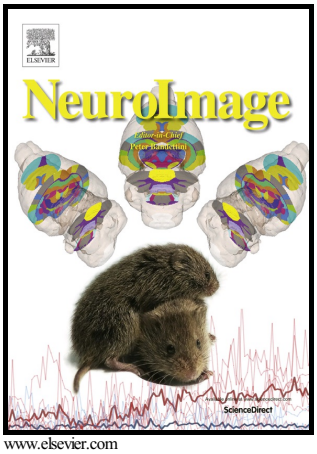
Characterization of neural entrainment to speech with and without slow spectral energy fluctuations in laminar recordings in monkey A1

Benedikt Zoefel, Jordi Costa-Faidella, Peter Lakatos, Charles E. Schroeder, Rufin VanRullen



PII: S1053-8119(17)30119-2
DOI: http://dx.doi.org/10.1016/j.neuroimage.2017.02.014
Reference: YNIMG13799

To appear in: *NeuroImage*

Received date: 30 September 2016
Accepted date: 6 February 2017

Cite this article as: Benedikt Zoefel, Jordi Costa-Faidella, Peter Lakatos, Charle E. Schroeder and Rufin VanRullen, Characterization of neural entrainment tc speech with and without slow spectral energy fluctuations in laminar recording in monkey A1, *NeuroImage* http://dx.doi.org/10.1016/j.neuroimage.2017.02.014

Characterization of neural entrainment to speech with and without slow spectral energy fluctuations in laminar recordings in monkey A1

Benedikt Zoefel[a,b,c*], Jordi Costa-Faidella[c,d,e], Peter Lakatos[c,f], Charles E. Schroeder[c,g], and Rufin VanRullen[a,b]

[a]Université Paul Sabatier, Toulouse, France

[b]Centre de Recherche Cerveau et Cognition (CerCo), CNRS, UMR5549, Pavillon Baudot CHU Purpan, BP 25202, 31052 Toulouse Cedex, France

[c]Nathan Kline Institute for Psychiatric Research, Orangeburg, NY, United States

[d]Institute of Neurosciences, University of Barcelona, Barcelona, Catalonia, 08035, Spain

[e]Brainlab – Cognitive Neuroscience Research Group, Department of Clinical Psychology and Psychobiology, University of Barcelona, Barcelona, Catalonia, 08035, Spain

[f]Department of Psychiatry, New York University School of Medicine, New York, NY, United States

[g]Departments of Neurosurgery and Psychiatry, Columbia University College of Physicians and Surgeons, New York, NY, United States

[*]Corresponding author: Benedikt Zoefel. Medical Research Council, Cognition and Brain Sciences Unit, 15 Chaucer Rd , Cambridge CB2 7EF , United Kingdom.
Email:benedikt.zoefel@mrc-cbu.cam.ac.uk

Abstract

Neural entrainment, the alignment between neural oscillations and rhythmic stimulation, is omnipresent in current theories of speech processing – nevertheless, the underlying neural mechanisms are still largely unknown. Here, we hypothesized that laminar recordings in non-human primates provide us with important insight into these mechanisms, in particular with respect to processing in cortical layers. We presented one monkey with human everyday speech sounds and recorded neural (as current-source density, CSD) oscillations in primary auditory cortex (A1). We observed that the high-excitability phase of neural oscillations was only aligned with those spectral components of speech the recording site was tuned to; the opposite, low-excitability phase was aligned with other spectral components. As low- and high-frequency components in speech alternate, this finding might

reflect a particularly efficient way of stimulus processing that includes the preparation of the relevant neuronal populations to the upcoming input. Moreover, presenting speech/noise sounds without systematic fluctuations in amplitude and spectral content and their time-reversed versions, we found significant entrainment in all conditions and cortical layers. When compared with everyday speech, the entrainment in the speech/noise conditions was characterized by a change in the phase relation between neural signal and stimulus and the low-frequency neural phase was dominantly coupled to activity in a lower gamma-band. These results show that neural entrainment in response to speech without slow fluctuations in spectral energy includes a process with specific characteristics that is presumably preserved across species.

Keywords: Neural oscillations, phase, entrainment, speech, A1, monkey, phase-amplitude coupling

1. Introduction

Many stimuli in the auditory environment – such as speech sounds – are rhythmic and alternate between important and less relevant events. Brain activity can be rhythmic as well: Neural oscillations reflect changes between high and low excitability phases of neuronal populations (Buzsáki and Draguhn, 2004). It has been proposed that these oscillations can be seen as alternations between open and closed "windows of opportunity" (more or less optimal moments) for input to be processed (Jensen et al., 2012). Thus, it is a reasonable assumption that the auditory system tries to align its oscillations with external rhythms (Calderone et al., 2014; Lakatos et al., 2008). Indeed, the alignment between neural oscillations and speech has been associated with an improved speech comprehension (Ahissar et al., 2001; Luo and Poeppel, 2007; Park et al., 2015). The precise neural

mechanisms underlying this phenomenon are currently debated (see Zoefel and VanRullen, 2015c, for discussion): Examples include a periodic adjustment of the phase of neural oscillations (McAuley, 1995), potentially via phase-reset (Schroeder and Lakatos, 2009), or the involvement of a balanced relation between excitatory and inhibitory neurons (Fries, 2005). In the present work, we therefore use the neutral term "neural entrainment" and define it as the alignment of an internal oscillatory (e.g., electrophysiological) signal to an external rhythm (e.g., speech).

Besides very few intracranial studies (Fontolan et al., 2014; Nourski et al., 2009), reports of neural entrainment are usually based on surface electrophysiological recordings with relatively low spatial resolution (using, e.g., electro-/magnetoencephalogram, EEG/MEG; e.g., Ahissar et al., 2001; Baltzell et al., 2016; Crosse et al., 2016, 2015; Di Liberto et al., 2015; Ding and Simon, 2013, 2012; Ding et al., 2016; Doelling et al., 2014; Gross et al., 2013; Horton et al., 2013; Kayser et al., 2015; Luo and Poeppel, 2007; Millman et al., 2015; Peelle et al., 2013; Zoefel and VanRullen, 2015b). The characterization of neural entrainment to speech sounds with respect to laminar processing in auditory cortex would thus represent a step forward in our understanding of the brain's processing of speech and, more generally, of rhythmic input. For this purpose, recordings in monkey auditory cortex are an important tool (Rauschecker and Scott, 2009): First, monkeys and humans are genetically closely related. Second, in contrast to humans, laminar profiles can be easily recorded in monkeys. Third, entrainment of neural oscillations has been demonstrated repeatedly in monkeys (Lakatos et al., 2005b, 2008, 2013a), indicating that humans and monkeys share a common mechanism of adaptation to rhythmicity.

We are aware of only one other study[1] showing neural responses in monkey A1 during the presentation of common human speech (Steinschneider et al., 2013). However, this study was limited to the presentation of words and focused on neural responses to phonemes. In the current study, we presented long (five one-minute) sequences of everyday speech and measured the entrainment of neural oscillations in the different cortical layers of A1. A1 is tonotopically organized (Merzenich and Brugge, 1973), and its spectral tuning determines how oscillations entrain to sound input: For instance, it has been shown that a given stimulus does not only affect the phase of neural oscillations in the area of A1 tuned to the stimulus frequency (defined as "best frequency" region, or BF region), but also in the rest of auditory cortex – by aligning the high- and low-excitability phase of oscillations in BF- and non-BF regions with the expected sound onset, respectively (Lakatos et al., 2013a; O'Connell et al., 2011). However, these results were reported using trains of pure tones as stimuli; we were therefore interested in how oscillations entrain to spectrally complex sounds such as speech, and how the tonotopic organization of A1 affects this mechanism.

Although everyday speech is an interesting stimulus, it contains large fluctuations in amplitude and spectral content (Fig. 1A, top; Fig. 1B, left) – any observed entrainment can thus be "biased", as it cannot be ruled out that it entails a passive "following" of these fluctuations at very early levels of auditory processing (e.g., a "ringing" or "frequency-following response" (FFR) of the cochlea; Dau, 2003; VanRullen et al., 2014). Recently, we reported the construction of speech/noise stimuli without systematic fluctuations in amplitude and spectral content (or "spectral energy"), i.e. features processed at the cochlear

[1] In the work by Kayser et al., 2009, the investigation of neural codes in monkey auditory cortex included the presentation of human speech, but results were generalized across a range of natural sounds.

level[2] (Fig. 1A, bottom; Fig. 1B, right; Zoefel and VanRullen, 2015a). Remaining features (including but not restricted to phonetic information) and intelligibility were conserved; we can therefore assume that neural entrainment in response to these constructed stimuli is produced at a level located beyond the earliest level of the auditory hierarchy.

In previous work, we reported that entrainment of EEG oscillations persists in the absence of systematic fluctuations in amplitude and spectral content; this entrainment, however, does not depend on linguistic features, as it is not disrupted when the speech/noise sound is reversed (Zoefel and VanRullen, 2015b; summarized in Zoefel and VanRullen, 2015c). Based on this finding, we hypothesized that we can find entrainment in response to our constructed stimuli in monkey A1 as well. Therefore, in addition to everyday speech, we presented these speech/noise sounds and measured the entrainment of neural oscillations to them. We were thus able to characterize neural entrainment to speech, both with (i.e. to everyday speech) and without slow spectral energy fluctuations (i.e. to our constructed speech/noise stimuli).

2. Materials and Methods

2.1 Subjects

In the present study, we analyzed the electrophysiological data recorded during nine penetrations of area A1 of the auditory cortex of one female rhesus macaque weighing ~9 kg, who had been prepared surgically for chronic awake electrophysiological recordings. Before surgery, the animal was adapted to a custom-fitted primate chair and to the

[2] Note that our constructed stimuli differ from common amplitude-modulated (AM) or frequency-modulated (FM) stimuli in that the latter only control for changes in amplitude *or* frequency but not in both.

recording chamber. All procedures were approved in advance by the Animal Care and Use Committee of the Nathan Kline Institute.

2.2 Surgery

Preparation of the subject for chronic awake intracortical recording was performed using aseptic techniques, under general anesthesia, as described previously (Schroeder et al., 1998). The tissue overlying the calvarium was resected and appropriate portions of the cranium were removed. The neocortex and overlying dura were left intact. To provide access to the brain and to promote an orderly pattern of sampling across the surface of the auditory areas, plastic recording chambers (Crist Instrument) were positioned normal to the cortical surface of the superior temporal plane for orthogonal penetration of area A1, as determined by preimplant MRI. Together with socketed Plexiglas bars (to permit painless head restraint), they were secured to the skull with orthopedic screws and embedded in dental acrylic. A recovery time of 6 weeks was allowed before we began data collection.

2.3 Electrophysiology

During the experiments, the animal sat in a primate chair in a dark, isolated, electrically shielded, sound-attenuated chamber with head fixed in position, and was monitored with infrared cameras. Neuroelectric activity was obtained using linear array multicontact electrodes (23 contacts, 100 um intercontact spacing, Plexon). The multielectrodes were inserted acutely through guide tube grid inserts, lowered through the dura into the brain, and positioned such that the electrode channels would span all layers of the cortex, which was determined by inspecting the laminar response profile to binaural broadband noise bursts.

