NeuroImage 297 (2024) 120675

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

On the speech envelope in the cortical tracking of speech

Mohamed F. Issa [a,c,*], Izhar Khan [a], Manuela Ruzzoli [a,b], Nicola Molinaro [a,b], Mikel Lizarazu [a]

[a] BCBL, Basque Center on Cognition, Brain and Language, San Sebastian, Spain
[b] Ikerbasque, Basque Foundation for Science, Bilbao, Spain
[c] Department of Scientific Computing, Faculty of Computers and Artificial Intelligence, Benha University, Benha, Egypt

ARTICLE INFO

Keywords:
Cortical tracking of speech
Envelope extraction methods
Functional neuroimaging techniques
Gammatone filterbanks

ABSTRACT

The synchronization between the speech envelope and neural activity in auditory regions, referred to as cortical tracking of speech (CTS), plays a key role in speech processing. The method selected for extracting the envelope is a crucial step in CTS measurement, and the absence of a consensus on best practices among the various methods can influence analysis outcomes and interpretation. Here, we systematically compare five standard envelope extraction methods the absolute value of Hilbert transform (*absHilbert*), gammatone filterbanks, heuristic approach, Bark scale, and vocalic energy), analyzing their impact on the CTS. We present performance metrics for each method based on the recording of brain activity from participants listening to speech in clear and noisy conditions, utilizing intracranial EEG, MEG and EEG data. As expected, we observed significant CTS in temporal brain regions below 10 Hz across all datasets, regardless of the extraction methods. In general, the gammatone filterbanks approach consistently demonstrated superior performance compared to other methods. Results from our study can guide scientists in the field to make informed decisions about the optimal analysis to extract the CTS, contributing to advancing the understanding of the neuronal mechanisms implicated in CTS.

1. Introduction

In recent years, there has been a growing interest in exploring the cortical tracking of speech (CTS) as a potential metric for evaluating acoustic, linguistic, and cognitive speech processing (Kösem and van Wassenhove, 2017; Obleser and Kayser, 2019; Meyer, 2018, 2020; Molinaro et al., 2021). CTS has been suggested to reflect the capacity of neural oscillations to synchronize, or phase-lock, with quasi-rhythmic information contained in slow amplitude modulations of speech (speech envelope). CTS is commonly observed in temporal brain regions and within the delta (<4 Hz) and theta (4 – 8 Hz) frequency bands, aligning with the prosodic and syllabic rhythms in the speech envelope, respectively (Gross et al., 2013; Peelle, Gross, and Davis, 2013; Doelling et al., 2014; Molinaro and Lizarazu, 2018; Destoky et al., 2019; Bourguignon et al., 2020; Ershaid et al., 2024). It has been suggested that CTS is an important part of speech processing because it helps separate and decode continuous speech signals into linguistic units at different timescales (Ahissar et al., 2001; Giraud and Poeppel, 2012; Peelle and Davis, 2012; Peelle et al., 2013; Zoefel and VanRullen, 2015; Ding et al., 2016; Keitel, Gross, and Kayser, 2018; Kosem et al., 2018; Meyer and Gumbert, 2018; Lizarazu, Carreiras and Molinaro, 2023). CTS can be observed throughout the lifespan (Bertels et al., 2023), from newborns (Menn et al., 2022; Ortiz-Barajas, Guevara and Gervain, 2023) to older adults (Henry et al., 2017). Furthermore, atypical CTS has been associated with language impairments, as evidenced by studies on hearing loss (Decruy et al., 2020; Gillis et al., 2022; Kurthen et al., 2021), stroke-related or dementia-related aphasias (Dial et al., 2021; Kries et al., 2023; Quique et al., 2023), dyslexia (Lizarazu et al., 2015, 2021a, Lizarazu et al., 2021b; Molinaro et al., 2016; Lallier et al., 2017, 2018; Rios-López et al., 2017; Schwarz et al., 2024) and specific language impairments (Kaganovich et al., 2014).

