i.e., peak latency).

the latency of HFA systematically distinguished an individual's conscious percept, the sign of this opposition varied across participants. For instance, the latency of the cross-correlation in the high-gamma range associated with the percept "plan" differed between *participants p04* and *p05* (Fig. 5*C*).

For each individual, we observed changes in the speech-brain cross-correlations in several frequency bands, but the latencies of the speech-brain correlations for a given condition were variable across participants (Fig. 5*C*). As a consequence, the grand-average analysis of the difference in cross-correlograms between conditions did not capture the effects we observed at the participant level. After statistical analysis of the contrast between the two percepts of a given sequence at the individual level, we assessed the proportion of participants with significant changes in cross-correlation per frequency band. This allowed us to obtain a descriptive profile of which frequencies accounted most for the differences between percepts across all individuals and conditions (Fig. 6*A*). Significant differences were concentrated across participants in the beta band (12–30 Hz; 12 of 15 participants presented significant clusters within this range for both monosyllabic sequences), gamma band (40–80 Hz; 12 participants presented significant clusters in "lampe" sequences and 13 in "sep" sequences), and high-gamma band (90–130 Hz; 14 participants presented significant clusters in "lampe" sequences and 11 in "sep" sequences; Fig. 6*A*).

The observed significant differences in cross-correlation could originate either from a change in the strength of speech-brain correlation or from a change in latency between speech-brain correlated dynamics. To test these two alternative accounts, we restricted the analyses to an individual's clusters in classical frequency ranges in the beta (12–30 Hz), gamma

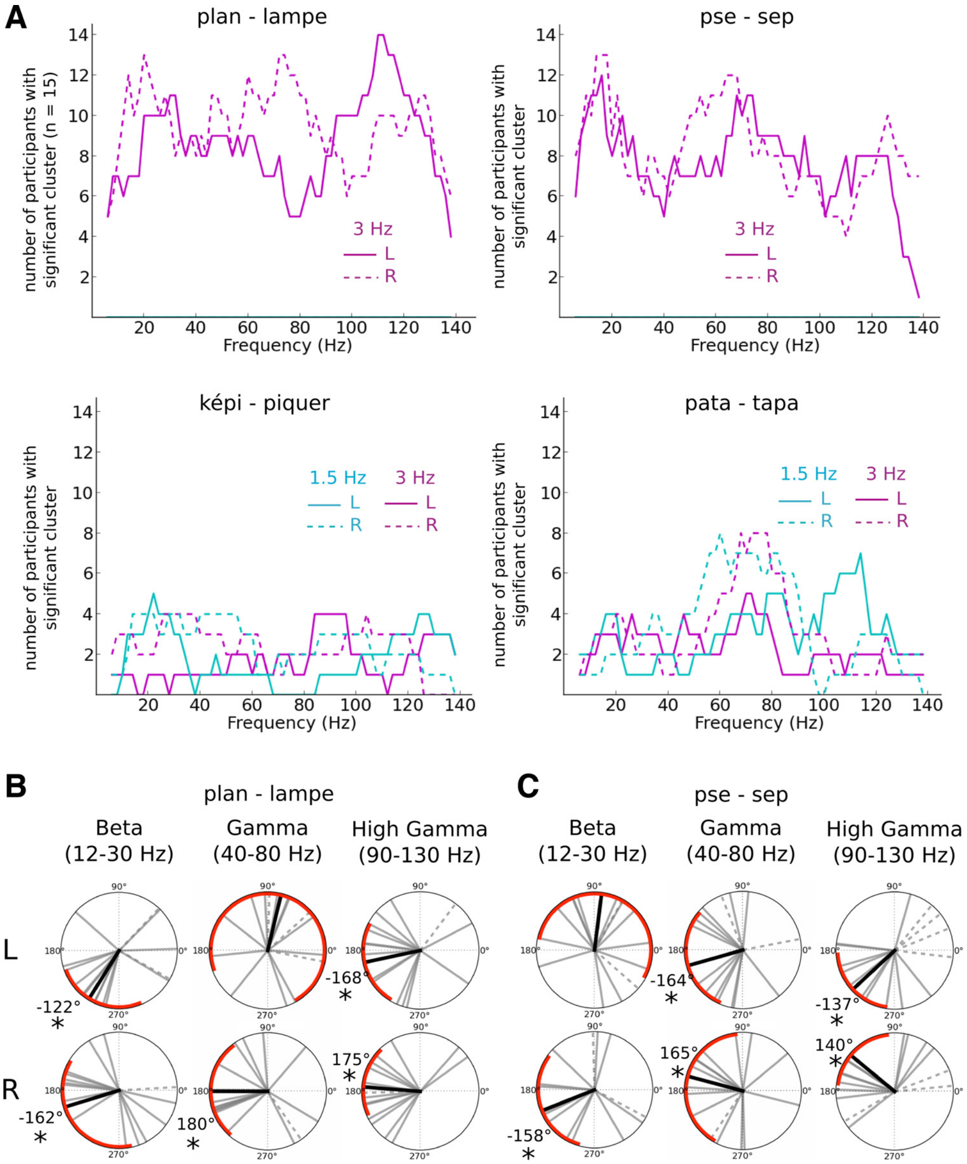

Fig. 6. HFA latency patterns during monosyllabic word sequences. *A*: number of participants with significant changes between percepts in 3-Hz speech component-neural response cross-correlation (magenta lines) and 1.5-Hz speech component-neural response cross-correlation (cyan lines) for each frequency band. Data are reported in both left (solid line) and right (dashed line) temporal sensors for each sequence. We observed significant changes in the cross-correlograms for a majority of participants in the monosyllabic word sequences for frequency bands in the beta (12–30 Hz), gamma (40–80 Hz), and high-gamma (90–130 Hz) ranges. *B* and *C*: phase shifts in speech envelope tracking between each percept condition in the beta (12–30 Hz), gamma (40–80 Hz), and high-gamma (90–130 Hz) range for ”lampe“ sequences (*B*) and ”sep“ sequences (*C*). Each gray line corresponds to the phase difference between one perceptual outcome and the other for one subject. The dashed lines refer to participants for whom significant clusters were not found within the target frequency range. The black line corresponds to the average phase across subjects who showed significant difference between the percept conditions; red arcs correspond to 95% CI. We show here that the reported differences in cross-correlation were related to strong shifts in the phase of neural-speech tracking.

