Semantic context enhances neural envelope tracking

Eline Verschueren [a], **Jonas Vanthornhout** [a] **and Tom Francart** [a,*]

[a] *Research Group Experimental Oto-rhino-laryngology (ExpORL), Department of Neurosciences, KU Leuven - University of Leuven, 3000 Leuven, Belgium*

Correspondence*:
Eline Verschueren
Herestraat 49, bus 721, 3000 Leuven, Belgium
eline.verschueren@kuleuven.be

Conflict of interest: The authors declare no competing financial interests.

Keywords: cortical entrainment, neural decoding, semantic processing, natural speech, EEG, speech understanding

Acknowledgements: This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 637424 to Tom Francart). Further support came from KU Leuven Special Research Fund under grant OT/14/119. Research of Jonas Vanthornhout and Eline Verschueren is funded by a PhD grant of the Research Foundation Flanders (FWO).

ABSTRACT

The speech envelope is known to be essential for speech understanding and can be reconstructed from the electroencephalography (EEG) signal in response to running speech. Today, the factors influencing this neural tracking of the speech envelope are still under debate. Is envelope tracking mainly related to the encoding of acoustic speech information or is it influenced by top-down processing of speech understanding and the availability of semantic context in the stimulus?

We recorded the EEG in 19 normal-hearing participants while they listened to two types of stimuli: concatenated Matrix sentences without contextual information and a coherent story. Each stimulus was presented with varying levels of background noise to vary speech understanding. The speech envelope was reconstructed from the EEG in both the delta (0.5-4 Hz) and the theta (4-8 Hz) band with the use of a linear decoder and then correlated with the real speech envelope in that band. We also conducted a spatiotemporal analysis using temporal response functions (TRFs).

For both stimulus types and filter bands the correlation between the speech envelope and the reconstructed envelope increased with increasing speech understanding. In addition, correlations were higher for the story compared to the Matrix sentences, indicating that neural envelope tracking may be enhanced by the availability of semantic context in the stimulus and speech understanding.

1 INTRODUCTION

Speech is characterized by temporal modulations. The modulations reflecting syllable, word and sentence boundaries are also called the envelope of speech. This envelope is known to be an essential cue for speech understanding (Shannon et al., 1995) and can be reconstructed from brain responses (Aiken and Picton, 2008; Luo and Poeppel, 2007; Ding and Simon, 2011). Previous studies showed that the brain tracks the acoustic features of speech (Luo and Poeppel, 2007; Peelle et al., 2013) and that envelope tracking can be enhanced by attention (Kerlin et al., 2010; Ding and Simon, 2012; Mesgarani and Chang, 2012), grammar knowledge (Ding et al., 2015; Meyer et al., 2017) or prior knowledge of the stimulus (Di Liberto et al., 2018). However, it is not entirely clear yet how speech understanding and semantic context relate to neural envelope tracking.

A number of researchers have investigated the relationship between neural envelope tracking and speech understanding. Molinaro and Lizarazu (2017) showed enhanced tracking in the delta band ($<$ 4 Hz) using speech versus non-speech stimuli, while Luo and Poeppel (2007) obtained similar results but in the theta band (4-8 Hz) using speech-noise chimeras. Howard and Poeppel (2010), in contrast, reported no envelope tracking difference (3-8 Hz) between intelligible and time-reversed sentences. Besides changing the speech signal itself, adding background noise to the signal can also vary speech understanding. Ding and Simon (2013) showed that envelope tracking in the theta band (4-8 Hz) gradually decreases with increasing noise level, while the delta band (1-4 Hz) is more robust to noise. In contrast, Ding et al. (2014) found increased envelope tracking in the delta band with increasing speech understanding. Vanthornhout et al. (2018) report similar results for the delta band (0.5-4 Hz). In their study the same sentences were used during a recall experiment and an electroencephalography (EEG) measurement, enabling the researchers to directly compare speech understanding to envelope tracking. With this study we aim to investigate if neural envelope tracking is related to speech understanding and which particular frequency band of the brain response is most associated with this.

In addition to the debate on neural envelope tracking and speech understanding, we also want to address the role of semantic context. Understanding speech relies on the active integration of two incoming information streams (Hickok and Poeppel, 2007; Gross et al., 2013). The acoustic information stream (bottom-up) processes the incoming acoustic features through the auditory pathway until the auditory cortex. The top-down stream, on the other hand, originates in different brain regions containing prior knowledge of the upcoming speech, for example, semantic context (Wild et al., 2012; Lewis and Bastiaansen, 2015). When speech is degraded because of background noise, the quality of the acoustic information (bottom-up) decreases. Access to semantic context (top-down) can compensate for this acoustic loss (Stickney and Assmann, 2001; Boothroyd et al., 1988). A recent study of Broderick et al. (2018) demonstrated that the brain encodes semantic dissimilarities of speech in a time-locked way (1-8 Hz). Di Liberto et al. (2018) confirmed these results and nuanced the use of the filter band: envelope tracking was more enhanced due to prior knowledge in the delta band, covering the slower modulations of meaningful phrases and sentences,

compared to the theta band. Therefore we hypothesize that the availability of semantic context will lead to enhanced envelope tracking, especially in the delta band.

In the current study we investigated neural envelope tracking using EEG. To study the effect of speech understanding we varied the level of background noise, similar to previous studies (Vanthornhout et al., 2018; Ding et al., 2014; Ding and Simon, 2013). To additionally analyze the influence of semantic context, we used two speech materials with a varying degree of semantic context. We hypothesize that neural envelope tracking will be enhanced with increasing speech understanding and semantic context, especially in the delta band.

2 MATERIAL AND METHODS

2.1 Participants

Nineteen participants aged between 18 and 28 years (3 men and 16 women) took part in the experiment after providing informed consent. Participants had Flemish as their mother tongue and were all normal-hearing, confirmed with pure tone audiometry (thresholds $\leq$ 25 dB HL at all octave frequencies from 125 Hz to 8 kHz). The study was approved by the Medical Ethics Committee UZ Leuven / Research (KU Leuven) with reference S57102. All participants were unpaid volunteers.

2.2 Auditory stimuli

During the experiment participants listened to three different speech materials: (1) Matrix sentences (no semantic context), (2) a story (semantic context) and (3) a story used to train the linear decoder on.

2.2.1 Matrix sentences (no semantic context)

Flemish Matrix sentences always contain 5 words spoken by a female speaker and have a fixed syntactic structure of 'proper name-verb-numeral-adjective-object', for example, 'Sofie sees ten blue socks' with a speech rate of 4.1 syllables/second, 2.5 words/second and 0.5 sentences/second. Each category of words has 10 options and each sentence consists of a random combination of these options, which reduces semantic context to a bare minimum. These sentences are gathered into standardized lists of 20 sentences. Speech was always fixed at 60 dBA and the noise level varied across trials. We used speech weighted noise (SWN) which has the long-term-average spectrum of the stimulus and therefore results in optimal energetic masking. We chose to use Matrix sentences because this is a validated speech material to measure speech understanding which allows us to directly compare EEG results with speech understanding. In addition, Matrix sentences are a random combination of words, but maintain the correct syntax. This enables us to present syntactical correct speech only lacking semantic context.

2.2.2 Coherent story (semantic context)

The coherent story we used is 'De Wilde Zwanen', written by Hans Christian Andersen and narrated in Flemish by Katrien Devos (female speaker) with a speech rate of 3.5 syllables/second, 2.5 words/second

and 0.2 sentences/second. Speech was always fixed at 60 dBA and the noise level of the SWN varied across trials. Using a fairy tale with maximal semantic context allows us, when comparing with the Matrix sentences, to study the influence of the availability of semantic context in the stimulus on neural envelope tracking. To measure speech intelligibility of the story, we cannot ask the participant to recall every word. Therefore we used a rating method, where the subjects were asked to rate their speech understanding.

