Articles in PresS. J Neurophysiol (September 7, 2016). doi:10.1152/jn.00074.2016

Title: High frequency neural activity predicts word parsing in ambiguous speech streams

Authors: Anne Kösem[1-3*], Anahita Basirat[1,4], Leila Azizi[1], Virginie van Wassenhove[1]

Affiliations:

[1] Cognitive Neuroimaging Unit, CEA DRF/I2BM, INSERM, Université Paris-Sud, Université Paris-Saclay, 91191 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

[4] SCALab, CNRS UMR 9193; Univ Lille, 59045 Lille France

*corresponding author :

Donders Institute for Cognitive Neuroimaging

Kapittelweg 29 6525 EN Nijmegen

kosem.anne@gmail.com

ABSTRACT

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 propose 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. While changes in low-frequency neural oscillations were compatible with the encoding of pre-lexical segmentation cues, high-frequency activity specifically informed on an individual's conscious speech percept.

New & Noteworthy: A critical problem the brain faces when analyzing speech is how to parse a continuous stream of information into relevant linguistic units. Using 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.

Keywords: speech segmentation, neural entrainment, bistability, MEG, phase

1. INTRODUCTION

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 neuro-computational models of speech processing, which describe brain oscillations as active parsers of the auditory speech signals (Ding & Simon 2014; Ghitza 2011; Giraud & Poeppel 2012; Hyafil et al. 2015; Peelle & 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; Luo and Poeppel, 2007, 2012; Peelle and Davis, 2012; Ding and Simon, 2013a; Gross et al., 2013; Millman et al., 2013; Zion Golumbic et al., 2013; Doelling et al., 2014; Rimmele et al., 2015). Second, high-frequency neural activities including the beta (20-30 Hz) and the gamma (> 40 Hz) bands, have been hypothesized to encode the fine-grained properties of the speech signal such as phonetic features (Poeppel, 2003; Poeppel et al., 2008; Ghitza, 2011; Giraud and Poeppel, 2012).

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; Giraud & Poeppel 2012; Ghitza 2011; Hyafil et al. 2015) 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 (Giraud and Poeppel, 2012; Doelling et al., 2014). 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 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 (Lakatos et al., 2008; Schroeder and Lakatos, 2009a; Stefanics et al., 2010; Besle et al., 2011; Gomez-Ramirez et al., 2011; Cravo et al., 2013). In complex auditory environments, the control of neural oscillations by attention has been shown to be beneficial for speech processing as well (Zion Golumbic et al., 2013; Rimmele et al., 2015), 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 like 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, 2013a; Gross et al., 2013; Peelle et al., 2013; Doelling et al., 2014; 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 (Peña and Melloni, 2012; Millman et al., 2015; 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 (Giraud and Poeppel, 2012; Ding and Simon, 2014; 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 & Kullmann 2014; Canolty et al. 2006; Canolty & Knight 2010; Hyafil et al. 2015; 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 (Giraud and Poeppel, 2012; Ding and Simon, 2014; 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' (Warren, 1968; Sato et al., 2006, 2007; Basirat et al., 2012; Billig et al., 2013) (Figure 1A). 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" ([lɑ̃p]) or "plan" ([plɑ̃]), 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 (Figure 1A). As 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 (Figure 1B). 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 et al. 2014; Giraud & Poeppel 2012; Hyafil et al. 2015), 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 (Figure 1B). 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 will be

discussed in detail.

FIGURE 1 ABOUT HERE

2. MATERIALS AND METHODS

2.1. Participants

20 participants (8 females, mean age: 23 years old) 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. Prior to 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 MEG recordings (rejection after visual inspection of MEG raw data, prior to MEG analysis); one participant did not finish the task; two participants did not perform correctly the 'volitional' verbal transformation task, as they could not voluntarily hear the required percept (<10 % report) in at least one of the sequences. Hence, 15 participants (5 females, mean age: 23 years old) were considered for the reported analysis.

2.2. Experimental paradigm

Stimuli

Four auditory sequences (adapted from (Sato et al., 2007; Basirat et al., 2012)) were presented binaurally to participants via Etymotic Earphones (Etymotic Research Inc., USA) 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” ([səp]) (leading to the alternative percept “pse” ([psə])), or bisyllabic “pata” ([pata]) (that 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.K.). 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 3Hz (1 syllable per 333 ms). All syllables had equalized sound-levels based on RMS. The selected syllables were assembled to form the word “képi” and the pseudo-word “pata”, 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 of 333 ms, leading to a repetition rate of 1.5 Hz in bisyllabic sequences and 3 Hz in monosyllabic sequences.

Procedure

First, and prior to the main experiment, participants where familiarized with the stimuli using 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 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 counter-balanced for each presentation. Hence, in one presentation, participants were asked to maintain the target “képi” or “pata” or “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 button 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 (labelled “other”) was used to report any other percepts participants might have experienced during the sequence.

2.3. 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 (Neuromag Elekta LTD, Helsinki) equipped with 102 triple-sensor elements (two orthogonal planar gradiometers and one magnetometer per sensor location). Shielding against environmental noise was provided by MaxShield (Neuromag Elekta LTD, Helsinki). 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 mastoïd areas. MEG recordings were sampled at 1 kHz and online band-pass filtered between 0.03 Hz 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 2s 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 insure 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, approximately 23% ± 9.6% S.D. 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, black dots). A spatial filter was used for channel averaging in each region of interest. The spatial filter was estimated based on 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 Hz and 40 Hz. The comparisons of evoked responses between conditions were computed using a non-parametric 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 Figure 2 the topography of the 3Hz component of the evoked response. To do so, 3Hz evoked amplitude was estimated 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 in order to compute the Power Spectrum Density (PSD) using a Welch's average periodogram method for each experimental condition (perceived speech), 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", "sep"), reported percept (2 levels: "percept 1" and "percept 2", e.g. "lampe" and "plan"), and hemisphere (2 levels: right, left).

Phase analysis

The phase of the 3 Hz and 1.5 Hz entrained oscillatory responses and the Phase- Locking Value (PLV, also called Phase Locking Factor or Inter Trial 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 θ(t,k) is the instantaneous phase at time t and trial k. Single trial MEG data (2s epochs centered on the plosive) were convolved with a 3-cycles Morlet wavelet at 1.5 Hz frequency for the computation of the 1.5 Hz preferential phase and a 4-cycles 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 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 which models the passage of the signal through the cochlea (Glasberg and Moore, 1990; Ghitza, 2011). The filter bank was designed as a set of parallel band-pass FIR filters, each tuned to a different frequency. The center frequencies of interest were chosen from 250 and 3000 Hz. The center frequency *f* and the critical bandwidth of each filter were computed following Glasberg & 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 Hz to 140 Hz (in 2 Hz steps) was computed on the 2s-length epochs on a per trial basis, by filtering the MEG signal for each frequency band (FIR filter, bandwidth of ±2 Hz for frequencies below 20 Hz, bandwidth of ±5 Hz for frequencies above or equal to 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 of [2.5, 3.5 Hz], in mono- and bisyllabic sequences) and 1.5 Hz (FIR filter, bandwidth of [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 Hz and 1.5 Hz speech envelopes over trials. For each individual, we tested if the cross-correlograms significantly differed between the percept conditions, using spectro-temporal 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 below 0.05 and were clustered based on spectro-temporal adjacency. The sum of the F-values within a cluster was used as the cluster-level statistics. The reference distribution for cluster level statistics was computed by performing 1000 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 below 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 either originate 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 Hz -30 Hz], gamma [40 Hz - 80 Hz] and high- gamma [90 Hz – 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.

3. RESULTS

3.1. 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 they heard a different percept. 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. 1C). Although post-hoc analysis showed that the percept "pse" was significantly harder to maintain as compared to other percepts (64% maintenance, against 84% to 96% in the other conditions), it remained significantly dominant compared to 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 Sato et al. 2007; Basirat et al., 2012).