(40–80 Hz), and high-gamma (90–130 Hz) frequency bands (Lopes da Silva 2013). The analyses showed significant latency differences between percepts that were consistent across participants (Fig. 6, *B* and *C*). In contrast, no significant changes in the maximum value of the cross-correlation were observed between percept conditions (Fig. 7). This suggests that the tracking of the speech envelope by HFA was operated at distinct latencies between percept conditions, whereas the amount of coupling between the speech envelope and HFA dynamics remained constant.

As discussed earlier, the distance between consonants is 80 ms for “lampe” and 50 ms for “sep.” The silence duration prior to the plosive is 90 ms for “lampe” and 70 ms for “sep.” Thus the switch from the percept “lampe” to the percept “plan” could be performed via a shift in the linguistic parsing window of 80 ms minimum and up to 170 ms (184°). The switch from the percept “sep” to the percept “pse” could occur through a temporal shift of the linguistic parsing window up to 120 ms (130°). These estimated shifts fall within the confidence intervals of the HFA phase data, suggesting that the reported phase shifts are consistent with shifts in linguistic parsing windows. In bisyllabic sequences, the significant clusters of the cross-correlograms contrasts were sparser (Fig. 6*A*) and inconsistent across participants.

Overall, although perception only weakly modulated the phase of entrained oscillations, it strongly impacted the dynamics of beta, gamma, and high gamma amplitude. Additional analyses were performed to assess the tracking of 1.5-Hz speech dynamics by HFA (this time by filtering the speech signal at 1.5 Hz) during bisyllabic parsing. Sparse significant changes were observed at the individual level (Fig. 6). Hence, high-frequency dynamics could predict the perceived word within one monosyllabic word sequence, but overall did not inform about the perceived word in bisyllabic word sequences.

LFO and HFA Effects are Not Driven by Volitional Control

So far, we have reported effects under volitional control: participants were asked to hear and maintain a specific percept during the presentation of the ambiguous sequences. Several of the effects that we interpret as the result of linguistic parsing

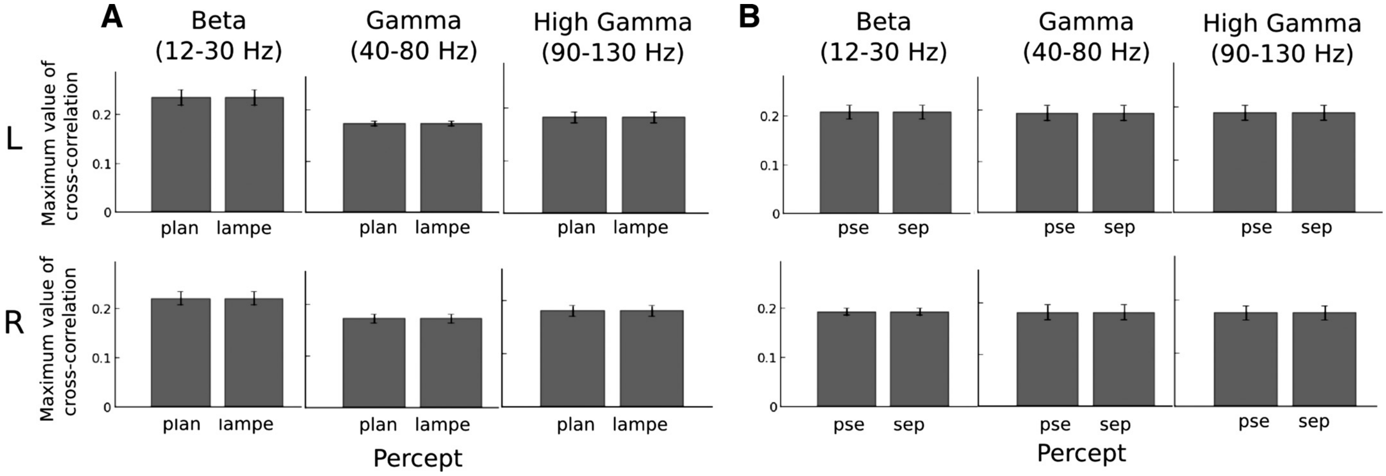

Fig. 7. Mean maximum value of cross-correlation across participants in beta, gamma, and high-gamma bands. The maximum value of cross-correlation did not significantly change between ”lampe“ and ”sep“ sequences [$F(1,14) < 1$], between hemispheres [$F(1,14) < 1$], and between target frequencies [$F(2,28) < 1$]. Crucially, the maximal value of cross-correlations did not differ between percept conditions [$F(1,14) < 1$].

also may have been influenced by the volitional control imposed by the task. To control for this, we analyzed the data from the familiarization task, in which participants listened to the same sequences and spontaneously reported what they heard, without trying to influence their percept in any way. The sequences and the reporting instructions were identical to the volitional task: participants depressed a button corresponding to their current percept (3 possibilities: “*percept 1*,” “*percept 2*,” or “other”). As in prior findings, participants did not report hearing the two percepts in equal proportions, but rather mostly reported hearing the initial veridical word repeated in the sequence (Basirat et al. 2012; Sato et al. 2007). Nevertheless, and with many cautionary steps, we performed a comparable analysis as for the volitional data. Data of participants that had too strong a perceptual bias (i.e., one of the percepts was reported less than 20% of the time) were rejected. Hence, 12 participants were included in the “képi”/“piquer” analyses, 8 participants in the “pata”/“tapa” analyses, and 8 participants in the “plan”/“lampe” analyses. Only 3 participants could be included in the “pse”/“sep” analyses, and thus we only report this condition for illustration purposes.