2.2.3 Decoder story

A children's story, 'Milan', written and narrated in Flemish by Stijn Vranken (male speaker), was presented to the participants with a speech rate of 3.7 syllables/second, 2.6 words/second and 0.3 sentences/second. This story is 14 minutes long and was presented at 60 dBA without noise. The purpose of this story was to have a continuous stimulus without background noise to train a linear decoder on (Vanthornhout et al., 2018) to reconstruct the speech envelope from the EEG. To maximize attention, a content question was asked.

2.3 Behavioral experiment

Speech understanding was measured behaviorally in order to compare envelope tracking results in terms of speech understanding. We need to measure speech understanding for both stimuli separately because they differ in content and acoustic parameters (speaker, speech rate, intonation). Adding a similar level of background noise will therefore not result in a similar level of speech understanding.

Before the EEG experiment we conducted a standardized Matrix test. This standardized test starts with 2 training lists followed by 3 testing lists of 20 sentences at different Signal-to-Noise Ratios (SNR): -9.5; -6.5 and -3.5 dB SNR. Subjects had to recall the sentence they heard. By counting the correctly recalled words, a percentage correct per presented SNR was calculated. Next, a psychometric function could be plotted through the data points to obtain estimated levels of speech understanding at different SNRs. To measure speech understanding for the story, we cannot ask the participants to recall every word, instead we used a rating method during the EEG experiment. Participants were asked to rate their speech understanding at the presented SNRs: -12.5; -9.5; -6.5; -3.5; -0.5 and 2.5 dB SNR.

2.4 EEG experiment

Ten subjects started the EEG experiment by listening to Matrix sentences followed by the coherent story. The remaining 9 subjects did this in the reversed order. The decoder story was presented in between. The coherent story was cut in 7 equal parts of approximately 4 minutes, which we presented in chronological order to optimize the use of semantic context. The first part was always presented in silence to optimize comprehension of the storyline. The following 6 parts were presented at 6 different SNRs in random order: -12.5; -9.5; -6.5; -3.5; -0.5 and 2.5 dB SNR. The Matrix sentences were concatenated into 7 lists of 40 sentences with a silent gap between the sentences randomly varying between 0.8 and 1.2 seconds. Each 2-minute trial, containing 40 sentences at a particular SNR, was presented twice to analyze test-retest reliability. The SNRs were the same SNRs as used for the story, also in random order. To maximize

attention and keep the subjects motivated, questions were asked about each SNR trial, for example, 'What happened after sunset?' or 'Which colors of boats were mentioned?' (answers were not used for further analysis). After the question, the subjects were asked to rate their speech understanding with the following question: 'Which percentage of the words did you understand?'.

2.5 Signal processing

In this study we measured neural envelope tracking and linked this to speech understanding and semantic context. Neural envelope tracking was calculated in two ways. (1) We correlated the acoustic speech envelope with the reconstructed speech envelope from the EEG response. The more accurate the envelope is encoded in the brain, the more similar it will be to the acoustic envelope en thus the higher the correlation. (2) In addition, we calculated TRFs for each electrode to analyze neural envelope tracking in a spatiotemporal way.

2.5.1 Acoustic envelope

The acoustic speech envelope was extracted from the stimulus according to Biesmans et al. (2017), using a gammatone filterbank followed by a power law. We used a filterbank containing 28 channels spaced by 1 equivalent rectangular bandwidth with center frequencies from 50 Hz until 5000 Hz. The absolute value of each sample in each channel was raised to the power of 0.6. All 28 channel envelopes were averaged which resulted in one single envelope. As a next step, the acoustic speech envelope was band-pass filtered, similar to the EEG signal, in the delta (0.5-4 Hz) or theta (4-8 Hz) frequency band with a Chebyshev filter with 80 dB attenuation at 10% outside the passband. Only these low frequencies were further processed, because they contain the information of interest of the slow modulating speech envelope.

2.5.2 Envelope reconstruction

As a first step the EEG data was downsampled from 8192 Hz to 256 Hz to reduce processing time and referenced to an average of the electrodes. Next, EEG artefact rejection was done using a multi-channel Wiener filter (MWF) (Somers et al., 2018). the MWF was calculated on the long decoder story without noise and applied on the shorter Matrix and coherent story SNR trials. After artefact rejection, the signal was bandpass filtered, similar to the acoustic speech envelope and the sample rate was further decreased from 256 Hz to 128 Hz. To enable reconstruction of the speech envelope from the neural data as a measure of neural envelope tracking, a linear decoder was created using the mTRF toolbox (Lalor et al., 2006, 2009). As speech elicits neural responses with some delay, the decoder not only attributes weights to each EEG channel (spatial filter), but it also takes the shifted neural responses of each channel into account (temporal filter), resulting in a matrix R containing the shifted neural responses of each channel. If g is the linear decoder and R is the shifted neural data, the reconstruction of the speech envelope $\hat{s}$(t) was obtained by $\hat{s}(t) = \sum_n \sum_\tau g(n,\tau)R(t+\tau,n)$, with t the time ranging from 0 to T, n the recording electrodes ranging from 1 to N and τ the post-stimulus samples used to reconstruct the envelope. The decoder was calculated using ridge regression by solving $g = (RR^T)^{-1}(RS^T)$ with R as the time-lagged matrix of the neural data

and S as the speech envelope. As we used an integration window from 0 until 250 ms post-stimulus, the decoder matrix g was a 64 (EEG channels) x 33 (time delays) matrix. The decoder was created using the Milan story (14 minutes) without any noise.

As a last step to reconstruct the envelope the decoder was applied to both test stimuli, the Matrix sentences (no semantic context) and the coherent story (semantic context), at various noise levels after normalization. Each SNR trial consisted of 2 presentations of 80 seconds of speech (silences excluded). To measure how similar this reconstructed envelope is to the acoustic envelope as a measure for neural envelope tracking, we calculated the bootstrapped Spearman correlation using Monte Carlo sampling after removing the silences in the stimulus and the corresponding part in the EEG. Removing the silences is indispensable as the Matrix sentences contain quasi-regular silent gaps between the sentences which would lead to suboptimal decoding. A schematic overview is shown in Figure 1.

The significance level of the correlation is calculated by correlating random permutations of the real and reconstructed envelope 1000 times and taking percentile 2.5 and 97.5 to obtain a 95% confidence interval.

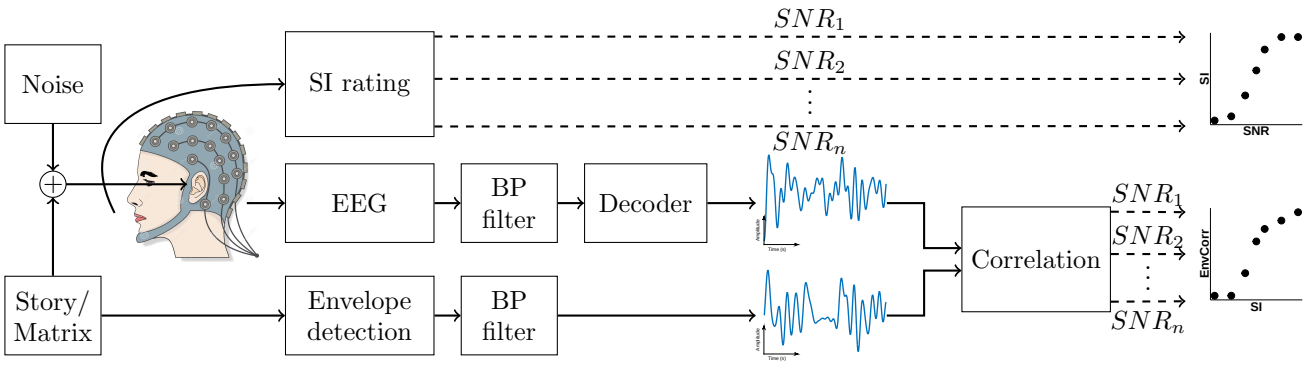

Figure 1. Overview of the experimental setup. We presented the Matrix sentences and a story. Participants listened to the speech with different levels of background noise and (1) rated their speech understanding while (2) their EEG was measured. To obtain a measure of envelope tracking we correlated the reconstructed envelope with the acoustic envelope. We compared the envelope tracking results with the rated speech understanding scores.