3.2. 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 in order 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. 2A).

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, as the acoustic signal was identical for both percept conditions. Alternatively, if linguistic parsing modulates LFO, substantial temporal shifts in the LFO should be observed. Here, 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. 1B). 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 in order 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. 3Hz 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”, $p = 0.04$).

TABLE 1 ABOUT HERE

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 % C.I. = [-15.5°, -0.3°]) as compared to “plan” (Fig. 2B). Similarly, a phase advance of -9° (95 % C.I. = [-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 3Hz were observed between the different percepts of the bisyllabic sequences (Fig. 2B-C). Hence, in both monosyllabic sequences, the perceptual outcomes were associated with phase shifts of the 3 Hz response (Fig. 2B). While these phase shifts were consistent across monosyllabic conditions, the confidence intervals of the differences were nevertheless close to zero and included zero when using Bonferroni correction 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 locked 50 ms and 100 ms before or after plosive onset and we replicated the main observations.

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 perceiving the word “plan” was delayed by 9 to 10 ms as compared to 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 to 80 ms corresponding to phase shifts in neural entrainment of at least 55 to 85°. As 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.

FIGURE 2 ABOUT HERE

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. 3A). Specifically, what appeared as the sub-harmonic 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 listening to bisyllabic speech utterances as compared to monosyllabic ones (Fig. 3A-B). An analysis of variance 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. 3B), 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 in (Nozaradan et al., 2011; Kösem et al., 2014).

FIGURE 3 ABOUT HERE

There are two possible origins for the observation that bisyllabic words induce significant 1.5 Hz auditory responses (Fig. 3 A-B). First, the 1.5 Hz response could be elicited by a sub-harmonic 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 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 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 may also be involved in the encoding of abstract linguistic structures (Buiatti et al., 2009; Ding et al., 2016).

FIGURE 4 ABOUT HERE

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 contrasting the two perceptual outcomes of the same bisyllabic sequences (main effect of percept $F[1,14] < 1$, interaction between percept and sequence $F[1,14] = 2.6$, $p = 0.12$, Table 2). Non-significant phase shifts were observed between the two perceptual outcomes (Fig. 3C). 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 as syllables composing the word were 333 ms apart (Fig. 1B) but this is not what we observed.

TABLE 2 ABOUT HERE

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 (3Hz) , but they were too small to account for a shift of the linguistic parsing window. Second, the sub-harmonic 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.

3.3. 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 et al. 2014; Giraud & Poeppel 2012; Hyafil et al. 2015), 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. In order 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 Hz to 140 Hz. Speech-neural response cross-correlograms have previously been used (Gross et al., 2013; Fontolan et al., 2014) 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. Here, 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 spectro-temporal 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. 5A) 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. 5A). Crucially, the significant changes in cross-correlation were related to strong latency differences as reflected by the phase shifts of the 3Hz cross-correlograms between perceiving “plan” and “lampe” (Fig. 5B-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. 5C). Additionally, while strong phase oppositions in 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 participant p04 and participant p05 (Fig. 5C).

FIGURE 5 ABOUT HERE

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 (Figure 5C). 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 (Figure 6A). Significant differences were concentrated

557 across participants in the beta band (12-30 Hz, 12 out of 15 participants presented significant
558 clusters within this range for both monosyllabic sequences), in the gamma band (40-80 Hz, 12
559 participants presented significant clusters for the “lampe” sequences, and 13 in for the “sep”
560 sequences) and the high gamma band (90-130 Hz, 14 participants with significant cluster in
561 “lampe” sequences, and 11 in “sep” sequences) (Fig. 6A).
562
563 **FIGURE 6 ABOUT HERE**
564
565 The observed significant differences in cross-correlation could either originate from a change
566 in the strength of speech-brain correlation, or from a change in latency between speech-brain
567 correlated dynamics. To test these two alternative accounts, we restricted the analyses to an
568 individual’s clusters in classical frequency ranges in the beta (12-30 Hz), gamma (40-80 Hz),
569 and high-gamma (90–130 Hz) frequency bands (Lopes da Silva, 2013). The analyses showed
570 significant latency differences between percepts that were consistent across participants (Fig.
571 6B-C). In contrast, no significant changes in the maximum value of the cross-correlation were
572 observed between percept conditions (Fig. 7). This suggests that the tracking of the speech
573 envelope by HFA was operated at distinct latencies between percept conditions, while the
574 amount of coupling between the speech envelope and HFA dynamics remained constant.
575 As discussed earlier, the distance between consonants is of 80 ms for “lampe”, 50 ms for
576 “sep”. The silence duration prior to the plosive is of 90 ms for “lampe”, 70 ms for “sep”.
577 Thus, the switch from the percept “lampe” to the percept “plan” could be performed via a
578 shift in the linguistic parsing window of 80 ms minimum and up to 170 ms (184 degrees). The
579 switch from the percept “sep” to the percept “pse” could occur through a temporal shift of the
580 linguistic parsing window up to 120 ms (130 degrees). These estimated shifts fall within the
581 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. 6A) and inconsistent across participants.

FIGURE 7 ABOUT HERE

Overall, while 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.

3.4. 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 may have also 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 (three possibilities: “percept 1”, “percept 2”, or

606 "other"). As in prior findings, participants did not report hearing the two percepts in equal
607 proportions but rather mostly reported hearing the initial veridical word repeated in the
608 sequence (Sato et al. 2007; Basirat et al. 2012). Nevertheless, and with many cautionary steps,
609 we performed a comparable analysis as for the volitional data. Data of participants that had
610 too strong a perceptual bias (i.e. one of the percepts were reported less than 20% of the time)
611 were rejected. Hence, 12 participants were included in the "képi"/ "piquer" analyses, 8
612 participants in the "pata"/ "tapa" analyses, and 8 participants in the "plan"/ "lampe" analyses.
613 Only 3 participants could be included for the "pse"/"sep" analyses and we thus only report
614 this condition for illustration purposes.

615 As previously, we analyzed the 3 Hz and 1.5 Hz phase shifts between percept conditions. We
616 also performed the cross-correlograms for the "plan"/"lampe" sequences in the spontaneous
617 conditions, which we illustrate below. As can be seen, we obtain similar results as with the
618 volitional task as reported in Fig. 2, 3, and 5. First, small 3 Hz phase shifts were observed for
619 the contrast "plan" – "lampe" (Fig. 8A). The phase shift was of the same amplitude and
620 direction than in the volitional task. This replication could then be interpreted in favor of a
621 consistent (but weak) top-down modulation 3 Hz oscillatory activity. Second, no significant
622 1.5 Hz phase shifts were observed for bisyllabic sequences (Fig. 8B). Third, significant
623 changes in 3Hz-modulated HFA activity were observed between "plan" and "lampe"
624 conditions for participants who were included in the analysis (p04 and p05, Fig. 8C). Though
625 the results should be interpreted with caution due to the small number of participants, our
626 results suggest that the main effects of LFO and HFA reported in the volitional task are
627 comparable with those seen in the spontaneous task. Hence, this control suggests that the
628 observed effects reflect genuine linguistic parsing processes, and cannot be easily confounded
629 by participants' cognitive strategy.

630

FIGURE 8 ABOUT HERE

4. DISCUSSION

Our results show an endogenous control of high-frequency activity (HFA) when listening 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 (Poeppel et al. 2008; Poeppel 2003; Ghitza 2011; Giraud & Poeppel 2012; Ding & Simon 2014) and show potential dissociable roles of HFA and LFO in the parsing of acoustic information into discrete linguistic content.

4.1. 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 (Rees et al., 1986; Hari, 1989; 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 are also subject to endogenous control. Specifically, the 3 Hz phase response in monosyllabic speech sequences differed between the two percepts.