Neuroelectric signals were impedance matched with a preamplifier (10X gain, bandpass dc 10 kHz) situated on the electrode, and after further amplification (500X) they were recorded continuously with a 0.01–8000 Hz bandpass digitized with a sampling rate of 20 kHz and precision of 16 bits using custom-made software in Labview. The signal was split into the local field potential (LFP; 0.1–300 Hz) and multiunit activity (MUA; 300–5000 Hz) range by zero phase shift digital filtering. Signals were downsampled to 2000 Hz and LFP data were notch-filtered between 59 and 61 Hz to remove electrical noise (transition bandwidth 2 Hz). MUA data were also rectified to improve the estimation of firing of the local neuronal ensemble (Legatt et al., 1980). One-dimensional CSD profiles were calculated from LFP profiles using a three-point formula for the calculation of the second spatial derivative of voltage (Freeman and Nicholson, 1975). The advantage of CSD profiles is that they are less affected by volume conduction like the LFP, and they also provide a more direct index of the location, direction, and density of the net transmembrane current flow (Mitzdorf, 1985; Schroeder et al., 1998).

At the beginning of each experimental session, after refining the electrode position in the neocortex, we established the BF of the recording site using a “suprathreshold” method (Lakatos et al., 2005a; Steinschneider et al., 1995). The method entails presentation of a stimulus train consisting of 100 random order occurrences of a broadband noise burst and pure tone stimuli with frequencies ranging from 353.5 Hz to 16 kHz in half-octave steps (duration: 100 ms, r/f time: 5 ms; inter-stimulus interval, ISI: 624.5 ms). Apart from anatomical information (e.g., preimplant MRI), an electrode position in A1 was inferred by a clear frequency tuning during our "suprathreshold" method. For two recordings, a position in A1 was likely but could not be ascertained (i.e. a recording in belt regions of auditory cortex could not be ruled out). These recordings were included in the present study, but did

not change results when excluded from data analysis. Auditory stimuli were produced using Matlab in-house scripts and delivered via Experiment Builder Software (SR Research Ltd., Mississauga, Ontario, Canada) at 50 dB SPL coupled with MF-1 free-field speakers.

2.4 Experimental paradigm

Stimuli of four experimental conditions (Fig. 1A) were presented to the monkey – all of them were based on five one-minute snippets of a male English speaker reading parts of a classic novel. In the first condition ("original"), these everyday speech snippets were presented. In the second condition ("original reversed"), these sounds were presented in reverse, a common procedure to test the influence of intelligibility on the observed results (e.g., Gross et al., 2013; Park et al., 2015). Although this test is not necessarily meaningful in our case of non-human primates, we were nonetheless interested in the outcome: The issue has been raised that the reversal of speech *per se* might destroy "acoustic edges" (acoustically abrupt landmarks; Doelling et al., 2014; Stevens, 2002), leading to a decline in neural entrainment that has nothing to do with the decrease in intelligibility (Millman et al., 2015; Peelle and Davis, 2012). If this assumption were true, we should see a decrease in entrainment in this condition as well.

In the third condition ("constructed"), we presented speech/noise sounds in which systematic spectral energy fluctuations (i.e. fluctuations in amplitude and spectral content; those are shown in Fig. 1B, left) were removed. The detailed construction of these stimuli was described in Zoefel and VanRullen (2015a) and MATLAB code is provided as Supplementary Material. In short, original speech was mixed with noise tailored to counterbalance fluctuations in amplitude and spectral content, so that – on *average* – the different phases of the speech envelope do not differ in these properties anymore (Fig. 1B,

right). Note that, although some remainder fluctuations are left in the stimulus (Fig. 1A, bottom), any deviation in one direction for one cycle must be compensated by opposite deviations during other cycles and cannot affect our analysis of neural entrainment explained below. Intelligibility – and therefore phonetic information – was preserved; neural entrainment to these stimuli is thus possible, assuming that the distinct features of speech and noise (which alternate rhythmically by construction) can be distinguished. Finally, it is important to note that the timing of remaining speech features in the *constructed* speech/noise snippets is reflected in the *original* speech envelope (as they co-vary with spectral energy fluctuations in everyday speech). Finally, for completeness, in the fourth condition ("constructed reversed"), the constructed speech/noise snippets were presented in reverse.

For each recording, the five one-minute snippets of all four conditions (i.e. 20 snippets) were presented in random order. The monkey listened passively to those stimuli. Snippets were separated by silence of 10 s. In the following, one "trial" is defined as the presentation of one snippet.

2.5 Data analyses

All analyses were done separately for supragranular, granular, and infragranular layers. For supragranular and infragranular layers, two channels (one sink and one source) were chosen based on the laminar profile obtained in the "suprathreshold method". The latter normally results in clear sinks and sources that can readily be assigned to these layers (e.g., Lakatos et al., 2013a). Only one channel – the sink – was chosen for the granular layer, as the corresponding source is sometimes difficult to define. For each trial, CSD and MUA signals were baseline corrected by subtracting the average of the 1-s window recorded before the

beginning of the respective trial. Moreover, the first and last 500 ms of each trial were rejected in order to avoid contamination by neural responses evoked by the onset or offset of the trial, respectively.

2.5.1 Neural entrainment: cross-correlation

In this work, neural entrainment is defined as an alignment between two oscillatory signals (here: speech or speech/noise sounds and recorded signal). Note that, in order to investigate this alignment, the *original* speech envelope (or its time-reversed version) was used as a reference for all conditions – but it reflects features that are processed both in the cochlea (i.e. amplitude and spectral content) and beyond in the two original conditions ("original" and "original reversed"), and only those processed at hierarchically higher stages in the two constructed conditions (as explained above; "constructed" and "constructed reversed"). In order to test neural entrainment to these features, the *broadband* speech envelope (extracted by Wavelet Transformation for 304 logarithmically spaced frequencies in the range between 0.59 and 21345 Hz, then summed across frequencies) was used as a reference, and neural entrainment to this envelope was compared across conditions, as outlined below. However, an additional aim of this study was to investigate the entrainment of different tonotopically organized regions to the various spectral components of speech. Therefore, for the original condition (only), neural entrainment to the envelope of different *filtered* versions of the speech signal was also analyzed. The following frequency bands were used for this analysis: 0.075-0.15 kHz, 0.225-0.45 kHz, 0.375-0.75 kHz, 0.6-1.2 kHz, 0.75-1.5 kHz, 1.5-3 kHz, 2.25-4.5 kHz, 3.75-7.5 kHz, 6-12 kHz, 9-18 kHz, 12-24 kHz, and 18-30 kHz.

In order to quantify neural entrainment, we calculated the cross-correlation between speech envelope and CSD/MUA signal, computed for time lags between -1 and 1s (Lalor et al., 2009; VanRullen and Macdonald, 2012):

$$cross-correlation\ (ch,t) = \sum_{T} env(T).signal(ch,T+t)$$

where *env(T)* and *signal(T)* denote the standardized (z-scored) speech envelope and the corresponding standardized (z-scored) CSD or MUA response at time T and channel *ch*, respectively, and *t* denotes the time lag between envelope and recorded signal. Cross-correlations were averaged across trials and recordings, but separately for each layer. Note that cross-correlation can also be used to determine the coherence between envelopes of different frequency bands of speech, if *signal* is replaced by *env2* in the above formula, and *env* and *env2* both represent envelopes of the speech signal, filtered into different frequency bands (cf. Fig. 4).

2.5.2 Neural entrainment: phase-dependent responses

We were able to use cross-correlation only to determine neural entrainment in response to the original speech sounds: Although cross-correlation is an important tool for the estimation of neural entrainment, it was necessary to develop an additional analysis. This is because, during the construction of our speech/noise stimuli, amplitude and spectral content were matched across binned phases of the original speech envelope. Cross-correlation analysis does not use phase bins, and might therefore be "contaminated" by spectral differences even in the case of the two constructed conditions ("constructed" and "constructed reversed").

We thus designed the following analysis: The original speech envelope (downsampled to 2000 Hz to match the sampling rate of the CSD/MUA signal and filtered between 2 and 8 Hz, the dominant frequency range of its spectrum) was divided into the same phase bins (i.e. 12 phase bins between –pi and pi) that were used for stimulus construction (Zoefel and VanRullen, 2015a). Each data point of the speech envelope corresponded to one data point of CSD/MUA signal that was recorded at the same time: Thus, we were able to calculate the average CSD/MUA amplitude as a function of phase (bin) of the original speech envelope. The CSD signal was filtered between 2 and 8 Hz before this procedure in order to match the dominant frequency of the original speech envelope (a prerequisite for neural entrainment). If there were neural entrainment, the CSD/MUA signal should be influenced by the speech envelope – logically thus, it should fluctuate as a function of original envelope phase. In order to test this, we fitted a sine wave to the CSD/MUA amplitude as a function of envelope phase: The amplitude of this sine wave reflects the strength of entrainment (but note that this variable can be influenced by the amplitude of the recorded signal; this is accounted for by the comparison with a surrogate distribution as explained below) and the phase of this sine wave reflects the phase relation between entrained signal and original speech envelope.

Two different hypotheses could be tested based on the thus extracted amplitudes values: (1) whether there is significant overall neural entrainment in the different conditions and layers and (2) whether there are significant differences in neural entrainment across conditions or layers. For (1), significance of neural entrainment was tested by a permutation procedure. Here, the analysis was repeated 1,000,000 times, but speech envelope and recorded signal were drawn from different trials. Thus, it was possible to obtain a range (i.e. a surrogate distribution) of (sine wave) amplitude values under the null hypothesis of no neural entrainment between original speech envelope and recorded signal. P-values were

obtained for the recorded data by comparing "real" amplitude values (averaged across recordings and either across layers – to test entrainment in the different conditions – or across conditions – to test entrainment in the different layers) with the surrogate distribution (averaged likewise). P-values were corrected for multiple comparisons using FDR (Benjamini and Hochberg, 1995). Note that the obtained amplitude values are necessarily positive; therefore, even under the null hypothesis of no entrainment, amplitude values above 0 are likely. This results in the fact that a simple t-test (e.g., test whether the obtained amplitudes are significantly different from 0) would not represent a valid approach to test the null hypothesis and speak in favor of the permutation procedure as a more appropriate test for neural entrainment. Moreover, as surrogate and "real" data contain the same CSD/MUA amplitude values, a significance difference between the two would reflect entrainment, independent of the amplitude of the recorded signal. For (2), the obtained amplitude values of the fitted sine waves were subjected to a two-factorial ANOVA (main factors condition and layer).