2.5.3 Temporal response function estimation

The analysis above integrates all neural activity over channels and time lags. To have a closer look at the spatiotemporal profile of the neural responses, we calculated TRFs. A TRF is a linear filter that describes how the acoustic speech envelope of the stimulus is transformed into neural responses. This is the inverse approach of the previously mentioned envelope reconstruction where analysis is done from EEG to stimulus.

We calculated a TRF for every electrode channel in every subject. The first signal processing steps are identical to the envelope reconstruction model starting with downsampling to 1028 Hz, artefact rejection

with MWF and filtering (0.5-8 Hz). Next, TRFs were calculated using the boosting algorithm (David et al., 2007; Brodbeck et al., 2018) with an l2 error norm (using the Eelbrain source code (Brodbeck, 2017)) as described in detail by David et al. (2007). After calculation, the TRFs were convolved with a rotationally symmetric Gaussian kernel of 5 samples long (SD = 2). To analyze the TRFs in the time domain, we investigate the latency and amplitude of the negative and positive peaks occurring directly after the stimulus onset (Ding and Simon, 2011; Ding et al., 2014; Obleser and Kotz, 2011; Ding and Simon, 2012).

2.6 Experimental design

Recordings were made in a soundproof and electromagnetically shielded room. Speech was presented bilaterally at 60 dBA and the setup was calibrated using a $2cm^3$ coupler of the artificial ear (Brüel & Kjær 4152, Denmark) per speech material. The stimuli were presented using APEX 3 (Francart et al., 2008), an RME Multiface II sound card (Germany) and Etymotic ER-3A insert phones (Illinois, USA). First the participants did a behavioral test to measure their speech understanding. Next, a 64-channel BioSemi ActiveTwo (the Netherlands) EEG recording system was used for the EEG recordings at a sample rate of 8192 Hz. Subjects sat in a comfortable chair and were asked to move as little as possible during the recordings. We inserted a small break between the behavioral and the EEG part and between the Matrix sentences and the story if necessary.

2.7 Statistical Analysis

Statistical analysis was performed using MATLAB (version R2016b) and R (version 3.3.2) software. The significance level was set at α=0.05 unless otherwise stated.

For the behavioral tests and envelope reconstruction we compared dependent samples (e.g. test-retest) using a nonparametric Wilcoxon signed-rank test. For every filter band and speech material we tested the correlation between envelope reconstruction and speech understanding using Spearman's rank correlation. Next, we assessed the relationship between speech understanding, envelope reconstruction, filter band and speech material by constructing a linear mixed effect (LME) model of envelope tracking in function of speech understanding (continuous variable) with interaction and main effects of speech material (2 factors) and filter band (2 factors) and a random intercept per participant with the following formula:

$$corr \sim SI + material + band + SI : band + SI : material + SI : band : material$$

where *corr* is defined as the Spearman correlation between the reconstructed and the acoustic envelope, with random effect of intercept of the participants and fixed and interaction effects of *SI* (speech intelligibility), *material* (the stimuli with differing level of semantic context) and *band* (the delta or theta filterband). To control for the effect of SNR, we constructed the exact same model, but in function of SNR instead of SI.

To control if every chosen fixed and random effect benefited the model the Akaike Information Criterion (AIC) was calculated. The model with the lowest AIC was selected and its residual plot was analyzed to

assess the normality assumption of the LME residuals. Unstandardized regression coefficients (beta) with 95% confidence intervals and p-value are reported in the results section.

To investigate which part of the TRF was significantly different from zero, we conducted a cluster-based permutation test. To explore significant differences between stimuli we conducted a positive and negative cluster-based analysis with a post hoc Bonferroni adjustment to correct for the positive and negative test. These tests are explained in detail by Maris and Oostenveld (2007). Spearman's rank correlation was used to investigate the possible change of amplitude and latency of the temporal-occipital peaks over time.

3 RESULTS

3.1 Speech understanding: Story versus Matrix sentences

During the experiment we measured speech understanding behaviorally at different SNRs for every participant. Figure 2 shows that the story (context) was significantly more difficult than the Matrix sentences (no context) ($p < 0.001$, CI(95%) = [15.99; 23.34], n=19, Wilcoxon signed-rank test). This indicates that the same SNR does not result in the same level of speech understanding for the different speech materials. To be able to compare the coherent story with the Matrix sentences in terms of semantic context and speech understanding, we need to account for this.

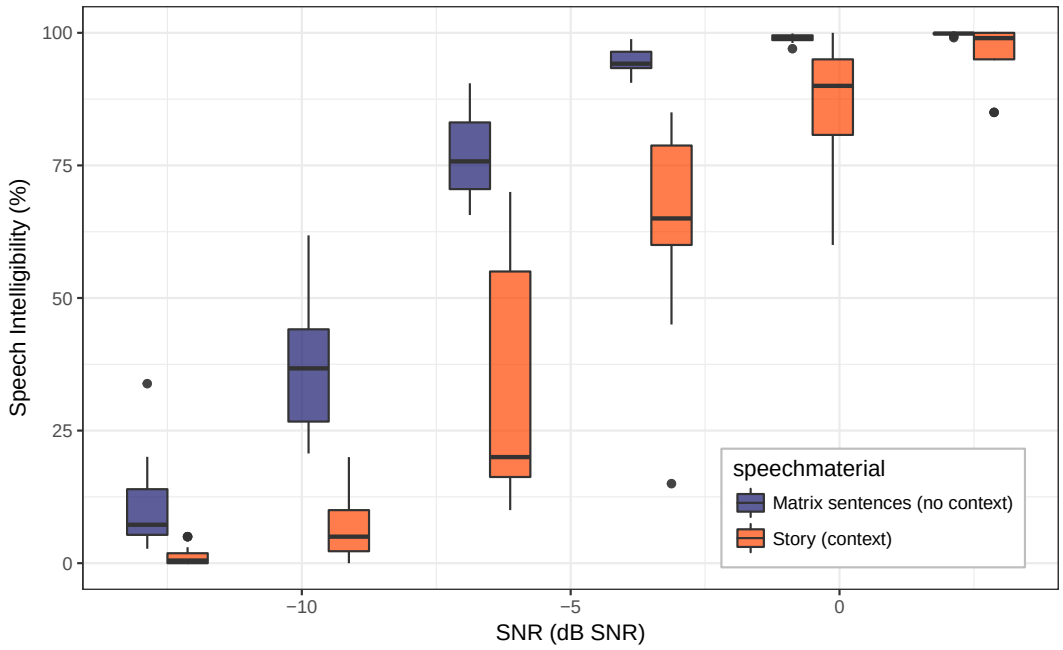

Figure 2. A comparison between the Matrix sentences and the story reveals that the story is more difficult to understand when adding background noise.

3.2 Envelope reconstruction

To measure neural envelope tracking, we calculated the Spearman correlation between the reconstructed envelope and the acoustic envelope. Conducting a test-retest analysis showed no significant difference between test and retest correlations (p=0.857, CI(95%) = [-0.005; 0.006], Wilcoxon signed-rank test), therefore we averaged the correlation of the test and retest conditions resulting in one correlation per participant per SNR per speech material. We also conducted a chance level analysis to investigate whether there is a difference in chance level between both stimuli. A difference in chance level would imply that the decoder would show a preference to one of the two stimuli. To obtain the chance level we reconstructed the envelope of the story similar to the standard analysis. Next we correlated the reconstructed envelope of each story trial with the acoustic envelope of all trials of both the story (except for the used trial) and the Matrix sentences. No significant difference was found between the chance level of the speech materials (p=0.534, CI(95%) = [-0.005; 0.003], Wilcoxon signed-rank test). In addition, the 95% confidence interval of the difference between the chance level of the stimuli is similar to the test-retest variability (CI(95%) = [-0.005; 0.006]), indicating that this non-significant difference is also a small difference.