653 The phase shifts were weak but consistent in direction and strength across the volitional and
654 spontaneous tasks. Consistent with this, recent findings have shown that the phase of LFO
655 entrainment in auditory cortices can be modulated when timing is relevant to the task (Kösem
656 et al., 2014; Ten Oever and Sack, 2015) and can be under top-down control (Lakatos et al.,
657 2008; Parkkonen et al., 2008; Stefanics et al., 2010; Gomez-Ramirez et al., 2011; Cravo et al.,
658 2013; Baldauf and Desimone, 2014; Park et al., 2015). Furthermore, our results and others
659 (Ten Oever and Sack, 2015) suggest that the phase of LFO correlates with perceptual speech
660 reports. The observed right hemispheric bias in our data could be linked to empirical
661 observations that the right hemisphere is more sensitive to slow speech fluctuations than the
662 left hemisphere (Poeppel, 2003; Boemio et al., 2005; Giraud et al., 2007), though the
663 lateralization of the phase effects has not been explicitly tested and is beyond the scope of this
664 study. During bisyllabic sequences, an additional 1.5 Hz neural response was found.
665 Although we cannot exclude the possibility that 1.5 Hz responses mainly reflect the acoustic
666 tracking of the speech signals, the presence of this oscillatory response is in line with previous
667 studies suggesting that delta power can be subject to top-down control during sound
668 processing (Nozaradan et al., 2011) and speech analysis (Buiatti et al., 2009; Ding et al.,
669 2016; Park et al., 2015) and does not solely reflect the temporal structure of the acoustic
670 signals.

671 Whereas the evidence for top-down modulation for LFO was rather weak, the temporal
672 alignment between the speech signals and the amplitude of HFA displayed systematic
673 latencies or phase shifts as a function of the perceived word. This observation was predicted
674 by speech models (Giraud & Poeppel 2012; Hyafil et al. 2015) as discussed in the
675 introductory section (cf. also Figure 1B). These effects were systematic within individuals,
676 concentrated in the beta, gamma and high-gamma frequency bands. Gamma oscillations are
677 markers of neural excitability (Lakatos et al., 2005) and HFA has more generally been shown

678 to track the dynamics of speech (Gross et al. 2013; Hyafil et al. 2015; Kubanek et al. 2013;
679 Mesgarani & Chang 2012; Millman et al. 2013; Nourski et al. 2009; Pasley et al. 2012;
680 Zion Golumbic et al. 2013). More specifically, the gamma band has been hypothesized to
681 encode speech information at the phonemic level (Poeppel, 2003; Poeppel et al., 2008). We
682 thus expected gamma activity to be largely indicative of the perceived word during bistable
683 speech perception. To the best of our knowledge, the implication of the beta band in speech
684 tracking has not yet been empirically reported although beta activity has been theoretically
685 posited for chunking dyads, i.e. speech units of 50 ms duration representing the transition
686 between pairs of phones (Ghitza, 2011). In addition, beta and gamma neural responses are
687 known dissociable markers of top-down and bottom-up communication (Arnal et al., 2011;
688 Arnal and Giraud, 2012; Bastos et al., 2014, 2015; Fontolan et al., 2014). Gamma responses
689 are typically reported as feedforward signals whereas beta activity has typically been
690 associated with feedback signaling. In our experiment, the fluctuations of beta (resp. gamma)
691 amplitude could potentially reflect the temporal alternation of feedback (resp. feedforward)
692 information transfer as shown in a recent report (Fontolan et al., 2014). In their study,
693 Fontolan and colleagues (2014) used natural speech stimuli and showed that the transition
694 between bottom-up information transfer via gamma activity and top-down communication via
695 beta channels occurred at 1-3 Hz during listening. Their observation was consistent with the
696 idea that speech information propagates along the processing hierarchy and back by units of
697 syllabic/word length. In this scenario, speech tracking by HFA may not only increase the
698 sensitivity to incoming acoustic information, but also reflect the linguistic parsing at the
699 syllabic scale.

700

701 **4.2. Brain oscillatory mechanisms of linguistic parsing**

702

Speech models suggest that LFO chunk the encoding of speech (indexed by HFA) into discrete informational units. Here, 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. 1B). Neither the changes in power nor the phase shifts of the 1.5Hz 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 could 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 (Schroeder & Lakatos 2009a; Lakatos et al. 2008; Gomez-Ramirez et al. 2011; Besle et al. 2011; Stefanics et al. 2010; Cravo et al. 2013; Zion-Golumbic et al., 2013 ; Rimmele et al. 2015). One possibility is that the observed top-down effects reflect the processing of non-lexical speech information relevant for speech segmentation. In particular, the neural tracking of acoustic rhythms in the 1-3 Hz range could be dedicated to the encoding of prosodic temporal fluctuations (Poeppel, 2003), known to give reliable cues for speech parsing (Greenberg et al., 2003; Ding and

Simon, 2013a). 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; Sunami et al., 2013; Strauß et al., 2014).

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, as 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 (Shahin and Pitt, 2012; Kayser et al., 2015) mechanisms may have an important role in multi-syllabic word and phrase chunking. Periodical enhancement of alpha power may in particular mark the inhibition of auditory cortex activity

753 at perceived word boundaries (Shahin and Pitt, 2012), and during speech silent gaps (Kayser
754 et al., 2015).

755

756 **4.3. Origins of the difference between HFA and LFO effects**

757

758 While the tracking of speech features by HFA was strongly influenced by perception, LFO
759 speech tracking only showed small modulations. It could be argued that the dissociation
760 between HFA and LFO behavior is mainly due to the coarse spatial resolution of MEG
761 analysis, and that our findings reflect the combined activity of distinct brain regions that serve
762 dissociable mechanisms. Distinct networks may reflect acoustic processing and linguistic
763 parsing: the neural tracking of fine-grained acoustic features would be restricted to primary
764 auditory cortices (Kubanek et al., 2013) whereas that of phonemic or lexical information
765 would take place in higher order regions (e.g. Superior Temporal Sulcus and Broca's areas)
766 specific to speech processing (Boemio et al., 2005; Kubanek et al., 2013; Zion Golumbic et
767 al., 2013; Liem et al., 2014; Overath et al., 2015) or attentional selection (Besle et al., 2011;
768 Zion Golumbic et al., 2013). In other words, MEG data reported here may at once capture
769 stimulus-tracking mechanisms in auditory cortices and cortical oscillators for speech parsing
770 in higher order auditory areas (Overath et al., 2015). Interestingly, a recent report suggests
771 that the top-down influences of HFA and LFO in linguistic processing are observed in
772 different areas (Ding et al., 2016). Top-down language-specific processing may mainly affect
773 HFO in Superior Temporal Gyri, and LFO in a more distributed network throughout frontal
774 and temporal lobes. Here, we selected activity from temporal sensors to focus our analysis on
775 auditory cortices' response to speech, we may thus have primarily captured activity from
776 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; Obleser et al., 2012; Peña and Melloni, 2012; Millman et al., 2015). The small phase differences observed in LFO contrast with the large phase reversals of low-frequency entrainment reported during attentional selection (Lakatos et al., 2008; Besle et al., 2011; Gomez-Ramirez et al., 2011) 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 (Lakatos et al., 2008; Besle et al., 2011; Gomez-Ramirez et al., 2011). 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. Here however, only one acoustic stream of information was provided to participants and the contribution of gain mechanisms may be much smaller since 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; Schroeder and Lakatos, 2009a, 2009b; Canolty and Knight, 2010). 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. 2015; Jensen et al. 2012; Jensen et al. 2014; Lisman & 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 (Montemurro et al., 2008; Kayser et al., 2009, 2012; Panzeri et al., 2010; Ng et al., 2013), and the relative phase of slow neural oscillations can predict perceptual features and attentional state (Bonnefond and Jensen, 2012; Agarwal et al., 2014; 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 (Scharnowski et al., 2013; Kösem et al., 2014). 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

This work was supported by an ERC-YStG-263584 and ANR10JCJC-1904 to V.vW, and by a DGA-CEA fellowship to A.K. The authors declare no competing financial interest. We are grateful to the NeuroSpin infrastructure. We thank Alexandre Gramfort for his advice on data analysis and the mne-python community.