As previously, we analyzed the 3- and 1.5-Hz phase shifts between percept conditions. We also performed the cross-correlograms for the “plan”/“lampe” sequences in the spontaneous conditions, which we illustrate below. As can be seen, we obtain similar results as with the volitional task as reported in Figs. 2, 3, and 5. First, small 3-Hz phase shifts were observed for the contrast “plan”–“lampe” (Fig. 8*A*). The phase shift was of the same amplitude and direction as in the volitional task. This replication could then be interpreted in favor of a consistent (but weak) top-down modulation of 3-Hz oscillatory activity. Second, no significant 1.5-Hz phase shifts were observed for bisyllabic sequences (Fig. 8*B*). Third, significant changes in 3-Hz-modulated HFA activity were observed between “plan” and “lampe” conditions for participants who were included in the analysis (*p04* and *p05*, Fig. 8*C*). Although the results should be interpreted with caution due to the small number of participants, our results suggest that the main effects of LFO and HFA reported in the volitional task are comparable with those seen in the spontaneous task. Hence, this control suggests that the observed effects reflect genuine linguistic parsing processes and cannot be easily confounded by participants’ cognitive strategy.

DISCUSSION

Our results show an endogenous control of high-frequency activity (HFA) when individuals listen to speech in the context of ambiguous acoustic information. Latency changes of HFA were indicative of the perceived segmented word in the speech streams. We also identified small changes in the phase of entrained low-frequency oscillatory (LFO) responses. Our findings help to shed light on the postulated roles of neuronal oscillations in speech processing (Ding and Simon 2014; Ghitza 2011; Giraud and Poeppel 2012; Poeppel 2003; Poeppel et al. 2008) and show potential dissociable roles of HFA and LFO in the parsing of acoustic information into discrete linguistic content.

Top-Down Control of LFO and HFA During Speech Processing

Both mono- and bisyllabic speech sequences elicited a significant LFO response akin to typical frequency tagging or low-frequency neural entrainment (Hari et al. 1989; Rees et al. 1986; Thut et al. 2011). Under the assumption of a passive entrainment of brain responses, LFO would be expected to remain phase-locked or stationary with respect to the temporal structure of entraining stimuli. Our results suggest that LFO may, to some extent, be not solely driven by the acoustics of the speech signals but also are subject to endogenous control. Specifically, the 3-Hz phase response in monosyllabic speech sequences differed between the two percepts. The phase shifts were weak but consistent in direction and strength across the volitional and spontaneous tasks. Consistent with this, recent findings have shown that the phase of LFO entrainment in auditory cortices can be modulated when timing is relevant to the task (Kösem et al. 2014; Ten Oever and Sack 2015) and can be under top-down control (Baldauf and Desimone 2014; Cravo et al. 2013; Gomez-Ramirez et al. 2011; Lakatos et al. 2008; Park et al. 2015; Parkkonen et al. 2008; Stefanics et al. 2010). Furthermore, our results and those of others (Ten Oever and Sack 2015) suggest that the phase of LFO correlates with

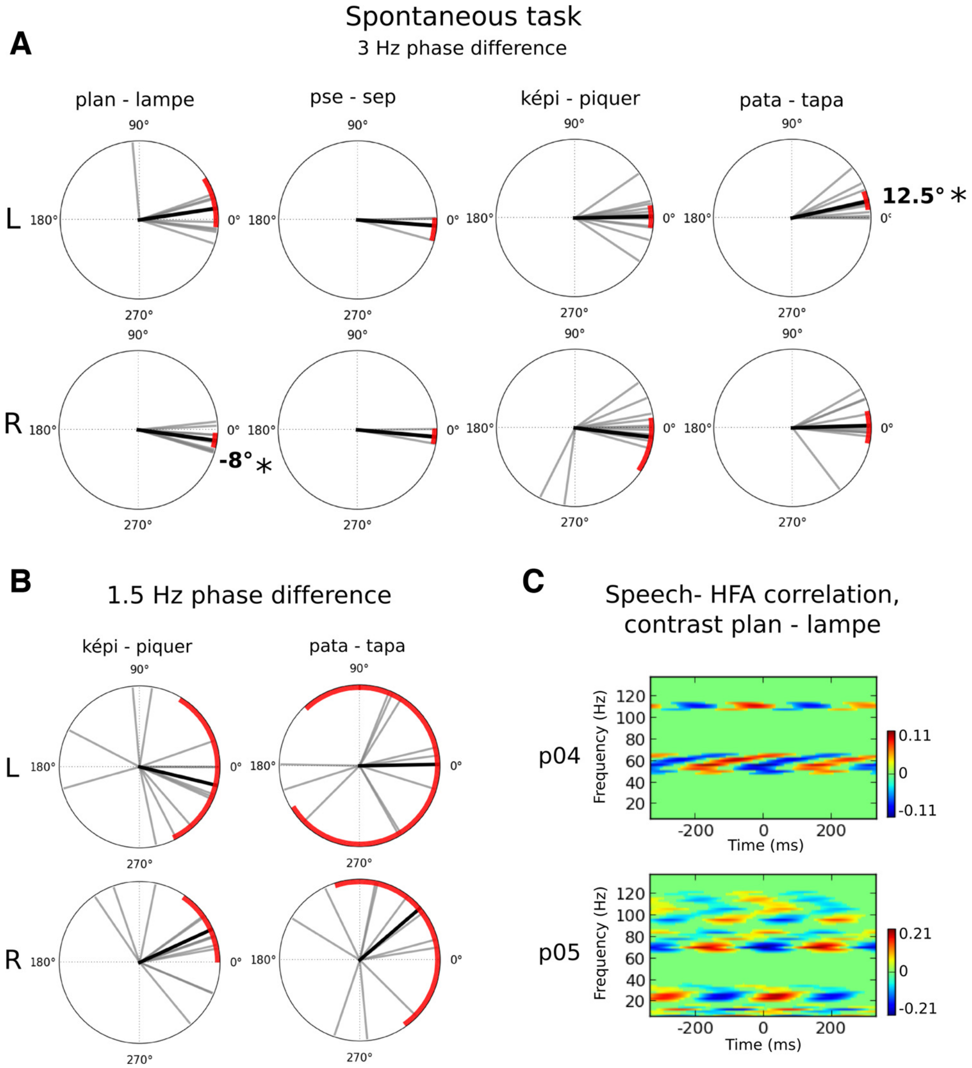

Fig. 8. LFO and HFA effects when participants spontaneously report their perception of the speech sequences. *A*: phase differences of the 3-Hz response contrasting the perceived speech utterances. Polar plots report the 3-Hz phase difference between the 2 perceptual outcomes (*top*, in the left hemispheric sensors; *bottom*, in the right hemispheric sensors). Each gray bar is an individual's phase difference between the 2 perceptual outcomes in a given condition. The black bar corresponds to the mean average phase difference across all participants. Red arcs are 95% CI. As during the volitional task, we observed significant phase shifts of −8° for the contrast "plan"–"lampe." The contrast "pse"–"sep" was not clearly interpretable in the spontaneous task due to a high rejection rate of participants' data. *B*: phase differences of the 1.5-Hz response contrasting perceived words in the bisyllabic sequences. As in the volitional task, the 1.5-Hz phase did not significantly differ between percept conditions. *C*: cross-correlations between the speech envelope and brain activity during the "lampe" sequences for 2 participants in the spontaneous task. Each plot shows the difference in cross-correlation between "plan" and "lampe" percept conditions. Significant differences are indicated by any patch not colored green.