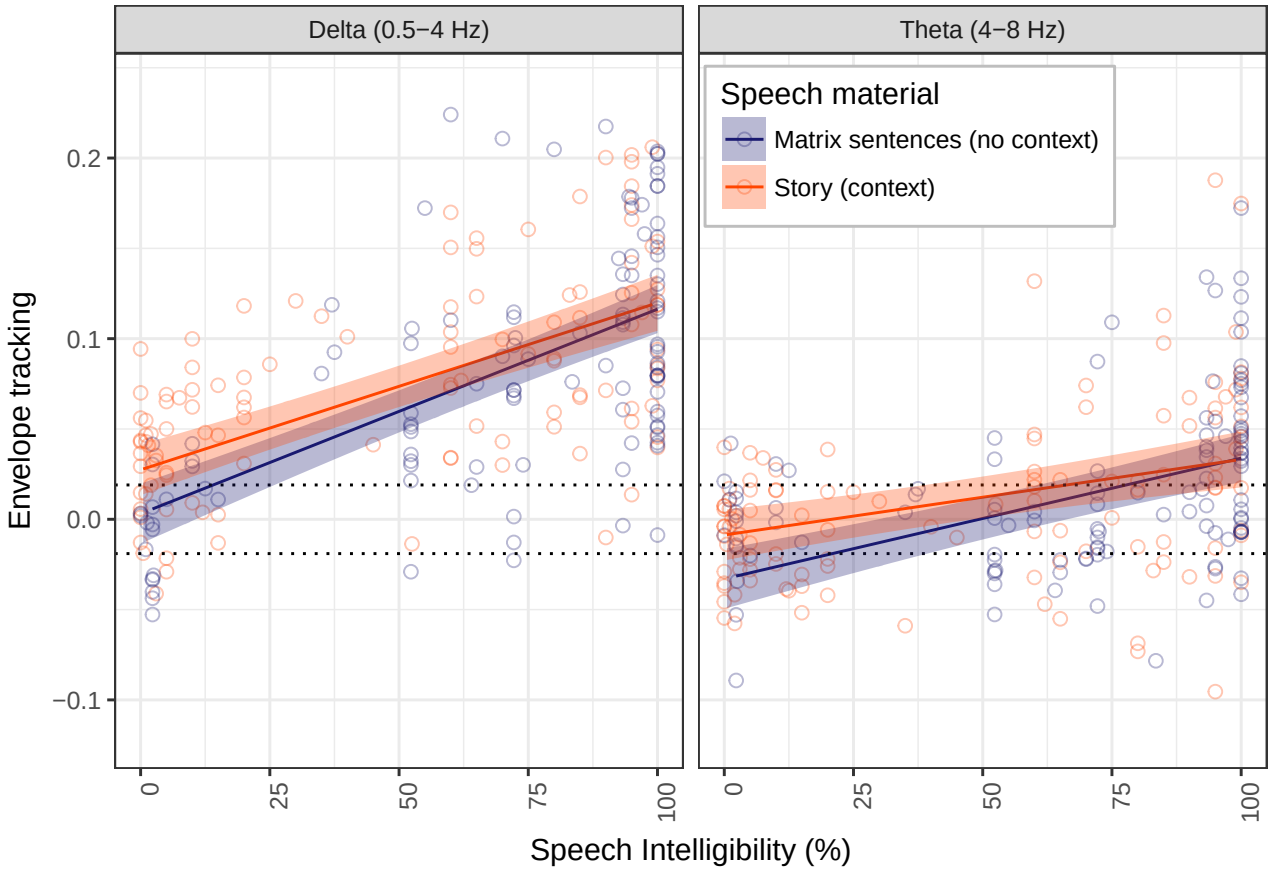

Figure 3. Neural envelope tracking increases with increasing speech intelligibility and by having semantic context available. The shading represents two times the standard error of the fit and the dotted line is the significance level of the correlation (±0.019).

Table 1. Spearman rank correlation between neural envelope tracking and speech understanding

Speechmaterial	Filter band	Correlation	p-value
Matrix sentences (no context)	Delta (0.5-4 Hz)	0.62	$p<0.001$
Story (context)	Delta (0.5-4 Hz)	0.59	$p<0.001$
Matrix sentences (no context)	Theta (4-8 Hz)	0.46	$p<0.001$
Story (context)	Theta (4-8 Hz)	0.41	$p<0.001$

We analyzed the decoding accuracy of the speech envelope in the delta (0.5-4 Hz) and the theta (4-8 Hz) band for two stimuli with a different level of semantic context and with various levels of speech understanding. Figure 3 shows that when speech understanding increases, the correlation between the acoustic and the reconstructed envelope, i.e. neural envelope tracking, also increases for every filter band and every stimulus tested (table 1, Spearman rank correlation).

Table 2. Linear Mixed Effect Model of envelope reconstruction in function of SI

Linear mixed effect model (factor)	beta value	CI(95%)	p-value
Fixed effect SI	1.08 x 10^{-3}	$\pm$ 1.90 x 10^{-4}	$p<0.001$
Fixed effect material	1.97 x 10^{-2}	$\pm$ 1.49 x 10^{-2}	p=0.010
Fixed effect band	-3.87 x 10^{-2}	$\pm$ 1.41 x 10^{-2}	$p<0.001$
Interaction effect SI:material	-1.74 x 10^{-4}	$\pm$ 2.39 x 10^{-4}	p=0.155
Interaction effect SI:band	-4.43 x 10^{-4}	$\pm$ 2.14 x 10^{-4}	$p<0.001$
Interaction effect SI:band:material	-1.28 x 10^{-5}	$\pm$ 2.25 x 10^{-4}	p=0.912

To additionally investigate the influence of semantic context, we created an LME model as a function of speech understanding. The analysis shows that neural envelope tracking is enhanced by a stimulus with more semantic context (fixed effect material, p=0.010, LME, table 2). This increase due to semantic context does not significantly depend on the level of speech understanding or filter band (interaction effect SI:material, p=0.155; interaction effect SI:band:material, p=0.912; LME, table 2). Further, envelope tracking in the delta band (0.5-4 Hz) is higher than in the theta band (4-8 Hz) (fixed effect band, $p<0.001$, LME, table 2) and the increase of envelope tracking with speech intelligibility is filter band dependent with a steeper slope in the delta band (0.5-4 Hz) (interaction effect SI:band, $p<0.001$, LME, table 2).

Because the same SI results in a different SNR per speech material, and neural envelope tracking could be influenced by SNR, we checked whether the obtained effect of semantic context was not just an effect of SNR by conducting the same analysis in function of SNR. The same fixed and interaction effects were

found to be significant as for the SI analysis (table 3), showing that even at the same SNR neural envelope tracking for the story is enhanced compared to the Matrix sentences.

Table 3. Linear Mixed Effect Model of envelope reconstruction in function of SNR

Linear mixed effect model (factor)	beta value	CI(95%)	p-value
Fixed effect SNR	7.75 x 10^{-3}	$\pm$ 1.39 x 10^{-3}	p<0.001
Fixed effect material	-1.25 x 10^{-2}	$\pm$ 1.06 x 10^{-2}	p=0.022
Fixed effect band	-8.10 x 10^{-2}	$\pm$ 1.06 x 10^{-2}	p<0.001
Interaction effect SNR:material	-1.01 x 10^{-3}	$\pm$ 1.83 x 10^{-3}	p=0.284
Interaction effect SNR:band	-3.20 x 10^{-3}	$\pm$ 1.83 x 10^{-3}	p<0.001
Interaction effect SNR:band:material	-1.40 x 10^{-6}	$\pm$ 2.13 x 10^{-3}	p=0.999

3.3 Temporal response function

The analysis above integrates all different time lags and channels to obtain a reconstruction of the envelope. In the following analysis we focus on how the neural responses follow the envelope in the time and spatial domain by investigating TRFs. TRFs were calculated on an individual level. This resulted in 868 TRFs per participant (64 channels x 2 speech materials x 7 SNRs). To visualize topoplots and TRF time courses, we averaged the TRFs per speech material per SNR over participants. Figure 4 shows the spatiotemporal activation profile of respectively the Matrix sentences (no semantic context) and the story (semantic context). In the no-noise condition both speech materials show positive central and negative parieto-occipital amplitudes over time. When only a little amount of noise is added and speech understanding remains almost unchanged (SNR = 2.5 dB SNR; Matrix sentences: median SI = 99.9%, sd = 0.2; Story: median SI = 99.0%, sd = 4.7), the amplitudes decrease between 0 to 150 ms, while amplitudes between 150 to 200 ms increase in both speech materials. Between 50 and 100 ms amplitudes even swap polarities. When more noise is added and speech understanding decreases all amplitudes decrease with decreasing speech understanding, except for the negative central activation between 50 and 100 ms that reaches a maximum at SNR = -3.5 dB SNR.

