One way or another: cortical language areas flexibly adapt processing strategies to perceptual and contextual properties of speech

Anastasia Klimovich-Gray[1], Ander Barrena[2], Eneko Agirre[2] & Nicola Molinaro[1,3]

[1] BCBL, Basque Center on Cognition, Brain and Language, Donostia/San Sebastian, Spain

[2] University of the Basque Country, Donostia/San Sebastian, Spain

[3] Ikerbasque, Basque Foundation for Science, Bilbao, Spain

Correspondence to:

Anastasia Klimovich-Gray
BCBL, Basque center on Cognition, Brain and Language Paseo Mikeletegi, 69
20009, Donostia/San Sebastian
Spain
email: a.klimovich@bcbl.eu

Abstract

Cortical circuits rely on the temporal regularities of speech to optimise signal parsing for sound-to-meaning mapping. Bottom-up speech analysis is accelerated by top-down predictions about upcoming words. In everyday communications, however, listeners are regularly presented with challenging input - fluctuations of speech rate or semantic content. In this study we asked how reducing speech temporal regularity affects its processing - parsing, phonological analysis and ability to generate context-based predictions. To ensure that spoken sentences were natural and approximated semantic constraints of spontaneous speech we built a neural network to select stimuli from large corpora. We analysed brain activity recorded with magnetoencephalography during sentence listening using evoked responses, speech-to-brain synchronization and representational similarity analysis. For normal speech theta band (6.5-8 Hz) speech-to-brain synchronization was increased and the left fronto-temporal areas generated stronger contextual predictions. The reverse was true for temporally irregular speech - weaker theta synchronization and reduced top-down effects. Interestingly, delta-band (0.5 Hz) speech tracking was greater when contextual/semantic predictions were lower or if speech was temporally jittered. We conclude that speech temporal regularity is relevant for (theta) syllabic tracking and robust semantic predictions while the joint support of temporal and contextual predictability reduces word and phrase-level cortical tracking (delta).

Key words: MEG, semantic predictions, phonological processing, coherence, representational similarity analysis, neural network

Introduction

Predictive processing approaches describe speech comprehension as a dynamic balance between bottom-up parsing of the speech stream and top-down predictions about the content of incoming linguistic information. This balance must be achieved quickly and efficiently in order for the listener to keep up with the temporal flow of information encoded in the speech signal. A mechanism thought to facilitate this is the temporal synchronisation of the language processing circuits to the temporally regular structures of speech. Speech stream encodes quasi-periodic acoustic and prosodic cues (Rosen 1992) that help cortical circuits to identify and extract meaningful linguistic features from the perceptual speech input more efficiently (Lakatos et al. 2008; Giraud and Poeppel 2012; Doelling et al. 2014). This is putatively done via cortical entrainment to these cues and arguably forming predictions about the temporal distribution of upcoming auditory objects (Arnal and Giraud 2012; Rimmele et al. 2018). This would ensure their preferential perceptual sampling and analysis (Zion Golumbic et al. 2012; Lakatos et al. 2019). Concurrently, the biasing pragmatic, sentential and phrasal context enables formation of semantic predictions (Freunberger and Roehm 2016; Maess et al. 2016; Willems et al. 2016; Wang et al. 2018; Klimovich-Gray et al. 2019) which facilitate integration of words into the emerging sentence representation. Yet little is known about the interdependence between the lower-level phonological analysis and semantic predictions and to what extent that interaction is supported by the inherent rhythmicity of the natural speech signal. In this study we asked whether reducing the temporal regularity of speech disrupts perceptual sampling and how this in turn affects the ability to make semantic predictions.

Background

Language comprehension mechanisms are typically studied in the context of normal speech comprehension with a syllabic rate of 2-10 Hz (Ding et al. 2017; Poeppel and Assaneo 2020). While cortical circuits can over-time adapt to quicker speech rates (Dupoux and Green 1997) by adjusting the frequencies of cortex-to-speech entrainment (Lizarazu et al. 2019), comprehension deteriorates quickly at faster rates (compression to 0.5 or less of original rate-Ahissar et al. 2001; Ghitza 2014). Such effects likely represent physiological constraints on the processing rate of the cortical areas involved in linguistic speech parsing (Giraud et al. 2007; Keitel and Gross 2016). Parsing in this context refers to the ability of the cortical circuits to identify salient linguistic features from continuous speech, enabled in part by the cortical tracking of these elements across the corresponding frequencies. It is, however, unclear how mechanisms underlying comprehension - bottom-up analysis and top-down contextual predictions - are able to adapt to the more taxing perceptual sampling conditions of temporally irregular speech input. Situations that require these adaptations occur regularly in everyday communication when the same speaker intermittently changes their speech rate or if multiple speakers in the ongoing conversation speak faster or slower due to either language proficiency or simply preference.

Phonetic and phonological analyses have consistently been shown to be facilitated by the temporal regularities of the speech signal (Ghitza 2011; Giraud and Poeppel 2012), perhaps most reliably by the cortical analysis of speech syllabic rhythms in the theta range (Hyafil et al. 2015; Monsalve et al. 2018). There is, however, very little evidence that context-driven semantic predictions are also facilitated if they are embedded in temporally predictable contexts. One of the few studies looking at these processes in naturalistic speech conditions is by Rothermich and colleagues (Rothermich et al. 2012). They have shown that the N400 effect (an index of contextual effects on a word's lexico-semantic analysis) for less predictable words is reduced in German sentences with metrically regular compared to irregular stress. Authors concluded that predictable rhythmicity of the stressed syllable locations facilitated semantic predictions and integrations. Consistently with this, during sentence reading Wlotko and Federmeier (2015) have shown reduced N400 effects at faster 4 Hz (250 ms SOA) compared to slower 2 Hz (500 ms SOA) word presentation rate. In contrast, Lau and Nguyen (2015) did not observe changes in contextual predictability effects in a prime-target paradigm - indexed by N400 amplitude - in temporally predictable (constant SOA) compared to temporally unpredictable (variable SOAs) written words. In summary, while there is an emerging understanding of how speech periodicity guides perceptual analysis and phonological processing, it is unclear whether higher-level top-down lexico-semantic predictions are equally dependent on the temporal regularities of speech signal. While there is limited literature exploring these effects in reading, studies that involve natural speech, where temporal information is critical for comprehension, are mostly missing.

In the study below we used a novel paradigm of temporally jittered auditory sentences (see Methods - Stimuli) to understand the role of the temporal regularity of speech on cortical tracking of the speech signal, analysis of phonological information and top-down semantic predictions. To disrupt the temporal regularity of the speech envelope we used random temporal jitter instead of a simple time-compression, so that participants would not over-time adapt to changes of the speech rate. Further, to ensure that our sentences were naturalistic and had a range of constraints representative of spontaneous speech, we built a neural network that enabled us to select sentences (from large online text corpora) that were more or less constraining to a set of target words, by estimating the probability of these target words given preceding context. To provide optimal temporal resolution of the brain response to our stimuli we collected MEG data from participants during natural listening. We analysed this data with a combination of different techniques - evoked responses, cortical entrainment to the speech envelope and representational similarity analysis. Overall our findings point towards two important conclusions. First, there is a close interdependence of perceptual speech parsing facilitated by cortext-to-speech entrainment, phonological information extraction and the ability to generate context-based predictions. Second, fronto-temporal language areas adapt their processing strategies to the perceptual properties of speech. Specifically, the presence of temporal jitter weakened contextual semantic predictions.

Methods

Participants

We collected data from 24 right-handed Spanish native speakers (15 female) with no known history of neurological disorders, no hearing issues and normal or corrected-to-normal vision. Average age was 27 years, SD=9 (min 18, max 49). All participants signed an informed consent form and were paid for their time. The experiment was approved by the BCBL ethics board.