perceptual speech reports. The observed right hemispheric bias in our data could be linked to empirical observations that the right hemisphere is more sensitive to slow speech fluctuations than the left hemisphere (Boemio et al. 2005; Giraud et al. 2007; Poeppel 2003), although the lateralization of the phase effects has not been explicitly tested and is beyond the scope of this study. During bisyllabic sequences, an additional 1.5-Hz neural response was found. Although we cannot exclude the possibility that 1.5-Hz responses mainly reflect the acoustic tracking of the speech signals, the presence of this oscillatory response is in line with previous studies suggesting that delta power can be subject to top-down control during sound processing (Nozaradan et al. 2011) and speech analysis (Buiatti et al. 2009; Ding et al. 2016; Park et al. 2015) and does not solely reflect the temporal structure of the acoustic signals.

Whereas the evidence for top-down modulation for LFO was rather weak, the temporal alignment between the speech signals and the amplitude of HFA displayed systematic latencies or phase shifts as a function of the perceived word. This observation was predicted by speech models (Giraud and Poeppel 2012; Hyafil et al. 2015a) as discussed in the Introduction (also cf. Fig. 1*B*). These effects were systematic within individuals, concentrated in the beta, gamma, and high-gamma frequency bands. Gamma oscillations are markers of neural excitability (Lakatos et al. 2005), and HFA has more generally been shown to track the dynamics of speech (Gross et al. 2013; Hyafil et al. 2015a; Kubanek et al. 2013; Mesgarani and Chang 2012; Mesgarani et al. 2014; Millman et al. 2013; Nourski et al. 2009; Pasley et al. 2012; Zion Golumbic et al. 2013). More specifically, the gamma band has been hypothesized to encode speech information at the phonemic level (Poeppel 2003; Poeppel et al. 2008). We thus expected gamma activity to be largely indicative of the perceived word during bistable speech perception. To the best of our knowledge, the implication of the beta band in speech tracking has not yet been empirically reported, although beta activity has been theoretically posited for chunking dyads, i.e., speech units of 50-ms duration representing the transition between pairs of phones (Ghitza 2011). In addition, beta and gamma neural responses are known dissociable markers of top-down and bottom-up communication (Arnal et al. 2011; Arnal and Giraud 2012; Bastos et al. 2014, 2015; Fontolan et al. 2014). Gamma responses are typically reported as feedforward signals, whereas beta activity has typically been associated with feedback signaling. In our experiment, the fluctuations of beta (respectively, gamma) amplitude could potentially reflect the temporal alternation of feedback (respectively, feedforward) information transfer as shown in a recent report (Fontolan et al. 2014). In their study, Fontolan et al. (2014) used natural speech stimuli and showed that the transition between bottom-up information transfer via gamma activity and top-down communication via beta channels occurred at 1–3 Hz during listening. Their observation was consistent with the idea that speech information propagates along the processing hierarchy and back by units of syllabic/word length. In this scenario, speech tracking by HFA may not only increase the sensitivity to incoming acoustic information but also reflect the linguistic parsing at the syllabic scale.

Brain Oscillatory Mechanisms of Linguistic Parsing

Speech models suggest that LFO chunk the encoding of speech (indexed by HFA) into discrete informational units. In this study, we tested whether these segmented units would directly inform on the perceived word segmentation of speech signals, i.e., whether LFO and HFA reflect linguistic parsing mechanisms. Specifically, the neural speech code parsed by the LFO should contain all acoustic features information necessary for the perceived word. Thus, in our experiment, the changes in percept during the bistable sequences should have been associated with temporal shifts of LFO and HFA of tens of milliseconds to capture the acoustic information of the distinct words. The tracking of speech by HFA showed latency shifts of ~ 80–150 ms, which is compatible with changes of linguistic parsing. In contrast, the observed modulations of LFO were insufficient to fully support a direct role of LFO in linguistic parsing. In monosyllabic sequences, the magnitude of the 3-Hz phase shifts was small and inconsistent with the expected extent of the phase delay that would have been expected if the acoustic speech signals were parsed on the basis of the oscillatory LFO duty cycle (Fig. 1*B*). Neither the changes in power nor the phase shifts of the 1.5-Hz neural responses could distinguish between conscious percepts in the bisyllabic conditions.

If the present LFO modulations cannot be explained by shifts in the parsing windows for speech segmentation, they may alternatively reflect an attentional modulation of acoustic processing, i.e., an enhanced neural excitability to particular acoustic features as has previously been reported for various kinds of sound stimuli (Besle et al. 2011; Cravo et al. 2013; Gomez-Ramirez et al. 2011; Lakatos et al. 2008; Rimmele et al. 2015; Schroeder and Lakatos 2009a; Stefanics et al. 2010; Zion-Golumbic et al. 2013). One possibility is that the observed top-down effects reflect the processing of nonlexical speech information relevant for speech segmentation. In particular, the neural tracking of acoustic rhythms in the 1- to 3-Hz range could be dedicated to the encoding of prosodic temporal fluctuations (Poeppel 2003), known to give reliable cues for speech parsing (Ding and Simon 2013; Greenberg et al. 2003). Alternatively or in addition to prosodic cues, delta-theta oscillations could reflect the processing of coarticulation (i.e., the overlap in the frequency spectrum of adjacent phonemes) that also provides relevant cues for word segmentation. In our design, monosyllabic sequences were composed of the word “lampe” and pseudo-word “sep”: both streams contained coarticulatory cues that are compatible with one of the two interpretations, i.e., a consonant vowel onset (“lampe” or “sep”), but not with the other interpretation in terms of a consonant cluster at the onset (“pse” or “plan”). The 3-Hz phase effects could then reflect the suppression of the irrelevant phonetic cues that would not be compatible with the perceived word. This would be consistent with recent findings showing that LFO encode phonemic information (Di Liberto et al. 2015) and that theta (3–5 Hz) oscillations are involved in phonemic restoration (Riecke et al. 2009, 2012; Strauss et al. 2014; Sunami et al. 2013).