To investigate the influence of semantic context we subtracted the TRFs of the story from the Matrix sentences. We did this analysis both in function of speech understanding and SNR to control for both factors. A positive and negative cluster analysis (Maris and Oostenveld, 2007) over all subjects revealed significant differences (α = 0.025) between both stimuli in the no-noise condition with larger amplitudes for the Matrix sentences, highlighted in red in Figure 4. In all conditions where noise was added to the speech signal, no significant differences could be found.

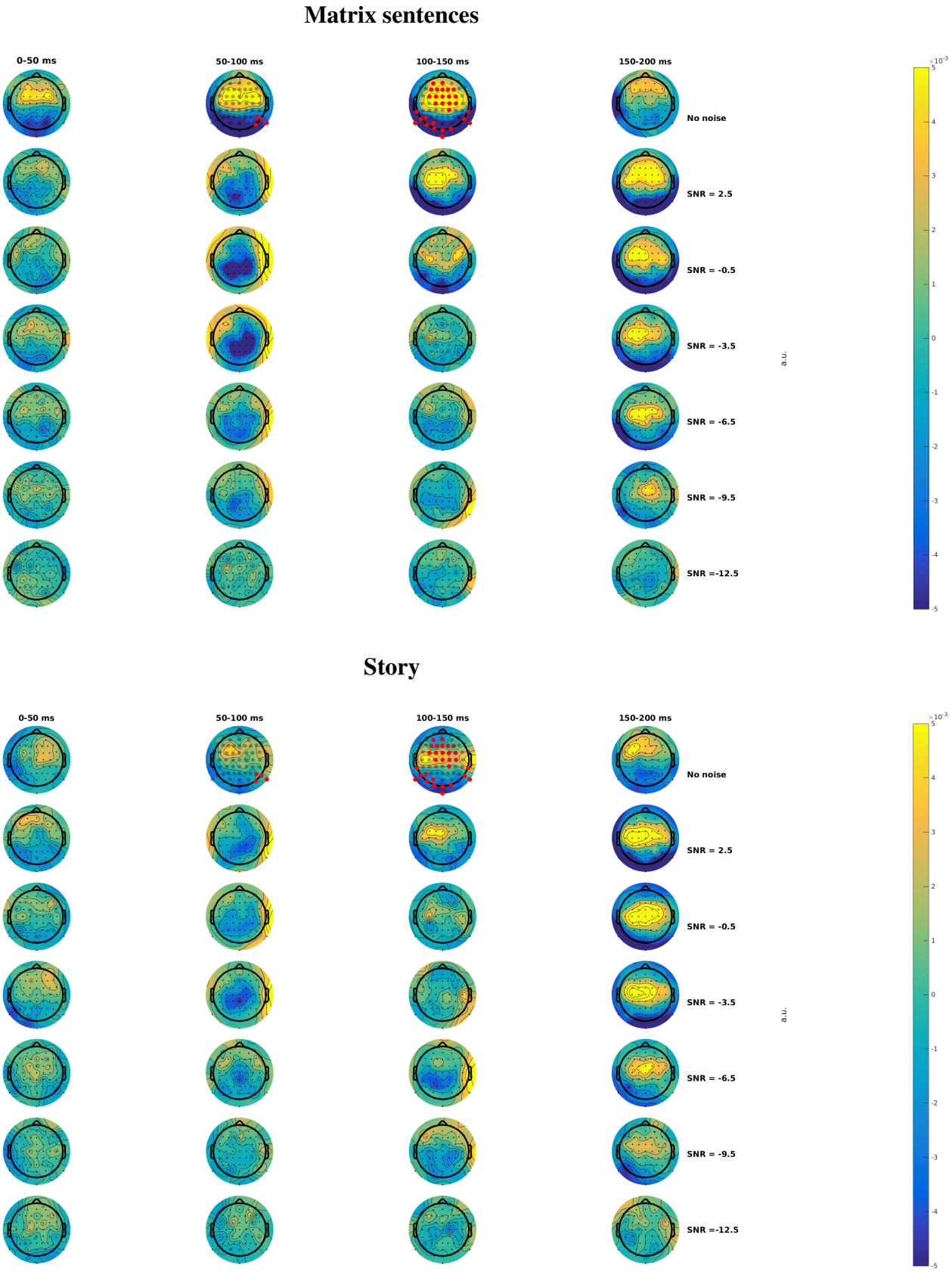

Figure 4. Topoplots for the story and the Matrix sentences at different SNRs and different time lags varying from 0 until 200 ms. Significant differences between the speech materials are highlighted in red.