825 # REFERENCES

826 **Agarwal G**, **Stevenson IH**, **Berényi A**, **Mizuseki K**, **Buzsáki G**, **Sommer FT**. Spatially
827 distributed local fields in the hippocampus encode rat position. *Science* 344: 626–30, 2014.

828 **Ahissar E**, **Nagarajan S**, **Ahissar M**, **Protopapas A**, **Mahncke H**, **Merzenich MM**. Speech
829 comprehension is correlated with temporal response patterns recorded from auditory cortex.
830 *Proc Natl Acad Sci U S A* 98: 13367–72, 2001.

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

833 **Arnal LH**, **Giraud A-L**. Cortical oscillations and sensory predictions. *Trends Cogn Sci* 16:
834 390–8, 2012.

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

837 **Baldauf D**, **Desimone R**. Neural mechanisms of object-based attention. *Science* 344: 424–7,
838 2014.

839 **Basirat A**, **Schwartz J-L**, **Sato M**. Perceptuo-motor interactions in the perceptual
840 organization of speech: evidence from the verbal transformation effect. *Philos Trans R Soc*
841 *Lond B Biol Sci* 367: 965–76, 2012.

842 **Bastos AM**, **Vezoli J**, **Bosman CA**, **Schoffelen J-M**, **Oostenveld R**, **Dowdall JR**,
843 **De Weerd P**, **Kennedy H**, **Fries P**. Visual Areas Exert Feedforward and Feedback Influences
844 through Distinct Frequency Channels. *Neuron* 85: 390–401, 2014.

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

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

850 **Billig AJ**, **Davis MH**, **Deeks JM**, **Monstrey J**, **Carlyon RP**. Lexical influences on auditory
851 streaming. *Curr Biol* 23: 1585–9, 2013.

852 **Boemio A**, **Fromm S**, **Braun A**, **Poeppel D**. Hierarchical and asymmetric temporal
853 sensitivity in human auditory cortices. *Nat Neurosci* 8: 389–95, 2005.

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

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

857 **Buiatti M**, **Peña M**, **Dehaene-Lambertz G**. Investigating the neural correlates of continuous
858 speech computation with frequency-tagged neuroelectric responses. *Neuroimage* 44: 509–19,
859 2009.

860 **Canolty RT**, **Edwards E**, **Dalal SS**, **Soltani M**, **Nagarajan SS**, **Berger MS**, **Barbaro NM**,
861 **Knight RT**. High gamma power is phase-locked to theta oscillations in human neocortex.
862 *Science (80-)* 313: 1626–1628, 2006.

863 **Canolty RT**, **Knight RT**. The functional role of cross-frequency coupling. *Trends Cogn Sci*
864 14: 506–15, 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: 4002–10, 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: 5728–35, 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 Pt 2: 761–8, 2014.

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: 1556–63, 2015.

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

Fontolan L, **Morillon B**, **Liegeois-Chauvel C**, **Giraud A-L**. 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 A-L, **Kleinschmidt A**, **Poeppel D**, **Lund TE**, **Frackowiak RSJ**, **Laufs H**. Endogenous cortical rhythms determine cerebral specialization for speech perception and production. *Neuron* 56: 1127–34, 2007.

Giraud A-L, **Poeppel D**. Cortical oscillations and speech processing: emerging computational principles and operations. *Nat Neurosci* 15: 511–7, 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, 2013a.

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*. .

Greenberg S, **Carvey H**, **Hitchcock L**, **Chang S**. Temporal properties of spontaneous

907 speech—a syllable-centric perspective. *J Phon* 31: 465–485, 2003.

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

911 **Hämäläinen M**, **Hari R**, **Ilmoniemi RJ**, **Knuutila J**, **Lounasmaa O V.**
912 Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies
913 of the working human brain. *Rev Mod Phys* 65: 413–497, 1993.

914 **Hari R**. Neuromagnetic steady-state responses to auditory stimuli. *J Acoust Soc Am* 86: 1033,
915 1989.

916 **Henry MJ**, **Obleser J**. Frequency modulation entrains slow neural oscillations and optimizes
917 human listening behavior. *Proc Natl Acad Sci U S A* 109: 20095–100, 2012.

918 **Howard MF**, **Poeppel D**. Discrimination of speech stimuli based on neuronal response phase
919 patterns depends on acoustics but not comprehension. *J Neurophysiol* 104: 2500–11, 2010.

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

922 **Hyafil A**, **Giraud A**, **Fontolan L**, **Gutkin B**. Neural Cross-Frequency Coupling: Connecting
923 Architectures, Mechanisms, and Functions. *Trends Neurosci.* .

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

926 **Jensen O**, **Gips B**, **Bergmann TO**, **Bonnefond M**. Temporal coding organized by coupled
927 alpha and gamma oscillations prioritize visual processing. *Trends Neurosci* 37: 357–369,
928 2014.

929 **Kayser C**, **Ince RAA**, **Panzeri S**. Analysis of slow (theta) oscillations as a potential temporal
930 reference frame for information coding in sensory cortices. *PLoS Comput Biol* 8: e1002717,
931 2012.

932 **Kayser C**, **Montemurro MA**, **Logothetis NK**, **Panzeri S**. Spike-Phase Coding Boosts and
933 Stabilizes Information Carried by Spatial and Temporal Spike Patterns. *Neuron* 61: 597–608,
934 2009.

935 **Kayser SJ**, **Ince RAA**, **Gross J**, **Kayser C**. Irregular Speech Rate Dissociates Auditory
936 Cortical Entrainment, Evoked Responses, and Frontal Alpha. *J Neurosci* 35: 14691–701,
937 2015.

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

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

942 **Lakatos P**, **Karmos G**, **Mehta AD**, **Ulbert I**, **Schroeder CE**. Entrainment of neuronal
943 oscillations as a mechanism of attentional selection. *Science (80-)* 320: 110–3, 2008.

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

947 **Liem F**, **Hurschler MA**, **J??ncke L**, **Meyer M**. On the planum temporale lateralization in
948 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: 1002–16, 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: 1001–10, 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 Work Pap 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, 2005.

Mesgarani N, **Chang EF**. Selective cortical representation of attended speaker in multi-talker speech perception. *Nature* 485: 233–6, 2012.

Mesgarani N, **Cheung C**, **Johnson K**, **Chang E**. Phonetic feature encoding in human superior temporal gyrus. *Science (80-.).* .

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: 533–45, 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: 185–96, 2013.

Montemurro MA, **Rasch MJ**, **Murayama Y**, **Logothetis NK**, **Panzeri S**. *Phase-of-Firing Coding of Natural Visual Stimuli in Primary Visual Cortex*. 2008.

Nadasdy Z. Binding by asynchrony: the neuronal phase code. *Front Neurosci* 4, 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: 12268–76, 2012.

Nourski K V, **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: 15564–74, 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.

Ten Oever S, **Sack AT**. Oscillatory phase shapes syllable perception. *Proc Natl Acad Sci U S A* 112: 15833–7, 2015.

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: 111–20, 2010.

Park H, **Ince RAA**, **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 U S A* 105: 20500–4, 2008.

Pasley BN, **David S V**, **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: 1378–87, 2013.

Peña M, **Melloni L**. Brain oscillations during spoken sentence processing. *J Cogn Neurosci* 24: 1149–64, 2012.

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: 1071–86, 2008.

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

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: 550–61, 2009.

Riecke L, **Vanbussel M**, **Hausfeld L**, **Başkent D**, **Formisano E**, **Esposito F**. Hearing an illusory vowel in noise: suppression of auditory cortical activity. *J Neurosci* 32: 8024–34, 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.* (January 7, 2015). doi: 10.1016/j.cortex.2014.12.014.

