Neural tracking as an objective measure of auditory perception and speech intelligibility

Jana Van Canneyt, Marlies Gillis*, Jonas Vanthornhout, Tom Francart

[a]ExpORL, Dept. of Neurosciences, KU Leuven, Herestraat 49 bus 721, Leuven, 3000, , Belgium

Abstract

The neural tracking framework enables the analysis of neural responses (EEG) to continuous natural speech, e.g., a story or a podcast. This allows for objective investigation of a range of auditory and linguistic processes in the brain during natural speech perception. This approach is more ecologically valid than traditional auditory evoked responses and has great potential for both research and clinical applications. In this article, we review the neural tracking framework and highlight three prominent examples of neural tracking analyses. This includes the neural tracking of the fundamental frequency of the voice (f0), the speech envelope and linguistic features. Each of these analyses provides a unique point of view into the hierarchical stages of speech processing in the human brain. f0-tracking assesses the encoding of fine temporal information in the early stages of the auditory pathway, i.e. from the auditory periphery up to early processing in the primary auditory cortex. This fundamental processing in (mostly) subcortical stages forms the foundation of speech perception in the cortex. Envelope tracking reflects bottom-up and top-down speech-related processes in the auditory cortex, and is likely necessary but not sufficient for speech intelligibility. To study neural processes more directly related to speech intelligibility, neural tracking of linguistic features can be used. This analysis focuses on the encoding of linguistic features (e.g. word or phoneme surprisal) in the brain. Together these analyses form a multi-faceted and time-effective objective assessment of the auditory and linguistic processing of an individual.

Keywords: Neural tracking, Speech intelligibility, EEG, f0 tracking, envelope tracking, linguistic features

Hearing loss is typically defined as a loss of perception of soft sounds, but hearing-impaired people tend to complain more about struggles to *understand* speech. To provide hearing-impaired people with appropriate rehabilitation, their hearing abilities need to be carefully evaluated in terms of both sound perception and speech intelligibility. The current golden standard methods for hearing evaluation, i.e. tone and speech audiometry, require active feedback from the tested person, which is not always obtainable (e.g. young children) or accurate (e.g. malingering). For this reason researchers are working towards new 'objective' methods, which rely on bodily signals, to asses hearing in clinical practice. One particularly promising objective measure is derived using the neural tracking framework, where

*Email address of corresponding author: marlies.gillis@kuleuven.be

electrical activity in the auditory pathway is measured with electroencephalography (EEG) while a participant listens to continuous speech, e.g. a story or a podcast. The use of continuous speech is a promising innovation as this type of stimulus is more relevant for communication in daily life than the tones and short speech samples used for behavioural audiometry. In this article, we discuss the neural tracking framework and its (dis)advantages, and review how it may be used to objectively asses auditory perception and predict speech intelligibility. We also discuss the opportunities and challenges for clinical implementation.

1. The neural tracking framework

1.1. Introduction

Traditional objective measures, like the auditory brainstem response (ABR), the auditory steady-state response (ASSR) or the frequency following response (FFR), require EEG measurement while a participant listens to repetitive presentations of a short sound stimulus (for a review, see Picton (2010)). Typical stimuli include clicks, tones, chirps and vowels. The repetitive stimulation is necessary as response instances need to be averaged to reduce measurement noise, but it is highly unnatural and demotivating for the listener (Theunissen et al., 2000; Hamilton and Huth, 2018). In recent years, technical advances have made it possible to analyse neural responses measured while a participant listens to continuous natural speech, without repetition (for a review, see Brodbeck and Simon, 2020). These neural responses to continuous speech are called neural tracking responses as they reflect how the auditory system of the listener 'tracks' the presented speech. They were originally proposed by Lalor et al. (Lalor et al., 2009; Lalor and Foxe, 2010) and the methods were further developed by, amongst others, Ding and Simon (2012a,b), O'Sullivan et al. (2015) and Crosse et al. (2016).

The possibility to investigate continuous speech processing with the neural tracking framework is an important innovation. Humans do not communicate with repetitive tones or clicks, as used for traditional objective measures. Context-rich continuous speech better approximates natural language use and as a result, research findings with these stimuli are more relevant for auditory processing in day-to-day communication (Kei et al., 1999; Pichora-Fuller et al., 2016; Hamilton and Huth, 2018; Keidser et al., 2020). Moreover, continuous speech is more comfortable and interesting for the listener. The stimulus can even be targeted towards the population of interest: e.g. a fairy-tale for young children or a podcast for adults. When a participant is interested in the content of the stimulus, they maintain attention for longer and as a result, the neural response measurement may be of higher quality. Finally, natural speech stimuli are better suited for research with hearing aids. Hearing aid signal processing is designed specifically for natural speech and may behave unpredictably with artificial sounds, corrupting the experiment.

In the neural tracking framework, neural responses to continuous speech are analysed without averaging over response instances. The most common approach to do so is based on linear encoding/decoding models. Other response analysis methods exist, including inter-trial coherence (ITC) (Zion Golumbic et al., 2013; Bourguignon et al., 2020), cross-correlation (Kong et al., 2014; Aiken and Picton, 2008; Petersen et al., 2016), mutual information (Gross et al., 2014;

Zan et al., 2020; Kaufeld et al., 2020) and neural networks (Katthi et al., 2020; Accou et al., 2021), but these will not be discussed further.

Linear modelling within the neural tracking framework requires two inputs: neural responses in the form of single-channel or multi-channel EEG (or MEG) and one or more features that represent the stimulus (see section 1.4). In the neural tracking framework, relations between the EEG and the stimulus feature are modelled, to investigate how well the stimulus information is encoded in the neural activity. The framework allows linear modelling in two directions: reconstructing the feature from the EEG (backward decoding, section 1.2) and conversely, reconstructing the EEG from the feature (forward encoding, section 1.3). As will be discussed below, the two analyses provide different but complementary information about the neural tracking responses. It is also possible to model in both directions at the same time with canonical correlation analysis (CCA), as described by de Cheveigné et al. (2019).

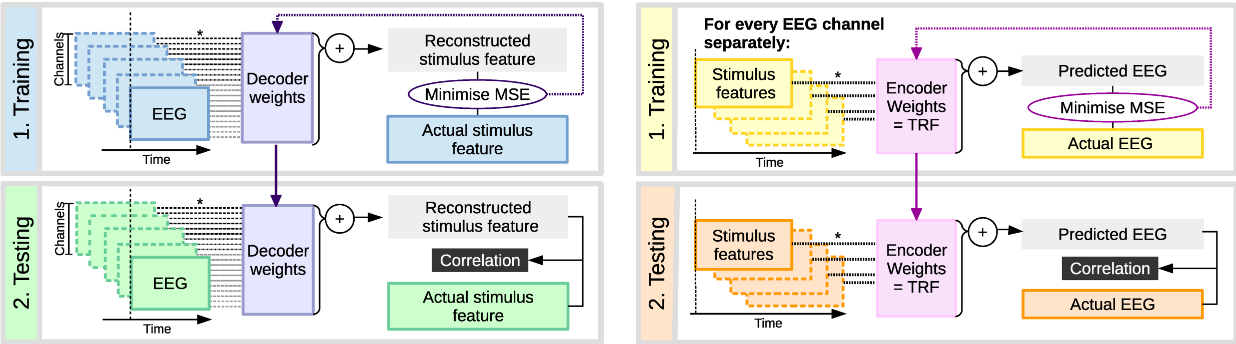
1.2. Backward modelling

In backward modelling, one reconstructs the stimulus feature from a weighted sum of the EEG signals from the different recording channels and their time-shifted versions. The time-shifted versions are included to account for neural processing delays. This delay or latency is estimated at about 5-10 ms for auditory processing in the upper brain stem and at least 12-30 ms for processes in the primary auditory cortex (Tichko and Skoe, 2017; Brugge et al., 2009). Higher-order cortical processes that modulate the neural response, like attention and interpretation of the speech, occur with delays of 200 ms or more (for a review, see Martin et al., 2008).