Stimuli

All stimuli were 320 audio sentences recorded by a female Spanish native speaker on a digital recorder (Marantz PMD). Each sentence was unique and did not contain complex syntactic constructions (such as open, complement or adverbial clauses) between the main verb and the target word (noun). All sentences were selected from a collection of online Spanish corpora (News Corpus - 63 M words Bojar et al. 2014; esCow - 150 M sentences Schäfe and Bildhauer 2012; Billion Word Corpus - 1.5 B words Cardellino, 2016; and Wikicorpus - 120M words Reese et al. 2010), and were assigned into one of the 4 conditions - Normal HP (high probability), Normal LP (low probability), Jitter HP, Jitter LP. The same target words n=80, embedded in sentences were repeated across 4 main conditions in different contexts. All targets were nouns and were either direct objects or obliques of the main verb and were never the last word of the sentence. All target words were frequent nouns (word frequency at least 8 per million) with high familiarity, imageability and concreteness ratings (min 5 out of 7), 2-3 syllables in length, taken from the ESPAL database (Duchon et al. 2013). Repeating the same target words across 4 conditions was done to ensure that any differences between conditions in the critical target-related epoch were not due to lexico-semantic differences between target words or their syntactic characteristics within the sentence. While the average position of the target within the sentence was similar across conditions (HP Normal - 11.4, LP Normal - 9.2, HP Jitter - 10.9, LP Jitter - 9.3) in HP sentences the target was on average 2 positions further into the sentence than in the LP sentences. An ANOVA on target position with factors Predictability (HP, LP) and Jitter (Jitter, Normal) confirmed that there was a significant main effect of Predictability (F(1,316)=34.29, p<0.001), but no effect of Jitter or, importantly, Jitter x Predictability interaction. HP vs LP differences of target position, however, were expected and unavoidable since more constraining HP contexts naturally tend to be longer. Finally, all sentences were manually checked by Spanish native speakers and were discarded if they contained strong emotional or graphic content.

Temporal Jitter

Prior to introducing the jitter manipulation, all audio files were normalised with respect to their loudness using the ffmpeg-normalise software (https://github.com/slhck/ffmpeg-normalize) and EBU R128 normalization method. This was done to ensure that all files have the same perceived audio volume.

The temporal jitter was implemented in a custom Matlab script, where random parts (min 200 ms, max 1000 ms) of audio files were temporally modulated - either compressed or expanded (rate selected randomly from 0.4 to 1.3 of the original) using the time-scale modification algorithm described in Driedger and Müller (2014), implemented in Matlab toolbox. This algorithm was designed to preserve the perceptual quality of the original signal to a high degree by applying distinct algorithms to the harmonic (phrase vocoder) versus punctuate (OLA) events in the audio. This is particularly important in our case since we do not want to introduce significant perceptual noise to our stimuli. The quality of the jittered audio files was further validated using manual inspection. Finally, to ensure that temporal modulation did not affect the overall length of the audio signal, expansion and compression of the audio segments was done randomly and repeatedly until the length of the overall file (in samples) matched the original non-jittered sentence length (allowed mismatch no more than 3% of overall length). This was done to exclude the auditory sentence length as a confound when constructing normal and jittered conditions.

Introducing a random interval temporal jitter deteriorates periodicity across the frequency spectrum and as a side-effect the power reduces across all frequency bins and consequently the audio loudness. We measured loudness in dB LUFS using the implementation of the integrated loudness measurement (according to EBU R128 / ITU-R BS.1770). The loudness for Jittered audio was -17.3 dB LUFS, while for Normal it was -15.3 dB LUFS, making the overall difference of 2 dB LUFS. We do not believe that this had an effect on the way that participants perceived or processed Jittered stimuli since this difference lies on the margin of barely perceptible difference in loudness in complex sounds (1 dB being the smallest detectable and 5 dB being clearly perceived change; Gray 2000).

Critically, we expected that random compression and expansion of the audio would deteriorate the quasi-periodicity of the speech envelope in the critical frequencies that encode syllabic and prosodic information. To show that periodicity was indeed affected at these frequencies, we report the amplitude of the envelope across the 0-10 Hz spectrum (Figure 1).

Spectral intensities of the audio files were computed as the square root of the power spectral density within a window length of two seconds and 50% overlap. Furthermore, to account for the minor differences in loudness across jittered and normal conditions we corrected the amplitude estimate by dividing measurements in each condition by the respective maximum amplitude value in the spectrum, producing normalised values bound between 0 and 1. Normal stimuli consistently showed more power across all selected frequencies. We tested this explicitly with a t-test comparing power between Jitter vs Normal conditions in the following frequency bins: (1) <4 Hz associated with prosodic word/phrase information and (2) 4 to 10 Hz range related to syllabic information (Giraud and Poeppel 2012; Myers et al. 2019). In both frequency bins Normal conditions had a significantly higher amplitude: for <4 Hz t=6.25, p<0.0001; for 4-10 Hz t=8.64, p<0.0001. This shows that signal periodicity was significantly deteriorated in those frequencies.

---- Figure 1 ----

Estimating target probability with the LSTM neural network

HP and LP conditions differed primarily in the probability of the target word given the preceding context. To derive target probabilities we chose to use a purpose-built long-short-term-memory (LSTM - Hochreiter and Schmidhuber 1997) neural network which was trained on a combination of the above-described text corpora. LSTMs are a type of a recurrent NN (RNN) typically used for sentence and text analysis that enable the representations in the hidden layer(s) to maintain information about arbitrarily long sequences of preceding words and accurately handle long-distance dependencies. They achieve this by learning to preferentially weight the preceding sentence content that is most conducive to the upcoming word prediction. This is an advantage over other methods such as n-gram or behavioural cloze judgments. N-grams can only take a fixed predefined window of context and the larger that window (n>5) the less accurate they become (due to collocation matrix sparsity). Behavioural cloze judgments typically generate shallow probability distributions with large variability due to the limited number of continuations produced by participants while LSTMs are used more and more in the cognitive neuroscience of language to derive more reliable stimuli-related probabilistic estimates and computationally model aspects of linguistic processing (Devereux et al. 2018; Donhauser and Baillet 2020).

Parameters of the neural network

We trained a word level AWD-LSTM language model (Merity et al. 2018) for predicting word probabilities given its previous context. In the experiments we use a two layer LSTM model with 2048 units and word embedding of size 400. For training, we used the non-monotonically triggered variant of the averaged stochastic gradient method (NT-AvSGD) and we trained the model for 13 epochs. We applied all the additional regularization techniques in Merity et al. 2018 to prevent overfitting the RNN model (see Supplementary materials, Appendix I for details).

For the HP conditions the average Surprisal (negative log2 of the probability - Hale 2001; 2016) of the same target was 2.5, SD=1.33 and for LP it was 14.4, SD =2.34. The distributions of the target Surprisal in HP and LP conditions did not overlap (see Supplementary Figure 1), ensuring there is a large difference in target probabilities in the two conditions. To validate that the Jitter and Normal conditions were matched on the target probability, we ran a Bayesian t-test of the null hypothesis (H0) that there was no difference between the corresponding target probability distributions (HP Normal versus HP Jitter, Bayesian factor BF of H0=7.5; LP Normal versus LP Jitter, BF of H0=7.9). Using the same method we also attempted to match corresponding conditions on the Surprisal of the pre target word. This was done to reduce the possibility that the effects present during the target word processing are driven by the probability of the preceding word (HP Normal versus HP Jitter, BF of H0=3.6; LP Normal versus LP Jitter, BF of H0=5). As seen from the H0 BFs that indicate the relative strength of our ability to match the corresponding distributions, matching was stronger for the target position, compared to the pre-target position. However, given our stimuli further and more stringent matching for the pre-target position was not possible without significantly reducing the number of acceptable trials.

Procedure and MEG data acquisition

The magnetoencephalography (MEG) data was acquired in a magnetically shielded room with a whole-scalp system (Elekta Neuromag, Helsinki, Finland) and the bandpass filter set to 0.03 – 330 Hz, 1 kHz sampling rate. Subjects' head positions were continuously monitored with four Head Position Indicator (HPI) coils. Coil position was digitised relative to the anatomical fiducials (nasion, left and right preauricular points) with a 3D digitizer (Fastrak Polhemus, Colchester, VA, USA). Subjects' horizontal and vertical eye movements and heart rate were monitored using bipolar electrodes.

Each participant performed 3 blocks of data acquisition. In the first block the resting state MEG activity was recorded as participants were instructed to look at the blank screen for 5 minutes with their eyes open. The second main block consisted of the experimental sentences, which participants were instructed to listen to attentively and occasionally (25 % of trials) perform a simple yes/no comprehension question about the immediately preceding sentence, while looking at the black fixation cross. Participants answered with an index finger button press and the hand (right vs left) used for yes/no response was counterbalanced across participants. The second block consisted of passive listening to 80 target words on their own. All auditory stimuli were delivered with a random inter-stimulus interval (ISI) (from 1 to 2.5 seconds) via non-magnetic plastic tubes.

Data Analysis

We performed three main analyses which are detailed below - event-related magnetic field (ERF) analysis on the average amplitude difference between conditions, entrainment analysis using coherence between cortical responses and the speech envelope and the RSA analysis for models of Phonological processing and Surprisal. ERF and RSA analysis were time-locked to the target word while for the coherence analysis the whole sentence epochs were used (see details below).