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

Sato M, **Schwartz J-L**, **Abry C**, **Cathiard M-A**, **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: 51–2, 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: 24–6, 2009b.

Shahin AJ, **Pitt MA**. Alpha activity marking word boundaries mediates speech segmentation. *Eur J Neurosci* 36: 3740–8, 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, 2002.

Strauß A, **Kotz SA**, **Scharinger M**, **Obleser J**. Alpha and theta brain oscillations index dissociable processes in spoken word recognition. *Neuroimage* 97: 387–95, 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: 164–73, 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.

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.

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

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

FIGURE CAPTIONS

Figure 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”), one button “other” if they perceived another utterance. (B) If 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”. As the phase of low-frequency neural oscillations may modulate the excitability of high-frequency activity, 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 s.e.m.

Figure 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 two perceptual outcomes (top panels in left hemispheric sensors; bottom panels in the right hemispheric sensors). Each grey bar is an individual's phase difference between the two 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. (C) Grand average auditory evoked response fields. ERFs in top panels correspond to left sensors, middle panels correspond to right sensors; bottom panels provide the speech waveforms for each sequence. No significant changes in the ERF were observed between percepts. Shaded areas denote s.e.m.

Figure 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" sequences. Color codes for the percept that was maintained. Grey areas highlight the frequencies of interest in the spectra, e.g. 1.5 Hz, 3 Hz, 4.5 Hz, 6 Hz, and alpha (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 as compared to monosyllabic sequences. The power did not significantly vary between the two bisyllabic sequences, or between the two monosyllabic sequences. Errors bars denote s.e.m. (C) No significant 1.5 Hz phase differences of the LFO when contrasting

perceived speech utterances. Each grey bar is an individual's 1.5 Hz phase difference observed when contrasting the two alternative percepts of a sequence. The black bar corresponds to the average phase shift across all participants. Red arcs correspond to 95 % confidence intervals.

Figure 4: Frequency power spectra of the envelopes of the acoustic stimuli. The 1.5 Hz sub-harmonic 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 to "pata" sequence.

Figure 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 two participants. Cross-correlations significantly changed over time according to the perceived word. Left panels provide the outcome of the cross-correlograms for each percept; the right panels provide 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 in spite of 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) where observed between the percept conditions despite inter-individual variability of absolute phase (i.e. peak latency).

Figure 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 is reported in both left (filled 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 grey 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 % confidence intervals. We show here that the reported differences in cross-correlation were related to strong shifts in the phase of neural-speech tracking.

Figure 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 ($F[1,14]<1$).

Figure 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 two perceptual outcomes (top panels in the left hemispheric sensors; bottom panels in the right hemispheric sensors). Each grey bar is an individual's phase difference between the two 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. As during the volitional task, we observed significant phase shifts of -8 degrees 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 pahse did not significantly differ between percept conditions. (C) Cross-correlations between the speech envelope and brain activity during the "lampe" sequences for two participants in the spontaneous task. Each plot shows the difference in cross-correlation between "plan" and "lampe" percept conditions. Significant differences are any patch differing from green