The backward modelling procedure, visualised in panel A of Figure 1, typically includes a training and a testing phase. First, the weights that provide the optimal reconstruction are determined based on a training data set (time-shifted EEG + corresponding stimulus feature). Then those weights are applied to the EEG from a separate testing data set, resulting in a reconstructed stimulus feature for the test data. The reconstructed feature is correlated with the actual stimulus feature of the test data to determine the reconstruction accuracy. This indicates how well the stimulus information can be reconstructed from the EEG, i.e., how well the speech is tracked by the brain. Note that this analysis is only reliable if the testing data is completely separated from the training data set. By training and testing on the same data, large reconstruction accuracies can be obtained, but the model has likely over-fitted on particularities of the data and will not generalise well to new data.

The backward modelling approach is a powerful analysis tool since the information of multiple EEG channels (often 32 or more) can be combined to predict a stimulus feature with often only one dimension (although multi-dimensional features are possible). However, this also means backward modelling is an ill-posed problem, complicated by linear dependency between EEG channels and their time-shifted versions, and therefore regularization is necessary to obtain a single solution (e.g. Hastie et al., 2001; Machens et al., 2004).

A disadvantage of backward modelling is that the weights are extraction patterns and these cannot and should not be interpreted to investigate the spatial pattern of the response (Haufe et al., 2014). One could assume that large



a) Backward modelling b) Forward modelling

Figure 1: *A. Schematic representation of backward modelling.* In backward modelling, the stimulus feature is reconstructed based on a linear combination of time-shifted EEG data. In the training phase, the model is estimated by optimizing the decoder weights to minimise the MSE (mean squared error) between the reconstructed stimulus feature and the actual stimulus feature for a training data set. Then, in the testing phase, the weights are applied to reconstruct the stimulus feature for the testing dataset. The final output is the correlation between the reconstructed stimulus feature and the actual stimulus feature for the testing dataset. *B. Schematic representation of forward modelling.* In forward modelling, the EEG data in each EEG channel is predicted based on a linear combination of time-shifted stimulus features. Again, the encoder weights or TRFs (temporal response functions) are estimated by minimizing the reconstruction MSE for a training data set. Then the TRFs can be studied as is, or they can be used to predict the EEG for a testing data set. The output of the testing phase is the correlation between predicted EEG and the actual EEG.

weights mean that the corresponding EEG channels contain a lot of response information. However, when an EEG channel captures information about a noise component, it can be used in the modelling process to 'subtract' the noise component from other EEG channels. As a result, some channels may receive large weights because they are helpful for noise reduction purposes and not because they contain response information (Montoya-Martínez et al., 2021).

1.3. Forward modelling

Forward modelling can be used to study the spatio-temporal properties of the response: the EEG signal in each channel is predicted from a weighted sum of the stimulus feature and its time shifted versions. Panel B of figure 1 schematically presents the forward modelling process. Note that for the forward modelling, the time-shifting occurs in the opposite direction than for backward modelling. Each EEG channel is considered separately, causing forward models to be less powerful, as they cannot combine information across channels. The advantage of this approach is that the weights are activation patterns and not extraction patterns and can thus be interpreted. Forward modelling may solely include a training phase, which results in interpretable weights (see below), or there may be a testing phase where the weights are applied to predict the EEG for a separate data set. In that case, similar to the backward modeling approach, the actual and predicted EEG responses can be correlated to obtain each EEG channel's prediction accuracy. Higher prediction accuracies can be related to better encoding of the speech features in the EEG, and therefore in the brain, but other factors that impact the SNR of the EEG could be at play as well.

For each channel, the weights estimated at the different time shifts form a temporal response function (TRF) that reflects response amplitude (~ weight) as a function of response latency (~ time shift). A TRF can be interpreted as the impulse response of the auditory system: the information in the input stimulus (~ the feature) is transformed with this impulse response to produce the output response (~ the EEG). The channel-specific TRFs tend to be noisy and are therefore often averaged over a selection of EEG channels and subjects. Based on the time shifts that receive large weights for many of the EEG channels/subjects, we can derive the dominant latencies of the response. These latencies (or delays) can then be used to estimate which stages of neural processing along the auditory pathway contribute to the response. The spatial properties of the response can be further investigated by looking at the spatial distribution of the magnitude of TRF weights over the scalp. This information is usually visualised on a topoplot. Examples of TRFs and topoplots are available in figure 3, which will be discussed further on. Note that such topoplots only allow for spatial information on scalp level, where the electrodes were located. To study the actual sources of the neural responses within the head, the inverse problem needs to be solved, i.e. transforming the information from electrode space to neural source space (e.g. Brodbeck et al., 2018c).

1.4. The stimulus feature

The stimulus feature is derived from the presented speech and reflects how a particular speech characteristic varies over time. Many stimulus features can be used, ranging from low-level acoustic characteristics (e.g. the acoustic envelope) to high-level linguistic information (e.g. word surprisal). This flexibility makes the neural tracking framework highly versatile. It also underlies one of the most prominent advantages of the framework: a single EEG measurement can be analysed with respect to various features of the stimulus and provides information on a range of auditory/language processes. This includes f0 tracking, envelope tracking, phoneme tracking, semantic tracking, etc. Since data collection is often time-intensive, this type of 'multi-functional' data and analysis can considerably speed up scientific progress and is also promising for clinical implementation.

We will focus on three prominent (groups of) stimulus features corresponding to three types of neural tracking analyses in the following sections. We discuss them following the hierarchical organisation of the auditory pathway: starting with auditory processing of the fundamental frequency (f0, section 2), which happens mostly in subcortical stages of the auditory pathway, then moving on to envelope processing (section 3) which happens in the auditory cortex and ending with linguistic processing (section 4) which happens in the language network of the brain. We focus on how these stimulus features can be used to investigate different aspects of speech processing and different parts of the auditory pathway. Moreover, we provide example results and review findings from relevant studies, including how the responses relate to important clinical measures like hearing thresholds and speech perception.