CTS can be measured using time-sensitive imaging methods in neuroscience, such as invasive and non-invasive EEG or MEG, and can be observed at the single-trial level (Horton et al., 2014; O'sullivan et al., 2015). Because of the unique rhythmic patterns in the speech signal, CTS is commonly evaluated using undirected connectivity measures in the frequency domain. The connectivity methods used to estimate the CTS vary significantly (Bastos and Schoffelen, 2016), and some of them have already been compared in previous studies (David, Cosmelli and Friston, 2004; Kreuz et al., 2007; Quiroga et al., 2002; Gross et al., 2021). For example, in Gross et al., 2021, various connectivity methods (such as phase-locking value, Gaussian-Copula mutual information, Rayleigh

* Corresponding author.
E-mail address: m.issa@bcbl.eu (M.F. Issa).

https://doi.org/10.1016/j.neuroimage.2024.120675
Received 26 February 2024; Received in revised form 5 June 2024; Accepted 6 June 2024
Available online 15 June 2024
1053-8119/© 2024 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

test, weighted pairwise phase consistency, magnitude squared coherence, and entropy) were contrasted for their impact on CTS estimation in participants engaged in speech perception during a MEG recording. The findings from Gross et al. (2021) highlighted that the weighted pairwise phase consistency (WPPC) and Gaussian-Copula mutual information (GCMI) consistently exhibited superior performance over the other metrics for assessing CTS, suggesting a substantial impact of the chosen method on CTS sensitivity. Yet, another crucial step in computing CTS is the method used for extracting the speech envelope. Among the frequently employed methods in CTS research, we can find the absolute value of Hilbert transform (*absHilbert*) (Hilbert, 1912), the *gammatone filterbanks* (Biesmans et al. 2017), the *heuristic approach* (Jarne, 2018), the *Bark scale* (Zwicker et al., 1979; Zwicker and Terhardt, 1980; Oganian and Chang, 2019), and the *vocalic energy* (Tilsen and Arvaniti, 2013). These methods utilize distinct mathematical and physical approaches to extract the speech envelope, therefore stressing different aspects of the speech signal. This decision is key in CTS studies as it involves aligning the speech envelope with the neural signals.

The Hilbert transform is a mathematical operation that, when applied to the speech signal, produces a complex analytic signal that contains information about both amplitude (or envelope) and phase. The simplicity of the Hilbert transform makes it the most used method in CTS studies when extracting the speech envelope (e.g., O'Sullivan et al., 2015; Assaneoet al., 2019; Braiman et al., 2018; Molinaro and Lizarazu, 2018). *Gammatone filterbanks* decompose the speech signal into spectral channels by using spaced filters within the human auditory range. The output signals from each gammatone filter are subsequently combined to derive the speech envelope (Biesmans et al. 2017). While the fundamental methodology of the *gammatone filterbanks* remains consistent, there is a notable degree of freedom in the chosen parameters across studies (e.g., number of filters, order, and weighting assigned to each filter) (Darling, 1991). *Gammatone filterbanks* provide a more biologically inspired method and demonstrate robustness to noise, but it involves a more complex computation compared to the *absHilbert*. The *heuristic approach* method, introduced by Jarne in 2018, relies on peak detection algorithms and also offers a faster and simpler alternative to *gammatone filterbanks* for extracting the speech envelope. It involves two steps: first, taking the absolute value of the speech signal; second, dividing the absolute signal into non-overlapping slots with a predefined window length and performing peak detection. This approach resembles simple signal rectification methods and is proposed to avoid attenuation typically observed when the *absHilbert* is applied to natural sounds like speech or music (Caetano and Rodet, 2011; Jarne, 2018). The *Bark scale* method implies filtering the signal according to the *Bark scale*, a frequency scale that mirrors the human ear's response (Zwicker and Terhardt, 1980) and subsequently averaging across all bands (Oganian and Chang, 2019). This methodology resembles the *gammatone filterbanks* approach, with the distinction that the *Bark scale* method directly employs filters based on perceptual frequency spacing, while *gammatone filterbanks* utilize filters inspired by the physiological structure of the cochlea. Ultimately, the method developed by Tilsen and Arvaniti in 2013, referred to as the *vocalic energy*, extracts the speech envelope by assigning greater importance to the presence of vocalic energy compared to consonantal energy. To achieve this, the speech signal undergoes bandpass filtering within the frequency range of 500 Hz to 4000 Hz. The lower cut-off frequency (500 Hz) significantly attenuates the impact of the fundamental frequency, making voiced and voiceless consonants more alike while distinguishing them from vowels, as vowel formant energy is retained within this frequency range. Simultaneously, the higher cut-off frequency (4000 Hz) reduces the contribution of high-frequency bands related to fricatives and bursts, preventing their representation as prominent peaks in the envelope. For all these methods, typically, the final step in calculating the speech envelope involves applying a low-pass filter with a cut-off frequency between 10 and 15 Hz.