Data pre-processing and epoching was done using the open source MNE Python platform and analysis pipeline (Gramfort et al. 2013, version 0.18) consisting of the following steps. First we used MaxFilter 2.2 to perform signal-noise separation and bad channel removal. Temporal extension of the signal space separation (Taulu et al. 2005) was applied to separate external noise from head-internal signal. Noisy and flat channels were detected automatically, cross-checked manually and subsequently interpolated (using field interpolation method, only good channels were used for interpolation). All sensor space subject-specific data was transformed to the second block of that subject's data. The data was further lowpass filtered at 40 Hz (finite impulse response filter with the hamming window) and blink and heart artifacts were removed with the independent component analysis implemented in MNE Python. Subsequently data was epoched from -100 to 600 ms aligned to the onset of the target word, removing noisy epochs with high sensor amplitudes (cut-off thresholds 4000 fT for magnetometers and 4000 fT/cm for gradiometers) and finally data epochs were baseline corrected using the pre-stimulus interval of -100 to 0 ms.

Coherence

Coherence between the MEG recorded brain activity in gradiometers and the speech envelope was computed with Fieldtrip (version 20200121). First we extracted the envelope of the audio signals by computing the absolute value of the Hilbert transform. Individual envelopes were then aligned to the MEG recordings and downsampled at the sampling frequency of the recordings (1 kHz). The resulting sentence-long epochs were segmented into 2 seconds-long epochs overlapping for half of their duration. Individual epochs with high variability were excluded from further analyses (automatic rejection procedure based on z-scores). Coherence was computed between each individual sensor and the audio envelope with the connectivity function available in Fieldtrip. Prior to that we extracted the cross-spectral density matrix with the Hann taper from 0 to 20 Hz in steps of 0.5 Hz. Signals from gradiometer pairs were linearly combined (see Molinaro and Lizarazu 2018) to obtain maximum coherence from a virtual gradiometer. We tested for significant group-level effects of Predictability (LP-HP contrast), Jitter (Jitter-Normal contrast) and the interaction between these two factors (second order subtraction: LP-HP subtraction in Jittered sentences minus LP-HP subtraction in Normal) with a cluster-permutation analysis (Maris and Oostenveld

2007) across sensors and frequency conditions.

ERF

For this analysis each participant's data was averaged within 4 main conditions - HP Jitter, LP Jitter, HP Normal and LP Normal. Then subtractions of interest were performed across all time-points within the epoch and the difference signals for all participants were subjected to a one-sample spatio-temporal permutation t-test across all timepoints and sensors (Maris and Oostenveld 2007) to identify significant sensors and time-points. This test was performed on RMS (root mean square) combined gradiometer pairs and magnetometers separately. Here we only report and plot significant gradiometer spatiotemporal clusters (the same effect was also present in the magnetometers). The contrasts of interest included the test of the main effect of Predictability (LP - HP conditions) and the interaction between Jitter and Predictability (second order subtraction: LP-HP subtraction in Jittered sentences minus LP-HP subtraction in Normal).

Sensor Space RSA Searchlight

RSA Searchlight analysis (Kriegeskorte et al. 2008) was performed on unaveraged single trial sensor space epochs with the open-source toolbox (https://github.com/wmvanvliet/mne-rsa). Only gradiometer data was used for the final analysis but similar results were found when performing RSA on both magnetometers and gradiometers. For each subject, data at each time-point across the epoch (from -100 to 600 ms, aligned to target word onset) was summarised with a symmetrical distance matrix where off-diagonal entries were pairwise standardized euclidean distances between data segments of individual trials. These data segments were sensor by time point arrays where dimensions corresponded to the searchlight spatial width (4 cm) and temporal length (30 ms). At each time step, these data distance matrices were correlated (spearman correlation) with the model distance matrix, which encoded the pairwise distances between data trials based on a theoretically derived measure. This procedure was repeated for every subject separately resulting in subject-specific model fit r-value maps across all sensors and time-points within the epoch. To test for group-level clusters of significant model-fit across sensors and time points, subject-level data was subjected to a one-sample spatiotemporal permutation t-test across subjects (same procedure as with the ERF analysis above - Maris and Oostenveld 2007). Here we report results based on joined magnetometer and gradiometer data but similar effects emerged when the analysis was run with the magnetometers or gradiometers separately.

We tested two models of interest both of which were expected to produce model-fits after the onset of the target word. The first model was Phonological processing and it was derived by producing IPA (international phonetic alphabet) transcriptions of the target words and converting those into binary vectors encoding phonological features of voicing, place and manner of articulation for the consonants and roundedness, backness and height for vowels. IPA to phonetic feature conversion was done with a PanPhon toolbox version 0.15 (Mortensen et al. 2016). With this model we tested bottom-up access of the phonological information associated with the target. The second model encoded the Surprisal i.e. negative log (base 2) of the probability of the target word given the preceding context. These were the same Surprisal values as used for sentence selection, derived with the neural network (see Stimuli section above). With this model we tested for the cognitive process of updating the current set of lexical context-based expectations based on the incoming perceptual information. Distance matrices for the Surprisal model were calculated using euclidean distances, while for Phonological processing the Jaccard distance was used (only non-zero features were considered during calculation).

Source localisation of the RSA model fit

To better explore the cortical localisation of the RSA analysis we source localised activity for each participant using the MNE procedure (MNE Python) (Gramfort et al. 2013) based on distributed source modeling (Lin et al. 2006), where sources of currents are localised by applying constraints and a priori assumptions about their distributions (dipole orientation and location summarized in the lead field matrix, derived from the structural MRI scans) and the noise estimates covariance matrix. For the majority of participants (19/24) individual T1 scans were used (3D MPRAGE sequence TR 2530 ms; TE 2.36 ms; flip angle 7; acceleration factor 2) acquired on a 3-T Trio scanner (Siemens) with 1 mm isotropic voxels. For 5 participants who did not have T1 scans, an MNI template brain was used. Anatomical images were processed and parcellated into surfaces (skull, gray and white matter) with FreeSurfer software (Fischl 2012) version 6. The MRI and MEG coordinate systems were coregistered using the MNE analysis interface, with respect to the anatomical locations marked during acquisition (the nasion and the left and right preauricular points) and additional ~300 head points. Subsequent steps were performed using the MNE Python environment. To derive the forward model (lead field matrix) a source space grid was set up on the white-gray matter boundary surface of 4098 sources per hemisphere and a 1-layer boundary element model was estimated. A regularized covariance matrix was derived from the epochs using the pre-trial (silent) period. The forward solution and the covariance matrix were used to estimate the linear inverse regularization parameter (inverse operator) for every source across all channels. To improve the spatial accuracy of the localization and correct for a bias toward assigning signals to superficial sources, a loose source orientation (0.2) constraint and a depth constraint (0.8) were applied (Lin et al. 2006). To derive the source estimates at every time point and for every trial, the inverse operator was applied to preprocessed data and the estimated activations were normalized with respect to signal noise by dividing the estimates by their predicted SE, thus producing unsigned dynamic statistical parametric maps (Dale et al., 2000).

For each subject source estimates across time and vertices were used for the source space Searchlight RSA analysis (https://github.com/wmvanvliet/mne-rsa). This was the same as described in the section above but now was run in the source spaces of individual subjects (30 ms time step, 2 cm source searchlight radius). The resulting subject-level spatiotemporal model-fit values were morphed to the Freesurfer average brain and averaged across subjects for visualisation. Since statistical analysis was already performed on the same data and for the same models in sensor space no further group-level statistical tests were done on the source maps. While the source space results map out the approximate cortical location of the main effects observed in the sensor space, the extent of these localisations must be interpreted with caution since they are not statistically thresholded.

Results

Behavioural analysis

To test whether participants were successful in comprehending all speech stimuli we examined the comprehension question accuracy across 4 conditions, taking out the first 5 questions of the experiment as practice ones. The percentage of correct answers averaged across participants was above 80% for all conditions, suggesting that overall all stimuli were accurately comprehended: HP Normal - 90 % ; LP Normal - 87 %; HP Jitter - 95 %; LP Jitter - 91 %. When contrasting comprehension accuracy across conditions with a repeated measures ANOVA (2 x 2 with Jitter and Predictability as main factors) we found a main effect of probability - responses to more constraining (HP) sentences were more accurate (F(1,23)=15.9, p<0.001); and a main effect of jitter - responses to temporally jittered sentences were more accurate compared to the normal ones (F(1,23)=16.6, p<0.001). While it was expected that more constraining sentences might be easier to comprehend and therefore to answer questions about, the effect of Jitter was not expected. A tentative explanation is that while jitter makes sentences more difficult to process, participants may also be more focused and try harder when answering questions about jittered sentences.