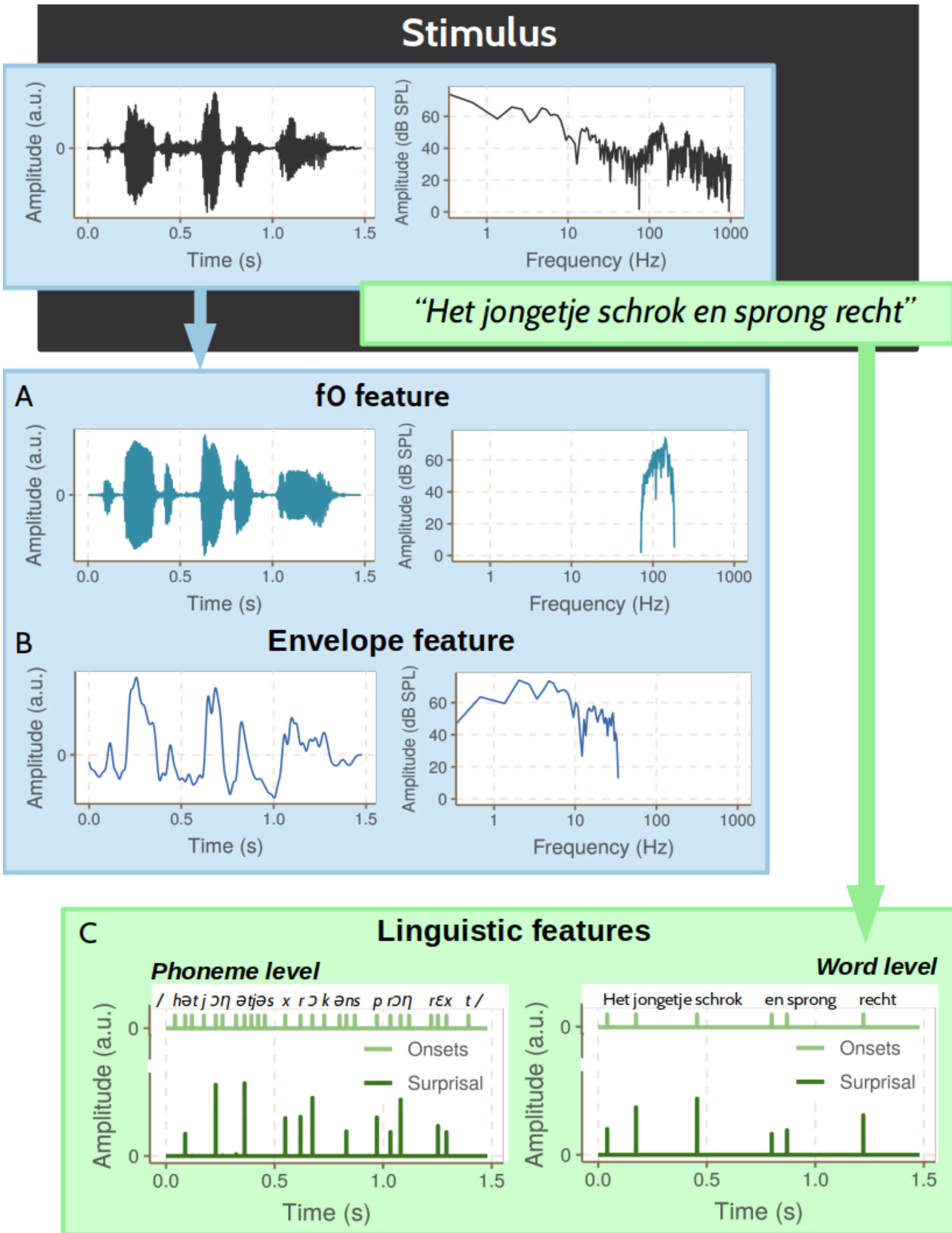


Figure 2: *Example of stimulus and derived features for an example sentence by a male speaker*. The f0 (panel A) and envelope feature (panel B) are derived from the stimulus waveform, whereas linguistic features (panel C) are derived from the stimulus transcription. The f0 and envelope features concern different spectral ranges with the envelope focusing on low frequencies (< 50 Hz) and the f0 focusing on higher frequencies (~ 85 – 300 Hz). Linguistic features can focus on different segmentation levels, including phoneme level and word level. Panel C visualises an example onset and surprisal feature for each level.

2. Neural tracking of the f0

Neural tracking of the fundamental frequency of the voice, or f0-tracking, is used to investigate how the f0 is represented in the brain activity (Forte et al., 2017; Etard et al., 2019; Van Canneyt et al., 2021c). The f0 is a periodic modulation in the speech signal generated by vocal fold vibration during speech production. It is related to the perception of pitch. The f0 of adult speakers typically ranges from 85 to 300 Hz, with male and female voices situated respectively at the lower and higher ends of the range. The f0 is an essential characteristic of the human voice and it is vital to convey intonation and emotion, but proper perception of the f0 is not required for speech intelligibility (e.g. cochlear implant listeners). Nevertheless, f0-tracking can provide information on the quality of fine temporal processing in the early stages of the auditory pathway, which is the foundation for proper speech processing in the brain.

Temporal processing of the f0 in the human auditory system happens through the synchronization of the activity of the neurons to the f0 modulations, i.e. phase-locking. Due to the relatively high frequency of the f0 modulations, this phase-locking occurs mainly in peripheral and subcortical stages of the auditory pathway, up to the upper brain stem. Neurons at cortical stages have poor phase-locking above about 100 Hz and are therefore less likely to contribute to f0-tracking (Joris et al., 2004). However, it has been shown that early cortical contributions to f0-tracking responses (and FFRs) can occur for low-frequency stimuli (85-100 Hz, e.g. low male voices) (Coffey et al., 2016, 2017; Van Canneyt et al., 2021c).

F0 tracking analysis requires an f0 feature that represents the f0 modulations in the presented speech. The f0 feature can be extracted from the speech stimulus in various ways. A simple yet effective way is to band-pass filter the stimulus in the range of the f0 (Etard et al., 2019; Van Canneyt et al., 2021c). An example of this type of feature is provided in panel A of figure 2. More complicated and computationally expensive techniques have been explored as well, including empirical mode decomposition (Etard et al., 2019; Forte et al., 2017) and auditory modelling (Van Canneyt et al., 2021b). Constructing an f0 feature that approximates the expected neural response using auditory modelling has proven particularly effective, nearly doubling the reconstruction accuracies obtained with the neural tracking analysis (Van Canneyt et al., 2021b).

Section 1 of figure 3 shows the results of a typical forward modelling analysis for f0-tracking, obtained using the methods described in Van Canneyt et al. (2021c). The data set used for this visualisation (and all others in figure 3) contained 64-channel EEG data from 32 young normal-hearing subjects measured in response to male-narrated speech (dataset from Accou et al., 2021). Panel A shows the mean TRF across subjects for the channel selection indicated in pink on panel B. The TRF for each subject is plotted as well to indicate the variance. The TRFs in this example are modified with a Hilbert transform to present the amplitude of the TRF without phase information resulting in only positive values. This technique suppresses the auto-correlative periodicity in the f0-tracking TRFs (see further) and aids with interpretation (for more information, see Van Canneyt et al. (2021c)). The TRF pattern indicates that the activity in the auditory system (~ EEG) best reflects the f0 information (~ the feature) at a latency of about 10-25 ms. Panel B of figure 3 presents an example f0 tracking topoplot with common-average rereferencing at 15 ms latency. The topoplot indicates strong response activity in the center of the head and across the back of the head. The temporal and spatial response patterns are consistent with dominant f0-related activity in the upper brain stem and early cortical regions. Saiz-Alia et al. (2020) has performed detailed computational modelling of the subcortical sources of the f0 tracking response, demonstrating important contributions from the cochlear nuclei and the inferior colliculus. Van Canneyt et al. (2021c) argues for additional contributions from the right primary auditory cortex for f0 tracking of low-frequency voices.

Although f0-tracking was only recently developed, the technique has led to several interesting findings. Forte et al. (2017) and Etard et al. (2019) have demonstrated that the f0 tracking response holds information on selective attention, possibly indicating that neural mechanisms for attention influence the brain stem. Kulasingham et al. (2020) and Van Canneyt et al. (2021a) have investigated how the age of the listener impacts f0 tracking. Kulasingham et al. (2020) found no age effects using MEG, which is most sensitive to cortical sources. In contrast, Van Canneyt et al. (2021a) found a significant reduction in response strength with advancing age using EEG (which is more sensitive to subcortical sources). This observation is in line with an age-related decrease in the phase-locking ability of the subcortical (and early cortical) auditory system. Van Canneyt et al. (2021a) also studied the effect of hearing loss and found increased f0-tracking responses in participants with hearing impairment compared to age-matched controls. The response enhancement was due to additional cortical activity phase-locked to the f0 (with latency of ~40 ms), likely compensating for the reduced quality of bottom-up auditory input due to diminished peripheral auditory sensitivity. Moreover, the amount of additional compensatory cortical activity was significantly related to the pure tone average (PTA) hearing loss of the participant. As such, a significant relation exists between the degree of hearing loss of an individual and the strength of their f0 tracking response.

At the moment, f0-tracking also has some limitations, which future advances may mitigate. One of the main issues is auto-correlative smearing in TRFs and topoplots because the f0 stays relatively steady over multiple f0 periods. This periodic smearing over latencies can be somewhat mitigated with Hilbert-transformed TRFs, which disregard phase information. However, TRF and topoplot interpretation are still limited to the most dominant peaks (see Van Canneyt et al. (2021c) for more details). A second limitation is that the f0 is only present in speech during voiced sounds (~ 50-60 % of the time) and not during unvoiced sounds (~ 40 % of the time), including silences. During analysis, these unvoiced sections in the speech stimulus (and corresponding sections in the EEG) are disregarded. As a result only about half of the measured data can be used to analyse f0-tracking, increasing the required measurement time. Another limitation is that the f0 tracking response is reduced for voices with higher and more variable f0, leading to weak and often non-significant responses for typical female voices. This occurs because neural phase-locking ability is decreased for higher and more variable f0s, especially for cortical sources. As such, the stimulus choice has a large impact on the f0 tracking response. A final limitation is that f0-tracking requires careful interpretation: f0-tracking reflects the capability of the auditory system to phase-lock to the f0, but it does not reflect the ability of a person to perceive pitch or speech in general. Fortunately, neural tracking analyses with other features help complete the picture.