Importantly, the method used for speech envelope extraction affects CTS itself, potentially influencing the study's conclusions and diminishing the comparability of results across different studies. Biesmans et al. (2017) conducted a comparative analysis of envelope extraction techniques within the context of auditory attention decoding. Their study revealed that classification performance increases with "auditory-inspired modifications," such as *gammatone filterbanks*, in contrast to more straightforward methods like half-wave rectification (e.g., Dellwo et al., 2015; Kolly and Dellwo, 2014) or the *absHilbert* for analytic signal extraction (e.g., Gervain and Geffen, 2019). The implications of Biesmans et al.'s (2017) findings suggest that not all speech envelope extraction procedures are equally sensitive in the investigation of the CTS.

The present study tackles the lack of consensus in speech envelope extraction methods, posing challenges for CTS research interpretation, replicability, and cross-study comparisons. Through a systematic investigation of diverse methods, this study aims to enhance CTS robustness and establish a standardized methodological foundation for future studies in the field. In our study, we conducted a comprehensive analysis of intracranial EEG (iEEG), MEG, and EEG data, which were recorded during continuous speech presentations. This extensive examination included EEG recordings with clear speech and varied background noise levels, encompassing a range of stimulus lengths, types, and languages. This approach allowed us to robustly test the method across diverse auditory environments and to evaluate the effects of different experimental conditions on CTS. We systematically assess five standard envelope extraction methods: absolute value of Hilbert transform (*absHilbert*), *gammatone filterbanks, heuristic approach, Bark scale,* and *vocalic energy*, analyzing their influence on CTS at both the group and individual levels. For transparency and reproducibility, our analysis scripts and data are publicly available on the Open Science Framework (https://osf.io/gtsa5/).

2. Methods

For our clear speech analysis, we utilized three datasets: intracranial electroencephalography (iEEG), accessible here: https://openneuro.org/datasets/ds004703/versions/1.1.0 (Mai et al., 2024), magnetoencephalography (MEG) from Destoky et al., 2019, and an EEG dataset detailed by Molinaro et al., 2021. For speech in noise, we used an EEG dataset available at https://osf.io/b9wdp/ (Mohammadi et al., 2023).

Participants

Clear speech

iEEG: Ten adult participants, including 4 women, were involved in the experiment, with an average age of 32 ± 11 years (mean ± SD). All participants from UC San Diego Health underwent intracranial stereo EEG and subdural electrode implantation as part of their treatment for refractory epilepsy or related conditions. All participants were native English speakers, reporting normal hearing (self-reported) and scored within the normal range on a series of neuropsychological language tests. The research protocol received approval from the UC San Diego Institutional Review Board, and all subjects provided written informed consent before surgery.

MEG: Ten healthy adult participants, including 5 women, were involved in the experiment, with an average age of 25 ± 4 years (mean ± SD). All participants were native French speakers and had no history of developmental, neurological, or psychiatric disorders. Normal hearing was confirmed through pure tone audiometry. The study received approval from the Ethics Committee of CUB Hôpital Erasme (Brussels, Belgium), and participants provided written informed consent.