Entrainment is reduced for temporally jittered sentences

To ensure that the temporal jitter substantially disrupted entrainment to the envelope we evaluated the phase synchronization between the speech envelope during the sentence context preceding the target words and the oscillatory brain activity: we used the coherence measure (Figure 2) to evaluate the degree to which activity in all sensors was entrained to the envelope of our sentences across different frequency bands (< 20 Hz).

The Predictability effect was not significant. The main effect of Jitter emerged in theta (6.5-8 Hz; Normal conditions: mean Coh=0.12, SD=0.04; Jitter: mean=0.10, SD=0.04) in the right temporal sensors - jittered sentences showed reduced coherence compared to the normal ones [p = 0.03]. We then tested how the Predictability effect was modulated by the Jitter manipulation by subtracting the HP from the LP condition separately for Jittered and Normal sentences. The direct comparison of these two differences revealed that the interaction was significant. The cluster resulting from this latter analysis emerged in delta (0.05 Hz) in the left frontal sensors [p = 0.02]. In the left frontal sensors where the interaction effect was larger, only normal and contextually more constraining sentences showed a reduction of speech-brain coherence in delta (mean Coh difference=0.016, SD=0.02) as compared to the normal/less constraining condition [p<0.01]. No effect was present (mean Coh difference = ~0, SD=0.02) for the two jittered conditions [p=0.14].

---- Figure 2 ----

Theta entrainment in the literature has been mainly interpreted as the brain's sensitivity to the temporal structure of the speech envelope, namely to the syllabic rhythms (Gross et al. 2013; Hyafil et al. 2015). Theta effects, therefore suggest that participants found jitter sentences more perceptually challenging. Delta entrainment, on the other hand, has been associated with tracking of the prosodic rhythms as a cue to accessing more high-level linguistic information and associated syntactic structures (Ding et al. 2016; Kösem et al. 2016; Molinaro and Lizarazu 2018). We further discuss the potential origins of both effects in the Discussion.

Contextually complex sentences show N400-like effects

Next, to test whether our contextual constraint manipulation was strong enough to elicit neuronal effects we conducted an ERF analysis on the epoch of the target word. We expected to replicate the well-documented N400 effect - reduction of the signal amplitude in the fronto-temporal sensors for more constraining/predictable words. It was important to confirm this since in this study the constraint differences between more and less predictable targets were much weaker (but also more naturalistic) than those usually derived through cloze measures. Separately, we wanted to explore whether contextual facilitation of the target word processing would also be present in the temporally jittered conditions, where online processing of the context was perceptually more demanding.

ERF analysis showed early robust differences between HP and LP conditions from 200 ms in the left frontotemporal sensors [p=0.005]. In both magnetometers and gradiometers they follow a typical N400-like topography (Lau et al. 2009) (Figure 3). Within the time-window of the main Predictability effect the magnitude of the difference between HP and LP conditions was smaller for Jittered (Mean=1.08, SD =2.8 ft/cm), compared to the Normal conditions (Mean=2.58, SD=3.67, see Supplementary Figure 2). However, this effect did not reach significance as there was no significant interaction between Jitter and Predictability, implying that the context affects the amplitude of the target word in both normal and temporally jittered speech.

---- Figure 3 ----

RSA analysis – target word neural responses across individual items

Phonological information analysis

First we explored listeners' ability to process the lower-level phonological information associated with the target words (which were the same words across all conditions) in both normal and perceptually jittered contexts with the Phonological model (see methods) which encoded only the first phoneme of the targets. We deliberately chose a small time-window (200 ms) since we expected the effects of the first phoneme processing to be transient. Only temporally jittered conditions have shown a model fit [p=0.05] in the left fronto-temporal sensors. To test if the model fit was significantly better in temporally jittered compared to normal speech we conducted a post-hoc temporal permutation t-test (1-tailed) on the r-values averaged across significant sensors and within the time-window of the significant model fit. The model fit was marginally [p=0.07] better for temporally jittered speech in the brief time-window of the first 50-90 ms of the epoch.

---- Figure 4 ----

This result implies that phonological processing of the target is enhanced when temporal jitter is present in the preceding context. These effects, however, need to be interpreted with caution since the difference between model fit in Normal and Jitter conditions is marginal and furthermore the source space maps show that in both cases the peaks of the model-fit are distributed across the right temporal, left middle frontal and precentral areas.

Generating context-specific semantic predictions

ERF analysis suggested that effects of predictive context were present both in temporally jittered and normal contexts. However, using amplitude effects alone we cannot evaluate to what extent context was effective in generating specific predictions about upcoming words. To do so we used the measure of Surprisal - the negative log of the probability of the word given the context - which quantifies the extent to which the word being processed diverges from the context-driven expectation. In the context of the information theory this measure quantifies the self-information encoded by the given stimulus.

Previously, in the context of predictive accounts of speech processing, this measure has been related to the updating of the current set of probabilistic context-dependent expectations (Willems et al. 2016). Instead of measuring Surprisal effects on the amplitude of cortical responses (as it is typically done), here we asked if the information about the Surprisal values of each target was encoded in the sensor space activity patterns (RSA analysis described in Methods). We expected this analysis to be more sensitive to the information present in the sensor data of different conditions compared to the amplitude-based ERF analysis. We therefore tested the Surprisal model separately on jittered and normal data. Only normal sentences have shown significant model fit [p=0.04] in the left fronto-temporal sensors, while jittered ones showed a similar but non-significant trend.The source level maps are consistent with this, showing a stronger averaged model-fit for normal speech in the left inferior frontal and middle frontal areas.

To confirm that the model fit was significantly better for normal versus temporally jittered trials, we conducted a temporal permutation post-hoc t-test (1-tailed) on the r-values averaged across significant sensors and within the time-window of the significant model fit. This test showed that a significantly better model fit [p=0.05] for the normal sentences emerged late in the epoch, from around 530 ms after the target onset. This suggests that if the context is perceptually challenging the ability to form and update context-specific predictions is diminished.

---- Figure 5 ----

Discussion:

In this study our goal was to understand the role of temporal regularity of speech in perceptual parsing, phonological information analysis and top-down semantic predictions. We manipulated the temporal periodicity of the speech signal by randomly compressing and expanding it, thus making it less periodic, more difficult to entrain to and more effortful to perceptually segment. Our first finding was that subjects were able to accurately comprehend both normal and temporally jittered sentences (accuracy for follow up questions above 80% for all conditions). However, while overall comprehension did not suffer in a major way, analysis of the neural data has shown that the strategies that participants used for normal versus temporally jittered sentence processing differed.

First we sought to confirm that our perceptual manipulation did in fact introduce a significant challenge for online perceptual parsing of the speech stream as quantified by the brain-to-speech entrainement. The entrainment analysis in the theta band (6.5-8 Hz) confirmed that participants did not entrain to the speech envelope in the right temporal sensors as efficiently if speech was temporally jittered, and this was true for both more and less constraining sentences. Entrainment in this band is associated with processing the syllabic structure (Gross et al. 2013; Hyafil et al. 2015), which is thought to facilitate online speech signal sampling, by aligning processing to the onsets of linguistically informative events. Right localisation of this effect is in line with the proposed role of the right hemisphere in perceptual tracking of slower oscillations (Giraud and Poeppel 2012), however, both bilateral temporal and right localisation bias for theta cortex-to-speech entertainment has been reported previously (Molinaro and Lizarazu 2018; Etard and Reichenbach 2019). Furthermore, a study by (Lam et al. 2016) using a large pool of participants has shown that lateralisation of theta effects is subject to considerable subject-to-subject variability therefore we interpret lateralisation of the theta effect cautiously. A recent study that is more compatible in design to the current study is Reichenbach and colleagues (2020) who manipulated speech clarity and has shown a similar effect - less perceptually clear sentences (with added noise) showed reduced envelope tracking in theta. They related this effect to the reduction of the ability to parse syllabic and phonetic information when noise is present. Here we propose a similar conclusion, but instead of noise these effects are caused by reduced temporal regularity and therefore predictability of the envelope. This suggests that our jitter manipulation indeed was sufficiently perceptually disruptive to the perceptual sentence parsing.