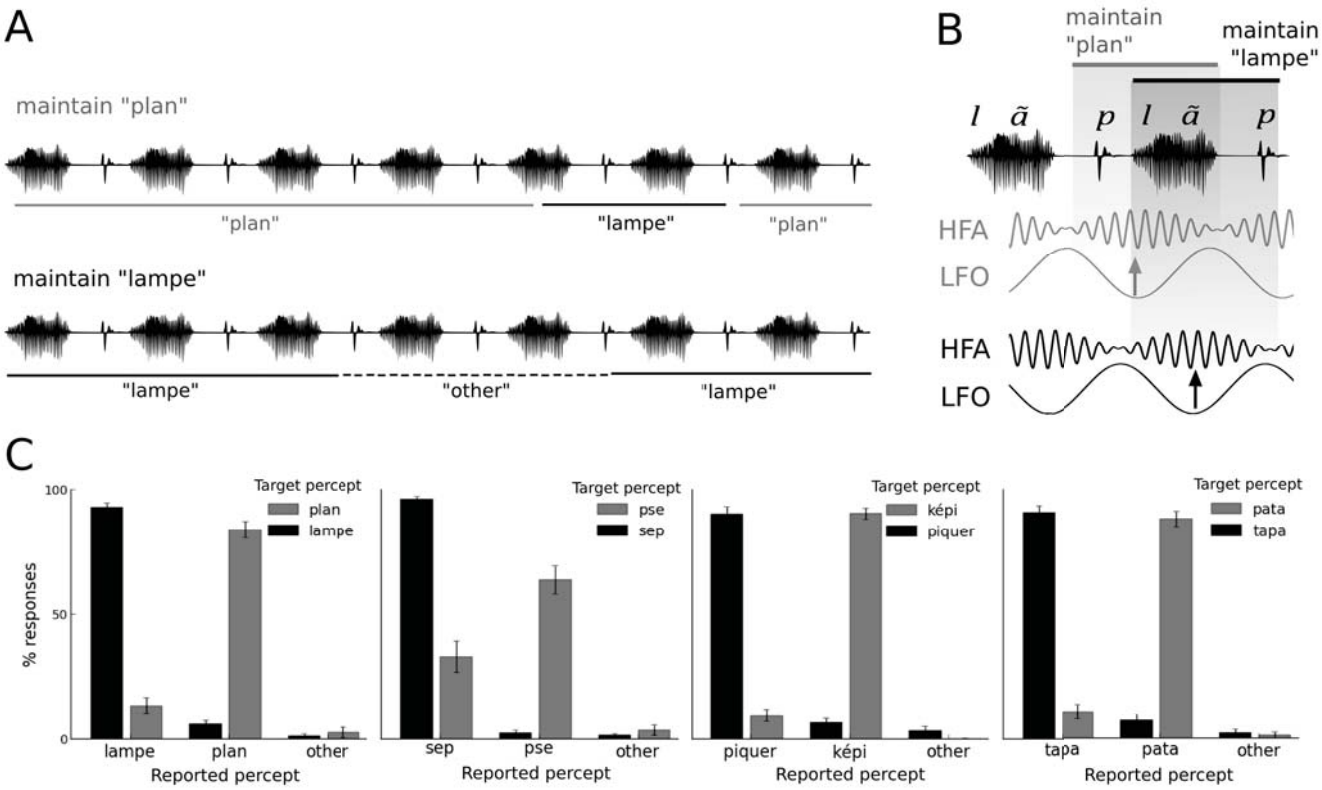
TABLES

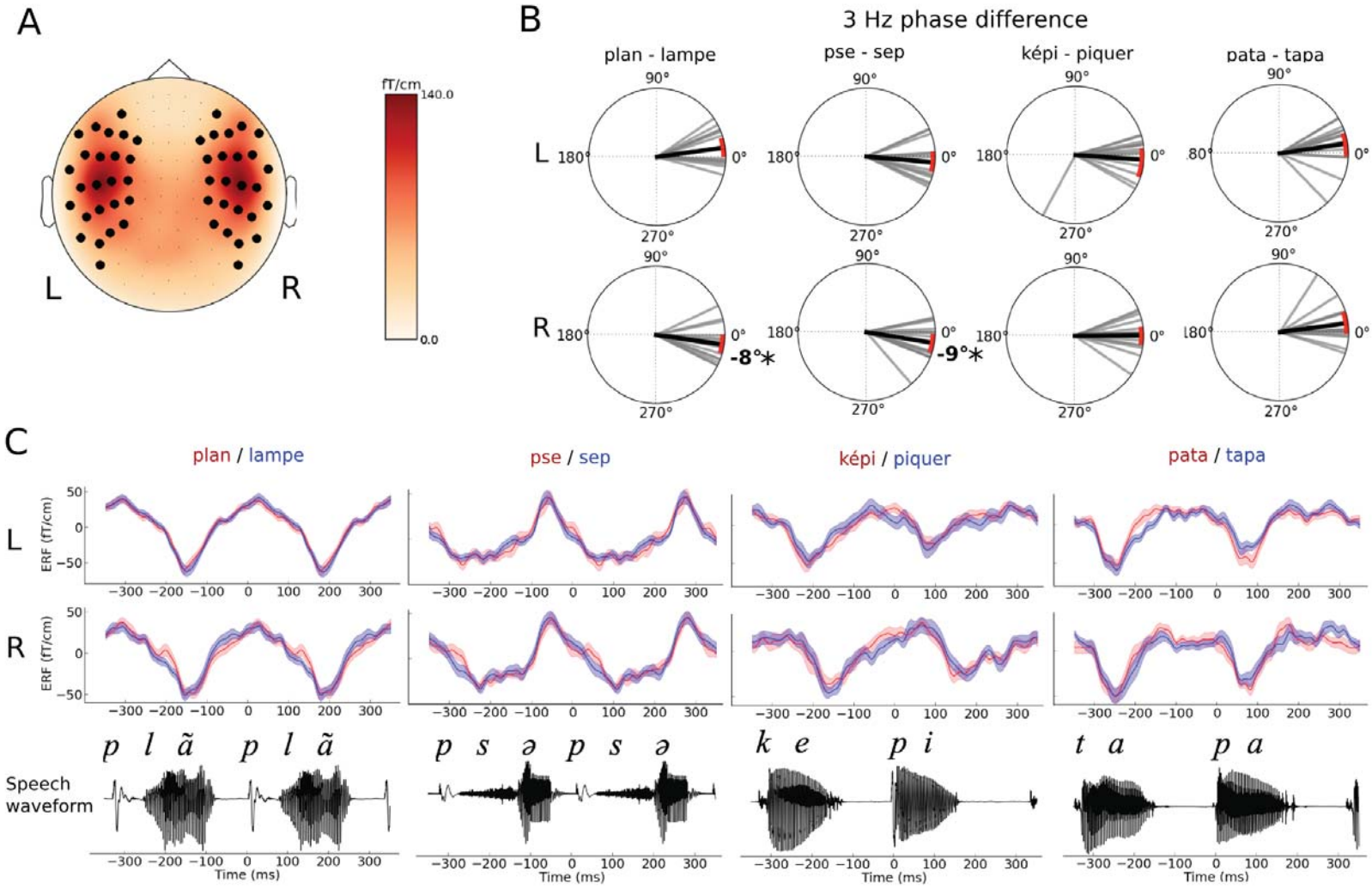
percept	PLV_L	percept	PLV_L
lampe	0.62	**plan**	0.65
sep	0.56	**pse**	0.60
képi	0.48	**piquer**	0.48
tapa	0.50	**pata**	0.54
	PLV_R		PLV_R
lampe	0.61	**plan**	0.62
sep	0.60	**pse**	0.56
képi	0.49	**piquer**	0.48
tapa	0.48	**pata**	0.55

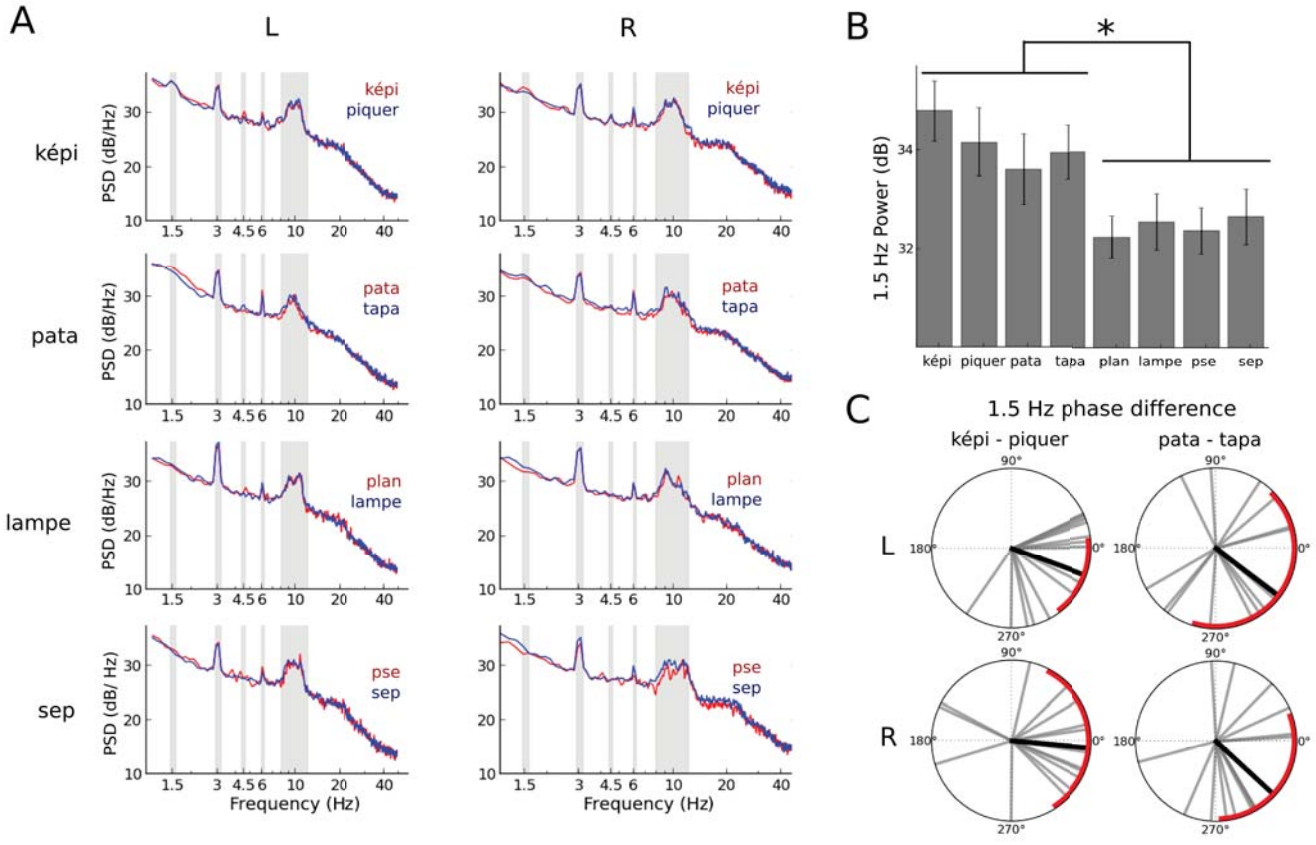
Table 1: Phase Locking Values (PLVs) at 3 Hz observed in the left and the right temporal sensors as a function of perceived speech for each speech sequence. PLVs did not significantly change between percept conditions (no main effect of percept $F[1,14] = 2.6$, $p = 0.13$, interaction between percept and sequence type $F<1$).