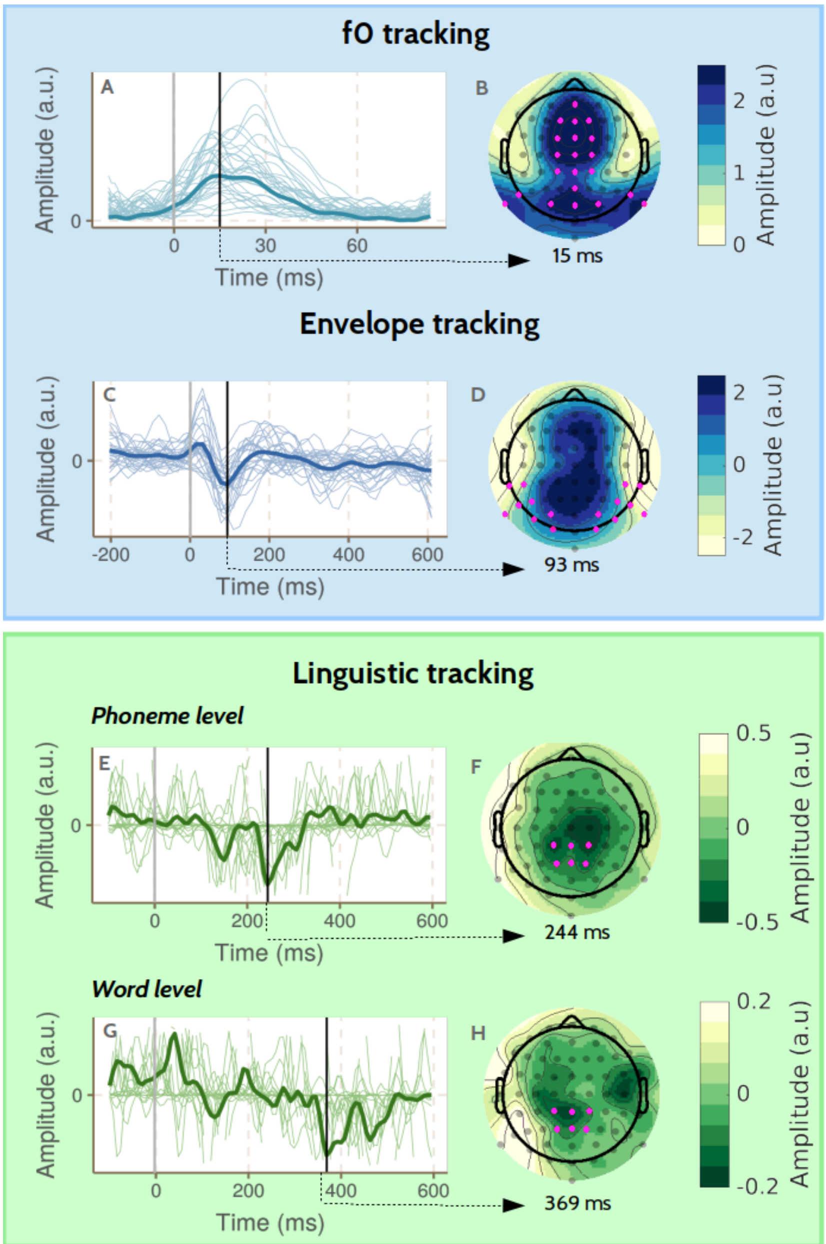


Figure 3: *Example of forward modelling results: TRFs and topoplots.* The figure is divided into three sections on f0-tracking, envelope tracking and linguistic tracking, respectively. For each type of tracking, an example mean TRF (+ individual TRFs) is presented (panel A, C, E and G), together with a corresponding topoplot at an important latency (panel B, D, F and H). The channels indicated with pink on the topoplot represent the channel selection used to obtain the corresponding TRF. Note the drastically different time scales in the TRFs, reflecting the presence of neural activity at different latencies for each feature.

3. Neural tracking of the speech envelope

The speech envelope consists of slow-varying modulations (< 50 Hz) in the speech signal. It contains acoustic temporal information (Rosen, 1992) but also reflects phonemes, syllables and word transitions (Peelle and Davis, 2012). Moreover, it also correlates with the area of the mouth opening during articulation (Chandrasekaran et al., 2009). Therefore it is not surprising that research indicates that the envelope is an essential acoustic cue for speech intelligibility (Shannon et al., 1995; Drullman et al., 1994a,b).

Envelope tracking is used to analyse the neural encoding of the speech envelope during speech perception (Ding and Simon, 2012a; O'Sullivan et al., 2015; Vanthornhout et al., 2018). From animal studies (Wang et al., 2008) and human studies with electrocochleography (ECoG), it is known that the speech envelope is processed in the primary auditory cortex, specifically in Heschl's Gyrus (Nourski et al., 2009). A growing body of evidence demonstrates envelope tracking is a requirement for speech understanding. Correspondingly, multiple studies show that neural tracking of the speech envelope is strongly correlated with behaviourally measured speech intelligibility (e.g. Ding et al. (2014); Vanthornhout et al. (2018); Lesenfants et al. (2019); Iotzov and Parra (2019); Verschueren et al. (2021)). As a specific example, Vanthornhout et al. (2018) found a significant correlation of 0.69 between the speech reception threshold (SRT) estimated based on envelope tracking and the SRT measured with behavioural speech audiometry.

Although the full-band envelope can be used, it is also possible to study the neural response to specific frequency bands of the envelope. Envelope tracking responses are most commonly investigated in the delta band (0.5-4 Hz), theta band (4-8 Hz) and gamma band (> 30 Hz) (Ding and Simon, 2013; Verschueren et al., 2021; Molinaro and Lizarazu, 2017). The lower envelope frequencies are often the main interest as they correspond with word onsets and the syllabic rate of the speech, which is hypothesised to be crucial for speech intelligibility. Some studies suggest that speech intelligibility is specifically related to the theta band (4-8 Hz) and not the delta band (1-4 Hz) (Ding and Simon, 2013). Other studies indicate the opposite (Verschueren et al., 2021; Molinaro and Lizarazu, 2017). In our opinion, the outcome may depend on the speech material. The syllabic rate is often very close to 4 Hz, and as such, envelope tracking to a slow speaker could be more dominant in the delta band while envelope tracking to a fast speaker could be more dominant in the theta band.

Envelope tracking responses can be analysed using a forward, backward or bidirectional model. In any case, the model requires an envelope feature that is extracted from the stimulus waveform. In essence, the envelope is just a curve outlining the peak values of the stimulus, which can be easily obtained by taking the absolute value of the Hilbert transform. Although this is a prevalent method, it is not the best choice as it disregards human perception. To better approximate human envelope perception, two important aspects of auditory processing need to be taken into account. First, the stimulus should be split into frequency bands before the actual envelope extraction process to mimic how the basilar membrane in the cochlea divides a sound stimulus into different auditory filters. Second, the compression and non-linear behaviour of the auditory system should be accounted for. To incorporate these factors in the envelope extraction process, complex computational models of the auditory periphery can be used (Yang et al., 2015; Bruce et al., 2018). However, Biesmans et al. (2017) evaluated various extraction methods in an auditory attention detection paradigm and proposed a simplified approach. They found that a combination of a gammatone filterbank, which simulates the auditory filters on the basilar membrane, followed by a power law to account for compression and non-linearity in the auditory system, performed equally well as the more complex and computationally expensive auditory models. Although AAD is not the same as envelope tracking, the underlying model is identical and the proposed technique is valid here as well. An example envelope feature obtained using this technique is provided in panel B of figure 2.