We were also interested in whether the phase relation between recorded signal and stimulus rhythm, reflected by the phases of the fitted sine waves, differ across conditions. This question was analyzed on a single-trial level and the fitted phases from the original condition were used as a reference: For each trial, layer and recording, the circular difference between fitted phases in the original condition and any other condition was determined. Note that trials are independent from each other: Any possible combination of trials could thus be compared (e.g., trial 1 of the original condition with trial 1 of the constructed condition, trial 1 of the original condition and trial 2 of the constructed condition etc.). For the original condition, phases of different trials could be compared as well – but, of course, phases of the same trials were excluded from the comparison. Phase

differences were determined separately for each layer and recording and then pooled across recordings, leading, for each layer, to 9 (recordings) * 5*5 (trial combinations) = 225 phase differences for original reversed, constructed and constructed reversed condition, and to 9 (recordings) * 5*4 (trial combination) = 180 phase differences for the original condition. For the original condition, this distribution of phase differences only serves control purposes and provides an estimation of the variability of the phase relation between recorded signal and stimulus across trials. For the other conditions, if there were a difference in phase relation compared to the original condition, the mean phase difference would be different from 0. We analyzed these phase differences by means of two statistical tests. First, using Rayleigh's Test, we tested whether phase differences are non-uniformly distributed – only in this case it would be appropriate to interpret the circular mean of the respective distribution of phase differences. Second, using a circular test equivalent to Student's t-test with specified mean direction, we tested whether the circular mean of the respective distribution of phase differences is significantly different from 0. A significant value in this test would indicate a difference in phase relation between the respective condition and the original condition.

2.5.3 Phase-amplitude coupling (PAC)

It has often been argued that the coupling of the phase of neural oscillations at frequencies corresponding to the dominant frequency range of the speech envelope (~2-8 Hz) with the amplitude of oscillations at higher frequencies (e.g., in the gamma-range, ~25-120 Hz) might play an important role for the "tracking" (or "parsing") of speech sounds, and rather theoretical argumentations (Ghitza, 2011, 2012; Giraud and Poeppel, 2012; Hyafil et al., 2015; Poeppel, 2003) have been underlined by practical findings in response to non-speech (Luo and Poeppel, 2012) and speech stimuli (Fontolan et al., 2014; Gross et al., 2013; Park et

al., 2015; Zion Golumbic et al., 2013). As the specific roles of different features of speech for phase-amplitude coupling (PAC) remained unclear, we were interested in differences in the observed coupling for our experimental conditions. Moreover, it has been shown convincingly that artificial PAC can be produced for different reasons, such as imperfect sinusoidal shapes of the recorded signal (Aru et al., 2015; Cole and Voytek, 2017; Dvorak and Fenton, 2014; Lozano-Soldevilla et al., 2016; but see Jensen et al., 2016). However, this concern might be reduced somewhat if a coupling between phase and amplitude of different layers can be demonstrated, as it has been done before (Spaak et al., 2012): In contrast to a common approach, where phase and amplitude are extracted from the same neural signal, in this case the signals from both layers would need to be imperfectly sinusoidal and synchronized in time in order to create artificial PAC. Thus, as a modified version of the phase-locking value proposed in Lachaux et al. (1999), we calculated PAC as follows:

$$PAC(l1, l2) = \frac{1}{R}\sum_{r=1}^{R} \left| \frac{1}{N}\sum_{n=1}^{N} \frac{1}{T}\sum_{t=1}^{T} e^{i\left(\varphi_{env\_csd}(t,n,r,l1) - \varphi_{csd}(t,n,r,l2)\right)} \right|$$

where T is the number of time points, N is the number of trials, R is the number of recordings, and l1 and l2 are the layers from which phase and amplitude were extracted (for the sake of simplification, only sinks were used for this analysis). PAC ranges between 0 (no coupling) and 1 (maximal coupling). φ_{csd} corresponds to the phase of the CSD signal, filtered between 2 and 8 Hz. φ_{env_csd} corresponds to the phase of the Hilbert envelope (filtered likewise) of the CSD signal, filtered at different gamma frequency bands. For the latter, center frequencies between 25 and 118 Hz were used (with the exception of 57-63 Hz, due to the application of a notch filter at those frequencies, see above), with the width of the band increasing with increased center frequency (smallest bandwidth 10 Hz, largest

bandwidth 42 Hz). Note that by averaging across time and trials in the complex domain, PAC is only high when the absolute phase difference between gamma envelope and the slower CSD oscillation does not vary across trials.

Again, two different hypotheses could be tested based on these PAC values: (1) whether there is significant overall PAC in the different conditions and (2) whether there are significant differences in PAC across conditions. For (1), as described above for the sine fit amplitude values, we tested significance of the obtained PAC by assigning φ_{csd} and φ_{env_csd} of a given trial to different simulated trials and re-calculating PAC 1,000,000 times. This was done in order to obtain a range of PAC values under the null hypothesis of no PAC across layers or conditions. P-values were calculated for the recorded data by comparing "real" PAC values with the surrogate distribution. P-values were corrected for multiple comparisons using FDR. For (2), the obtained PAC values were subjected to a three-factorial ANOVA (main factors gamma-frequency, spectral energy (original vs constructed) and linguistic information (forward vs reverse)). The variability across recording sessions was taken into account by including "recording session" as an additional (fourth) factor which was ignored for all subsequent steps. As we did not find a 3-way interaction in our data (and therefore differences in PAC across conditions are common to all gamma-frequencies; see Results), we further characterized our results by averaging across certain conditions (e.g., across linguistic information to compare original and constructed conditions). Here, using additional permutation tests, the PAC difference between averaged conditions (e.g., original vs constructed) was tested for significance for each gamma-frequency: A surrogate distribution was constructed for which the assignment "condition" was assigned randomly and, for each gamma-frequency, the PAC difference between (e.g., "pseudo"-original and "pseudo"-constructed) conditions was re-calculated 1,000,000 times. This was done in order to obtain

a range of PAC difference values under the null hypothesis of no PAC difference between conditions for any given gamma-frequency. P-values were calculated for the recorded data by comparing “real” PAC difference values with the surrogate distribution. P-values were corrected for multiple comparisons using FDR.

All analyses were performed in MATLAB, using the toolbox for circular statistics (Berens, 2009) where appropriate.

3. Results

In this study, we characterized neural entrainment to human speech as observed in laminar recordings in monkey A1. We presented one monkey (nine recordings) with four different experimental categories (Fig. 1A): Original (everyday) speech sounds in the “original” condition; speech/noise sounds, for which the original speech was mixed with noise to counterbalance fluctuations of features processed at the cochlear level (i.e. amplitude and spectral content) in the "constructed" condition" (Fig. 1B; Zoefel and VanRullen, 2015a); and the time-reversed version of both conditions in the “original reversed” and “constructed reversed” condition, respectively. The layer- and frequency-specific adjustment (i.e. neural entrainment) to sounds in these four conditions is described in the following.

3.1 Neural entrainment to speech in monkey A1

We used cross-correlation to characterize the alignment between everyday speech sounds and the recorded CSD and MUA signals. Exemplary results from one recording in the original condition are shown in Fig. 2B: Prominent peaks of cross-correlation at different time lags and polarity are visible. Importantly, these peaks correspond to the different cortical layers (Fig. 2A), indicating that they entrain to the speech sounds with specific delays and polarity.

This finding can be seen again in Fig. 2C, in which cross-correlation signals are shown for the different layers as chosen based on the laminar profile (obtained in response to pure tones; see Materials and Methods) in Fig. 2A. For the CSD (Fig. 2, top), in all layers, one positive and one negative peak of cross-correlation can be seen. Rather trivially, the polarity of these peaks is opposite when comparing sinks and sources of current flow. Interestingly though, one complete "cycle" of cross-correlation (i.e. one positive peak followed by one negative peak or vice versa) has a duration of ~200 ms, which corresponds to one cycle of the speech envelope. This result reflects entrainment of the CSD signal to the envelope of speech, as expected. Neural entrainment can be seen in all layers, but is strongest in supragranular layers. Importantly, the entrainment goes along with a change in neuronal firing (reflected in MUA; Fig. 2, bottom): A strong decrease, followed by an increase, in neuronal firing can be seen in response to speech sounds, and this effect is strongest in granular and adjacent layers. Note, however, that this pattern can be reversed, depending on the BF of the recording site, as we will see in the following paragraph.

Next, we analyzed whether the precise alignment between the high- and low-excitability phase of brain oscillations and speech with a given spectral content depends on the BF of the recording site (only for the original condition). First, we filtered the speech signal into different frequency bands (see Material and Methods for the exact definition of those bands). We then cross-correlated their envelope and the recorded (unfiltered) signal in different tonotopical regions (note that the same frequency bands were used for all tonotopical regions). Cross-correlation results were compared (subtracted) between regions

tuned to frequencies typical for the spectral content of vowels[3] (<= 1000 Hz; 4 recordings) and those tuned to frequencies associated with the spectral content of consonants (e.g., fricatives, such as /s/; >= 8000 Hz; 3 recordings). In Fig. 3, these differences are shown as a function of the frequency bands of speech used for cross-correlation analysis. Importantly, if neural oscillations in low- and high-frequency regions were entrained to sounds in a similar way (i.e. with a similar phase lag or "polarity", color-coded in Fig. 3), the cross-correlation difference between these regions would be approximately 0. Moreover, if they entrained differently (i.e. with different phase lags) but irrespective of input sound frequency, results (or "colors") shown in Fig. 3 would be similar across sound frequencies (or "y-axis"). As it can be seen, neither is the case: There are pronounced differences in the phase of CSD entrainment between low- and high-frequency sites (Fig. 3A) which, as expected, depend both on the stimulus frequency and on the polarity of the layer of interest (i.e. sink vs. source). We made sure that these differences resulted from opposite patterns (and not from a similar pattern that is more pronounced in one case) between sites of different BF (Supplemental Figs. 1 and 2 show separate results for low- and high-frequency sites, i.e. before making the contrast in Fig. 3, for two typical recording sites and for the average across sites, respectively). Regardless of layer, the entrainment to stimulus frequencies corresponding to the BF of the recording site always goes along with an increase in MUA (assumedly reflecting neuronal firing) at a time lag close to 0, and a decrease otherwise (Fig. 3B). This finding indicates that, indeed, the high-excitability phase (associated with an increase in MUA) of neural oscillations is aligned with speech events, but only in parts of A1 where the spectral content of the presented sounds matches their BF; in the rest of A1 (i.e.

[3] We acknowledge here that our spectral classification of vowels and consonants is oversimplified and probably insufficient for phonological purposes. However, we argue that it is sufficient for a demonstration of frequency-specific entrainment to speech sounds.

in non-BF regions), the opposite, low-excitability phase is aligned with these events . For sinks, the high-excitability phase corresponds approximately to the trough of a CSD oscillation and the low-excitability phase corresponds approximately to the peak, as it can be extracted from Fig. 3 (e.g., an increase in MUA, red in Fig. 3B, goes along with the oscillatory trough, blue in Fig. 3A).