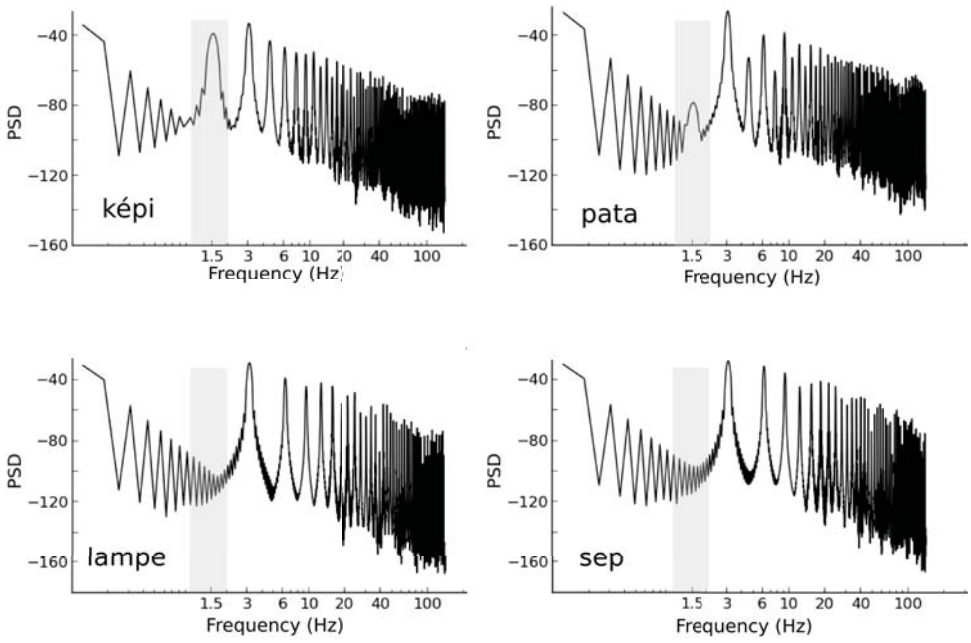
percept	PLV_L	percept	PLV_L
képi	0.30	**piquer**	0.25
tapa	0.22	**pata**	0.19
	PLV_R		PLV_R
képi	0.27	**piquer**	0.22
tapa	0.19	**pata**	0.19

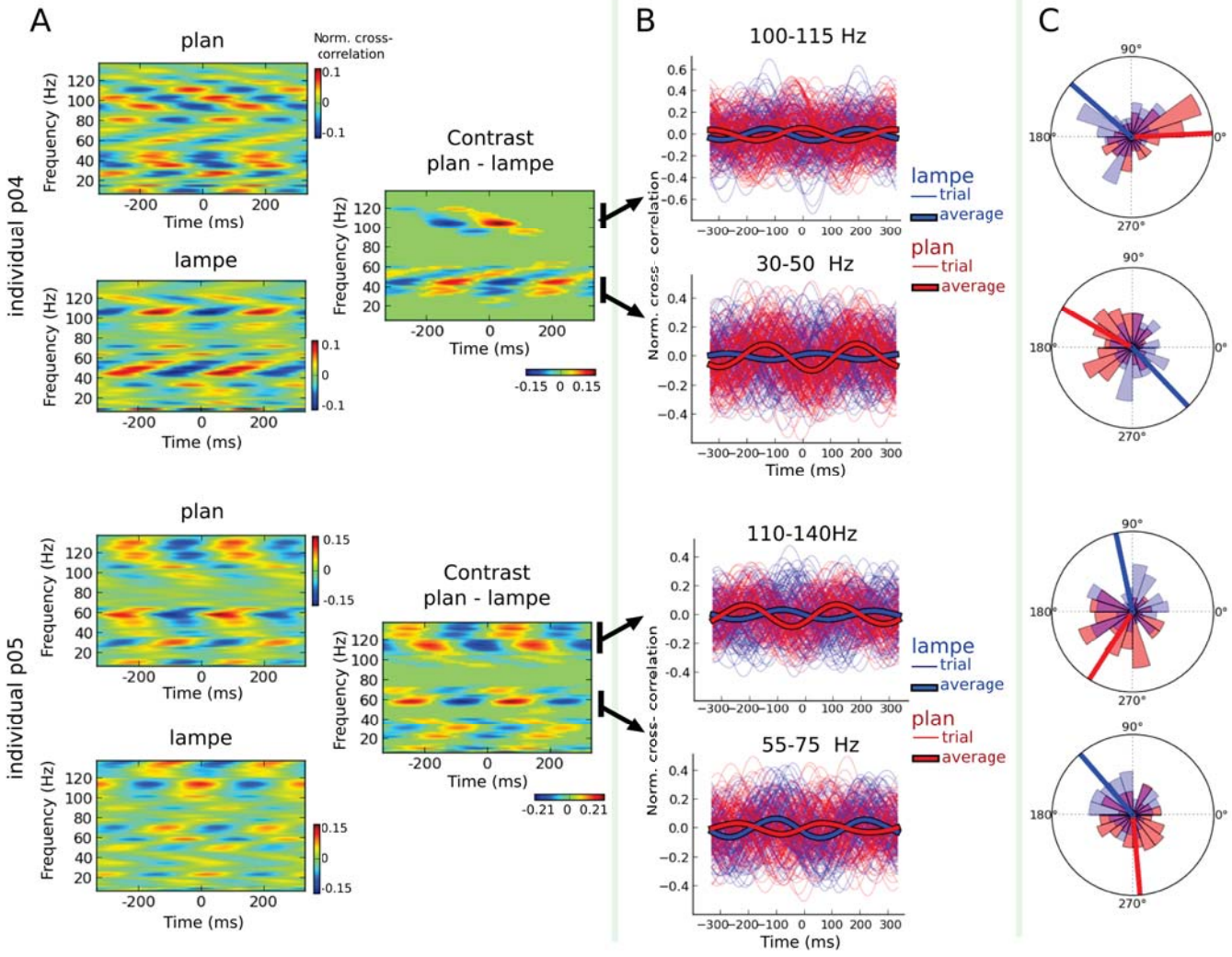
Table 2: Phase Locking Values (PLVs) at 1.5 Hz observed in the left and the right temporal sensors as a function of the perceived speech in bisyllabic speech sequences. No significant changes of the 1.5 Hz PLV were observed when contrasting the two perceptual outcomes of the same bisyllabic sequences (main effect of percept $F[1,14] < 1$, interaction between percept and sequence $F[1,14] = 2.6$, $p = 0.12$).