Apart from a main effect in theta we also observed an interaction in the delta band (~0.5 Hz) - for the normal and contextually more constraining sentences (HP Normal condition) there was a reduction of coherence between left frontal sensors and the envelope. Delta band in speech processing has been related to the tracking of larger, higher-level linguistic units such as words and syntactic phrases (Ding et al. 2016; Kösem et al. 2016). This interaction therefore implies that under normal prosodic conditions (no temporal jitter) semantically constraining contexts - that are likely to generate stronger top-down lexico-semantic predictions - require less effort in tracking lexical and syntactic information. Recent studies have also argued that, unlike theta band, delta is not driven primarily by the perceptual properties of the auditory input but by internally generated predictions (Park et al., 2015) about the timing of upcoming auditory objects such as learned auditory scenes (Breska and Deouell) or contextually expected syntactic structures (Meyer et al. 2017; Kaufeld et al. 2020). The goal of such temporal predictions in speech comprehension is perceptual sampling optimization and subsequent facilitation of information integration (for discussion see Meyer et al. 2020; Obleser and Kayser 2019). From this perspective, reduced delta coherence in the HP Normal condition can be interpreted as a shift in the comprehension strategy: reduced tracking of lexico-syntactic units in speech due to greater reliance on the contextual/semantic predictions. On the other hand, if contextual/semantic predictions are lower (LP Normal condition), or if the speech input is difficult to parse (Jitter conditions), stronger delta-band speech tracking is required. Our data thus underline how delta-band speech tracking is qualitatively different from the theta-band effect (note the opposite direction of the Jitter effect in the two frequency bands), being the former more sensitive to cognitive demands. This interpretation is also partially supported by the RSA analysis results, showing greater sensitivity to contextual predictions in normal, compared to temporally modulated speech - effects further discussed below.

Next we used RSA analysis to explore if the information encoded by the phonological features of the first phoneme of the targets was being accessed shortly after the targets' perceptual onset. Interestingly, only for temporally jittered sentences there was significant evidence of the phonological feature processing in the frontotemporal sensors. While the phonological model did not reach significance for the normal speech, the direct model-fit comparison between normal and jittered speech was only marginally significant. Furthermore, source localisation of the uncorrected group level model-fit in the first 200 ms of target processing showed that in both cases the peaks of the phonological model-fit were in the left interior frontal areas (extending to middle frontal and precentral areas) and right temporal areas (encompassing most of the right temporal lobe). This localization is consistent with neuro-cognitive models of phonological feature access and word-initial phonological cohort competition in the left inferior frontal areas (Poldrack et al. 1999; Hickok and Poeppel 2004; Liebenthal et al. 2005; Vigneau et al. 2006; Zhuang and Devereux 2017). While the marginal increase of phonological processing in temporally jittered, compared to normal speech must be interpreted with caution, this result would be consistent with the idea that when contextual predictions are weaker (as evidenced by the Surprisal model below) the weighting on the perceptual analysis is increased and vice versa (Blank and Davis 2016; Sohoglu and Davis 2016; Cope et al. 2017). This intriguing hypothesis should be further evaluated to understand to what extent contextual predictions can reduce the load onto the ongoing phonological analysis of words.

The ability to use context to generate predictions was enhanced in the normal, compared to temporally jittered speech but this was only evident through a more sensitive multivariate analysis. Previous literature (Rothermich et al. 2012; Wlotko and Federmeier 2015) found that the context-driven N400 effect was weaker for temporally misaligned linguistic information. In the present dataset while there was a trend towards weaker predictability effects in the N400 window in temporally jittered sentences, it was not statistically significant. This suggests that some effects of context on the target processing were present in both normal and temporally jittered speech. The RSA analysis, however, has shown that the ability to make and update context-specific semantic predictions is indeed deteriorated in temporally jittered sentences and enhanced in normal sentences. We observed significantly stronger Surprisal model-fit in the left fronto-temporal sensors in normal compared to jittered conditions from 530 ms post target onset. The cortical distribution of the peaks of this effect (source localisation analysis

Figure 5) in the left IFG, middle frontal and middle temporal areas, right temporal and left tempo-parietal areas is consistent with previous findings, where the Surprisal effect has been found to be an index of prediction-updating of the phonological and lexico-semantic levels in the left frontal and temporal areas (Ettinger et al. 2014; Willems et al. 2016). As indicated above, it is interesting to observe the parallel sensitivity of (i) Surprisal RSA, which is time-locked to the target words, and (ii) delta coherence estimates, observed for the listening of the previous context, to our predictability manipulations in normal speech (but not in jittered sentences). It should be added that in both cases the effects involved left-frontal brain regions. While it is premature to draw a conclusion from this parallelism, it is however striking to observe activity in the left hemisphere frontal regions being modulated by contextual predictability at multiple time-scales: from single-word lexical Surprisal to sentence-level phrasal tracking.

Overall our results show that the natural quasi-periodic and therefore predictable structure of the speech signal plays a critical role in regulating variable processing strategies that cortical circuits use for successful linguistic comprehension. Presence of temporal regularity improves speech perceptual parsing in the theta band and concurrently facilitates the use of contextual information for generating semantic predictions. Conversely, deterioration of temporal cues degrades contextual effects and recruits increased delta-band cortical tracking resources possibly reflecting greater effort for the higher-order linguistic processes involving syntactactic tracking and integration. Together these complementary findings suggest that optimisation of natural speech comprehension is in a constant dynamic balance, where speech processing circuits flexibly adjust their processing strategies based on both the perceptual and linguistic properties of the speech stimulus.

Acknowledgments

Authors would like to thank Jordi Martorell for help with stimuli selection and checking, Piermatteo Morucci and Manex Lete for help with data collection and Mikel Lizarazu for theoretical feedback.

Funding

This work was supported by the European Union’s Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 798971 awarded to AKG. NM was supported by the Spanish Ministry of Science, Innovation and Universities (grant RTI2018-096311-B-I00), the Agencia Estatal de Investigación (AEI), the Fondo Europeo de Desarrollo Regional (FEDER). The authors acknowledge financial support from the Basque Government through the BERC 2018-2021 program, by the Spanish State Research Agency through BCBL Severo Ochoa excellence accreditation SEV-2015-0490, DeepText project (KK-2020/00088) and Ixa excellence research group (IT1343-19). AB enjoys a postdoctoral grant ESPDOC18/101 from the UPV/EHU and also acknowledges the support of the NVIDIA Corporation with the donation of a Titan V GPU used for this research.

References

Ahissar E, Nagarajan S, Ahissar M, Protopapas A, Mahncke H, Merzenich MM. 2001. Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. Proc Natl Acad Sci U S A. 98(23):13367–13372.

Arnal LH, Giraud A-L. 2012. Cortical oscillations and sensory predictions. Trends in Cognitive Sciences. 16(7):390–398.

Blank H, Davis MH. 2016. Prediction Errors but Not Sharpened Signals Simulate Multivoxel fMRI Patterns during Speech Perception. PLOS Biology. 14(11):e1002577.

Bojar O, Buck C, Federmann C, Haddow B, Koehn P, Leveling J, Monz C, Pecina P, Post M, Saint-Amand H, et al. 2014. Findings of the 2014 Workshop on Statistical Machine Translation. Proceedings of the Ninth Workshop on Statistical Machine Translation.

Breska A, Deouell LY. 2017. Neural mechanisms of rhythm-based temporal prediction: Delta phase-locking reflects temporal predictability but not rhythmic entrainment. PLoS biology. 15(2):e2001665.

Cardellino, Christian. 2016. Spanish Billion Words Corpus and Embeddings. https://crscardellino.github.io/SBWCE/

Cope TE, Sohoglu E, Sedley W, Patterson K, Jones PS, Wiggins J, Dawson C, Grube M, Carlyon RP, Griffiths TD, et al. 2017. Evidence for causal top-down frontal contributions to predictive processes in speech perception. Nat Commun. 8(1):2154.

Devereux BJ, Clarke A, Tyler LK. 2018. Integrated deep visual and semantic attractor neural networks predict fMRI pattern-information along the ventral object processing pathway. *Scientific reports*, *8*(1):1-12.

Ding N, Melloni L, Zhang H, Tian X, Poeppel D. 2016. Cortical tracking of hierarchical linguistic structures in connected speech. Nat Neurosci. 19(1):158–164.

Ding N, Patel AD, Chen L, Butler H, Luo C, Poeppel D. 2017. Temporal modulations in speech and music. Neurosci Biobehav Rev. 81(Pt B):181–187.

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

Donhauser PW, Baillet S. 2020. Two distinct neural timescales for predictive speech processing. Neuron. 105(2):385-93.

Duchon A, Perea M, Sebastián-Gallés N, Martí A, Carreiras M. 2013. EsPal: one-stop shopping for Spanish word properties. Behav Res Methods. 45(4):1246–1258.

Dupoux E, Green K. 1997. Perceptual adjustment to highly compressed speech: effects of talker and rate changes. J Exp Psychol Hum Percept Perform. 23(3):914–927.

Etard O, Reichenbach T. 2019. Neural Speech Tracking in the Theta and in the Delta Frequency Band Differentially Encode Clarity and Comprehension of Speech in Noise. The Journal of Neuroscience. 39(29):5750–5759. doi:10.1523/jneurosci.1828-18.2019. http://dx.doi.org/10.1523/jneurosci.1828-18.2019.