We were also interested in how our findings above relate to the spectrotemporal organization of speech. We therefore filtered the speech signal around its fundamental frequency (here ~100 Hz, defined as the frequency with the largest amount of energy, cf. Fig. 1B) and cross-correlated its envelope with the envelopes of the speech signal filtered into other frequency bands. As the fundamental frequency of speech is of particular importance for features composed of relatively low spectral components (e.g., vowels; Barreda and Nearey, 2012), this frequency range seemed to be a good reference for a comparison with other (higher) frequency bands. Results are shown in Fig. 4, with two important points that are worth emphasizing: First, a negative correlation at time lag 0 can be seen for all frequency bands above ~3000 Hz whereas lower frequencies show a positive correlation at this time lag. Second, essentially all cross-correlation signals change sign at a certain delay (~ 150 ms), a characteristic for oscillatory signals. Together, both findings are sufficient to conclude that lower frequencies (important, e.g., for vowels) and higher frequencies (important, e.g., for certain fricatives, such as /s/) alternate in speech – it is also reminiscent of the data reported in the preceding paragraph showing a negative correlation in excitability (reflected in the oscillatory phase) between regions of A1 processing low and high sound frequencies, respectively (see Discussion for a hypothetical explanation of how these findings might be related).

3.2 Neural entrainment to speech with and without slow spectral energy fluctuations in monkey A1

We developed an additional analysis in order to compare neural entrainment across the four experimental conditions and the different layers, respectively (see Materials and Methods). In short, CSD and MUA amplitudes were binned as a function of original envelope phase. A modulation of CSD and/or MUA amplitude by envelope phase would indicate neural entrainment, and the amplitude of a sine wave fitted to this function (CSD/MUA amplitude as a function of original envelope phase) represents our measurement of entrainment strength. Moreover, this amplitude can be averaged across layers (in order to compare entrainment across conditions) or across conditions (in order to compare entrainment across layers) and the outcome is shown in Fig. 5.

First, we tested whether there is significant overall neural entrainment in the different conditions and/or layers. To do so, the observed sine fit amplitude values were compared with a surrogate distribution, the latter obtained by repeating the extraction of those values multiple times, but with speech envelope and recorded signal drawn from different trials. For both CSD and MUA, we found that the amplitude values for all conditions (Fig. 5A) and layers (Fig. 5B) are significantly ($p < 10^{-6}$) different from the surrogate distributions (the significance thresholds for alpha = 0.05, corrected for multiple comparisons, are shown as dashed lines); thus, there was significant entrainment for all conditions and layers, indicating that entrainment of neural oscillations to speech persists in monkey A1 even in the absence of systematic fluctuations in amplitude and spectral content.

Second, we used a two-factorial ANOVA in order to compare neural entrainment across conditions and layers, respectively. For CSD, this test revealed a main effect of condition

(F(3,24) = 4.62, p = 0.004), but no main effect of layer (F(4,32) = 0.61, p = 0.655) and no interaction (F(12,96) = 0.41, p = 0.959). Post-hoc tests resulted in a significant difference in CSD entrainment between original and constructed condition as well as between original and constructed reversed condition. This is expected, as in the original (but not in the constructed or constructed reversed) condition the sounds entail large amplitude fluctuations that might result in pronounced regular evoked neural activity and thus bias our measurement of entrainment. We tested this by restricting our analysis to sites tuned to (relatively high) sound frequencies whose spectral energy is low in the presented sounds (Fig. 1B) and therefore do not produce large evoked potentials. Indeed, we found that CSD entrainment is still significant in all layers and conditions ($p < 10^{-6}$) but, critically, not significantly different between conditions (F(3,24) = 0.64, p = 0.591) (data not shown). This shows that, when the frequency tuning of A1 is ignored, the entrainment observed in response to everyday speech is most likely a mixture of "real" alignment between neural oscillations and speech sounds and regular evoked potentials (not necessarily involving neural oscillations). For MUA, the ANOVA revealed a main effect of both condition (F(3,24) = 3.18, p = 0.026) and layer (F(4,32) = 4.92, p = 0.001), but no interaction (F(12,96) = 0.65, p = 0.794). Post-hoc tests resulted in a significant difference in MUA entrainment between original and constructed reversed condition. Moreover, MUA entrainment was significantly stronger for the granular layer than for the sink of the supragranular layer and the source of the infragranular layer; this is expected, due to the highest firing rates in the granular layer.

Finally, for the CSD, we compared the phases of the fitted sine waves (reflecting the phase relation between entrained CSD signal and original speech envelope) across conditions. This was done on a single-trial level by using phases from the original condition as a reference (see Materials and Methods) – in Fig. 5C, circular histograms show the observed phase

differences between original condition and any experimental condition (including those obtained in different trials for the original condition). Only results from the supragranular sink are shown as effects were strongest for this layer. All distributions of phase differences in Fig. 5C are significantly non-uniform (Rayleigh's Test; $p < 10^{-5}$); the mean direction of these distributions can thus be interpreted. As shown in Fig. 5C, the phase relations between recorded signal and speech envelope were very consistent across trials in the original condition: The mean circular difference between phases from different trials is not significantly different from 0 and the confidence interval of the mean phase difference includes 0 ([-0.20 – 0.20]; shown in blue in Fig. 5C). Although the mean circular phase difference between original and original reversed condition is significantly different from 0, it is nevertheless obvious that phases are very similar between these conditions: The confidence interval of the mean phase difference, with the lower limit very close to 0 ([0.18 – 0.52]), suggests that the aligned phase was very similar and that the observed phase difference might functionally not be relevant. This result was virtually unchanged when restricting the analysis to areas tuned to frequencies that are only marginally present in the energy profile of the speech stimulus, indicating that our results are not only a trivial reflection of evoked potentials. Interestingly, for both constructed conditions, there is a phase shift with respect to the original condition, indicating that the removal of slow fluctuations in spectral energy of speech also results in a change of the phase relation between neural oscillations and speech sounds (confidence interval of mean phase difference, constructed condition: [-2.42 – -1.42], constructed reversed condition: [-2.72 – -1.67].

3.3 Phase-amplitude coupling in response to speech with and without slow spectral energy fluctuations

Phase-amplitude coupling (PAC) in response to speech/noise sounds without systematic fluctuations in amplitude and spectral content has not been tested yet; also, results for PAC in monkey A1 in response to everyday speech sounds are still lacking. Thus, we quantified PAC in our study (see Materials and Methods) and compared results across conditions. For these analyses, the CSD phase was computed in a band between 2 and 8 Hz, and the band for which the amplitude was extracted was varied between ~25 and 120 Hz ("gamma-frequencies"). PAC was tested systematically with the phase taken from one layer and the amplitude taken from another (thereby reducing spurious PAC; see Materials and Methods); however, reliable effects were found only when phase extracted from supragranular layer was coupled with amplitude extracted from granular layer and the following description is based on this combination (note that this result would not be expected for spurious PAC where strongest coupling would be observed for phase and amplitude extracted from the same signal (i.e. layer)). Results are shown, separately for the four conditions, in Suppl. Fig. 3. For all conditions and gamma-frequencies, we found significant PAC (by means of a permutation test; $p < 0.001$, FDR-corrected). Subjecting results to a 3-way ANOVA with factors gamma-frequency, spectral energy (original vs constructed) and linguistic information (forward vs reverse), we found a main effect of gamma-frequency ($F(37,296) = 4.45$, $p < 0.0001$), a main effect of linguistic information (forward > reverse; $F(1,8) = 18.38$, $p < 0.0001$) and an interaction between spectral energy and linguistic information (main effect of linguistic information is stronger for constructed than original conditions; $F(1,8) = 15.72$, $p = 0.0001$), but no other effect. Fig. 6A, showing PAC results averaged across all conditions, illustrates the main effect of gamma-frequency. Two prominent peaks are visible at ~30 Hz and ~95 Hz. Post-hoc tests revealed significantly larger PAC for a cluster of frequencies around 90 Hz (+/- 10 Hz) than for a cluster of frequencies around 50 Hz (+/- 5 Hz). As no 3-

way interaction was found (and therefore differences in PAC across conditions are common to all frequencies), we further explored our results by averaging across the factor "spectral energy" (to illustrate frequency differences between forward and reverse conditions; Fig. 6B) as well as "linguistic information" (to illustrate frequency differences between original and constructed conditions; Fig. 6C). We ran further permutation tests in order to test at which gamma-frequencies these averaged conditions differ (see Materials and Methods). For the comparison between forward and reverse conditions, a significant difference at ~55 Hz (forward > reverse) was found which did not survive FDR-correction. For the comparison between original and constructed conditions, we found stronger PAC for the original conditions at ~95 Hz; again, this difference did not survive FDR-correction. For the constructed conditions, there was a significantly higher PAC in a range of ~31-34 Hz, even after FDR-correction. As for the entrainment described above, we found very similar PAC results when the analysis was restricted to regions tuned to sound frequencies that are less prominent in the energy profile of the presented sounds (not shown; note that this finding would not be expected in the case of spurious PAC introduced by typical evoked potentials; Aru et al., 2015).

4. Discussion

Brain activity can be rhythmic, a phenomenon that has been observed across species (Buzsáki et al., 2013). Importantly, this rhythm – reflected in neural oscillations – can be aligned with rhythms in the environment (e.g., Schroeder and Lakatos, 2009), a process termed "neural entrainment" in this paper. As the detailed mechanisms underlying neural entrainment – in particular with respect to laminar processing of spectrotemporally complex signals by tonotopically organized neuronal ensembles in A1 – are still unclear, we tried to

fill this gap by presenting one monkey with speech with and without slow spectral energy fluctuations (i.e. with and without systematic modulations of features processed at the cochlear level). We report key findings that are discussed in the following paragraphs: (1) neural entrainment to everyday speech is a sophisticated but well-organized process structured by the relation between the frequency tuning of the neural site and spectral content of the stimulus input; (2) neural entrainment can be found in all layers of monkey A1 and seems to be a process that is conserved across species; (3) neural oscillations in monkey A1 can entrain to speech even if it does not entail slow fluctuations in sound amplitude or spectral content; (4) the latter entrainment is characterized by specific properties (detailed below).

4.1 Neural entrainment to speech as a means for efficient stimulus processing

One aim of this study was a detailed characterization of how the different cortical layers and tonotopically organized regions entrain to such a complex stimulus as speech. We found that all regions of A1 entrain to (everyday) speech, independent of their BF – but the latter influences the phase relation between speech envelope and (CSD) oscillation (Fig. 3): Whereas the high-excitability phase of oscillations is aligned with spectral components of speech corresponding to the BF of a given region, the low-excitability phase is aligned with the same sounds in other (non-BF) regions (as visible in the differences in MUA in Fig. 3B, assumedly reflecting differences in neuronal firing). In line with findings by O'Connell et al. (2011), using pure tones, and theoretical considerations of the same group (O'Connell et al., 2015), these results suggest that neural entrainment might act like a spectrotemporal filter that can be easily applied to speech sounds (but note that this interpretation remains speculative, as the interaction between low- and high-frequency tuned regions was not

tested in this study): Due to its intrinsic structure, low- and high-frequency components of speech alternate (Fig. 4). Whenever low sound frequencies prevail in the input, tonotopic regions in A1 dominantly processing (speech) sounds with these frequencies reset other regions, responding primarily to high frequencies, to their low-excitability phase. This phase reset automatically results in a convergence of the high-excitability phase in these regions and their preferred stimulation (i.e. the dominance of higher sound frequencies in the input). This input then initiates the same mechanism again: This time, regions processing lower sound frequencies are set to their low-excitability phase. By completing two distinct "tasks" simultaneously, this hypothesized "dual timing mechanism" (O'Connell et al., 2015) would represent a particularly efficient and energy-saving way of speech processing: A given input would not only inhibit regions not tuned to its principal frequency (by setting their oscillations to the low-excitability phase) but also preparing them to their preferred input half a "speech cycle" later.