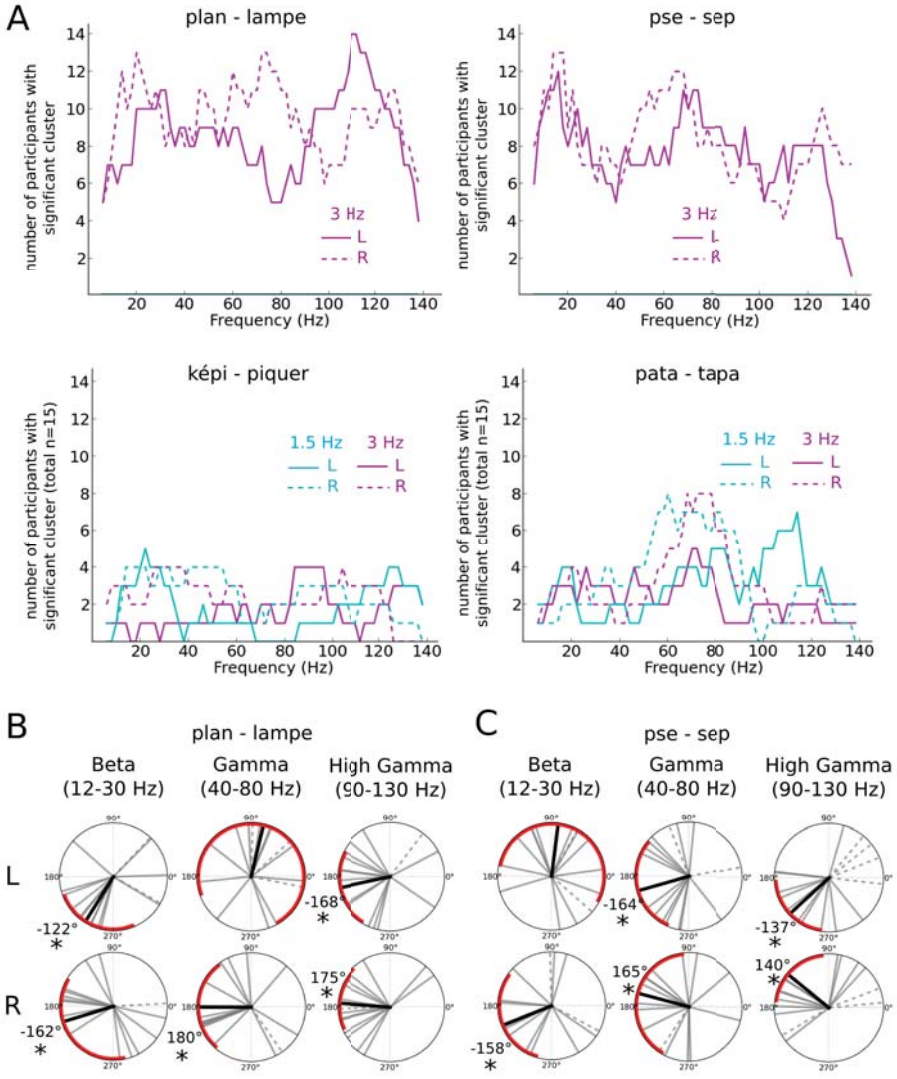

A
maintain "plan"
"plan"
"lampe"
"plan"
maintain "lampe"
"lampe"
"other"
"lampe"
B
maintain "plan"
maintain "lampe"
l ã p l ã p
HFA
LFO
HFA
LFO
C
% responses
100
50
0
Target percept
plan
lampe
lampe
plan
other
Reported percept
Target percept
pse
sep
sep
pse
other
Reported percept
Target percept
képi
piquer
piquer
képi
other
Reported percept
Target percept
pata
tapa
tapa
pata
other
Reported percept

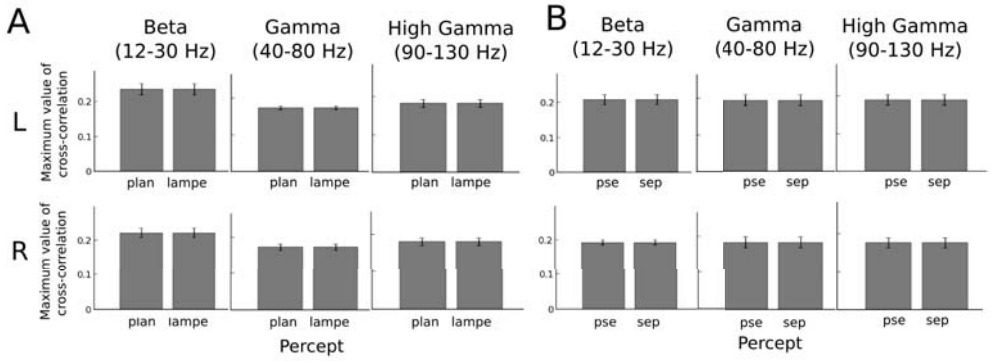

A
fT/cm
140.0
0.0
L
R
B
3 Hz phase difference
plan - lampe
pse - sep
képi - piquer
pata - tapa
L
R
90°
180°
0°
270°
-8°*
-9°*
C
plan / lampe
pse / sep
képi / piquer
pata / tapa
ERF (fT/cm)
Speech waveform
p l ã p l ã
p s ə p s ə
k e p i
t a p a
−300 −200 −100 0 100 200 300
Time (ms)

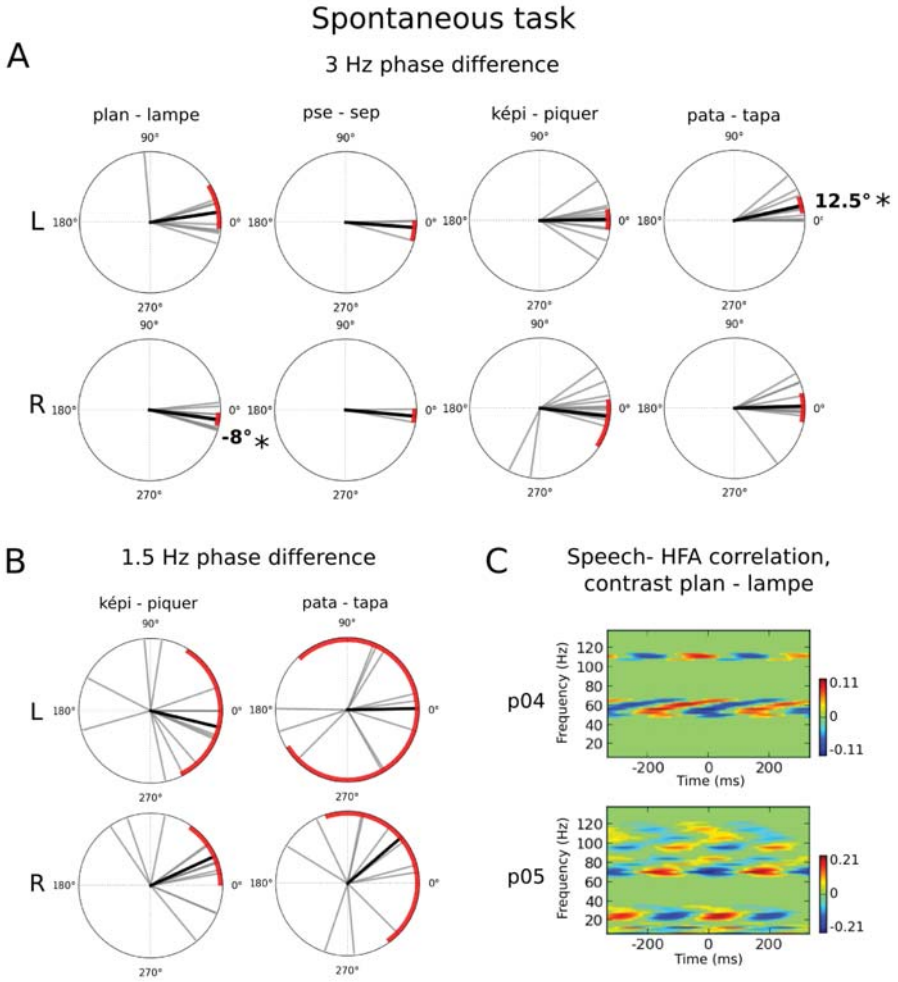

A
L
R
képi
pata
lampe
sep
PSD (dB/Hz)
PSD (dB/ Hz)
képi
piquer
pata
tapa
plan
lampe
pse
sep
1.5 3 4.5 6 10 20 40
10 20 30
Frequency (Hz)
B
*
1.5 Hz Power (dB)
32
34
képi piquer pata tapa plan lampe pse sep
C
1.5 Hz phase difference
képi - piquer
pata - tapa
L
R
90°
180°
0°
270°


képi
pata
lampe
sep
PSD
−40
−80
−120
−160
1.5
3
6
10
20
40
100
Frequency (Hz)


A
B
C
individual p04
individual p05
plan
lampe
Contrast
plan - lampe
Norm. cross-
correlation
Frequency (Hz)
Time (ms)
100-115 Hz
30-50 Hz
110-140Hz
55-75 Hz
Norm. cross- correlation
lampe
trial
average
plan
trial
average
90°
180°
0°
270°


A
plan - lampe
pse - sep
képi - piquer
pata - tapa
number of participants with significant cluster
number of participants with significant cluster (total n=15)
Frequency (Hz)
3 Hz
1.5 Hz
L
R
B
C
Beta (12-30 Hz)
Gamma (40-80 Hz)
High Gamma (90-130 Hz)
-122°
-168°
-164°
-137°
-162°
180°
175°
165°
140°
-158°


A
Beta
(12-30 Hz)
Gamma
(40-80 Hz)
High Gamma
(90-130 Hz)
L
R
Maximum value of
cross-correlation
0
0.1
0.2
plan
lampe
Percept
B
Beta
(12-30 Hz)
Gamma
(40-80 Hz)
High Gamma
(90-130 Hz)
pse
sep
Percept


Spontaneous task
A
3 Hz phase difference
plan - lampe
pse - sep
képi - piquer
pata - tapa
L
R
12.5° ∗
-8° ∗
B
1.5 Hz phase difference
képi - piquer
pata - tapa
C
Speech- HFA correlation,
contrast plan - lampe
p04
p05
Frequency (Hz)
Time (ms)
0.11
0
-0.11
0.21
-0.21