A visualisation of the results of a typical forward modelling analysis for envelope tracking is visualised in section 2 of figure 3. These results were obtained by applying the methods described in Vanthornhout et al. (2019) and Lesenfants et al. (2019) to the data set described earlier. Panel C presents the mean TRF, averaged over subjects and a channel selection (indicated in pink on panel D). The TRFs of the individual subjects are visualised with a thin line to indicate the variance. The TRF displays three distinct peaks. The P1 peak (50 ms), the N1 peak (93 ms) and the P2 peak (170 ms). This typical P1-N1-P2 complex is also found in AEP studies with impulse-like stimuli and can thus be used to infer the neural source of the peaks. The P1 peak originates in Heschl's Gyrus, and the N1 peak originates in the Superior Temporal Gyrus (O'Sullivan et al., 2019b; Steinschneider et al., 2011). The origin of the P2 peak is less clear but is probably in the (higher) auditory cortex (Godey et al., 2001). The topoplot shows negative weights for the temporal channels and positive weights for the central channels. This distribution is an indication of a dipole located near the auditory cortex. Without analyses in source space, the exact location is difficult to pinpoint.

Over the past decade, envelope tracking has been used to study, among others, how cortical speech processing is affected by individual factors like age and hearing status. Decruy et al. (2019) and Brodbeck et al. (2018b) found stronger envelope tracking for older participants compared to younger participants, even though older adults typically have more difficulty understanding speech. Similarly, Decruy et al. (2020b) and Fuglsang et al. (2020) found increased envelope tracking for hearing-impaired listeners compared to age-matched normal-hearing listeners. The enhanced tracking in older listeners or listeners with a hearing impairment may be explained by a compensatory central gain mechanism (Parthasarathy et al., 2019; De Villers-Sidani et al., 2010; Chambers et al., 2016), recruitment of additional cortical resources (Brodbeck et al., 2018b; Gillis et al., 2021a) and increased listening effort and attention (Decruy et al., 2020a; Vanthornhout et al., 2019; Lesenfants and Francart, 2020). With an innovative artefact removal technique, Somers et al. (2019) succeeded to analyse envelope tracking for cochlear implant listeners as well. For both hearing-impaired listeners (with simulated amplification) (Decruy et al., 2020b) and cochlear implant listeners (Verschueren et al., 2019) the tracking strength was significantly correlated to behaviourally-measured speech intelligibility, indicating a similar relation with speech intelligibility as observed for normal hearing listeners (Vanthornhout et al., 2018).

One challenge with envelope tracking is that its functional interpretation is unclear. The main complicating factor is that the envelope itself is highly correlated with linguistic cues, like the onsets of words and syllables. As such, the envelope represents multiple unique features that all may contribute to the observed neural tracking response and are hard to disentangle. In addition, the interpretation of envelope tracking is complicated by the fact that it is modulated by top-down effects, such as attention and audio-visual integration (O'Sullivan et al., 2019a). A final challenge is that the exact relation between envelope tracking and speech intelligibility remains a point of discussion (Ding and Simon, 2014; Brodbeck and Simon, 2020). Multiple studies have shown that envelope tracking reflects experimental changes in speech intelligibility (Vanthornhout et al., 2018; Lesenfants et al., 2019; Verschueren et al., 2021), even in the case of degraded speech with an intact envelope (Ding et al., 2014). However, it is unlikely that envelope tracking is a direct reflection of successful speech intelligibility as neural tracking responses have been observed for non-speech signals (Zuk et al., 2021) and foreign languages (Etard and Reichenbach, 2019). As such, envelope tracking is likely necessary but not sufficient for speech intelligibility. To gain further insight into how the brain processes the meaning of speech, i.e. speech intelligibility, linguistic features can be used.

4. Neural tracking of linguistic features

In pursuit of an accurate neural marker of speech intelligibility, recent studies focus on linguistic speech features. While the f0 and speech envelope are derived from the acoustic waveform of the speech, linguistic features are derived from the content of the speech. Proper encoding of these features in the brain requires accurate linguistic processing and not mere acoustic processing.

Linguistic features can be divided in two categories. Features in the first category denote lexical segmentation. They represent (aspects of) a sequence of small building blocks that make up spoken language, e.g., sequences of phonemes, phonetic features, words, or specific word categories like content and function words (Di Liberto et al., 2015; Lesenfants et al., 2019). These features are arrays consisting of zeros with a fixed, non-zero entry (~ spike) at the onset of each lexical building block (see features in light green on Panel C of figure 2). Features in the second category reflect higher-level linguistic aspects of the speech, e.g., how familiar, predictable or surprising a word or phoneme is in its context (Weissbart et al., 2019; Brodbeck et al., 2018a; Koskinen et al., 2020). These features can be applied on three levels, which require different amounts of linguistic context: (1) at the level of a phoneme (e.g., phoneme surprisal or cohort entropy), (2) at the level of a word (e.g., word frequency or word surprisal), and (3) at a semantic contextual level (e.g., semantic dissimilarity). These features consist of arrays of zeroes and ones, similar to lexical segmentation features. However, in this case the spike amplitude at each onset is not fixed but modulated by the linguistic information of the specific phoneme or word (see features in dark green on Panel C of figure 2).

The fact that linguistic features are sparse arrays consisting of mostly zeroes with some non-zero entries (~ spikes), makes them different from the continuous f0 and envelope features and poses challenges for response analysis. In backward modelling the reconstructed feature needs to be compared to the actual feature but traditional measures to do so, like MSE or correlation, are not well-behaved with sparse inputs. These problems do not occur for forward modelling, where the non-sparse reconstructed and actual EEG are compared. Therefore the forward model is a more common choice for analysis with linguistic features.

Panels E-H of figure 3 present a visualisation of the results of a typical forward modelling analysis for linguistic tracking with phoneme suprisal and word suprisal features (see Brodbeck et al., 2018a; Gillis et al., 2021b, for detailed methods). The TRFs at both phoneme (panel E) and word level (panel G) show a negative response, situated centrally in the topography (panel F and H), around respectively 250 and 350 ms. The earlier response peak for phonemes compared to words is consistent with the hierarchy of the language processing of these linguistic building blocks, i.e., the phonemes making up a word are processed before the word's surprisal can be estimated. Moreover, the response to word surprisal resembles the N400 response, which is classically observed in ERP paradigms (Lau et al., 2008). These congruent topographic responses indicate that this small and specific language response can also be observed when listening to natural running speech rather than stand-alone sentences.

Measuring neural tracking of linguistic features is an exciting avenue to test psycho-linguistic theories of speech understanding. It is accepted that listeners use linguistic context to continuously adapt expectations of upcoming concepts, words and phonemes, but it is unclear how these expectations are integrated with what is actually being perceived. Brodbeck et al. (2021) showed that the neural prediction of an upcoming phoneme or word relies on contextual processing in a parallel manner, combining both bottom-up and top-down processing. Additional evidence of the presence of top-down processing comes from Heilbron et al. (2020) who observed that higher-level predictions influence the predictions at lower levels (i.e., word prediction affects the predictions at phoneme level).

Another exciting research path is the disentanglement of acoustic and linguistic neural processing. Verschueren et al. (2022) disentangled acoustic and linguistic neural processing by changing the speech rate, which kept the linguistic content the same while varying the acoustic properties and the intelligibility of the speech. As the speech rate became higher, the neural tracking of acoustic properties increased. This means that better neural encoding was observed, even though the speech became harder to understand. In contrast, neural tracking of linguistic properties decreased with increasing speech rate. This indicates that linguistic tracking provides a more accurate objective measure for speech intelligibility.

Linguistic speech representations can also provide insight into age-related speech intelligibility deficits. We are aware of two studies that study the speech intelligibility deficits in older adults. Although Mesik et al. (2021) did not report differences, Broderick et al. (2021) reported that older adults rely less on pre-activated semantic representations than younger adults. Furthermore, they showed that older adults who relied more on this semantic pre-activation mechanism showed higher verbal fluency. Please note that due to the novelty of linguistic tracking, many of the studies mentioned here have not yet passed peer review.