4.2 Characterization of neural entrainment to speech in laminar recordings

For the first time, it was possible to investigate the processing of long sequences of human speech with a spatial resolution corresponding to the different cortical layers. We found neural entrainment in all layers of A1. In line with previous studies on entrainment of neural oscillations to trains of pure tones (e.g., Lakatos et al., 2005b, 2013a), we found that the alignment between CSD and speech envelope is strongest in supragranular layers (although not significant), whereas alignment between MUA and this envelope is most pronounced in granular layers (Fig. 2C, 5B). Importantly, slower oscillations (such as the alpha band in the visual system or the theta band in the auditory system) are often associated with top-down processing in extragranular layers whereas faster oscillations might reflect bottom-up

processing in granular layers (Bastos et al., 2015; Bonnefond and Jensen, 2015; Jensen et al., 2015; van Kerkoerle et al., 2014). Neural entrainment reflects predictions about upcoming events (Schroeder and Lakatos, 2009) and therefore top-down processes: Our findings are thus well in line with the current literature, in that they suggest that the entrainment of slower, supragranular CSD oscillations is an important tool of the brain (in particular of the auditory system; VanRullen et al., 2014) to control (or "gate") feedforward stimulus processing (reflected in gamma activity) via top-down mechanisms.

4.3 Neural entrainment as a conserved process across species

Steinschneider et al. (2013) recorded evoked neural responses to human speech in A1 of both humans and non-human primates and found similar neural patterns in response to changes in amplitude envelope, voice-onset time or fundamental frequency. Their conclusion is related to the one obtained in the current study: Neural processing of human speech in A1 of monkeys is similar to that in humans. Moreover, entrainment of neural activity in primary auditory cortex to animal vocalizations has been reported for several species (Grimsley et al., 2012; Schnupp et al., 2006; Wang et al., 1995). Also, the dominant frequency of the envelope of humans speech (~2-8 Hz) is similar to the rhythm of other animal calls (see, e.g., Fig. 1 in Wang et al., 1995), and A1 of different species (including humans) seems to be tuned to these relatively slow frequencies (Edwards and Chang, 2013; Eggermont, 1998; Oshurkova et al., 2008). Thus, it is likely that the adjustment to human speech, as it has been observed frequently (Ding and Simon, 2012, 2013; Ding et al., 2016; Doelling et al., 2014; Gross et al., 2013; Park et al., 2015; Peelle et al., 2013; Peelle and Davis, 2012; Zion Golumbic et al., 2012, 2013), is only one specific occurrence of neural

entrainment to rhythmic stimuli (including vocalization calls) as a general mechanism of efficient stimulus processing across species.

4.4 Neural entrainment to speech without slow spectral energy fluctuations and its characteristics

In the present study, we were able to show that neural entrainment to speech persists even after the removal of systematic fluctuations in amplitude and spectral content. This entrainment was characterized by several distinct properties: First, we observed a change in the phase relation between the recorded neural signal and the presented stimulus as compared to everyday speech (Fig. 5C). Interestingly, a similar phase shift has been observed when the same stimuli were presented to human subjects (Zoefel and VanRullen, 2015b). It is possible that neural oscillations align with "acoustic edges" (Doelling et al., 2014; Ghitza, 2012; Giraud and Poeppel, 2012) if large amplitude fluctuations are present in the stimulus (i.e. in our original conditions) but – necessarily – to different features (such as those differentiating speech and noise in our constructed speech/noise snippets, see below) when they are absent. This change in "landmark" for entrainment might be reflected in the observed phase shift. We emphasize that, of course, several "landmarks" might co-exist in speech (potentially even tracked by different neuronal populations); our assumptions are therefore in line with the current literature cited above.

Second, the aligned phase was coupled to the amplitude of higher-frequency gamma-oscillations in all conditions, a finding that is well in line with a multitude of previous studies (e.g., Lakatos et al., 2005b; Fell and Axmacher, 2011; Spaak et al., 2012; Lisman and Jensen, 2013). However, whereas this phase was coupled most prominently (although not significantly) to oscillatory amplitudes around 95 Hz for the original conditions, it was

dominantly coupled to amplitudes around 30 Hz for the constructed conditions. It has been shown before that CSD high-gamma activity (~50-150 Hz) is correlated with neuronal firing (and MUA; for a review, see Lachaux et al., 2012). Thus, we assume that the observed 95-Hz effect might reflect the modulation of neuronal activity by the phase of slow neural oscillations, a finding that has been described before (Whittingstall and Logothetis, 2009). It has been argued that theta-gamma coupling of neural oscillations is of particular importance for the processing of speech sounds, because the frequency of both theta- and gamma-bands correspond to important characteristics of speech (e.g., to the syllabic and phonetic rhythm, respectively; Poeppel, 2003). In most theoretical models (Giraud and Poeppel, 2012; Hyafil et al., 2015; Poeppel, 2003; Poeppel et al., 2008), the frequency of this gamma-band is usually set to 25-40 Hz (sometimes also called beta; Ghitza, 2011). Thus, it is possible that the phase-amplitude coupling we observed in the constructed conditions is related to speech processing, and that it might be enhanced (as compared with the original conditions) due to an improved signal-to-noise ratio in the conditions where slow fluctuations in spectral energy have been removed.

Importantly, our findings are in line with two of the few studies examining neural entrainment to speech with high spatial resolution (by means of intracranial recordings in humans; cortical layers could not be resolved in these data): Nourski et al. (2009) showed that high-gamma power (> 70 Hz) is aligned with the speech envelope in human auditory cortex and that this entrainment is not abolished when speech intelligibility is disrupted (by time-compression of the sound; see below). Fontolan et al. (2014) demonstrated neural entrainment and phase-amplitude coupling in response to speech sounds and that it involves top-down and bottom-up processes at different neural frequencies. Strikingly, and partly in agreement with our assumptions made above (but see next paragraph), they

observed peaks at 30 and 90 Hz as well and related them to top-down and bottom-up processes, respectively. In order to emphasize the commonalities between their study and ours, we reproduced (Fig. 6D) parts of their findings (one panel of their Fig. 3), next to the corresponding results from our study (Fig. 6A-C). Nevertheless, due to its presence in granular layers and its coupling to slower oscillations in extragranular layers, it is likely that the component in the lower gamma range (~30 Hz) rather reflects a feedforward than a top-down process in our study.

4.5 The role of intelligibility of neural entrainment to speech sounds

Of note is the finding in our study that time-reversal of the stimuli (removing linguistic content) does not abolish significant neural entrainment. Again, this result supports the findings from our recent work on human subjects, showing that reversing our constructed speech/noise stimuli does not disrupt the entrainment of EEG oscillations (Zoefel and VanRullen, 2015b). Moreover, it is reasonable to assume that the monkey does not understand human speech so that the role of linguistic features (i.e. intelligibility) is rather weak for the entrainment that was measured here. However, we emphasize that it is possible to distinguish speech and noise based on features that are unrelated to abstract linguistic content (see Zoefel and VanRullen, 2015b, for a detailed discussion): They enable the listener to recognize speech without understanding what is being said (for instance, a foreign language can be identified as such without understanding a single word). Thus, we assume that these speech-related but linguistic-independent features drive the entrainment in our constructed conditions. Nevertheless, it should be mentioned that some studies did report a decrease in neural entrainment when speech sounds were rendered unintelligible by time-reversal (Gross et al., 2013; Park et al., 2015) or noise-vocoding (Peelle et al., 2013).

One important suggestion was that the time-reversal *per se* might have destroyed acoustic edges, important landmarks for neural entrainment (Doelling et al., 2014), leading to a decrease in both neural entrainment and intelligibility without a link between the two (Millman et al., 2015; Peelle and Davis, 2012). Our results might be seen as an argument against this point, as, if this assumption were true, even in monkey A1, a decrease in entrainment should have been observed (which was not the case). Nevertheless, we did see a decrease in entrainment when comparing everyday speech (original condition) with speech/noise sounds where abrupt changes in amplitude have been removed (constructed condition). This result suggests that acoustic edges do play a role for entrainment. In addition, we observed a significantly higher PAC for forward conditions as compared with their time-reversed versions, indicating that the time-reversal of our stimuli (and the associated removal of linguistic information) did have an effect on neural processing. Thus, the question why some studies report a correlation between neural entrainment and speech intelligibility, and some do not (and the role of PAC for this relationship), is still open and further studies are necessary for its answer (for a detailed discussion, see Kösem and Van Wassenhove, 2016; Zoefel and VanRullen, 2015c).

4.6 Implications for EEG/MEG recordings

Our results have implications for the entrainment recorded using methods with lower spatial resolution, such as EEG/MEG. In particular, our data suggest that there are at least two neural populations entraining to speech in A1 (one in BF regions, the other in non-BF regions), and the two are shifted by 180 degrees, so they might cancel each other out when summed. Interestingly, this is not the case in practice, as entrainment by speech is readily detectable in macroscopic recordings in humans (e.g., Crosse et al., 2016, 2015; Di Liberto et

al., 2015; Ding and Simon, 2013, 2012; Ding et al., 2016; Doelling et al., 2014; Gross et al., 2013; Kayser et al., 2015; Luo and Poeppel, 2007; Millman et al., 2015; Peelle et al., 2013; Zoefel and VanRullen, 2015b). We speculate that this is the consequence of "low frequency dominance" in A1: a significantly larger proportion of A1 neuronal ensembles is tuned to relatively low compared to relatively high frequencies (due to the logarithmic relation between BF and area in A1; see Fig. 6 in Merzenich et al., 1975, for an example in cats). Future studies are clearly necessary to investigate this further, for instance by combining intracortical measurements with superficial recordings and correlating the two, similarly to what has been done in recordings of spontaneous activity or responses to simple attention tasks (e.g., Snyder et al., 2015). These studies will give us an important idea on what we actually measure with methods that are commonly used to quantify entrainment to speech, such as EEG/MEG.