Driedger, Jonathan and Meinard Müller. “TSM Toolbox: MATLAB Implementations of Time-Scale Modification Algorithms.” DAFx (2014).

Ettinger A, Linzen T, Marantz A. 2014. The role of morphology in phoneme prediction: evidence from MEG. Brain Lang. 129:14–23.

Freunberger D, Roehm D. 2016. Semantic prediction in language comprehension: evidence from brain potentials. Lang Cogn Neurosci. 31(9):1193–1205.

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

Ghitza O. 2014. Behavioral evidence for the role of cortical Î¸ oscillations in determining auditory channel capacity for speech. Frontiers in Psychology. 5. doi:10.3389/fpsyg.2014.00652. http://dx.doi.org/10.3389/fpsyg.2014.00652.

Giraud A-L, Poeppel D. 2012. Cortical oscillations and speech processing: emerging computational principles and operations. Nat Neurosci. 15(4):511–517.

Giraud A-L, Kleinschmidt A, Poeppel D, Lund TE, Frackowiak RS, Laufs H. 2007. Endogenous cortical rhythms determine cerebral specialization for speech perception and production. Neuron. 56(6):1127-34.

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

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(12):e1001752.

Hale J. 2003. The information conveyed by words in sentences. Journal of Psycholinguistic Research, 32(2):101–123.

Hale J. 2016. Information-theoretical complexity metrics. Language and Linguistics Compass. 10(9):397–412.

Hickok G, Poeppel D. 2004. Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. Cognition. 92(1-2):67–99.

Hochreiter S, Schmidhuber J. 1997. Long short-term memory. Neural Comput. 9(8):1735–1780.

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

Keitel A, Gross J. 2016. Individual Human Brain Areas Can Be Identified from Their Characteristic

Spectral Activation Fingerprints. PLOS Biology 14(6): e1002498.

Klimovich-Gray A, Tyler LK, Randall B, Kocagoncu E, Devereux B, Marslen-Wilson WD. 2019. Balancing Prediction and Sensory Input in Speech Comprehension: The Spatiotemporal Dynamics of Word Recognition in Context. The Journal of Neuroscience. 39(3):519–527. doi:10.1523/jneurosci.3573-17.2018. http://dx.doi.org/10.1523/jneurosci.3573-17.2018.

Kösem A, Basirat A, Azizi L, van Wassenhove V. 2016. High-frequency neural activity predicts word parsing in ambiguous speech streams. J Neurophysiol. 116(6):2497–2512.

Kriegeskorte N, Mur M, Bandettini P. A. 2008. Representational similarity analysis-connecting the branches of systems neuroscience. Frontiers in systems neuroscience, 2, 4.

Lakatos P, Gross J, Thut G. 2019. A New Unifying Account of the Roles of Neuronal Entrainment. Curr Biol. 29(18):R890–R905.

Lakatos P, Karmos G, Mehta AD, Ulbert I, Schroeder CE. 2008. Entrainment of neuronal oscillations as a mechanism of attentional selection. Science. 320(5872):110–113.

Lam NHL, Schoffelen J-M, Uddén J, Hultén A, Hagoort P. 2016. Neural activity during sentence processing as reflected in theta, alpha, beta, and gamma oscillations. Neuroimage. 142:43–54.

Lau E, Almeida D, Hines PC, Poeppel D. 2009. A lexical basis for N400 context effects: evidence from MEG. Brain Lang. 111(3):161–172.

Lau EF, Nguyen E. 2015. The role of temporal predictability in semantic expectation: An MEG investigation. Cortex. 68:8–19.

Liebenthal E, Binder JR, Spitzer SM, Possing ET, Medler DA. 2005. Neural substrates of phonemic perception. Cereb Cortex. 15(10):1621–1631.

Lizarazu M, Lallier M, Molinaro N. 2019. Phase−amplitude coupling between theta and gamma oscillations adapts to speech rate. Annals of the New York Academy of Sciences. 1453(1):140–152. doi:10.1111/nyas.14099.

Maess B, Mamashli F, Obleser J, Helle L, Friederici AD. 2016. Prediction Signatures in the Brain: Semantic Pre-Activation during Language Comprehension. Front Hum Neurosci. 10:591.

Maris E, Oostenveld R. 2007. Nonparametric statistical testing of EEG- and MEG-data. J Neurosci Methods. 164(1):177–190.

Marslen-Wilson WD, Welsh A. 1978. Processing interactions and lexical access during word recognition in continuous speech. Cognitive Psychology. 10(1):29–63.

Meyer L, Henry MJ, Gaston P, Schmuck N, Friederici AD. 2017. Linguistic bias modulates interpretation of speech via neural delta-band oscillations. Cerebral Cortex. 27(9):4293-302.

Meyer L, Sun Y, Martin AE. 2020. Synchronous, but not entrained: Exogenous and endogenous cortical rhythms of speech and language processing. Language, Cognition and Neuroscience. 35(9):1089-99.

Merity, Stephen, Nitish Shirish Keskar, and Richard Socher. "Regularizing and Optimizing LSTM Language Models." International Conference on Learning Representations. 2018.

Molinaro N, Lizarazu M. 2018. Delta(but not theta)-band cortical entrainment involves speech-specific processing. Eur J Neurosci. 48(7):2642–2650.

Monsalve IF, Bourguignon M, Molinaro N. 2018. Theta oscillations mediate pre-activation of highly expected word initial phonemes. Sci Rep. 8(1):9503.

Mortensen DR, Littell P, Bharadwaj A, Goyal K, Dyer C, Levin L. 2016. PanPhon: A Resource for Mapping IPA Segments to Articulatory Feature Vectors. Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 3475–3484.

Myers BR, Lense MD, Gordon RL. 2019. Pushing the Envelope: Developments in Neural Entrainment to Speech and the Biological Underpinnings of Prosody Perception. Brain Sci. 9(3).

Obleser J, Kayser C. 2019. Neural entrainment and attentional selection in the listening brain. Trends in cognitive sciences. 23(11):913-26.

Park H, Ince RA, Schyns PG, Thut G, Gross J. 2015. Frontal top-down signals increase coupling of auditory low-frequency oscillations to continuous speech in human listeners. Current Biology. 25(12):1649-53.

Poeppel D, Florencia Assaneo M. 2020. Speech rhythms and their neural foundations. Nature Reviews Neuroscience. 21(6):322–334.

Poldrack RA, Wagner AD, Prull MW, Desmond JE, Glover GH, Gabrieli JD. 1999. Functional specialization for semantic and phonological processing in the left inferior prefrontal cortex. Neuroimage. 10(1):15–35.

Reese, Samuel, Gemma Boleda, Montse Cuadros, Lluís Padró, German Rigau. 2010. Wikicorpus: A Word-Sense Disambiguated Multilingual Wikipedia Corpus. In Proceedings of 7th Language Resources and Evaluation Conference (LREC'10).

Rimmele JM, Morillon B, Poeppel D, Arnal LH. 2018. Proactive Sensing of Periodic and Aperiodic Auditory Patterns. Trends Cogn Sci. 22(10):870–882.

Rothermich K, Schmidt-Kassow M, Kotz SA. 2012. Rhythm’s gonna get you: Regular meter facilitates semantic sentence processing. Neuropsychologia. 50(2):232–244. doi:10.1016/j.neuropsychologia.2011.10.025. http://dx.doi.org/10.1016/j.neuropsychologia.2011.10.025.

Schäfer, Roland, and Felix Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain." In LREC, pp. 486-493.

Sohoglu E, Davis MH. 2016. Perceptual learning of degraded speech by minimizing prediction error.

Proc Natl Acad Sci U S A. 113(12):E1747–56.

Taulu S, Simola J, Kajola M. 2005. Applications of the signal space separation method. IEEE Transactions on Signal Processing. 53(9):3359–3372. doi:10.1109/tsp.2005.853302. http://dx.doi.org/10.1109/tsp.2005.853302.

Temporal information in speech: acoustic, auditory and linguistic aspects. 1992. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences. 336(1278):367–373. doi:10.1098/rstb.1992.0070. http://dx.doi.org/10.1098/rstb.1992.0070.

Tyler LK. 1984. The structure of the initial cohort: evidence from gating. Percept Psychophys. 36(5):417–427.

Vigneau M, Beaucousin V, Hervé PY, Duffau H, Crivello F, Houdé O, Mazoyer B, Tzourio-Mazoyer N. 2006. Meta-analyzing left hemisphere language areas: phonology, semantics, and sentence processing. Neuroimage. 30(4):1414–1432.