Linguistic tracking is an up-and-coming research technique but it also has a few difficulties. Firstly, the linguistic representations coincide with the boundaries of phonemes and words. These boundaries are often associated with high acoustic power, and therefore, it is necessary to carefully control for acoustic properties of the speech when evaluating linguistic representations. If not, the speech tracking analysis might be biased to find spurious significant linguistic representations due to its correlations with acoustic representations (Daube et al., 2019). One way to overcome this issue is by investigating the added value of linguistic representations (as used in e.g. Broderick et al., 2018; Gillis et al., 2021b). Such an approach requires two model fits: a baseline model, accounting for acoustic and lexical segmentation features, and a more complex model that includes linguistic information on top of the baseline model. The performance (i.e., prediction accuracy) of the baseline model is then subtracted from the performance of the more complex model to obtain the added value of linguistic representations. Note that this approach is conservative and restrictive: it only quantifies unique information contributed to the model by the linguistic features, neglecting acoustic and linguistic information that is shared with other features in the model.

Secondly, due to sparse features, the analyses are often based on forward modelling. Prediction accuracies, i.e. correlations, obtained with forward models are typically small in magnitude: only around 3 to 7% of the variance in the EEG signal can be explained by neural responses time-locked to the presented stimulus. Moreover, most of this variance is explained by acoustic characteristics of the speech, as these lower-level acoustic representations evoke responses over large parts of the auditory system. In contrast, linguistic tracking targets the neural response from a precisely localized neural process related to intelligibility. Therefore, the associated magnitudes of these neural processes measured at the scalp level are much smaller. As the prediction accuracies of the forward model are small in magnitude, finding a significant improvement of the linguistic representation over and beyond acoustic representations is statistically challenging (e.g. an improvement of $\sim$1% corresponds to an increase in prediction accuracy of 3.4×10^{-4} using the conservative and restrictive approach as described above Gillis et al. (2021b)).

5. Clinical applications of neural tracking responses

To provide people with hearing problems with evidence-based and innovative health care, it is useful to review the merits and limitations of all (objective) audiological measures and investigate how the measures may be combined to form a complete assessment.

The current gold standard methods, i.e. tone and speech audiometry, have proven their worth but they are challenging in key patient populations like young children. To remedy this, objective measures for sound perception like the ABR and the ASSR have been introduced in the clinical toolset. However, there is no clinically available objective measure of speech intelligibility. Since speech intelligibility is the basis for human communication, this is a significant gap to fill. Various populations may benefit from such a measure, including young children, stroke patients (especially those with aphasia) and people with dementia.

The versatile neural tracking paradigm is highly promising for this purpose: based on a single twenty-minute long EEG recording, a wide range of speech processing abilities may be assessed (incl. phase locking to the f0, envelope tracking, phonetic processing, phonemic processing and even linguistic processing). This versatility may lead to a highly time-effective objective assessment of both auditory and language abilities. Moreover, neural tracking is easily automated, paving the way to improved automated screening, diagnostics, and automatic fitting of auditory prostheses, or even auditory prostheses that continuously adapt themselves to the listener based on their brain activity (Geirnaert et al., 2021).

Future studies preparing for clinical implementation may need to shift focus from group-level analyses towards subject-specific analyses. Moreover, they may focus on which combination of neural tracking features provides the most information and how these can be optimally analysed. As the features are highly correlated with each other, special care needs to be taken to investigate the effect of each feature (Gillis et al., 2021b). Subsequent research efforts are also required to decide on the best speech stimuli (required to work well for all types of tracking) and the best EEG measurement set-up, including the number of EEG electrodes and their position (Montoya-Martínez et al., 2021). It is also essential to validate the measures in a comprehensive sample of the population, including participants of all ages and with various audiological and non-audiological pathologies. Furthermore, the neural tracking results need to be transformed into an easy-to-interpret set of scores and visualisations, to allow for intuitive use by clinicians.

6. Acknowledgements

The authors would like to thank Bernd Accou, Wendy Verheijen and their students for collecting the dataset used for the examples in this article. The research received funding from the European Research Council under the European Unions Horizon 2020 research and innovation programme (grant agreement No. 637424, ERC starting grant to Tom Francart). Jana Van Canneyt and Marlies Gillis are both supported by a PhD grant for Strategic Basic research by the Research Foundation Flanders (FWO): project number 1S83618N and project number 1SA0620N, respectively. Jonas Vanthornhout is funded by a postdoctoral grant from FWO, project number 1290821N.

7. Declaration of interest

The authors declare that author Tom Francart is involved in translational research which may lead to the commercialisation of a product related to the presented research. Besides this, there are no conflicts of interest, financial, or otherwise.

References

Accou, B., Jalilpour Monesi, M., Montoya, J., Van hamme, H., and Francart, T. (2021). Modeling the relationship between acoustic stimulus and EEG with a dilated convolutional neural network. In *2020 28th European Signal Processing Conference (EUSIPCO)*, pages 1175–1179.

Aiken, S. J. and Picton, T. W. (2008). Human cortical responses to the speech envelope. *Ear and Hearing*, 29(2):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(5):402–412.

Bourguignon, M., Molinaro, N., Lizarazu, M., Taulu, S., Jousmäki, V., Lallier, M., Carreiras, M., and De Tiège, X. (2020). Neocortical activity tracks the hierarchical linguistic structures of self-produced speech during reading aloud. *NeuroImage*, 216(September 2019).

Brodbeck, C., Bhattasali, S., Heredia, A. C., Resnik, P., Simon, J. Z., and Lau, E. (2021). Parallel processing in speech perception: Local and global representations of linguistic context. *bioRxiv*, page 2021.07.03.450698.

Brodbeck, C., Hong, L. E., and Simon, J. Z. (2018a). Rapid Transformation from Auditory to Linguistic Representations of Continuous Speech. *Current Biology*, 28(24):3976–3983.e5.

Brodbeck, C., Presacco, A., Anderson, S., and Simon, J. Z. (2018b). Over-representation of speech in older adults originates from early response in higher order auditory cortex. *Acta Acustica united with Acustica*, 104(5):774–777.

Brodbeck, C., Presacco, A., and Simon, J. Z. (2018c). Neural Source Dynamics of Brain Responses to Continuous Stimuli: {{Speech}} Processing from Acoustics to Comprehension. *NeuroImage*, 172:162–174.

Brodbeck, C. and Simon, J. Z. (2020). Continuous speech processing. *Current Opinion in Psychology*, 18:25–31.

Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., and Lalor, E. C. (2018). Electrophysiological Correlates of Semantic Dissimilarity Reflect the Comprehension of Natural, Narrative Speech. *Current Biology*, 28(5).

Broderick, M. P., Di Liberto, G. M., Anderson, A. J., Rofes, A., and Lalor, E. C. (2021). Dissociable electrophysiological measures of natural language processing reveal differences in speech comprehension strategy in healthy ageing. *Scientific Reports*, 11(1):1–12.

Bruce, I. C., Erfani, Y., and Zilany, M. S. (2018). A phenomenological model of the synapse between the inner hair cell and auditory nerve: Implications of limited neurotransmitter release sites. *Hearing Research*, 360:40–54.

Brugge, J. F., Nourski, K. V., Oya, H., Reale, R. A., Kawasaki, H., Steinschneider, M., and Howard, M. A. (2009). Coding of Repetitive Transients by Auditory Cortex on Heschl's Gyrus. *Journal of Neurophysiology*, 102(4):2358–2374.

Chambers, A. R., Resnik, J., Yuan, Y., Whitton, J. P., Edge, A. S., Liberman, M. C., and Polley, D. B. (2016). Central Gain Restores Auditory Processing following Near-Complete Cochlear Denervation. *Neuron*, 89(4):867–879.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., and Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology*, 5(7).

Coffey, E. B., Musacchia, G., and Zatorre, R. J. (2017). Cortical Correlates of the Auditory Frequency-Following and Onset Responses: EEG and fMRI Evidence. *The Journal of Neuroscience*, 37(4):830–838.