4.7 Limitations

It has been shown repeatedly that attention is an important modulatory variable for neural entrainment (for a review, see Zoefel and VanRullen, 2015c): For instance, neural entrainment is abolished when the stimulus is unattended (e.g., Lakatos et al., 2013a) and the attended speaker in a "cocktail-party scenario" can be decoded from the pattern of neural oscillations (e.g., Zion Golumbic et al., 2013). Unfortunately, in the present study, we did not control the monkey's attention (the monkey listened passively to the stimuli). Thus, it is possible that the monkey did not pay attention to the presented sounds, reducing entrainment and significance of our results. Nevertheless, we emphasize that, although we cannot conclude that the monkey attended to the stimuli, we also cannot conclude that the monkey did *not* attend to them. Indeed, in most studies testing the role of attention for

neural entrainment, a competing stimulus stream is present, reducing attention and entrainment in response to other input (e.g., Lakatos et al., 2008; Zion Golumbic et al., 2013). In our experimental paradigm, however, the presented sounds were the only stimulus input available. It therefore seems logical that the monkey attended to them; the fact that we were able to observe significant neural entrainment in all conditions seems to confirm this notion. In follow-up experiments, the role of attention in our experimental paradigm could and should be tested systematically. Based on the current literature (e.g., Ding and Simon, 2012; Horton et al., 2013; Lakatos et al., 2008, 2009, 2013; Zion Golumbic et al., 2013; Zoefel and VanRullen, 2015c), we would expect a modulation of the reported neural entrainment effects by attentional processes..

It is currently still debated whether phase entrainment is more than a simple "regular repetition of evoked potentials" (Capilla et al., 2011; Keitel et al., 2014; Lakatos et al., 2013a; VanRullen et al., 2014; Zoefel and Heil, 2013). We acknowledge that, although the removal of systematic fluctuations in amplitude and spectral content of the stimulus is certainly a first step towards an answer on this question, the stimuli used here cannot fully resolve this issue. One reason is that different regions in auditory cortices can respond to different auditory features: There are regions that prefer noise to pure tones or vice versa (Wang et al., 2005; Wessinger et al., 2001). It is also plausible that certain features that are characteristic for speech (e.g., frequency sweeps) are preserved in our speech/noise sounds (indeed, by construction, some features needed to be conserved in order to enable a differentiation between speech and noise) and recognized by regions or neurons tuned to them. Thus, it cannot be ruled out that the entrainment we observed in response to our speech/noise stimuli is not merely a succession of regular neural activity evoked by certain speech-related features. Nevertheless, several important points should be made here. First,

we observed significant entrainment at recording sites tuned to sound frequencies that are not very prominent in the spectrum of the presented stimuli (and therefore do not evoke a pronounced neural response). Second, there is evidence for entrainment being more than a reactive process: Entrained oscillations can be observed even after the offset of the entraining stimulus (Lakatos et al., 2013a) and in response to auditory stimuli that are too weak to be perceived (and therefore do not evoke neural responses; Zoefel and Heil, 2013). Third, even when assuming a fully reactive mechanism, in principle, the spectrotemporal filter mechanism as described in the previous paragraph would not be rendered unfeasible: It is likely that evoked responses entail a phase-reset of neural oscillations (Makeig et al., 2002; Sauseng et al., 2007). As this phase-reset is the only prerequisite to "prepare" oscillations for the upcoming stimulation (e.g., their "preferred" spectral components of speech), even the absence of an active entrainment mechanism (although unlikely, see above) would not contradict the mechanism of spectrotemporal filtering as described above.

Finally, we note that the alignment between neural activity and speech rhythm we observed in A1 does not necessarily have to be *generated* there: Indeed, it has been suggested that neural entrainment in cortical regions might be "inherited" from preceding regions in the auditory hierarchy (Henry and Obleser, 2012). Nevertheless, we suggest that A1, as a hub between early and late auditory processing (Nelken, 2008), might be a promising location for the role of neural entrainment as a "spectrotemporal filter" (O'Connell et al., 2011). Further studies are necessary, systematically comparing mechanisms of entrainment to speech at different levels of the auditory hierarchy.

4.8 Conclusion

Several conclusions can be drawn from our study: (1) Based on its strong dependence on the relation between frequency tuning of the neural site and spectral content of the stimulus input, entrainment of neural oscillations to rhythmic input is a highly efficient "spectrotemporal filter" (Lakatos et al., 2013a) that helps to reliably process events of high relevance within a continuous stream of input. (2) This mechanism was observed in monkey A1 – thus, it seems to be preserved across species and includes but is not restricted to oscillations in the human brain as "filter" and human speech as the to-be-filtered input. (3) As neural oscillations in monkey A1 entrained to speech sounds without systematic fluctuations in amplitude and spectral content, neural entrainment entails a process operating at a hierarchical level beyond the cochlea. This process is present in non-human primates and therefore most likely based on linguistic-independent features of speech. Did the entrainment of neural oscillations evolve as an adaptation to the rhythmic structure of the auditory environment (including communication sounds) – or was it evolutionary successful to adapt communication sounds to the rhythmic structure of the brain (indeed, oscillations are present even in species that do not exhibit rhythmic calls; Buzsáki et al., 2013)? Although this study cannot answer this question, its exciting answer might help us understand the brain's adjustment to rhythm and, ultimately, the origin of human speech.

Conflict of Interest:

The authors declare no competing financial interests.

Acknowledgements:

This study was supported by a Studienstiftung des deutschen Volkes (German National Academic Foundation) scholarship to BZ, a Marie Curie International Outgoing Fellowship within the 7th European Community Framework Programme (FP7-PEOPLE-2012-IOF; PIOF-

GA-2012-331251) to JCF, NIH R01DC012947 to PL, NIH DC011490 to CES, and an ERC Consolidator grant P-CYCLES under grant agreement 614244 to RV.

Supplementary Figure 1. Cross-correlation between CSD (A,B) or MUA (C,D) and the speech envelope filtered into narrow frequency bands (original condition) for a typical low-frequency (0.8 Hz; A,C) and a typical high-frequency site (8 kHz; B,D). Note the opposite pattern between low- and high-frequency pattern that often appears (e.g., compare the pattern for IG sink in A and B), indicating that the differences we observed (cf. Fig. 3) did not merely result from a similar pattern of cross-correlation that is more pronounced in one case. SG: supragranular, G: granular, IG: infragranular.

Supplementary Figure 2. Same as in Supplementary Figure 1, but for the average across low-frequency (A,C) and high-frequency (B,D) recording sites. Note that the difference between the depicted cross-correlations (A-B and C-D) is shown in Fig. 3.

Supplementary Figure 3. Phase-amplitude coupling as shown in Figure 6A-C, but separately for the four conditions. Data variability not shown to improve visibility of the results. For other conventions, see description of Figure 6.

References

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

Aru, J., Aru, J., Priesemann, V., Wibral, M., Lana, L., Pipa, G., Singer, W., Vicente, R., 2015. Untangling cross-frequency coupling in neuroscience. Curr. Opin. Neurobiol. 31, 51–61. doi:10.1016/j.conb.2014.08.002

Baltzell, L.S., Horton, C., Shen, Y., Richards, V.M., D'Zmura, M., Srinivasan, R., 2016. Attention selectively modulates cortical entrainment in different regions of the speech spectrum. Brain Res. 1644, 203–212. doi:10.1016/j.brainres.2016.05.029

Barreda, S., Nearey, T.M., 2012. The direct and indirect roles of fundamental frequency in vowel perception. J. Acoust. Soc. Am. 131, 466–477. doi:10.1121/1.3662068

Bastos, A.M., Vezoli, J., Bosman, C.A., Schoffelen, J.-M., Oostenveld, R., Dowdall, J.R., De Weerd, P., Kennedy, H., Fries, P., 2015. Visual areas exert feedforward and feedback influences through distinct frequency channels. Neuron 85, 390–401. doi:10.1016/j.neuron.2014.12.018

Benjamini, Y., Hochberg, Y., 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B Methodol 57 289–300.

Berens, P., 2009. CircStat: A Matlab Toolbox for Circular Statistics. J. Stat. Softw. 31 1–31.

Bonnefond, M., Jensen, O., 2015. Gamma activity coupled to alpha phase as a mechanism for top-down controlled gating. PloS One 10, e0128667. doi:10.1371/journal.pone.0128667

Buzsáki, G., Draguhn, A., 2004. Neuronal oscillations in cortical networks. Science 304, 1926–1929. doi:10.1126/science.1099745

Buzsáki, G., Logothetis, N., Singer, W., 2013. Scaling brain size, keeping timing: evolutionary preservation of brain rhythms. Neuron 80, 751–764. doi:10.1016/j.neuron.2013.10.002

Calderone, D.J., Lakatos, P., Butler, P.D., Castellanos, F.X., 2014. Entrainment of neural oscillations as a modifiable substrate of attention. Trends Cogn. Sci. 18, 300–309. doi:10.1016/j.tics.2014.02.005

Capilla, A., Pazo-Alvarez, P., Darriba, A., Campo, P., Gross, J., 2011. Steady-state visual evoked potentials can be explained by temporal superposition of transient event-related responses. PloS One 6, e14543. doi:10.1371/journal.pone.0014543

Cole, S.R., Voytek, B., 2017. Brain Oscillations and the Importance of Waveform Shape. Trends Cogn. Sci. doi:10.1016/j.tics.2016.12.008

Cousineau, D., 2005. Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson method. Tutor. Quant. Methods Psychol. 1, 4–45.

Crosse, M.J., Butler, J.S., Lalor, E.C., 2015. Congruent Visual Speech Enhances Cortical Entrainment to Continuous Auditory Speech in Noise-Free Conditions. J. Neurosci. 35, 14195–14204. doi:10.1523/JNEUROSCI.1829-15.2015

Crosse, M.J., Di Liberto, G.M., Lalor, E.C., 2016. Eye Can Hear Clearly Now: Inverse Effectiveness in Natural Audiovisual Speech Processing Relies on Long-Term Crossmodal Temporal Integration. J. Neurosci. 36, 9888–9895. doi:10.1523/JNEUROSCI.1396-16.2016

Dau, T., 2003. The importance of cochlear processing for the formation of auditory brainstem and frequency following responses. J. Acoust. Soc. Am. 113, 936–950. doi:10.1121/1.1534833

Davis, M.H., Johnsrude, I.S., 2003. Hierarchical processing in spoken language comprehension. J. Neurosci. 23, 3423–3431.

Di Liberto, G.M., O'Sullivan, J.A., Lalor, E.C., 2015. Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing. Curr. Biol. 25, 2457–2465. doi:10.1016/j.cub.2015.08.030

Ding, N., Melloni, L., Zhang, H., Tian, X., Poeppel, D., 2016. Cortical tracking of hierarchical linguistic structures in connected speech. Nat. Neurosci. 19, 158–164. doi:10.1038/nn.4186

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

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

Doelling, K.B., Arnal, L.H., Ghitza, O., Poeppel, D., 2014. Acoustic landmarks drive delta-theta oscillations to enable speech comprehension by facilitating perceptual parsing. NeuroImage 85 Pt 2, 761–768. doi:10.1016/j.neuroimage.2013.06.035

Dvorak, D., Fenton, A.A., 2014. Toward a proper estimation of phase-amplitude coupling in neural oscillations. J. Neurosci. Methods 225, 42–56. doi:10.1016/j.jneumeth.2014.01.002

Edwards, E., Chang, E.F., 2013. Syllabic (~2-5 Hz) and fluctuation (~1-10 Hz) ranges in speech and auditory processing. Hear. Res. 305, 113–134. doi:10.1016/j.heares.2013.08.017

Eggermont, J.J., 1998. Representation of spectral and temporal sound features in three cortical fields of the cat. Similarities outweigh differences. J. Neurophysiol. 80, 2743–2764.