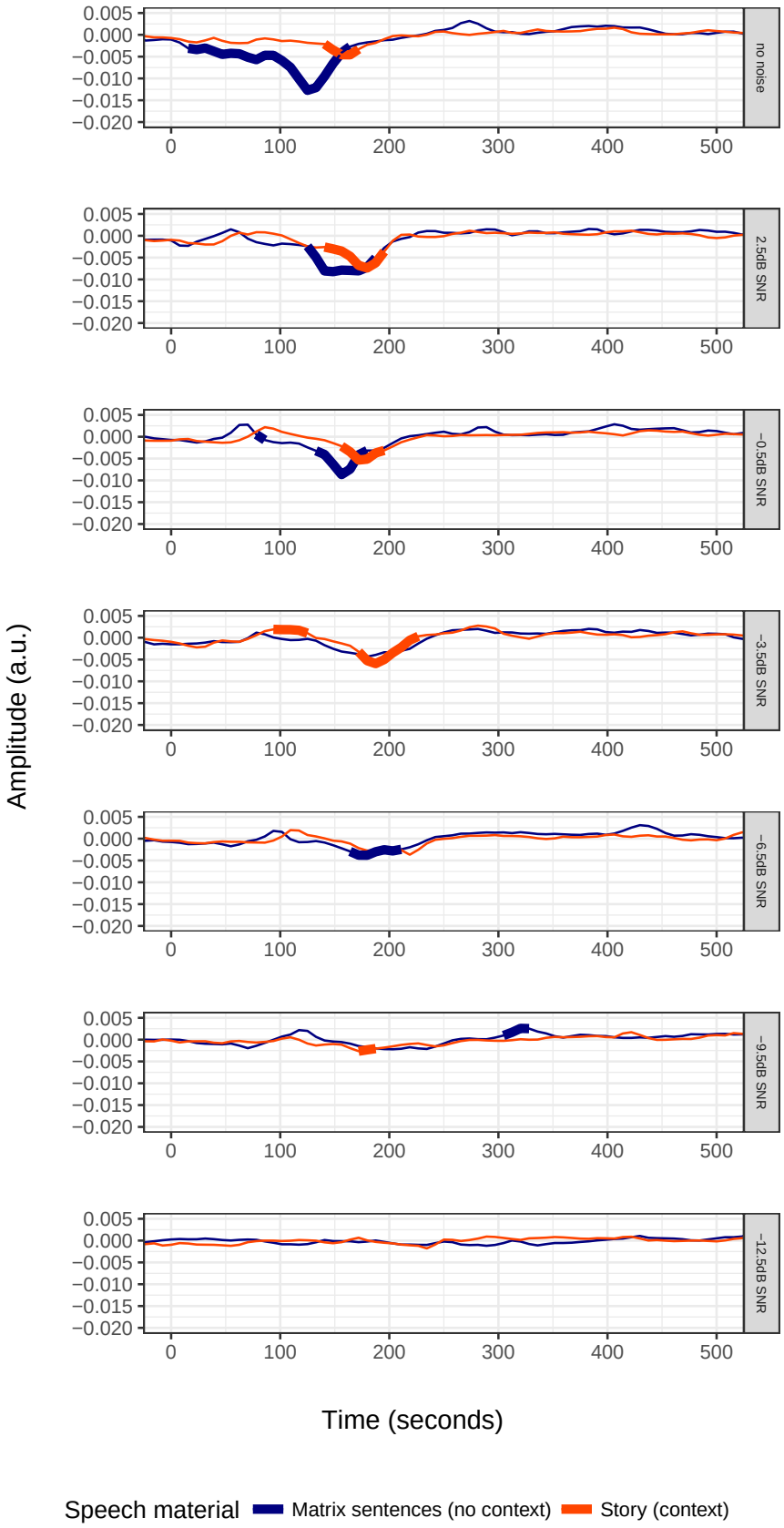

Figure 5. Significant TRF samples are highlighted in bold. The temporal-occipital channels show a negative peak between 100 and 200 ms for both speech materials shifting in latency and amplitude as speech understanding varies.

To investigate the time-course of the TRFs, we calculated TRFs for a temporal-occipital channel selection (Figure 7). This selection is data driven, based on the TRF results shown in Figure 4. A cluster-based permutation test (Maris and Oostenveld, 2007) shows the TRF samples significantly different from zero. These samples are highlighted in bold in Figure 5. Figure 5 shows that when SI is very low (SNR $<$ -12.5 dB SNR) both speech materials have very low responses over time. When speech is understood a negative peak can be found for both speech materials. Figure 6 shows the latency and amplitude results of this peak on a subject level over SI. It was determined individually by selecting the lowest amplitude of the TRF between 50 and 300 ms. With decreasing SI the amplitude of the negative peak per subject decreases for both speech materials (Matrix sentences: Spearman rank correlation = 0.49, $p<0.001$; Story: Spearman rank correlation = 0.26, $p = 0.005$). Besides the amplitudes, the latency also changes over SI for the Matrix sentences (Spearman rank correlation = 0.46, $p<0.001$), while for the story latency remains the same (Spearman rank correlation = 0.02, $p = 0.835$)). Next to the prominent peak between 100 and 200 ms, a positive significant peak arises around 300 ms for the Matrix sentences at -9.5 dB SNR (Figure 5).

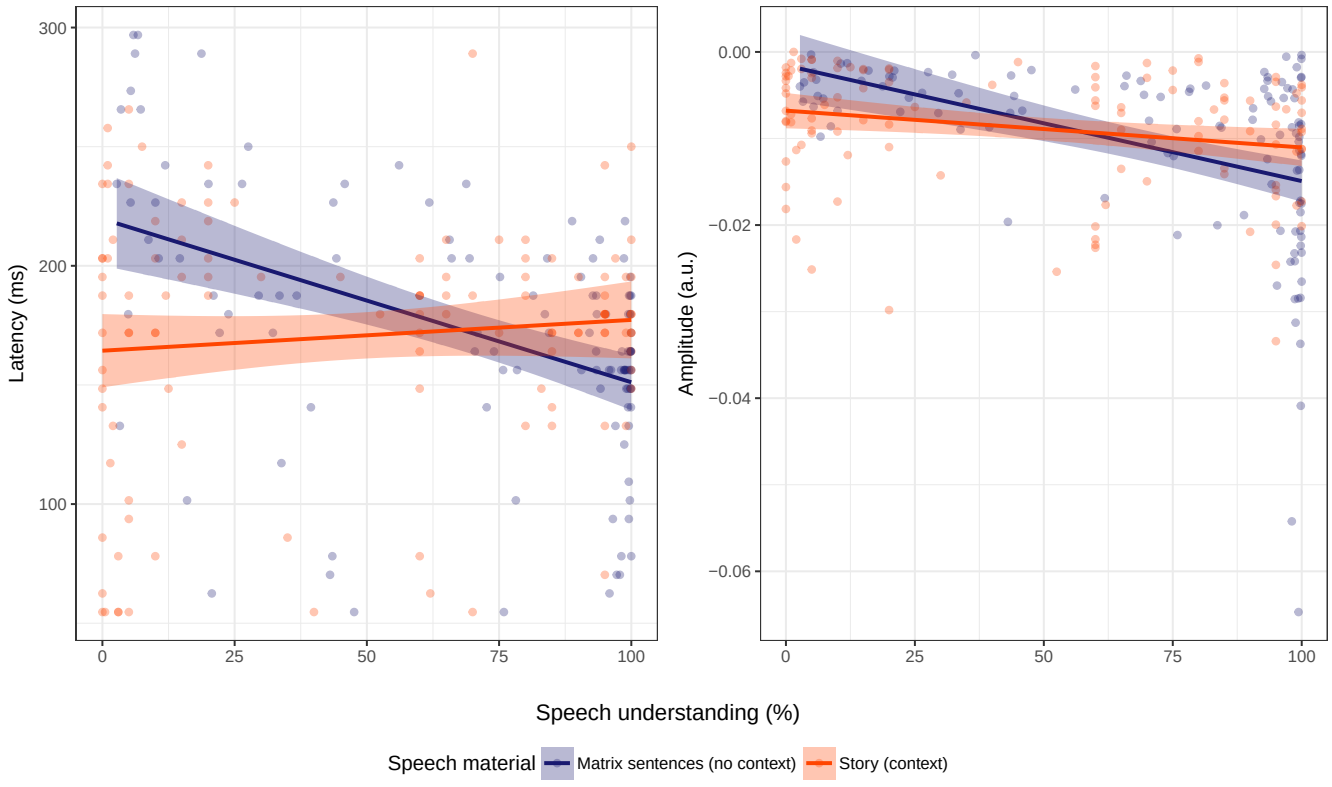

Figure 6. The negative peak in the temporal-occipital channels varies per subject and shows a general decrease in amplitude with a decrease in speech understanding. Latency increases with decreasing speech understanding for the Matrix sentences, but remains stable for the story.

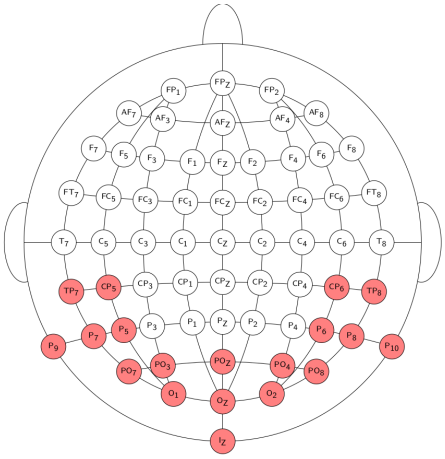

Figure 7. Electrode selection: 64 active electrodes placed according to the 10-20 electrode system. The locations of the electrodes that were selected for the calculation of the occipital-temporal TRF are indicated in red.

4 DISCUSSION

In order to gain more insight in the mechanisms underlying neural speech processing, we investigated whether speech understanding and the availability of semantic context in the stimulus influence neural envelope tracking. To that end, we tested 19 normal-hearing subjects. They listened to stimuli with two degrees of semantic context at varying levels of background noise while their EEG was recorded. Afterwards, we conducted an envelope reconstruction and TRF analysis and compared these results per speech material to their speech understanding. We found an effect of speech understanding and speech material which supports the hypothesis that neural envelope tracking is influenced by speech understanding and the availability of semantic context in the stimulus.