EEG: The experiment involved twenty-five participants, among whom thirteen were females, with an average age of 41 ± 5 years. All participants were native Spanish speakers and had no documented history of developmental, neurological, or psychiatric disorders, and exhibited normal hearing abilities. The Ethical Committee of the Basque Center on Cognition Brain and Language (BCBL) granted approval for the experiment, adhering to the principles outlined in the Declaration of

Helsinki, and all participants provided written informed consent.

Speech in noise

EEG: Thirty-one healthy native Danish speakers, including 13 women, with an average range of 24 ± 3 were recruited, all with normal hearing and no history of neurological, psychiatric illness, or use of psychotropic medication. Written informed consent was obtained from all participants, who were also financially compensated for their involvement. The study adhered to the principles of the Declaration of Helsinki and received approval from the ethics committee of Northern Jutland, Denmark.

Stimuli and task

Clear speech

iEEG: Participants were instructed to attentively listen to succinct excerpts of conversational American English speech obtained from the Buckeye Corpus (Fosler-Lussier et al., 2007). To assess task engagement, participants orally responded to a two-alternative question concerning the content of each passage immediately after listening to it. The passages, from 27 native speakers (12 women, 15 men), lasted 25–76 s (mean 38 s). Monophonic recordings were captured using a head-mounted microphone (Crown CM-311A) and directed to a DAT recorder (Tascam DA-30 MKII) at a 48 kHz sampling rate via an amplifier (Yamaha MV 802). Subsequent to each excerpt, participants were presented with a two-alternative choice question regarding the heard passage, prompting an oral response to confirm attentiveness. The questions were recorded using a Blue Yeti USB microphone, sampling at a rate of 48 kHz, by a native American English speaker. Experimental instructions and stimuli were presented to participants in their hospital rooms using PsychoPy for Python 2.7 (Peirce, 2009) on a Windows 10 desktop PC (Dell XPS 8910). All stimuli were presented at −20 dB (dB) full scale. The task comprised six blocks, each containing eight English trials, with each trial featuring a short spoken passage, a content question, and an oral response.

MEG: The speech stimuli comprised six 5-minute French stories, read aloud by different speakers (3 males and 3 females). Stimuli were presented at approximately 60 dB through a MEG-compatible 60 × 60 cm^2 high-quality flat panel loudspeaker (Panphonics SSH sound shower, Panphonics), located around 2.7 m away and facing the subjects. Following the story, participants were asked 16 questions about the story they attended.

EEG: We obtained a 14-minute recording of a text being read by a Spanish native male speaker, which was digitized at a sampling rate of 44.1 kHz using a digital recorder (Marantz PMD670). The audio files (*.wav) were segmented using Praat (Boersma, 2007). The reader was unaware of the experiment's purpose. Participants were directed to assume a comfortable seated position facing the computer screen and were instructed to minimize movements during recording. They were instructed to focus their gaze on a fixation cross while attentively listening to the stimuli, without the need to engage in any additional tasks. Stimuli were delivered via loudspeakers at a sound pressure level of 80 dB Sound Pressure Level using Psychopy (Peirce, 2009).

Speech in noise

EEG: The stimuli were a series of disconnected sentences. The speech stimuli utilized in this study were sourced from the Dantale II database (Wagener et al., 2003), which comprises 150 sentences. Each sentence was generated by a random combination of the alternatives of a base list. The base list consisted of ten sentences, each structured with a subject, verb, numeral, adjective, and object, ensuring syntactical consistency while introducing semantic unpredictability (e.g., "Ulla owns five red jackets" in English). All sentences were audibly recorded by a female native Danish speaker at a sampling rate of 44.1 kHz. Their durations ranged from 1.85 s to 2.52 s, with an average of 2.22 s ± 0.12 s.