Fell, J., Axmacher, N., 2011. The role of phase synchronization in memory processes. Nat. Rev. Neurosci. 12, 105–118. doi:10.1038/nrn2979

Fontolan, L., Morillon, B., Liegeois-Chauvel, C., Giraud, A.-L., 2014. The contribution of frequency-specific activity to hierarchical information processing in the human auditory cortex. Nat. Commun. 5, 4694. doi:10.1038/ncomms5694

Freeman, J.A., Nicholson, C., 1975. Experimental optimization of current source-density technique for anuran cerebellum. J. Neurophysiol. 38, 369–382.

Fries, P., 2005. A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. Trends Cogn. Sci. 9, 474–480.

Ghitza, O., 2012. On the role of theta-driven syllabic parsing in decoding speech: intelligibility of speech with a manipulated modulation spectrum. Front. Psychol. 3, 238. doi:10.3389/fpsyg.2012.00238

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

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

Grimsley, J.M.S., Shanbhag, S.J., Palmer, A.R., Wallace, M.N., 2012. Processing of communication calls in Guinea pig auditory cortex. PloS One 7, e51646. doi:10.1371/journal.pone.0051646

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

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

Horton, C., D'Zmura, M., Srinivasan, R., 2013. Suppression of competing speech through entrainment of cortical oscillations. J. Neurophysiol. 109, 3082–3093. doi:10.1152/jn.01026.2012

Hyafil, A., Fontolan, L., Kabdebon, C., Gutkin, B., Giraud, A.-L., 2015. Speech encoding by coupled cortical theta and gamma oscillations. eLife 4. doi:10.7554/eLife.06213

Jensen, O., Bonnefond, M., Marshall, T.R., Tiesinga, P., 2015. Oscillatory mechanisms of feedforward and feedback visual processing. Trends Neurosci. 38, 192–194. doi:10.1016/j.tins.2015.02.006

Jensen, O., Bonnefond, M., VanRullen, R., 2012. An oscillatory mechanism for prioritizing salient unattended stimuli. Trends Cogn. Sci. 16, 200–206. doi:10.1016/j.tics.2012.03.002

Jensen, O., Spaak, E., Park, H., 2016. Discriminating valid from spurious indices of phase-amplitude coupling. ENeuro ENEURO.0334-16.2016. doi:10.1523/ENEURO.0334-16.2016

Kayser, S.J., Ince, R.A.A., Gross, J., Kayser, C., 2015. Irregular Speech Rate Dissociates Auditory Cortical Entrainment, Evoked Responses, and Frontal Alpha. J. Neurosci. 35, 14691–14701. doi:10.1523/JNEUROSCI.2243-15.2015

Keitel, C., Quigley, C., Ruhnau, P., 2014. Stimulus-driven brain oscillations in the alpha range: entrainment of intrinsic rhythms or frequency-following response? J. Neurosci. 34, 10137–10140. doi:10.1523/JNEUROSCI.1904-14.2014

Kösem, A., Wassenhove, V. van, 2016. Distinct contributions of low- and high-frequency neural oscillations to speech comprehension. Lang. Cogn. Neurosci. 0, 1–9. doi:10.1080/23273798.2016.1238495

Lachaux, J.-P., Axmacher, N., Mormann, F., Halgren, E., Crone, N.E., 2012. High-frequency neural activity and human cognition: past, present and possible future of intracranial EEG research. Prog. Neurobiol. 98, 279–301. doi:10.1016/j.pneurobio.2012.06.008

Lachaux, J.P., Rodriguez, E., Martinerie, J., Varela, F.J., 1999. Measuring phase synchrony in brain signals. Hum. Brain Mapp. 8, 194–208.

Lakatos, P., Karmos, G., Mehta, A.D., Ulbert, I., Schroeder, C.E., 2008. Entrainment of neuronal oscillations as a mechanism of attentional selection. Science 320, 110–113. doi:10.1126/science.1154735

Lakatos, P., Musacchia, G., O'Connel, M.N., Falchier, A.Y., Javitt, D.C., Schroeder, C.E., 2013a. The spectrotemporal filter mechanism of auditory selective attention. Neuron 77, 750–761. doi:10.1016/j.neuron.2012.11.034

Lakatos, P., O'Connell, M.N., Barczak, A., Mills, A., Javitt, D.C., Schroeder, C.E., 2009. The leading sense: supramodal control of neurophysiological context by attention. Neuron 64, 419–430. doi:10.1016/j.neuron.2009.10.014

Lakatos, P., Pincze, Z., Fu, K.-M.G., Javitt, D.C., Karmos, G., Schroeder, C.E., 2005a. Timing of pure tone and noise-evoked responses in macaque auditory cortex. Neuroreport 16, 933–937.

Lakatos, P., Schroeder, C.E., Leitman, D.I., Javitt, D.C., 2013b. Predictive suppression of cortical excitability and its deficit in schizophrenia. J. Neurosci. 33, 11692–11702. doi:10.1523/JNEUROSCI.0010-13.2013

Lakatos, P., Shah, A.S., Knuth, K.H., Ulbert, I., Karmos, G., Schroeder, C.E., 2005b. An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. J. Neurophysiol. 94, 1904–1911. doi:10.1152/jn.00263.2005

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

Legatt, A.D., Arezzo, J., Vaughan, H.G., 1980. Averaged multiple unit activity as an estimate of phasic changes in local neuronal activity: effects of volume-conducted potentials. J. Neurosci. Methods 2, 203–217.

Lisman, J.E., Jensen, O., 2013. The θ-γ neural code. Neuron 77, 1002–1016. doi:10.1016/j.neuron.2013.03.007

Luo, H., Poeppel, D., 2012. Cortical oscillations in auditory perception and speech: evidence for two temporal windows in human auditory cortex. Front. Psychol. 3, 170. doi:10.3389/fpsyg.2012.00170

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

Makeig, S., Westerfield, M., Jung, T.P., Enghoff, S., Townsend, J., Courchesne, E., Sejnowski, T.J., 2002. Dynamic brain sources of visual evoked responses. Science 295, 690–694. doi:10.1126/science.1066168

McAuley, J.D., 1995. Perception of Time As Phase: Toward an Adaptive-oscillator Model of Rhythmic Pattern Processing. PhD thesis, Indiana University, Indianapolis, IN, USA.

Merzenich, M.M., Brugge, J.F., 1973. Representation of the cochlear partition of the superior temporal plane of the macaque monkey. Brain Res. 50, 275–296.

Merzenich, M.M., Knight, P.L., Roth, G.L., 1975. Representation of cochlea within primary auditory cortex in the cat. J. Neurophysiol. 38, 231–249.

Millman, R.E., Johnson, S.R., Prendergast, G., 2015. The Role of Phase-locking to the Temporal Envelope of Speech in Auditory Perception and Speech Intelligibility. J. Cogn. Neurosci. 27, 533–545. doi:10.1162/jocn_a_00719

Mitzdorf, U., 1985. Current source-density method and application in cat cerebral cortex: investigation of evoked potentials and EEG phenomena. Physiol. Rev. 65, 37–100.

Moerel, M., De Martino, F., Formisano, E., 2012. Processing of natural sounds in human auditory cortex: tonotopy, spectral tuning, and relation to voice sensitivity. J. Neurosci. 32, 14205–14216. doi:10.1523/JNEUROSCI.1388-12.2012

Nelken, I., 2008. Processing of complex sounds in the auditory system. Curr. Opin. Neurobiol. 18, 413–417. doi:10.1016/j.conb.2008.08.014

Nourski, K.V., Reale, R.A., Oya, H., Kawasaki, H., Kovach, C.K., Chen, H., Howard, M.A. 3rd, Brugge, J.F., 2009. Temporal envelope of time-compressed speech represented in the human auditory cortex. J Neurosci 29, 15564–15574.

O'Connell, M.N., Barczak, A., Ross, D., McGinnis, T., Schroeder, C.E., Lakatos, P., 2015. Multi-Scale Entrainment of Coupled Neuronal Oscillations in Primary Auditory Cortex. Front. Hum. Neurosci. 655. doi:10.3389/fnhum.2015.00655

O'Connell, M.N., Barczak, A., Schroeder, C.E., Lakatos, P., 2014. Layer specific sharpening of frequency tuning by selective attention in primary auditory cortex. J. Neurosci. 34, 16496–16508. doi:10.1523/JNEUROSCI.2055-14.2014

O'Connell, M.N., Falchier, A., McGinnis, T., Schroeder, C.E., Lakatos, P., 2011. Dual mechanism of neuronal ensemble inhibition in primary auditory cortex. Neuron 69, 805–817. doi:10.1016/j.neuron.2011.01.012

Oshurkova, E., Scheich, H., Brosch, M., 2008. Click train encoding in primary and non-primary auditory cortex of anesthetized macaque monkeys. Neuroscience 153, 1289–1299. doi:10.1016/j.neuroscience.2008.03.030

Park, H., Ince, R.A.A., Schyns, P.G., Thut, G., Gross, J., 2015. Frontal Top-Down Signals Increase Coupling of Auditory Low-Frequency Oscillations to Continuous Speech in Human Listeners. Curr. Biol. 25, 1649–1653. doi:10.1016/j.cub.2015.04.049

Peelle, J.E., Davis, M.H., 2012. Neural Oscillations Carry Speech Rhythm through to Comprehension. Front. Psychol. 3, 320. doi:10.3389/fpsyg.2012.00320

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

Petkov, C.I., Kayser, C., Steudel, T., Whittingstall, K., Augath, M., Logothetis, N.K., 2008. A voice region in the monkey brain. Nat. Neurosci. 11, 367–374. doi:10.1038/nn2043

Poeppel, D., 2003. The analysis of speech in different temporal integration windows: cerebral lateralization as "asymmetric sampling in time." Speech Commun. 41, 245–255. doi:10.1016/S0167-6393(02)00107-3

Poeppel, D., Idsardi, W.J., van Wassenhove, V., 2008. Speech perception at the interface of neurobiology and linguistics. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 363, 1071–1086. doi:10.1098/rstb.2007.2160

Rauschecker, J.P., 2015. Auditory and visual cortex of primates: a comparison of two sensory systems. Eur. J. Neurosci. 41, 579–585. doi:10.1111/ejn.12844

Rauschecker, J.P., Scott, S.K., 2009. Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nat. Neurosci. 12, 718–724. doi:10.1038/nn.2331

Sauseng, P., Klimesch, W., Gruber, W.R., Hanslmayr, S., Freunberger, R., Doppelmayr, M., 2007. Are event-related potential components generated by phase resetting of brain oscillations? A critical discussion. Neuroscience 146, 1435–1444. doi:10.1016/j.neuroscience.2007.03.014

Schnupp, J.W.H., Hall, T.M., Kokelaar, R.F., Ahmed, B., 2006. Plasticity of temporal pattern codes for vocalization stimuli in primary auditory cortex. J. Neurosci. 26, 4785–4795. doi:10.1523/JNEUROSCI.4330-05.2006

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

Schroeder, C.E., Mehta, A.D., Givre, S.J., 1998. A spatiotemporal profile of visual system activation revealed by current source density analysis in the awake macaque. Cereb. Cortex 8, 575–592.