The reported top-down effects on oscillatory activity were mostly observed in the monosyllabic word sequences. The involvement of LFO and HFA in the encoding of coarticulation could provide a first explanation for the absence of endogenous phase effects in the bisyllabic conditions, because syllabic items were pronounced independently and no coarticulation cues were favoring one or the other interpretation of these sequences. Our results could also highlight the importance of syllabic analyses (Greenberg et al. 2003; Poeppel 2003) and support the hypotheses that the brain specifically computes syllabic-like speech primitives for perception (Poeppel et al. 2008). Additional mechanisms might be required for the building up of bigger temporal speech units, e.g., when two syllabic units have to be concatenated or segregated to form a word. Frontal delta activity (Ding et al. 2016; Park et al. 2015) and fronto-parietal alpha mechanisms (Kayser et al. 2015; Shahin and Pitt 2012) may have an important role in multisyllabic word and phrase chunking. Periodical enhancement of alpha power may in particular mark the inhibition of auditory cortex activity at perceived word boundaries (Shahin and Pitt 2012) and during speech silent gaps (Kayser et al. 2015).

Origins of the Difference Between HFA and LFO Effects

Whereas the tracking of speech features by HFA was strongly influenced by perception, LFO speech tracking only showed small modulations. It could be argued that the dissociation between HFA and LFO behavior is mainly due to the coarse spatial resolution of MEG analysis and that our findings reflect the combined activity of distinct brain regions that serve dissociable mechanisms. Distinct networks may reflect acoustic processing and linguistic parsing: the neural tracking of fine-grained acoustic features would be restricted to primary auditory cortices (Kubanek et al. 2013), whereas that of phonemic or lexical information would take place in higher order regions (e.g., superior temporal sulcus and Broca’s areas) specific to speech processing (Boemio et al. 2005; Kubanek et al. 2013; Liem et al. 2014; Overath et al. 2015; Zion Golumbic et al. 2013) or attentional selection (Besle et al. 2011; Zion Golumbic et al. 2013). In other words, MEG data reported in this study may at once capture stimulus-tracking mechanisms in auditory cortices and cortical oscillators for speech parsing in higher order auditory areas (Overath et al. 2015). Interestingly, a recent report suggested that the top-down influences of HFA and LFO in linguistic processing are observed in different areas (Ding et al. 2016). Top-down language-specific processing may mainly affect HFO in superior temporal gyri and LFO in a more distributed network throughout frontal and temporal lobes. In this study, we selected activity from temporal sensors to focus our analysis on auditory cortices’ response to speech, and we may thus have primarily captured activity from regions having stronger top-down HFA effects.

Alternatively, our results suggest that during speech listening, low-frequency neural tracking may be weakly modulated by top-down word segmentation processing (Howard and

Poeppel 2010; Millman et al. 2015; Obleser et al. 2012; Peña and Melloni 2012). The small phase differences observed in LFO contrast with the large phase reversals of low-frequency entrainment reported during attentional selection (Besle et al. 2011; Gomez-Ramirez et al. 2011; Lakatos et al. 2008) and cocktail party effects (Zion Golumbic et al. 2013). These differences may be accounted for by fundamental differences in stimuli and task. In previous experiments, two distinct rhythmic inputs were competing for attentional selection, and the phase of slow oscillations reflected the dynamics of the selected sensory input (Besle et al. 2011; Gomez-Ramirez et al. 2011; Lakatos et al. 2008). The modulations of neural dynamics by attention were based on existing external temporal information, and changes in oscillatory phase may thus result from the amplification of the evoked responses to stimuli of distinct temporal profiles. Hence, the slow dynamics may have primarily reflected the gain of relevant sensory information as opposed to fundamentally providing endogenous temporal parsing mechanisms. In the present study, however, only one acoustic stream of information was provided to participants, and the contribution of gain mechanisms may be much smaller because no competing sensory inputs were physically provided to participants.

Second, and perhaps more controversial, the power fluctuations of HFA question the hypothesis of a fixed phase-amplitude coupling between slow and fast brain oscillations (Canolty et al. 2006; Canolty and Knight 2010; Schroeder and Lakatos 2009a, 2009b). A fixed phase-amplitude coupling would predict similar temporal shifts according to the conscious percept in both slow and fast oscillatory speech tracking. However, we found that speech tracking in beta-gamma amplitude predicted perception, whereas speech tracking in delta-theta phase only weakly changed as a function of the perceived speech percept. This suggests that the position of maximal beta-gamma amplitude is variable with respect to the low-frequency phase but systematic with respect to a participant's percept. As such, the relative phase of coupling could constitute a valuable code to partition neural activity for sensory processing (Hyafil et al. 2015b; Jensen et al. 2012, 2014; Lisman and Jensen 2013; Nadasdy 2010; Panzeri et al. 2010). Consistent with this, the phase of firing according to low-frequency oscillations has been shown to be a reliable decoder of sensory content (Kayser et al. 2009, 2012; Montemurro et al. 2008; Ng et al. 2013; Panzeri et al. 2010), and the relative phase of slow neural oscillations can predict perceptual features and attentional state (Agarwal et al. 2014; Bonnefond and Jensen 2012; Kösem et al. 2014; van Ede et al. 2015). Low-frequency neural oscillations could thus provide temporal metrics for sensory processing, and the entrainment of neural oscillations to external rhythms could support the extraction of timing information without a priori knowledge of external timing (Kösem et al. 2014; Scharnowski et al. 2013). We conjecture that this mechanism applies for speech processing, as well: the position of high-frequency neural oscillations in the cycle of the entrained neural oscillation may be a crucial cue for delineating temporal windows for syllabic segmentation.