The experiment comprised four blocks, each assigned a randomly determined Signal-to-Noise Ratio (SNR) level (−9 dB, −6 dB, −3 dB, 0 dB). This was achieved by adjusting the intensity of the speech while maintaining a constant background noise level. The speech-shaped noise was tailored to mimic the long-term power spectrum of speech. SNR was computed as the ratio of the power of the speech signal to the power of the background noise. Intensity adjustments for the speech at different SNRs were performed using MATLAB, which also served as the platform for audio presentation. The volume levels were calibrated based on the comfort assessments of a small group of normal-hearing individuals. Notably, none of the participants reported discomfort with the volume levels. Within each block, 25 trials were conducted. Each trial commenced with a 3-second background noise segment, followed by a random interval of 0–1 second, during which participants focused on a fixation cross displayed on a screen in front of them. Subsequently, a stimulus was presented wherein speech was delivered amidst background noise. Following the speech presentation, the fixation cross remained on screen while background noise persisted for approximately 3 s. Subsequently, a response interval ensued, during which all items from the base list appeared on the screen in a 10 × 5 grid (word × category). Participants utilized a mouse to select the words in the order they were heard. After each block of 25 trials, participants rated their level of listening effort on a scale of 1 to 10 using the NASA Task Load Index (Hart and Staveland, 1988), followed by a 3-minute rest period. The experiment was programmed using custom code in MATLAB (R2021b, MathWorks Inc.). Sound playback was facilitated through a soundcard (Scarlett 2i2 2nd Gen), and presentation was controlled using the Psychophysics Toolbox (PTB-3). The audio signal was delivered diotically via insert-earphones (a-JAYS Three). Prior to commencing the main experiment, participants were exposed to sample speech in each condition and were familiarized with all procedures.

Data acquisition and preprocessing

Clear speech

***iEEG*:** Intracranial EEG during speech listening was amplified utilizing a multi-channel amplifier system (Natus Quantum) and recorded through Natus NeuroWorks software. For all patients, a scalp electrode was used for referencing and ground. Depth electrodes were manufactured by Ad-Tech and were Spencer Probe depth electrodes. Each electrode has 10 leads evenly spaced 3–7 mm apart. Simultaneous recording of auditory stimuli and oral responses was achieved by incorporating the output of a Zoom H2n microphone as an additional input channel to the Natus Quantum amplifier.

Pre-operative T1-weighted magnetic resonance (MR) sequences were co-registered with post-operative axial non-contrast CT scans slices using Statistical Parametric Mapping (SPM) 12.2 (Friston, 2003). Macro- and microelectrodes were automatically localized on the fused image and manually adjusted using LeGUI software (Davis et al., 2021). Electrode positions were warped into Montreal Neurological Institute (MNI) 152 space for assignment of electrodes to the nearest region of interest in the Automated Anatomical Labeling (AAL) atlas. We used the BrainView software (Xia, Wand and He, 2013) for the spatial visualization of electrodes in the MNI space.

After recording, neural data were exported from the clinical NeuroWorks system in .edf (European Data Format) format. Pre-processing was conducted using the Python package MNE Python (Gramfort et al., 2013). A total of 45 channels displaying excessive artifacts or line noise were removed. The remaining channels were common average referenced, notch filtered at 60 Hz and its harmonics and bandpass filtered in the range of 0.1–170 Hz.