Coffey, E. B. J., Herholz, S. C., Chepesiuk, A. M. P., Baillet, S., and Zatorre, R. J. (2016). Cortical contributions to the auditory frequency-following response revealed by MEG. *Nature Communications*, 7:11070.

Crosse, M. J., Di Liberto, G. M., Bednar, A., and Lalor, E. C. (2016). The multivariate temporal response function (mTRF) toolbox: A MATLAB toolbox for relating neural signals to continuous stimuli. *Frontiers in Human Neuroscience*, 10(NOV2016).

Daube, C., Ince, R. A., and Gross, J. (2019). Simple Acoustic Features Can Explain Phoneme-Based Predictions of Cortical Responses to Speech. *Current Biology*, 29(12):1924–1937.e9.

de Cheveigné, A., Di Liberto, G. M., Arzounian, D., Wong, D. D., Hjortkjær, J., Fuglsang, S., and Parra, L. C. (2019). Multiway canonical correlation analysis of brain data. *NeuroImage*, 186(November 2018):728–740.

De Villers-Sidani, E., Alzghoul, L., Zhou, X., Simpson, K. L., Lin, R. C., and Merzenich, M. M. (2010). Recovery of functional and structural age-related changes in the rat primary auditory cortex with operant training. *Proceedings of the National Academy of Sciences of the United States of America*, 107(31):13900–13905.

Decruy, L., Lesenfants, D., Vanthornhout, J., and Francart, T. (2020a). Top-down modulation of neural envelope tracking: The interplay with behavioral, self-report and neural measures of listening effort. *European Journal of Neuroscience*, European J(October 2019):3375–3393.

Decruy, L., Vanthornhout, J., and Francart, T. (2019). Evidence for enhanced neural tracking of the speech envelope underlying age-related speech-in-noise difficulties. *Journal of Neurophysiology*, 122(2):601–615.

Decruy, L., Vanthornhout, J., and Francart, T. (2020b). Hearing impairment is associated with enhanced neural tracking of the speech envelope. *Hearing Research*, 393:107961.

Di Liberto, G. M., O'Sullivan, J. A., and Lalor, E. C. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing.

Current Biology, 25(19):2457–2465.

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.

Ding, N. and Simon, J. Z. (2012a). Emergence of Neural Encoding of Auditory Objects While Listening to Competing Speakers. *PNAS*, 109(29):11854–11859.

Ding, N. and Simon, J. Z. (2012b). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of Neurophysiology*, 107(1):78–89.

Ding, N. and Simon, J. Z. (2013). Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. *Journal of Neuroscience*, 33(13):5728–5735.

Ding, N. and Simon, J. Z. (2014). Cortical entrainment to continuous speech: Functional roles and interpretations. *Frontiers in Human Neuroscience*, 8(MAY):1–7.

Drullman, R., Festen, J. M., and Plomp, R. (1994a). Effect of Reducing Slow Temporal Modulations on Speech Reception. *The Journal of the Acoustical Society of America*, 95(5):2670–2680.

Drullman, R., Festen, J. M., and Plomp, R. (1994b). Effect of Temporal Envelope Smearing on Speech Reception. *The Journal of the Acoustical Society of America*, 95(2):1053–1064.

Etard, O., Kegler, M., Braiman, C., Forte, A. E., and Reichenbach, T. (2019). Decoding of selective attention to continuous speech from the human auditory brainstem response. *NeuroImage*, 200(May):1–11.

Etard, O. and 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 : the official journal of the Society for Neuroscience*, 39(29):5750–5759.

Forte, A. E., Etard, O., and Reichenbach, T. (2017). The human auditory brainstem response to running speech reveals a subcortical mechanism for selective attention. *eLife*, 6:1–13.

Fuglsang, S. A., Märcher-Rørsted, J., Dau, T., and Hjortkjær, J. (2020). Effects of sensorineural hearing loss on cortical synchronization to competing speech during selective attention. *Journal of Neuroscience*, 40(12):2562–2572.

Geirnaert, S., Vandecappelle, S., Alickovic, E., de Cheveigne, A., Lalor, E., Meyer, B., Miran, S., Francart, T., and Bertrand, A. (2021). Electroencephalography-based Auditory Attention Decoding : Toward Neuro-Steered Hearing Devices. *Ieee Signal Processing Magazine. Special issue on Signal Processing for Neurorehabilitation and Assistive Technologies*, 38(4):89–102.

Gillis, M., Decruy, L., Vanthornhout, J., and Francart, T. (2021a). Hearing loss is associated with delayed neural responses to continuous speech. *bioRxiv*, 2021.01.21.

Gillis, M., Vanthornhout, J., Simon, J. Z., Francart, T., and Brodbeck, C. (2021b). Neural markers of speech comprehension: measuring EEG tracking of linguistic speech representations, controlling the speech acoustics. *The Journal of Neuroscience*, (October):JN–RM–0812–21.

Godey, B., Schwartz, D., de Graaf, J. B., Chauvel, P., and Liégeois-Chauvel, C. (2001). Neuromagnetic source localization of auditory evoked fields and intracerebral evoked potentials: a comparison of data in the same patients. *Clinical Neurophysiology*, 112(10):1850–1859.

Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., and Garrod, S. (2014). Speech Rhythms and Multiplexed Oscillatory Sensory Coding in the Human Brain. *PLOS Biology*, 11(12):1–14.

Hamilton, L. S. and Huth, A. G. (2018). The revolution will not be controlled: natural stimuli in speech neuroscience. *Language, Cognition and Neuroscience*, 35(5):573–582.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer, New York.

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D. D., Blankertz, B., and Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, 87:96–110.

Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., and De Lange, F. P. (2020). A hierarchy of linguistic predictions during natural language comprehension. *bioRxiv*, page 2020.12.03.410399.

Iotzov, I. and Parra, L. C. (2019). EEG can predict speech intelligibility. *Journal of Neural Engineering*, 16(3):036008.

Joris, P. X., Schreiner, C. E., and Rees, A. (2004). Neural Processing of Amplitude-Modulated Sounds. *Physiological Reviews*, 84(2):541–577.

Katthi, J. R., Ganapathy, S., Kothinti, S., and Slaney, M. (2020). Deep Canonical Correlation Analysis for Decoding the Auditory Brain. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2020-July:3505–3508.

Kaufeld, G., Bosker, H. R., Alday, P. M., Meyer, A. S., and Martin, A. E. (2020). Structure and meaning organize neural oscillations into a content-specific hierarchy. *bioRxiv*, 40(49):9467–9475.

Kei, J., Smyth, V., Murdoch, B., and McPherson, B. (1999). Measuring the Understanding of Connected Discourse: An Overview of Methodology and Clinical Applications in Rehabilitative Audiology. *Asia Pacific Journal of Speech, Language and Hearing*, 4(1):13–37.

Keidser, G., Naylor, G., Brungart, D. S., Caduff, A., Campos, J., Carlile, S., Carpenter, M. G., Grimm, G., Hohmann, V., Holube, I., Launer, S., Lunner, T., Mehra, R., Rapport, F., Slaney, M., and Smeds, K. (2020). The Quest for Ecological Validity in Hearing Science: What It Is, Why It Matters, and How to Advance It. *Ear and hearing*, 41:5S–19S.

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*, 316:73–81.

Koskinen, M., Kurimo, M., Gross, J., Hyvärinen, A., and Hari, R. (2020). Brain activity reflects the predictability of word sequences in listened continuous speech: Brain activity predicts word sequences. *NeuroImage*, 219(May).

Kulasingham, J. P., Brodbeck, C., Presacco, A., Kuchinsky, S. E., Anderson, S., and Simon, J. Z. (2020). High gamma cortical processing of continuous speech in younger and older listeners. *NeuroImage*, 222(June):117291.

Lalor, E. C. and Foxe, J. J. (2010). Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. *European Journal of Neuroscience*, 31(1):189–193.

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(1):349–359.

Lau, E. F., Phillips, C., and Poeppel, D. (2008). A cortical network for semantics: (de)constructing the N400.