Wang L, Kuperberg G, Jensen O. 2018. Specific lexico-semantic predictions are associated with unique spatial and temporal patterns of neural activity. Elife. 7. doi:10.7554/eLife.39061. http://dx.doi.org/10.7554/eLife.39061.

Willems RM, Frank SL, Nijhof AD, Hagoort P, van den Bosch A. 2016. Prediction During Natural Language Comprehension. Cerebral Cortex. 26(6):2506–2516. doi:10.1093/cercor/bhv075. http://dx.doi.org/10.1093/cercor/bhv075.

Wlotko EW, Federmeier KD. 2015. Time for prediction? The effect of presentation rate on predictive sentence comprehension during word-by-word reading. Cortex. 68:20–32.

Zhuang J, Devereux BJ. 2017. Phonological and syntactic competition effects in spoken word recognition: evidence from corpus-based statistics. Lang Cogn Neurosci. 32(2):221–235.

Zion Golumbic EM, Poeppel D, Schroeder CE. 2012. Temporal context in speech processing and attentional stream selection: a behavioral and neural perspective. Brain Lang. 122(3):151–161.

Figures

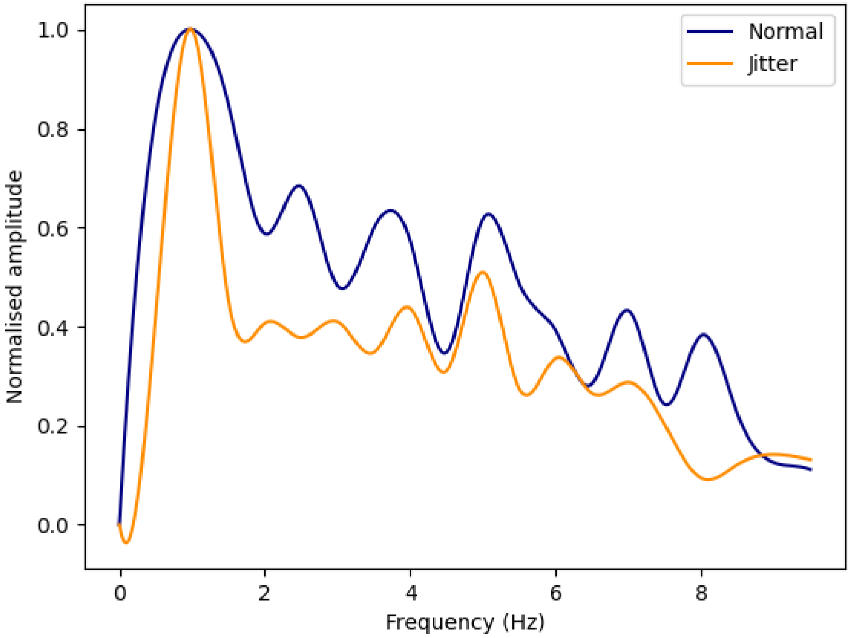

Figure 1- Normalised amplitude of the spectral power in Normal vs Jitter conditions Plotted for frequencies from 0 to 10 Hz for Normal (purple) and Jitter (orange) conditions.

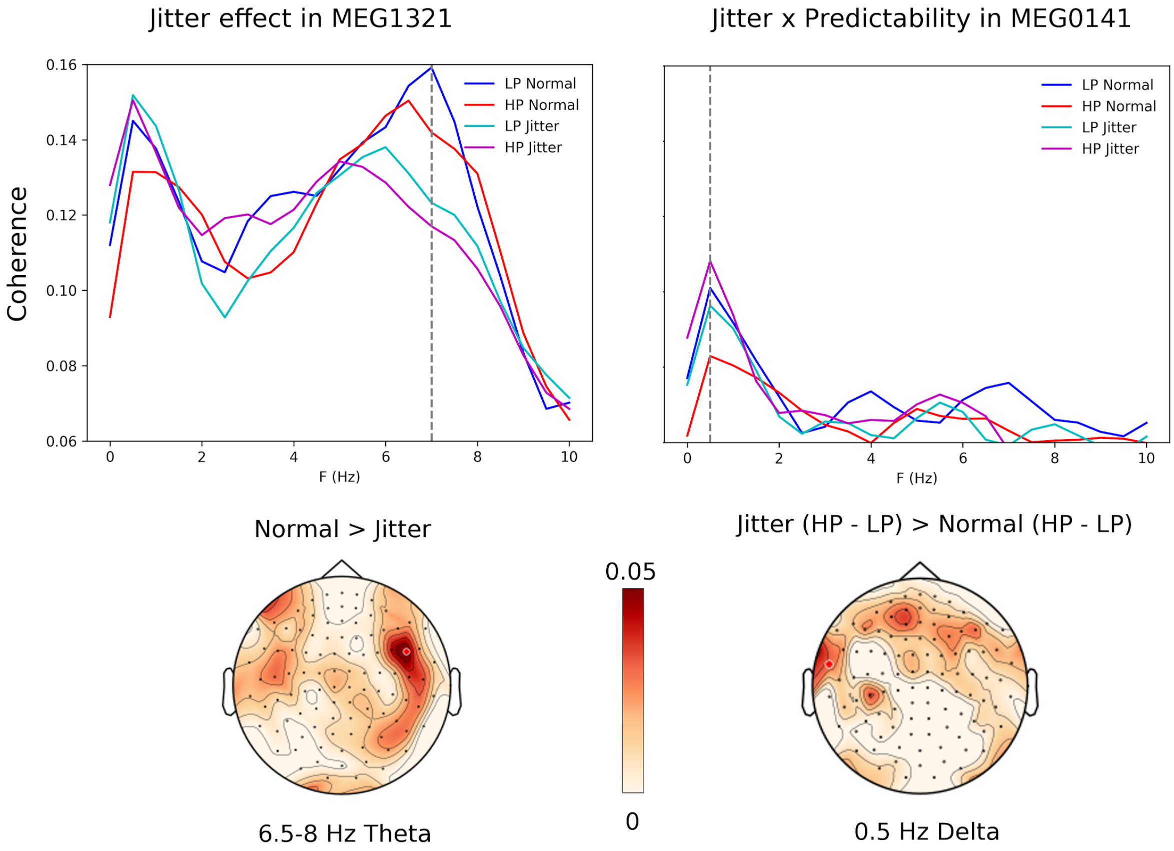

Figure 2 - Envelope coherence

Top panel - estimates of coherence between the spoken sentence envelope and the cortical signals in 4 conditions across 0-10 Hz frequency in sensors showing the highest effects. Bottom panel - topographies of the coherence values for two contrasts that showed significant effects. Significant differences between normal versus temporally jittered conditions emerged in the theta range 6.5-8 Hz, showing strongest effect in the right fronto-temporal sensors. A significant interaction between temporal jitter and context predictability was found in the delta band 0.5 Hz, with peak effects in the left frontal sensors.

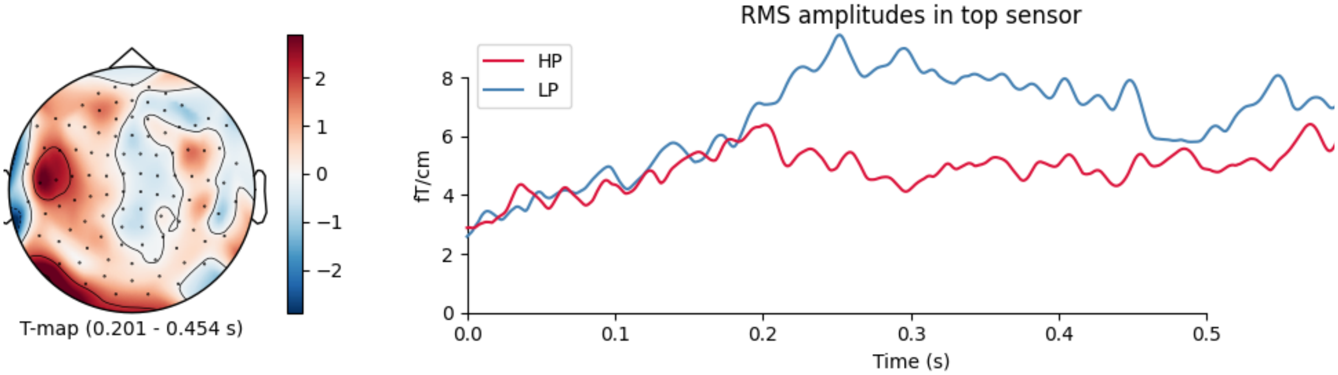

Figure 3 - N400 effect in Gradiometers (RMS)

Left - time-averaged topographical t-value map of the LP - HP contrast showing the main effect of target Predictability emerging in the frontotemporal sensors. Right - RMS signal plotted in the gradiometer which showed the peak of this main effect (in the 200 to 450 ms time window).