MEG: During the presentation of the story, participants' brain activity was recorded using MEG at the CUB Hôpital Erasme. The MEG system used was a whole-scalp-covering system called Triux, manufactured by Elekta. The MEG sensor array consisted of 306 sensors arranged in 102 triplets, with each triplet comprising one magnetometer and two orthogonal planar gradiometers. The recordings took place in a lightweight magnetically shielded room known as Maxshield, also manufactured by Elekta. MEG signals were band-pass filtered within the frequency range of 0.1 – 330 Hz and sampled at a rate of 1000 Hz. EEG signals, on the other hand, underwent a low-pass filtering at 450 Hz and were also sampled at 1000 Hz. The participants' head position throughout the experiment was monitored using four head-position

indicator coils. Before the MEG session, an electromagnetic tracker (Fastrack, Polhemus) was employed to digitize the location of these coils and a minimum of 300 points on the head's surface (including the scalp, nose, and face) concerning anatomical fiducials. The MEG data underwent initial offline preprocessing using the temporal signal space separation method within the MaxFilter software (MaxFilter, Neuromag, Elekta; correlation limit 0.9, segment length 20 s) to eliminate external interferences and rectify head movements (Taulu and Simola, 2006). To mitigate physiological artifacts in MEG data, 30 independent components underwent evaluation from data band-pass filtered within the 0.1 – 25 Hz range and reduced to a rank of 30 using principal component analysis. We successfully identified and isolated the components associated with electrocardiogram (ECG) and electrooculogram (EOG) artifacts. The corresponding MEG signals were reconstructed by subtracting them from the complete-rank data through the mixing matrix. Across subjects and conditions, an average of 2.3 ± 1.0 components (mean ± SD) were rejected. Sections of data within a 1-second timing around remaining artifacts were marked as bad. Data were flagged as contaminated by artifacts when MEG amplitude reached at least 5 pT in a magnetometer or 1 pT/cm in a gradiometer.

EEG: EEG data were collected using a BrainAmp amplifier and BrainVision Recorder software (Brain Products, Germany). EEG signals were recorded from 32 electrodes positioned according to the international 10–20 system. To maintain high-quality EEG recordings, scalp-electrode impedance was kept below 5 kΩ for scalp electrodes and under 10 kOhm for reference and EOG electrodes.The experiment was conducted in a room equipped with electromagnetic shielding. Data were sampled at a rate of 1000 Hz and band-pass filtered online from 0.1 to 1000 Hz. The recording reference was electrode FCz, and offline re-referencing was performed to the average of the left and right mastoids. Electrode AFz served as the ground. Additionally, two electrodes placed at the outer canthi of the eyes recorded horizontal eye movements, while two electrodes positioned above and below the left eye monitored vertical eye movements.

ECG and EOG artifacts were identified using Independent Component Analysis (ICA) and subsequently subtracted from the recordings in a linear fashion. The ICA decomposition was conducted utilizing the Infomax algorithm as implemented in the Fieldtrip toolbox (Oostenveld et al., 2011). The number of removed components associated with heartbeat and ocular artifacts varied among participants, ranging from 1 to 2 for heartbeat components and 1 to 3 for ocular components. Additionally, visual inspection of the recordings was performed to detect bad channels, which were substituted with interpolated values computed as the average of the neighboring electrodes obtained through the triangulation method implemented in Fieldtrip.

Speech in noise

EEG: The EEG data were collected using a g.HIamp biosignal amplifier (g.tec medical engineering GmbH, Austria), equipped with 64 channels. Electrodes were positioned on a cap following the 10–20 international system. Sampling of EEG signals occurred at a rate of 1200 Hz, with the left earlobe (A1) serving as the reference point. Throughout the recording process, electrode impedances were maintained below 5 kOhm. The experiment took place within an electromagnetically shielded room.

ECG and EOG artifacts were detected using ICA and then subtracted from the recordings linearly. ICA decomposition was conducted using the Infomax algorithm within the Fieldtrip toolbox. The number of heartbeat and ocular components removed varied across participants, ranging from 1 to 4 and 1 to 3 components, respectively. Trials were visually inspected to remove residual artifacts and faulty channels were substituted with interpolated values derived from neighboring electrodes using the triangulation method in Fieldtrip. Following the inspection of EEG data, six participants were removed from the analysis due to poor data quality, leaving 25.