Snyder, A.C., Morais, M.J., Willis, C.M., Smith, M.A., 2015. Global network influences on local functional connectivity. Nat. Neurosci. 18, 736–743. doi:10.1038/nn.3979

Spaak, E., Bonnefond, M., Maier, A., Leopold, D.A., Jensen, O., 2012. Layer-specific entrainment of γ-band neural activity by the α rhythm in monkey visual cortex. Curr. Biol. 22, 2313–2318. doi:10.1016/j.cub.2012.10.020

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

Steinschneider, M., Reser, D., Schroeder, C.E., Arezzo, J.C., 1995. Tonotopic organization of responses reflecting stop consonant place of articulation in primary auditory cortex (A1) of the monkey. Brain Res. 674, 147–152.

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

Tian, B., Reser, D., Durham, A., Kustov, A., Rauschecker, J.P., 2001. Functional specialization in rhesus monkey auditory cortex. Science 292, 290–293. doi:10.1126/science.1058911

van Kerkoerle, T., Self, M.W., Dagnino, B., Gariel-Mathis, M.-A., Poort, J., van der Togt, C., Roelfsema, P.R., 2014. Alpha and gamma oscillations characterize feedback and feedforward processing in monkey visual cortex. Proc. Natl. Acad. Sci. U. S. A. 111, 14332–14341. doi:10.1073/pnas.1402773111

VanRullen, R., Macdonald, J.S.P., 2012. Perceptual echoes at 10 Hz in the human brain. Curr. Biol. 22, 995–999. doi:10.1016/j.cub.2012.03.050

VanRullen, R., Zoefel, B., Ilhan, B., 2014. On the cyclic nature of perception in vision versus audition. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 369, 20130214. doi:10.1098/rstb.2013.0214

Wang, X., Lu, T., Snider, R.K., Liang, L., 2005. Sustained firing in auditory cortex evoked by preferred stimuli. Nature 435, 341–346. doi:10.1038/nature03565

Wang, X., Merzenich, M.M., Beitel, R., Schreiner, C.E., 1995. Representation of a species-specific vocalization in the primary auditory cortex of the common marmoset: temporal and spectral characteristics. J. Neurophysiol. 74, 2685–2706.

Wessinger, C.M., VanMeter, J., Tian, B., Van Lare, J., Pekar, J., Rauschecker, J.P., 2001. Hierarchical organization of the human auditory cortex revealed by functional magnetic resonance imaging. J. Cogn. Neurosci. 13, 1–7.

Whittingstall, K., Logothetis, N.K., 2009. Frequency-band coupling in surface EEG reflects spiking activity in monkey visual cortex. Neuron 64, 281–289. doi:10.1016/j.neuron.2009.08.016

Zion Golumbic, E.M., Ding, N., Bickel, S., Lakatos, P., Schevon, C.A., McKhann, G.M., Goodman, R.R., Emerson, R., Mehta, A.D., Simon, J.Z., Poeppel, D., Schroeder, C.E., 2013. Mechanisms underlying selective neuronal tracking of attended speech at a “cocktail party.” Neuron 77, 980–991. doi:10.1016/j.neuron.2012.12.037

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

Zoefel, B., Heil, P., 2013. Detection of Near-Threshold Sounds is Independent of EEG Phase in Common Frequency Bands. Front. Psychol. 4, 262. doi:10.3389/fpsyg.2013.00262

Zoefel, B., VanRullen, R., 2015a. Selective perceptual phase entrainment to speech rhythm in the absence of spectral energy fluctuations. J. Neurosci. 35, 1954–1964. doi:10.1523/JNEUROSCI.3484-14.2015

Zoefel, B., VanRullen, R., 2015b. EEG oscillations entrain their phase to high-level features of speech sound. NeuroImage 124, 16–23. doi:10.1016/j.neuroimage.2015.08.054

Zoefel, B., VanRullen, R., 2015c. The Role of High-Level Processes for Oscillatory Phase Entrainment to Speech Sound. Front. Hum. Neurosci. 9, 651. doi:10.3389/fnhum.2015.00651

Figure 1. A. Three-second excerpts from the stimuli used for the four conditions in this study. For all conditions, the actual stimulus is shown in black, and the corresponding envelope in gray. In the first condition (“original”), everyday speech was presented. These stimuli were time-reversed for the second condition (“original reversed”). Speech/noise stimuli were used for the other two conditions – again, played forwards (“constructed” condition) and

backwards (‘‘constructed reversed” condition). The reasoning underlying the construction of these stimuli was described in detail in Zoefel and VanRullen (2015a). In short, as shown in B, in everyday speech sounds (left panel), the different phases of the speech envelope strongly differ in their spectral energy (color-coded in B) and might trivially entrain early levels (e.g., the cochlea) of the auditory system. We therefore constructed speech/noise stimuli (right panel) which do not exhibit systematic fluctuations in amplitude and spectral content anymore. Entrainment to these stimuli thus requires a process located beyond the cochlear level, as speech and noise have to be distinguished based on remaining properties (i.e. other features than amplitude or spectral content). Reproduced with permission from Zoefel and VanRullen (2015a).

Figure 2. Layer-specific entrainment of CSD (top) and MUA (bottom) signals to the broadband envelope of speech in one exemplary recording (original condition). Using cross-correlation between speech envelope and recorded signals, a clear laminar pattern of entrainment can be revealed (B): The time lag and polarity of entrainment depends on the layer in which the signal is recorded (A). Choosing layers based on the laminar profile obtained in response to pure tones at the BF of the recording site (A) results in cross-correlations fluctuating at ~5 Hz (C, top), reflecting entrainment to the speech envelope (whose dominant frequency is in the same range). This neural entrainment goes along with a change in neural firing (reflected in MUA), as visible in C (bottom). Note that, for the MUA and to a smaller degree also for the CSD, during peaks and troughs of the cross-correlations, ripples at higher frequencies are visible that might reflect the phase-amplitude coupling of neural oscillations that is often described in the literature (cf. Fig. 6; e.g., Lakatos et al., 2005b). SG: supragranular, G: granular, IG: infragranular.

Figure 3. Difference of cross-correlation (recording sites tuned to frequencies <= 1000 Hz minus recording sites tuned to frequencies >= 8000 Hz) between CSD (A) or MUA (B) and the speech envelope filtered into narrow frequency bands (original condition). Cross-correlation patterns – as those shown in Fig. 2 – are visible again. However, for CSD, their "polarity" (i.e. the phase lag, reflected by the sign of cross-correlation) depends on both the BF of the recording site and on the polarity of the layer (i.e. sink vs. source). Regardless of the latter, entrainment in response to sound frequencies corresponding to the BF/non-BF of the recording site always results in an increase/decrease (respectively) of MUA with a time lag close to 0. SG: supragranular, G: granular, IG: infragranular.

Figure 4. The original speech, used in this study, was filtered into different frequency bands and the envelope of these bands was computed. The speech envelope filtered around its fundamental frequency (here 100 Hz; thick black line) was used as a reference and its coherence (using cross-correlation) with the envelopes of other frequency bands is shown here. At time lag 0, this cross-correlation is positive for most frequency bands of relatively low frequency (up to 2.25 kHz center frequency; black lines), including the broadband envelope (thick blue line). Importantly, this changes for higher frequencies, which are (at this time lag) negatively correlated with the envelope of speech around 100 Hz (red lines). Note also that these correlations change their sign at a time lag of ~±one cycle of the broadband speech envelope (~±200 ms). This finding indicates that low- and high-frequency components in speech (e.g., vowels and certain fricatives) alternate.

Figure 5. Amplitude values of sine waves fitted to the CSD or MUA amplitude as a function of envelope phase (separately for each recording) – reflecting the strength of the measured entrainment – and averaged across recordings, layers (A) or conditions (B), respectively. Significance thresholds of entrainment (obtained by comparison with a surrogate distribution) are shown as dashed lines. Standard error of mean is shown as errorbars. The circular differences between phases of the fitted sine waves are shown in C (for supragranular sink). These phases reflects the phase relation between recorded signal and speech envelope. They are shown using the original condition as a reference; circular histograms thus show, on a single-trial level (see Materials and Methods), distributions of phase differences between original and any experimental condition (differences in the phase of different trials in the case of original vs. original condition, topmost plot). As it can be seen, fitted phases are very similar between different trials in the original condition and between the two original conditions. However, there is a pronounced phase difference between original and the two constructed conditions. Phase differences falling into a given phase bin are shown in percent; the confidence interval of the mean phase difference is shown as blue lines. SG: supragranular, G: granular, IG: infragranular.

Figure 6. A-C. Phase-amplitude coupling (PAC) between the phase of slower CSD oscillations (delta/theta; filtered between 2 and 8 Hz) and the amplitude of oscillations at higher ("gamma") frequencies. Different combinations of layers were tested; here we show results for the combination of supragranular phase and granular amplitude as this was the only combination that produced reliable peaks at meaningful frequencies. PAC results are shown as averaged across all conditions (A), averaged across the factor "spectral energy" (original vs constructed) to illustrate frequency differences between forward and reverse conditions

(B), and averaged across the factor “linguistic information” (forward vs reverse) to illustrate frequency differences between original and constructed conditions (C). Shaded areas show standard error of mean (SEM) where variability between recordings was removed (by subtracting the mean of each recording before computing the SEM; Cousineau, 2005) in order to improve visibility of the results. Frequencies between 57 and 63 Hz were excluded from the analyses (and are masked in A-C), due to the application of a notch filter at those frequencies (to remove line noise). D. Fontolan et al. (2014) measured oscillatory activity in human introcortical recordings from different parts of the auditory cortical pathway. Using sophisticated signal analysis techniques, they demonstrated bottom-up and top-down components around 95 and 30 Hz, respectively. Both frequency components are strikingly similar to the prominent frequencies measured in our primate data (A-C). Reproduced with permission from Fontolan et al. (2014).

Highlights

- Speech entrains oscillations in monkey A1
- Aligned phase depends on frequency tuning of recording site
- Reflects highly efficient mechanism of speech processing
- Neural entrainment persists in the absence of slow energy fluctuations
- This entrainment modulates aligned phase and phase-amplitude coupling