Lesenfants, D. and Francart, T. (2020). The interplay of top-down focal attention and the cortical tracking of speech. *Scientific Reports*, 10(1):1–11.

Lesenfants, D., Vanthornhout, J., Verschueren, E., Decruy, L., and Francart, T. (2019). Predicting individual speech intelligibility from the cortical tracking of acoustic- and phonetic-level speech representations. *Hearing Research*, 380:1–9.

Machens, C. K., Wehr, M. S., and Zador, A. M. (2004). Linearity of Cortical Receptive Fields Measured with Natural Sounds. *Journal of Neuroscience*, 24(5):1089–1100.

Martin, B. A., Tremblay, K. L., and Korczak, P. (2008). Speech evoked potentials: From the laboratory to the clinic. *Ear and Hearing*, 29(3):285–313.

Mesik, J., Ray, L., and Wojtczak, M. (2021). Effects of Age on Cortical Tracking of Word-Level Features of Continuous Competing Speech. *Frontiers in Neuroscience*, 15(April):1–21.

Molinaro, N. and Lizarazu, M. (2017). Delta(but Not Theta)-band Cortical Entrainment Involves Speech-specific Processing. *European Journal of Neuroscience*, 48(7).

Montoya-Martínez, J., Vanthornhout, J., Bertrand, A., and Francart, T. (2021). Effect of number and placement of EEG electrodes on measurement of neural tracking of speech. *PLoS ONE*, 16(2 Febuary):1–18.

Nourski, K. V., Reale, R. A., Oya, H., Kawasaki, H., Kovach, C. K., Chen, H., Howard, M. A., and Brugge, J. F. (2009). Temporal Envelope of Time-Compressed Speech Represented in the Human Auditory Cortex. *Journal of Neuroscience*, 29(49):15564–15574.

O'Sullivan, A. E., Lim, C. Y., and Lalor, E. C. (2019a). Look at me when I'm talking to you: Selective attention at a multisensory cocktail party can be decoded using stimulus reconstruction and alpha power modulations. *European Journal of Neuroscience*, 50(8):3282–3295.

O'Sullivan, J., Herrero, J., Smith, E., Schevon, C., McKhann, G. M., Sheth, S. A., Mehta, A. D., and Mesgarani, N. (2019b). Hierarchical Encoding of Attended Auditory Objects in Multi-talker Speech Perception. *Neuron (Cambridge, Mass.)*, 104(6):1195—-1209.e3.

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., Slaney, M., Shamma, S. A., and Lalor, E. C. (2015). Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG. *Cerebral Cortex*, 25(7):1697–1706.

Parthasarathy, A., Bartlett, E. L., and Kujawa, S. G. (2019). Age-related Changes in Neural Coding of Envelope Cues: Peripheral Declines and Central Compensation. *Neuroscience*, 407:21–31.

Peelle, J. E. and Davis, M. H. (2012). Neural Oscillations Carry Speech Rhythm through to Comprehension. *Frontiers in Psychology*, 3(September):1–17.

Petersen, E. B., Wöstmann, M., Obleser, J., and Lunner, T. (2016). Neural Tracking of Attended versus Ignored Speech Is Differentially Affected by Hearing Loss. *Journal of Neurophysiology*, 117(1):18–27.

Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W., Humes, L. E., Lemke, U., Lunner, T., Matthen, M., Mackersie, C. L., Naylor, G., Phillips, N. A., Richter, M., Rudner, M., Sommers, M. S., Tremblay, K. L., and Wingfield, A. (2016). Hearing Impairment and Cognitive Energy: The Framework for Understanding Effortful Listening (FUEL). *Ear & Hearing*, 37(1):5S–27S.

Picton, T. W. (2010). *Human Auditory Evoked Potentials*. Plural Pub.

Rosen, S. (1992). Temporal Information in Speech: Acoustic, Auditory and Linguistic Aspects. *Phil. Trans. R. Soc. Lond. B*, 336(1278):367–373.

Saiz-Alia, M., Reichenbach, T., Saiz-Alía, M., and Reichenbach, T. (2020). Computational modeling of the auditory brainstem response to continuous speech. *Journal of Neural Engineering*, in press(3):0–31.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., Ekelid, M., Series, N., and Oct, N. (1995). Speech Recognition with Primarily Temporal Cues. *Source: Science, New Series*, 270(5234):303–304.

Somers, B., Verschueren, E., and Francart, T. (2019). Neural tracking of the speech envelope in cochlear implant users. *Journal of Neural Engineering*, 16(1).

Steinschneider, M., Liégeois-Chauvel, C., and Brugge, J. F. (2011). *Auditory Evoked Potentials and Their Utility in the Assessment of Complex Sound Processing*, pages 535–559. Springer US, Boston, MA.

Theunissen, F. E., Sen, K., and Doupe, A. J. (2000). Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. *Journal of Neuroscience*, 20(6):2315–2331.

Tichko, P. and Skoe, E. (2017). Frequency-dependent fine structure in the frequency-following response: The byproduct of multiple generators. *Hearing Research*, 348:1–15.

Van Canneyt, J., Wouters, J., and Francart, T. (2021a). Cortical compensation for hearing loss, but not age, in neural tracking of the fundamental frequency of the voice. *Journal of Neurophysiology*, 126(3):791–802.

Van Canneyt, J., Wouters, J., and Francart, T. (2021b). Enhanced neural tracking of the fundamental frequency of the voice. *IEEE Transactions on Biomedical Engineering (Early Access)*, x:1–1.

Van Canneyt, J., Wouters, J., and Francart, T. (2021c). Neural tracking of the fundamental frequency of the voice: the effect of voice characteristics. *European Journal of Neuroscience*, 00(January):1–14.

Vanthornhout, J., Decruy, L., and Francart, T. (2019). Effect of Task and Attention on Neural Tracking of Speech. *Frontiers in Neuroscience*, 13.

Vanthornhout, J., Decruy, L., Wouters, J., Simon, J. Z., and Francart, T. (2018). Speech Intelligibility Predicted from Neural Entrainment of the Speech Envelope. *JARO - Journal of the Association for Research in Otolaryngology*, 19(2):181–191.

Verschueren, E., Gillis, M., Decruy, L., Vanthornhout, J., and Francart, T. (2022). The effect of speech rate on acoustic and linguistic neural speech processing.

Verschueren, E., Somers, B., and Francart, T. (2019). Neural envelope tracking as a measure of speech understanding in cochlear implant users. *Hearing Research*, 373:23–31.

Verschueren, E., Vanthornhout, J., and Francart, T. (2021). The effect of stimulus intensity on neural envelope tracking. *Hearing Research*, 403:108175.

Wang, X., Lu, T., Bendor, D., and Bartlett, E. (2008). Neural coding of temporal information in auditory thalamus and cortex. *Neuroscience*, 154(1):294–303.

Weissbart, H., Kandylaki, K. D., and Reichenbach, T. (2019). Cortical Tracking of Surprisal during Continuous Speech Comprehension. *Journal of Cognitive Neuroscience*, pages 1–12.

Yang, M., Sheth, S. A., Schevon, C. A., II, G. M. M., and Mesgarani, N. (2015). Speech Reconstruction from Human Auditory Cortex with Deep Neural Networks. *Interspeech*, page 5.

Zan, P., Presacco, A., Anderson, S., and Simon, J. Z. (2020). Exaggerated cortical representation of speech in older listeners: mutual information analysis. *Journal of Neurophysiology*, 124(4):1152–1164.

Zion Golumbic, E. M., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., Goodman, R. R., Emerson, R., Mehta, A. D., Simon, J. Z., Poeppel, D., and Schroeder, C. E. (2013). Mechanisms Underlying Selective Neuronal Tracking of Attended Speech at a "Cocktail Party". *Neuron*, 77(5):980–991.

Zuk, N. J., Murphy, J. W., Reilly, R. B., and Lalor, E. C. (2021). Envelope reconstruction of speech and music highlights stronger tracking of speech at low frequencies. *PLOS Computational Biology*, 17(9):e1009358.