Speech envelope extraction method

In this study, we utilized five methods for extracting the speech envelope, chosen for their common use in disciplines like linguistics, neuroscience, and other speech sciences. The following is a concise description of each method:

1- *Absolute value of Hilbert transform* (*absHilbert*): The speech envelope was obtained by applying the hilbert transform to the broadband speech signal. The amplitude envelope, which represents the instantaneous amplitude, is then obtained by taking the absolute value of the analytic signal. (e.g., O'Sullivan et al., 2015; Assaneoet al., 2019; Braiman et al., 2018; Molinaro and Lizarazu, 2018).
2- *Gammatone filterbanks*: Speech envelopes were derived using *gammatone filterbanks*, followed by a power-law operation that emulated the compressive response of the inner ear (Biesmans et al., 2027). The filter bank comprised 15 perceptually uniform gammatone filters, each having an equivalent rectangular bandwidth of 1.5, and center frequencies spanning from 150 Hz to 4 kHz. The output from each filter was subjected to full-wave rectification and power-law compression (i.e., taking the absolute value and raising it to the power of 0.6). The resulting sub-band envelopes were then averaged, yielding a consolidated single envelope.
3- *Heuristic approach*: The initial step involved computing the absolute value of the signal, which was segmented into non-overlapping intervals using a 250 ms moving window. Peak detection was then executed by replacing all values within each signal interval with the maximum value found within that specific interval (Jarne, 2018).
4- *Bark scale*: The process involves square-rectifying the signal within specific filter banks, defined according to the Bark scale (Zwicker et al., 1979; Zwicker and Terhardt, 1980). Subsequently, the computation of the averaged signal across these bands follows (Oganian and Chang, 2019).
5- *Vocalic energy*: To derive the speech envelope, we filtered the speech signal with a fourth-order bandpass Butterworth filter set at [400, 4000] Hz, corresponding to the estimated locus of vocalic energy (Tilsen and Arvaniti, 2013).

In each method, the obtained speech envelopes underwent smoothing through zero-phase low-pass filters with a 10 Hz cutoff frequency and were subsequently rescaled to a standardized range for ease of comparison. In our study, we adopted a shared code from MacIntyre et al. (2022), available at the following link: (https://github.com/alexisdmacintyre/AcousticLandmarks).

Cortical tracking of speech analysis

Cortical Tracking of Speech (CTS) between neural activity from iEEG, MEG and EEG data and the output of different envelope extraction methods followed a standardized procedure (Gross et al., 2021). Numerous neuroimaging studies demonstrate that CTS is a neuronal mechanism primarily present in auditory regions (Gross et al., 2013; Molinaro et al., 2016; Lizarazu et al., 2021c). Therefore, in the present study, we computed CTS from artifact-free data obtained from sensors located in temporal brain regions for each technique (Supplementary Fig. 1 illustrates the layout of the sensors).

For iEEG, within the MNI space, we selected electrodes located in the superior temporal and mid-temporal gyrus. One patient was excluded from the subsequent analyses due to the absence of electrodes implanted in temporal brain regions. The analysis of CTS involved a total of 272 electrodes across the 9 patients, with an average of 27.9 ± 12.1 (mean ± SD) electrodes. In MEG, the analysis centered on temporal sensors located in both the left (0131, 0221, 0141, 1511, 0231, 1541, 1621, 1521, 1531, and 1641) and right (1311, 1441, 1431, 1341, 2611, 2411, 2621, 2641, 2431 and 2631) hemispheres.

Similarly, in both the EEG experiments for clear speech and noisy speech, we focused on temporal sensors located in the left hemisphere (FT7, FC5, C3, C5, T7, TP7, CP5, and CP3) and right hemisphere (FT8, FC6, C4, C6, T8, TP8, CP6, and CP4).