4.1 The same SNR does not result in similar speech understanding for different stimuli

As a first step we measured speech understanding for both speech materials at different noise levels. The results show that the same SNR does not result in similar speech understanding for different speech materials. The story was found to be more difficult to understand than Matrix sentences. Although we controlled for the sex of the speaker and chose speech materials with similar speech rates and spectrum, the difference could still be due to different acoustic features such as prosody. The Matrix sentences belong to a standardized speech material where every word is spoken at the same intensity. The story, on the other hand, is narrated for children and has more variations. An additional reason to explain this difference is lexical prediction. Even though the permutations of the words are different in each Matrix sentence, the words themselves are becoming more familiar to the participants during the course of the experiment, in contrast to the story where lexical prediction remains more fixed. Perhaps drawing from a larger pool of words for the Matrix sentences might have led to more similar intelligibility ratings between speech

materials. These diverging intelligibility results between speech materials also confirm the findings of Decruy et al. (2018), who emphasized the importance of measuring speech understanding for all stimuli used.

4.2 The interplay between neural envelope tracking and speech understanding

4.2.1 Envelope reconstruction

We found that the correlation between the reconstructed and the acoustic envelope increased with speech understanding. This finding supports the results of Molinaro and Lizarazu (2017); Luo and Poeppel (2007); Ding and Simon (2013); Ding et al. (2014); Vanthornhout et al. (2018) where an increase in speech understanding was also found to accompany an increase in envelope tracking.

Secondly, the tracking results in the delta band were significantly higher than in the theta band, while the significance levels remain the same. The slope of envelope tracking as a function of speech understanding was steeper in the delta band. These filter band differences might mean that the delta band is more sensitive to speech understanding, or it can just be a reflection of the specific modulation frequencies of the presented speech (Aiken and Picton, 2008; Luo and Poeppel, 2007). The story and Matrix sentences have modulation frequencies for sentence, word and syllable rate that can mainly be found within the delta band, possibly explaining the imbalance in correlation magnitude between the frequency bands.

4.2.2 Temporal response function

In addition to envelope reconstruction, we conducted a TRF analysis to gain more insight in the spatiotemporal profile of the neural responses. The topoplots in Figure 4 show a negative-positive interaction between the TRFs from the temporal-occipital channels and central channels. This is a typical topography of auditory evoked far-field potentials (Picton, 2011). The large negative peak within the 100 to 200 ms time lag (Figure 5) could be the so-called N100, usually occurring at a latency between 70-150 ms (Picton, 2011).

Generally we found, similar to envelope reconstruction, high TRF amplitudes when speech understanding was high (SI=100%) and reduced amplitudes when speech understanding decreased for both speech materials. When a small amount of noise was added and speech understanding remained almost unchanged (SNR = 2.5 dB SNR; Matrix sentences: SI = 99.9%; Story: SI = 99.0%), TRF amplitudes decreased between 0 to 150 ms, while amplitudes between 150 to 200 ms increased, revealing noise induced changes possibly related to enhanced attention and listening effort (Ding and Simon, 2012; Petersen et al., 2017; Kong et al., 2014; Obleser and Kotz, 2011). Most remarkable are TRF amplitudes between 50 and 100 ms which switch polarities in the presence of noise and show maximal activation at an SNR of $\pm$ -3.5 dB SNR. When more noise was added and speech understanding does decrease, amplitudes between 150 to 200 ms consistently decreased over subjects, perhaps indicating a time window sensitive to speech understanding.

4.3 Semantic context influences neural envelope tracking

4.3.1 Envelope reconstruction

This study demonstrated that a stimulus with semantic context available enhanced neural envelope tracking which is consistent with the dual stream model of speech processing (Hickok and Poeppel, 2007; Gross et al., 2013). Following this model, processing of Matrix sentences would mainly rely on the auditory acoustic information stream (bottom-up), while processing of the story also extensively uses the top-down stream. A potential confound is that we used different SNRs for the two stimulus types (to control for intelligibility). This means that the differences in envelope tracking could be related simply to SNR rather than other stimulus properties. To investigate this, we ran the same analysis, but now with SNR as predictor instead of intelligibility, and again found significantly increased envelope tracking for the story stimulus. This shows that SNR by itself does not account for the full difference between the two stimulus types. However, these results still have to be interpreted with care as some confounding factors cannot be controlled for. First, although the acoustics of the stimuli were matched in terms of sex and speech rate of the speaker and spectrum of the stimulus, acoustic differences like prosody are still present, as discussed in paragraph 4.1. Second, despite the questions asked to motivate the participants, the reduced correlations for the Matrix sentences could be linked to attention. Because listening to concatenated sentences can be boring, attention loss could occur which reduces neural envelope tracking (Ding and Simon, 2012; Petersen et al., 2017; Kong et al., 2014). For the coherent story, on the other hand, attention could be less of an issue as attending this speech is entertaining resulting in higher correlations.

Furthermore, the enhancement in neural envelope tracking due to the availability of semantic context was similar in both frequency bands. These results do not confirm our hypothesis of enhanced tracking primarily in the delta band. A possible explanation for this result could be the fixed syntactical 5-word structure of the Matrix sentences. Consequently this stimulus has a more rigid word and sentence rate compared to the story, possibly resulting in stronger neural tracking at these word- and sentence frequencies, respectively 2.5 Hz and 0.5 Hz, occurring within the delta band (0.5-4 Hz). As a result, this purely acoustic phenomenon might mask the interaction effect between semantic context processing and the filter bands.

4.3.2 Temporal response Function

In the no-noise condition significantly higher amplitudes were found for the Matrix sentences compared to the story. These enhanced TRF amplitudes could be caused by the previously mentioned rigid word and sentence rate of the Matrix sentences resulting in stronger neural tracking. However, when a small amount of noise is added (SNR = 2.5 dB SNR; Matrix sentences: SI = 99.9%; Story: SI = 99.0%), this significant difference disappears and the TRF amplitudes, including N100, increase for the story but decrease for the Matrix sentences. These apparently opposite results could, similarly to envelope reconstruction results, be explained by either the dual model of speech processing (Hickok and Poeppel, 2007; Gross et al., 2013), attention research (Ding and Simon, 2012; Petersen et al., 2017; Kong et al., 2014) as discussed

in paragraph 4.3.1 or the effort it takes to understand the stimulus (Kong et al., 2014; Obleser and Kotz, 2011).

Next, the latency of the N100 peak shows a difference depending on the stimulus. The latency decreases with increasing SI for the Matrix sentences, while latency remains the same over SI for the story. A latency decrease with increasing SNR, similar to the Matrix sentences, has been reported in literature by Petersen et al. (2017) and is also supported by research of (Kong et al., 2014), but not by Ding and Simon (2012). However, it is difficult to directly compare our results with these studies as 2 out of 3 only reported SNRs and not speech understanding scores. The difference in latency pattern we find for both speech materials could again be modulated by the dual model of speech processing, attention or listening effort.

A last result to point out is the positive peak around 300 ms for the Matrix sentences at -9.5 dB SNR (SI=49%) (Figure 5). The increased TRF amplitudes for the Matrix sentences could be related to the P300 peak. P300, like N100, is a deflection in the human event related potential. It can occur when a participant tries to detect a target stimulus (Picton, 1992, 2011). As the Matrix sentences do not contain semantic context, which makes content questions not possible, counting questions were asked at every SNR trial, for example, 'Which colors of boats were mentioned?'. We hypothesize that the question type, content questions for the story versus counting questions for the Matrix sentences, account for this P300 difference. As a consequence, the type of questions is also an important factor to take into account for event related potential research.

4.4 Implications for applied research

We found an effect of speech understanding and the amount of semantic context in the stimulus, indicating that neural envelope tracking might be more than the encoding of acoustic information. When developing an objective measure of speech understanding, like for example Vanthornhout et al. (2018), it is important to select the speech material based on the intended purpose. For example, to conduct research and investigate neural speech processing in noise, a story could be an interesting choice as neural envelope tracking is more pronounced. However, when comparing speech understanding outcomes in a clinical setting with for example hearing aids, top-down processing effects are undesired and should be ruled out and the Matrix sentences could be used instead.