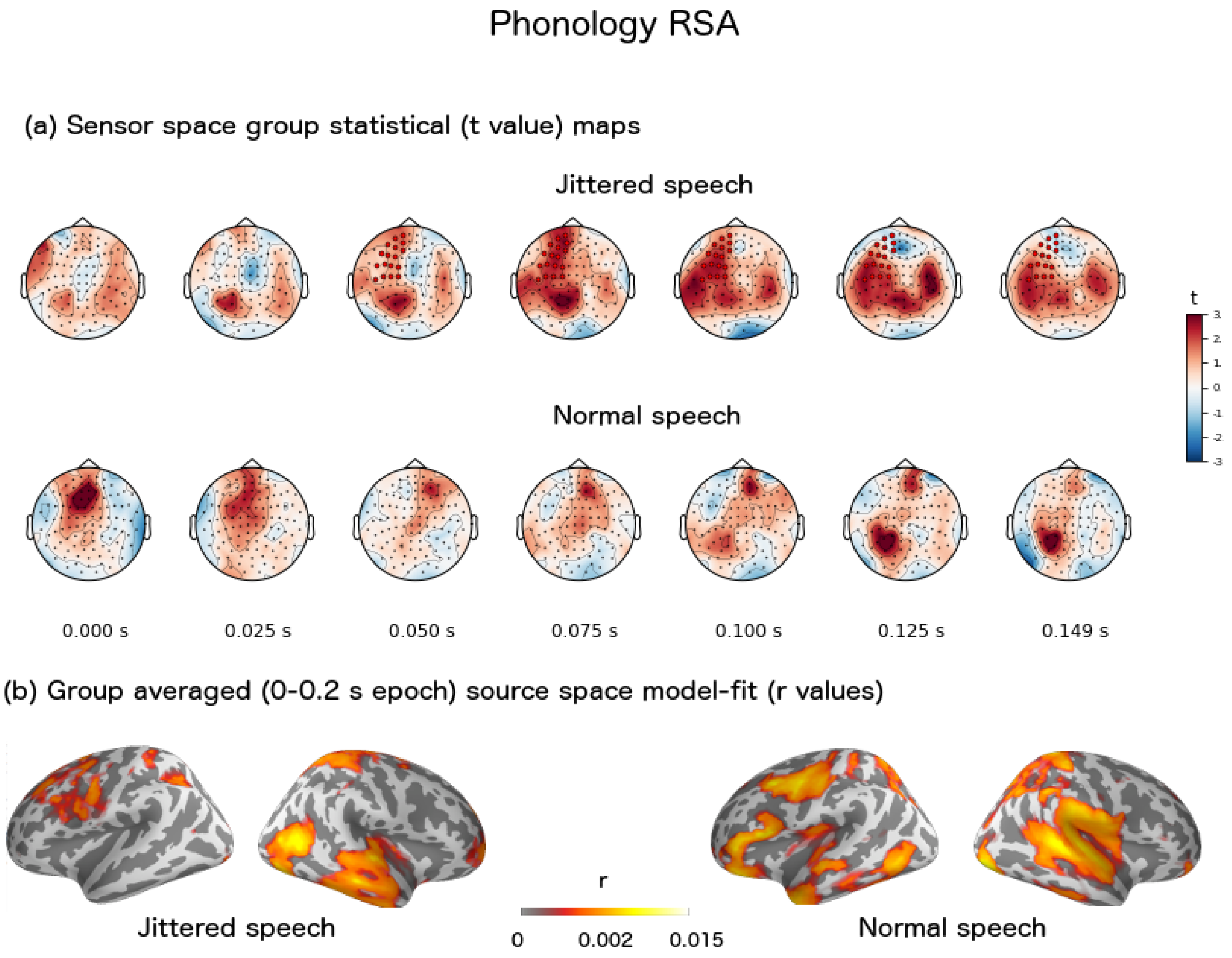

Figure 4 - RSA Phonology model-fit (a) Results of the spatiotemporal Searchlight analysis for the Phonology model. Over-time topographies of the t-vales derived from a one-sample t-test over individual subjects' model-fit r-value maps. The sensors that are part of the spatiotemporal cluster contributing to the significant model fit are marked in red. (b) Group-average source space model fit r-maps, averaged over 0-200ms epoch aligned to the target word onset.

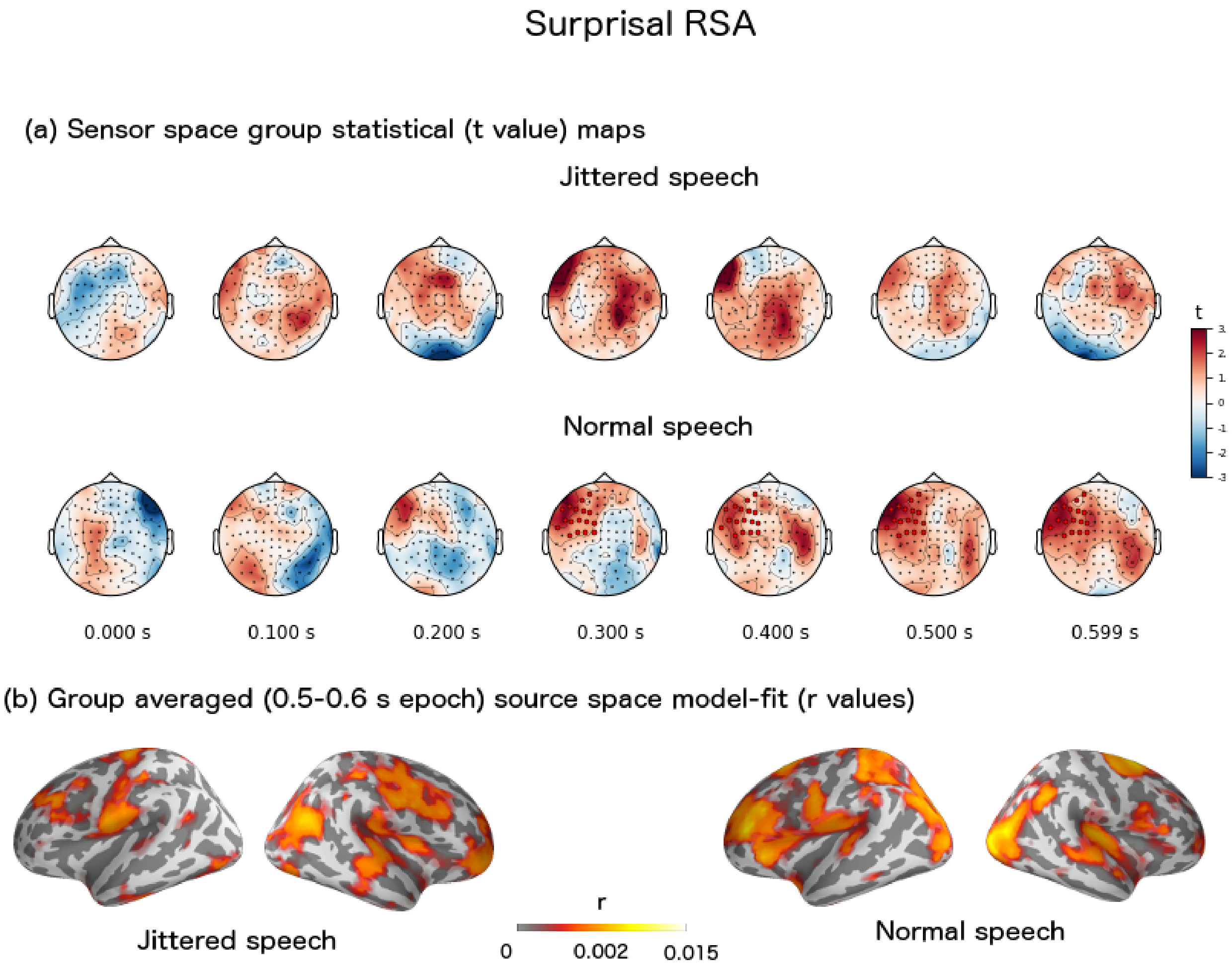

Figure 5 - RSA Surprisal model-fit (a) Results of the spatiotemporal Searchlight analysis for the Surprisal model. Over-time topographies of the t-vales derived from a one-sample t-test over individual subjects' model-fit r-value maps. The sensors that are part of the spatiotemporal cluster contributing to the significant model fit are marked in red. (b) Group-average source space model fit r-maps, averaged over 0.5-0.6 s epoch aligned to the target word onset.