ACKNOWLEDGMENTS

We thank Alexandre Gramfort for advice on data analysis and the MNE-Python community. We are grateful to the NeuroSpin infrastructure.

GRANTS

This work was supported by European Research Council Starling Grant ERC-YStG-263584 and French National Research Agency Grant ANR10JCJC-1904 (to V. van Wassenhove) and by a Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA)/Direction Génerale des Armées (DGA) fellowship (to A. Kösem).

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the authors.

AUTHOR CONTRIBUTIONS

A.K., A.B., and V.v.W. conceived and designed research; A.K. and L.A. performed experiments; A.K. and L.A. analyzed data; A.K. and V.v.W. interpreted results of experiments; A.K. and V.v.W. prepared figures; A.K. and V.v.W. drafted manuscript; A.K. and V.v.W. edited and revised manuscript; A.K., A.B., and V.v.W. approved final version of manuscript.

REFERENCES

Agarwal G, Stevenson IH, Berényi A, Mizuseki K, Buzsáki G, Sommer FT. Spatially distributed local fields in the hippocampus encode rat position. *Science* 344: 626–630, 2014.

Ahissar E, Nagarajan S, Ahissar M, Protopapas A, Mahncke H, Merzenich MM. Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proc Natl Acad Sci USA* 98: 13367–13372, 2001.

Akam T, Kullmann DM. Oscillatory multiplexing of population codes for selective communication in the mammalian brain. *Nat Rev Neurosci* 15: 111–122, 2014.

Arnal LH, Giraud AL. Cortical oscillations and sensory predictions. *Trends Cogn Sci* 16: 390–398, 2012.

Arnal LH, Wyart V, Giraud AL. Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. *Nat Neurosci* 14: 797–801, 2011.

Baldauf D, Desimone R. Neural mechanisms of object-based attention. *Science* 344: 424–427, 2014.

Basirat A, Schwartz JL, Sato M. Perceptuo-motor interactions in the perceptual organization of speech: evidence from the verbal transformation effect. *Philos Trans R Soc Lond B Biol Sci* 367: 965–976, 2012.

Bastos AM, Vezoli J, Bosman CA, Schoffelen JM, Oostenveld R, Dowdall JR, De Weerd P, Kennedy H, Fries P. Visual areas exert feedforward and feedback influences through distinct frequency channels. *Neuron* 85: 390–401, 2014.

Bastos AM, Vezoli J, Fries P. Communication through coherence with inter-areal delays. *Curr Opin Neurobiol* 31: 173–180, 2015.

Besle J, Schevon CA, Mehta AD, Lakatos P, Goodman RR, McKhann GM, Emerson RG, Schroeder CE. Tuning of the human neocortex to the temporal dynamics of attended events. *J Neurosci* 31: 3176–3185, 2011.

Billig AJ, Davis MH, Deeks JM, Monstrey J, Carlyon RP. Lexical influences on auditory streaming. *Curr Biol* 23: 1585–1589, 2013.

Boemio A, Fromm S, Braun A, Poeppel D. Hierarchical and asymmetric temporal sensitivity in human auditory cortices. *Nat Neurosci* 8: 389–395, 2005.

Boersma P. Praat, a system for doing phonetics by computer. *Glot Int* 5: 341–345, 2002.

Bonnefond M, Jensen O. Alpha oscillations serve to protect working memory maintenance against anticipated distracters. *Curr Biol* 22: 1969–1974, 2012.

Buiatti M, Peña M, Dehaene-Lambertz G. Investigating the neural correlates of continuous speech computation with frequency-tagged neuroelectric responses. *Neuroimage* 44: 509–519, 2009.

Canolty RT, Edwards E, Dalal SS, Soltani M, Nagarajan SS, Berger MS, Barbaro NM, Knight RT. High gamma power is phase-locked to theta oscillations in human neocortex. *Science* 313: 1626–1628, 2006.

Canolty RT, Knight RT. The functional role of cross-frequency coupling. *Trends Cogn Sci* 14: 506–515, 2010.

Cravo AM, Rohenkohl G, Wyart V, Nobre AC. Temporal expectation enhances contrast sensitivity by phase entrainment of low-frequency oscillations in visual cortex. *J Neurosci* 33: 4002–4010, 2013.

Di Liberto GM, O'Sullivan JA, Lalor EC. Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Curr Biol* 25: 2457–2465, 2015.

Ding N, Melloni L, Zhang H, Tian X, Poeppel D. Cortical tracking of hierarchical linguistic structures in connected speech. *Nat Neurosci* 19: 158–164, 2016.

Ding N, Simon JZ. Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. *J Neurosci* 33: 5728–5735, 2013.

Ding N, Simon JZ. Cortical entrainment to continuous speech: functional roles and interpretations. *Front Hum Neurosci* 8: 311, 2014.

Doelling KB, Arnal LH, Ghitza O, Poeppel D. Acoustic landmarks drive delta-theta oscillations to enable speech comprehension by facilitating perceptual parsing. *Neuroimage* 85: 761–768, 2014.

Fisher NI. *Statistical Analysis of Circular Data*. Cambridge, UK: Cambridge University Press, 1995.

Fontolan L, Morillon B, Liegeois-Chauvel C, Giraud AL. The contribution of frequency-specific activity to hierarchical information processing in the human auditory cortex. *Nat Commun* 5: 4694, 2014.

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

Giraud AL, Kleinschmidt A, Poeppel D, Lund TE, Frackowiak RSJ, Laufs H. Endogenous cortical rhythms determine cerebral specialization for speech perception and production. *Neuron* 56: 1127–1134, 2007.

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

Glasberg BR, Moore BC. Derivation of auditory filter shapes from notched-noise data. *Hear Res* 47: 103–138, 1990.

Gomez-Ramirez M, Kelly SP, Molholm S, Sehatpour P, Schwartz TH, Foxe JJ. Oscillatory sensory selection mechanisms during intersensory attention to rhythmic auditory and visual inputs: a human electrocorticographic investigation. *J Neurosci* 31: 18556–18567, 2011.

Gramfort A, Luessi M, Larson E, Engemann D, Strohmeier D, Brodbeck C, Goj R, Jas M, Brooks T, Parkkonen L, Hämäläinen M. MEG and EEG data analysis with MNE-Python. *Front Neurosci* 7: 267, 2013.