5 CONCLUSION

We investigated whether speech understanding and/or the availability of semantic context in the stimulus influence neural envelope tracking in order to gain more insight in the mechanisms underlying neural speech processing. We found increasing neural envelope tracking with increasing speech understanding and an additional enhancement with semantic context in the stimulus.

REFERENCES

Aiken, S. J. and Picton, T. W. (2008). Human Cortical Responses to the Speech Envelope. *Ear & Hearing* 29, 139–157

Biesmans, W., Das, N., Francart, T., and Bertrand, A. (2017). Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection in a cocktail party scenario. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25, 402–412. doi:10.1109/TNSRE.2016.2571900

Boothroyd, A., Nittrouer, S., and Boothroyd, A. (1988). Mathematical treatment of context effects in phoneme and word recognition Mathematical treatment of context effects in phoneme and word recognition. *The Journal of the Acoustical Society of America* 84, 101–114

Brodbeck, C. (2017). Eelbrain: 0.25. Zenodo.

Brodbeck, C., Presacco, A., and Simon, J. Z. (2018). Neural source dynamics of brain responses to continuous stimuli: Speech processing from acoustics to comprehension. *NeuroImage* 172, 162–174. doi:10.1016/j.neuroimage.2018.01.042

Broderick, M. P., Anderson, A. J., Liberto, G. M. D., Crosse, M. J., and Edmund, C. (2018). Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural , narrative speech . *Current Biology* 28, 803–809

David, S. V., Mesgarani, N., and Shamma, S. A. (2007). Estimating sparse spectro-temporal receptive fields with natural stimuli. *Network: Computation in Neural Systems* 18, 191–212. doi:10.1080/09548980701609235

Decruy, L., Das, N., Verschueren, E., and Francart, T. (2018). The self-assessed Békesy procedure: validation of a method to measure intelligibility of connected discourse. *Trends in Hearing* , (Accepted).

Di Liberto, G. M., Lalor, E. C., and Millman, R. E. (2018). Causal cortical dynamics of a predictive enhancement of speech intelligibility. *NeuroImage* 166, 247–258. doi:10.1016/j.neuroimage.2017.10.066

Ding, N., Chatterjee, M., and Simon, J. Z. (2014). Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. *NeuroImage* 88, 41–46. doi:10.1016/j.neuroimage.2013.10.054

Ding, N., Melloni, L., Zhang, H., Tian, X., and Poeppel, D. (2015). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience* 19, 158–64. doi:10.1038/nn.4186

Ding, N. and Simon, J. Z. (2011). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of Neurophysiology* 107, 78–89. doi:10.1152/jn.00297.2011

Ding, N. and Simon, J. Z. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences of the United States of America* 109, 11854–9. doi:10.1073/pnas.1205381109

Ding, N. and Simon, J. Z. (2013). Adaptive Temporal Encoding Leads to a Background-Insensitive Cortical Representation of Speech. *The Journal of Neuroscience* 33, 5728–5735. doi:10.1523/JNEUROSCI.5297-12.2013

Francart, T., van Wieringen, A., and Wouters, J. (2008). APEX 3: a multi-purpose test platform for auditory psychophysical experiments. *Journal of Neuroscience Methods* 172, 283–293. doi:http://dx.doi.org/10.1016/j.jneumeth.2008.04.020

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

Hickok, G. and Poeppel, D. (2007). The cortical organization of speech processing. *Nature reviews. Neuroscience* 8, 393–402. doi:10.1038/nrn2113

Howard, M. F. and Poeppel, D. (2010). Discrimination of Speech Stimuli Based on Neuronal Response Phase Patterns Depends on Acoustics But Not Comprehension. *Journal of Neurophysiology* 104, 2500–2511. doi:10.1152/jn.00251.2010

Kerlin, J. R., Shahin, A. J., and Miller, L. M. (2010). Attentional gain control of ongoing cortical speech representations in a "cocktail party". *The Journal of neuroscience : the official journal of the Society for Neuroscience* 30, 620–8. doi:10.1523/JNEUROSCI.3631-09.2010

Kong, Y.-Y., Mullangi, A., and Ding, N. (2014). Differential modulation of auditory responses to attended and unattended speech in different listening conditions. *Hearing Research* 0, 73–81. doi:10.1002/ana.22528.Toll-like

Lalor, E. C., Pearlmutter, B. A., Reilly, R. B., McDarby, G., and Foxe, J. J. (2006). The VESPA: A method for the rapid estimation of a visual evoked potential. *NeuroImage* 32, 1549–1561. doi:10.1016/j.neuroimage.2006.05.054

Lalor, E. C., Power, A. J., Reilly, R. B., and Foxe, J. J. (2009). Resolving Precise Temporal Processing Properties of the Auditory System Using Continuous Stimuli. *Journal of Neurophysiology* 102, 349–359. doi:10.1152/jn.90896.2008

Lewis, A. G. and Bastiaansen, M. (2015). A predictive coding framework for rapid neural dynamics during sentence-level language comprehension. *Cortex* 68, 155–168. doi:10.1016/j.cortex.2015.02.014

Luo, H. and Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron* 54, 1001–10. doi:10.1016/j.neuron.2007.06.004

Maris, E. and Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods* 164, 177–190. doi:10.1016/j.jneumeth.2007.03.024

Mesgarani, N. and Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature* 485, 233–6. doi:10.1038/nature11020

Meyer, L., Henry, M. J., Gaston, P., Schmuck, N., and Friederici, A. D. (2017). Linguistic Bias Modulates Interpretation of Speech via Neural Delta-Band Oscillations. *Cerebral Cortex* 27, 4293–4302. doi:10.1093/cercor/bhw228

Molinaro, N. and Lizarazu, M. (2017). Delta(but not theta)-band cortical entrainment involves speech-specific processing. *European Journal of Neuroscience* 9, 1–9. doi:10.1111/ejn.13811

Obleser, J. and Kotz, S. A. (2011). Multiple brain signatures of integration in the comprehension of degraded speech. *NeuroImage* 55, 713–723. doi:10.1016/j.neuroimage.2010.12.020

Peelle, J. E., Gross, J., and Davis, M. H. (2013). Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cerebral cortex (New York, N.Y. : 1991)* 23, 1378–87. doi:10.1093/cercor/bhs118

Petersen, E. B., Wöstmann, M., Obleser, J., and Lunner, T. (2017). Neural tracking of attended versus ignored speech is differentially affected by hearing loss. *Journal of Neurophysiology* 117, 18–27. doi:10.1152/jn.00527.2016

Picton, T. W. (1992). The P300 Wave of the Human Event-Related Potential. *Journal of clinical Neurpohisiology* 9, 456–479

Picton, T. W. (2011). *Human Auditory Evoked Potentials* (San Diego: Plural Publishing inc.)

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (1995). Speech Recognition with Primarily Temporal Cues. *Science* 270, 303–304. doi:10.1126/science.270.5234.303

Somers, B., Francart, T., and Bertrand, A. (2018). A generic EEG artifact removal algorithm based on the multi-channel Wiener filter. *Journal of neural engineering* 15. doi:10.1088/1741-2552/aaac92

Stickney, G. S. and Assmann, P. F. (2001). Acoustic and Linguistic Factors in the Perception of Bandpass-Filtered Speech. *Journal of the Acoustical Society of America* 109, 1157–65

Vanthornhout, J., Decruy, L., Wouters, J., Simon, J. Z., and Francart, T. (2018). Speech intelligibility predicted from neural entrainment of the speech envelope. *JARO* 19, 181–191

Wild, C. J., Yusuf, A., Wilson, D. E., Peelle, J. E., Davis, M. H., and Johnsrude, I. S. (2012). Effortful Listening: The Processing of Degraded Speech Depends Critically on Attention. *Journal of Neuroscience* 32, 14010–14021. doi:10.1523/JNEUROSCI.1528-12.2012