Gramfort A, Luessi M, Larson E, Engemann DA, Strohmeier D, Brodbeck C, Parkkonen L, Hämäläinen MS. MNE software for processing MEG and EEG data. *Neuroimage* 86: 446–460, 2014.

Greenberg S, Carvey H, Hitchcock L, Chang S. Temporal properties of spontaneous speech-a syllable-centric perspective. *J Phon* 31: 465–485, 2003.

Gross J, Hoogenboom N, Thut G, Schyns P, Panzeri S, Belin P, Garrod S. Speech rhythms and multiplexed oscillatory sensory coding in the human brain. *PLoS Biol* 11: e1001752, 2013.

Hämäläinen M, Hari R, Ilmoniemi RJ, Knuutila J, Lounasmaa OV. Magnetoencephalography-theory, instrumentation, and applications to non-invasive studies of the working human brain. *Rev Mod Phys* 65: 413–497, 1993.

Hari R, Hämäläinen M, Joutsiniemi SL. Neuromagnetic steady-state responses to auditory stimuli. *J Acoust Soc Am* 86: 1033–1039, 1989.

Henry MJ, Obleser J. Frequency modulation entrains slow neural oscillations and optimizes human listening behavior. *Proc Natl Acad Sci USA* 109: 20095–20100, 2012.

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

Hyafil A, Fontolan L, Kabdebon C, Gutkin B, Giraud AL. Speech encoding by coupled cortical theta and gamma oscillations. *Elife* 4: e06213, 2015a.

Hyafil A, Giraud A, Fontolan L, Gutkin B. Neural cross-frequency coupling: connecting architectures, mechanisms, and functions. *Trends Neurosci* 38: 725–740, 2015b.

Jensen O, Bonnefond M, VanRullen R. An oscillatory mechanism for prioritizing salient unattended stimuli. *Trends Cogn Sci* 16: 200–206, 2012.

Jensen O, Gips B, Bergmann TO, Bonnefond M. Temporal coding organized by coupled alpha and gamma oscillations prioritize visual processing. *Trends Neurosci* 37: 357–369, 2014.

Kayser C, Ince RAA, Panzeri S. Analysis of slow (theta) oscillations as a potential temporal reference frame for information coding in sensory cortices. *PLoS Comput Biol* 8: e1002717, 2012.

Kayser C, Montemurro MA, Logothetis NK, Panzeri S. Spike-phase coding boosts and stabilizes information carried by spatial and temporal spike patterns. *Neuron* 61: 597–608, 2009.

Kayser SJ, Ince RAA, Gross J, Kayser C. Irregular speech rate dissociates auditory cortical entrainment, evoked responses, and frontal alpha. *J Neurosci* 35: 14691–14701, 2015.

Kösem A, Gramfort A, van Wassenhove V. Encoding of event timing in the phase of neural oscillations. *Neuroimage* 92: 274–284, 2014.

Kubanek J, Brunner P, Gunduz A, Poeppel D, Schalk G. The tracking of speech envelope in the human cortex. *PLoS One* 8: e53398, 2013.

Lakatos P, Karmos G, Mehta AD, Ulbert I, Schroeder CE. Entrainment of neuronal oscillations as a mechanism of attentional selection. *Science* 320: 110–113, 2008.

Lakatos P, Shah AS, Knuth KH, Ulbert I, Karmos G, Schroeder CE. An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. *J Neurophysiol* 94: 1904–1911, 2005.

Liem F, Hurschler MA, Jäncke L, Meyer M. On the planum temporale lateralization in suprasegmental speech perception: Evidence from a study investigating behavior, structure, and function. *Hum Brain Mapp* 35: 1779–1789, 2014.

Lisman JE, Jensen O. The θ-γ neural code. *Neuron* 77: 1002–1016, 2013.

Lopes da Silva F. EEG and MEG: relevance to neuroscience. *Neuron* 80: 1112–1128, 2013.

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

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

Maddieson I. Phonetic cues to syllabification. *UCLA Working Papers in Phonetics* 59: 85–101, 1984.

Maris E, Oostenveld R. Nonparametric statistical testing of EEG- and MEG-data. *J Neurosci Methods* 164: 177–190, 2007.

Mattys SL, White L, Melhorn JF. Integration of multiple speech segmentation cues: a hierarchical framework. *J Exp Psychol Gen* 134: 477–500, 2005.

Mesgarani N, Chang EF. Selective cortical representation of attended speaker in multi-talker speech perception. *Nature* 485: 233–236, 2012.

Mesgarani N, Cheung C, Johnson K, Chang E. Phonetic feature encoding in human superior temporal gyrus. *Science* 343: 1006–1010, 2014.

Millman RE, Johnson SR, Prendergast G. The role of phase-locking to the temporal envelope of speech in auditory perception and speech intelligibility. *J Cogn Neurosci* 27: 533–545, 2015.

Millman RE, Prendergast G, Hymers M, Green GGR. 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, 2013.

Montemurro MA, Rasch MJ, Murayama Y, Logothetis NK, Panzeri S. Phase-of-firing coding of natural visual stimuli in primary visual cortex. *Curr Biol* 18: 375–380, 2008.

Nadasdy Z. Binding by asynchrony: the neuronal phase code. *Front Neurosci* 4: 51, 2010.

Ng B, Logothetis N, Kayser C. EEG phase patterns reflect the selectivity of neural firing. *Cereb Cortex* 23: 389–398, 2013.

Ng B, Schroeder T, Kayser C. A precluding but not ensuring role of entrained low-frequency oscillations for auditory perception. *J Neurosci* 32: 12268–12276, 2012.

Nourski KV, Reale RA, Oya H, Kawasaki H, Kovach CK, Chen H, Howard MA, Brugge JF. Temporal envelope of time-compressed speech represented in the human auditory cortex. *J Neurosci* 29: 15564–15574, 2009.

Nozaradan S, Peretz I, Missal M, Mouraux A. Tagging the neuronal entrainment to beat and meter. *J Neurosci* 31: 10234–10240, 2011.

Obleser J, Herrmann B, Henry MJ. Neural oscillations in speech: don't be enslaved by the envelope. *Front Hum Neurosci* 6: 250, 2012.

Overath T, McDermott JH, Zarate JM, Poeppel D. The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nat Neurosci* 18: 903–911, 2015.

Panzeri S, Brunel N, Logothetis NK, Kayser C. Sensory neural codes using multiplexed temporal scales. *Trends Neurosci* 33: 111–120, 2010.

Park H, Ince RA, Schyns PG, Thut G, Gross J. Frontal top-down signals increase coupling of auditory low-frequency oscillations to continuous speech in human listeners. *Curr Biol* 25: 1649–1653, 2015.

Parkkonen L, Andersson J, Hämäläinen M, Hari R. Early visual brain areas reflect the percept of an ambiguous scene. *Proc Natl Acad Sci USA* 105: 20500–20504, 2008.

Pasley BN, David SV, Mesgarani N, Flinker A, Shamma SA, Crone NE, Knight RT, Chang EF. Reconstructing speech from human auditory cortex. *PLoS Biol* 10: e1001251, 2012.

Peelle JE, Davis MH. Neural oscillations carry speech rhythm through to comprehension. *Front Psychol* 3: 320, 2012.

Peelle JE, Gross J, Davis MH. Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cereb Cortex* 23: 1378–1387, 2013.

Peña M, Melloni L. Brain oscillations during spoken sentence processing. *J Cogn Neurosci* 24: 1149–1164, 2012.

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

Poeppel D, Idsardi WJ, van Wassenhove V. Speech perception at the interface of neurobiology and linguistics. *Philos Trans R Soc Lond B Biol Sci* 363: 1071–1086, 2008.

Rees A, Green GGR, Kay RH. Steady-state evoked responses to sinusoidally amplitude-modulated sounds recorded in man. *Hear Res* 23: 123–133, 1986.

Riecke L, Esposito F, Bonte M, Formisano E. Hearing illusory sounds in noise: the timing of sensory-perceptual transformations in auditory cortex. *Neuron* 64: 550–561, 2009.

Riecke L, Vanbussel M, Hausfeld L, Bas¸kent D, Formisano E, Esposito F. Hearing an illusory vowel in noise: suppression of auditory cortical activity. *J Neurosci* 32: 8024–8034, 2012.

Rimmele JM, Zion Golumbic E, Schröger E, Poeppel D. The effects of selective attention and speech acoustics on neural speech-tracking in a multi-talker scene. *Cortex* 68: 144–154, 2015.

Sato M, Basirat A, Schwartz JL. Visual contribution to the multistable perception of speech. *Percept Psychophys* 69: 1360–1372, 2007.

Sato M, Schwartz JL, Abry C, Cathiard MA, Loevenbruck H. Multistable syllables as enacted percepts: a source of an asymmetric bias in the verbal transformation effect. *Percept Psychophys* 68: 458–474, 2006.

Scharnowski F, Rees G, Walsh V. Time and the brain: neurorelativity: the chronoarchitecture of the brain from the neuronal rather than the observer's perspective. *Trends Cogn Sci* 17: 51–52, 2013.

Schroeder CE, Lakatos P. Low-frequency neuronal oscillations as instruments of sensory selection. *Trends Neurosci* 32: 9–18, 2009a.

Schroeder CE, Lakatos P. The gamma oscillation: master or slave? *Brain Topogr* 22: 24–26, 2009b.

Shahin AJ, Pitt MA. Alpha activity marking word boundaries mediates speech segmentation. *Eur J Neurosci* 36: 3740–3748, 2012.

Stefanics G, Hangya B, Hernadi I, Winkler I, Lakatos P, Ulbert I. Phase entrainment of human delta oscillations can mediate the effects of expectation on reaction speed. *J Neurosci* 30: 13578–13585, 2010.

Stevens KN. Toward a model for lexical access based on acoustic landmarks and distinctive features. *J Acoust Soc Am* 111: 1872–1891, 2002.

Strauss A, Kotz SA, Scharinger M, Obleser J. Alpha and theta brain oscillations index dissociable processes in spoken word recognition. *Neuroimage* 97: 387–395, 2014.

Sunami K, Ishii A, Takano S, Yamamoto H, Sakashita T, Tanaka M, Watanabe Y, Yamane H. Neural mechanisms of phonemic restoration for speech comprehension revealed by magnetoencephalography. *Brain Res* 1537: 164–173, 2013.

Tallon-Baudry C, Bertrand O, Delpuech C, Pernier J. Stimulus specificity of phase-locked and non-phase-locked 40 Hz visual responses in human. *J Neurosci* 16: 4240–4249, 1996.

Taulu S, Kajola M, Simola J. Suppression of interference and artifacts by the signal space separation method. *Brain Topogr* 16: 269–275, 2003.

Ten Oever S, Sack AT. Oscillatory phase shapes syllable perception. *Proc Natl Acad Sci USA* 112: 15833–15837, 2015.

Tesche CD, Uusitalo MA, Ilmoniemi RJ, Huotilainen M, Kajola M, Salonen O. Signal-space projections of MEG data characterize both distributed and well-localized neuronal sources. *Electroencephalogr Clin Neurophysiol* 95: 189–200, 1995.

Thut G, Schyns PG, Gross J. Entrainment of perceptually relevant brain oscillations by non-invasive rhythmic stimulation of the human brain. *Front Psychol* 2: 170, 2011.

Uusitalo MA, Ilmoniemi RJ. Signal-space projection method for separating MEG or EEG into components. *Med Biol Eng Comput* 35: 135–140, 1997.

van Ede F, van Pelt S, Fries P, Maris E. Both ongoing alpha and visually induced gamma oscillations show reliable diversity in their across-site phase-relations. *J Neurophysiol* 113: 1556–1563, 2015.

Vrba J. Magnetoencephalography: the art of finding a needle in a haystack. *Physica C Supercond* 368: 1–9, 2002.

Warren RM. Verbal transformation effect and auditory perceptual mechanisms. *Psychol Bull* 70: 261–270, 1968.

Zion Golumbic EM, Ding N, Bickel S, Lakatos P, Schevon CA, McKhann GM, Goodman RR, Emerson R, Mehta AD, Simon JZ, Poeppel D, Schroeder CE. Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". *Neuron* 77: 980–991, 2013.

Zoefel B, VanRullen R. EEG oscillations entrain their phase to high-level features of speech sound. *Neuroimage* 124: 16–23, 2015.