Subsequently, the weighted pairwise phase consistency (WPPC,

Vinck et al., 2010) was employed to assess raw CTS. WPPC was calculated as the mean of the circular correlation between the phase of neural activity in temporal regions and the phase of the corresponding speech envelope for each method. Evaluation of WPPC spanned the 0.5 – 10 Hz frequency range with 0.5 Hz frequency resolution (following Bourguignon et al. 2013; Molinaro and Lizarazu 2018; Gross et al., 2021). The time-frequency representation of the data (iEEG/MEG/EEG + envelope) and the phase angles for WPPC computation were obtained using Fast Fourier Transform (FFT). To accomplish this, the multi-taper discrete prolate spheroidal sequences (DPSS) with ±2 Hz smoothing were employed in time windows of 2 s with 50 % overlap. We chose this combination of spectral estimation method and connectivity measure, as it enables the optimization of CTS calculation (Gross et al., 2021). Following this systematic procedure, WPPC values were obtained for each (i) participant, (ii) technique, (iii) hemisphere, and (iv) frequency bin below 10 Hz.

For each technique and envelope extraction method, surrogate CTS values were generated by randomly shifting the spectral estimates of the envelope signals relative to the M/EEG data, employing circular wrapping around the edges of the time series. Temporal shifting of data was a well-established technique for surrogate data generation, as it eradicated any inherent synchronization in the data (Andrzejak et al., 2003; Gross et al., 2021) while preserving the autocorrelation structure of the signals. The shifting procedure was iterated 200 times, and for each iteration, CTS was calculated using the WPPC in the same manner as it was employed for calculating the raw CTS. Subsequently, the raw CTS values were normalized (z-scored) by subtracting the mean and dividing by the standard deviation of the surrogate distribution corresponding to each frequency (Lancaster et al., 2018; Schreiber and Schmitz, 2000). This beneficial normalization ensured comparability among CTS values associated with different speech envelopes. Finally, we computed the mean of normalized CTS values within the delta (< 4 Hz) and theta (4 – 8 Hz) frequency bands, as we expected that CTS would be maximal in these bands (Gross et al., 2013; Molinaro and Lizarazu, 2018; Destoky et al., 2019; Lizarazu et al., 2021c).

Statistical analysis

Clear speech

We conducted repeated measures ANOVA on the CTS values to explore differences between speech envelope methods. The within-subject factors included speech Envelope method (absHilbert, gammatone filterbanks, heuristic approach, Bark scale, and vocalic energy), while the between-subject factor was Neuroimaging technique (iEEG, MEG, and EEG).

Speech in noise

Separate repeated measures ANOVAS were performed on the intelligibility and listening effort scores, with SNE level (−9 dB, −6 dB, −3 dB, 0 dB) as the within-subject factor. Additionally, we computed repeated measures ANOVA on the CTS values, with speech Envelope method (absHilbert, gammatone filterbanks, heuristic approach, Bark scale, and vocalic energy) and SNR level (−9 dB, −6 dB, −3 dB, 0 dB) as the within-subject factors. Finally, correlations between the behavioral scores and CTS values were computed for each SNR condition separately and in combination, using Pearson correlation coefficients. For both clear speech and speech in noise conditions, post-hoc analyses were conducted employing two-tailed *t*-tests. To account for multiple comparisons, the p-values were adjusted using the Bonferroni correction method.

3. Results

The first line of Fig. 1 depicts the time evolution of a speech signal (a 4.5 s phrase); the next lines represent its corresponding speech envelope calculated using different methods (i.e., *absHilbert, gammatone filterbanks, heuristic approach, Bark scale, vocalic energy*). Significant correlations were observed across all combinations of methods (all $rs > 0.69$, all $ps < 0.01$). The highest correlation was identified between the *gammatone filterbanks* and the *heuristic approach* ($r = 0.95$), while the lowest correlation occurred between envelopes obtained with the *heuristic approach* and *vocalic energy* ($r = 0.69$). See Supplementary Fig. 2 for all the other combinations and the correlation matrix among speech envelopes derived from various methods.

Clear speech

The CTS was obtained for each neuroimaging technique (iEEG, MEG and EEG) and envelope extraction method across all participants ($N_{iEEG} = 9$; $N_{MEG} = 10$; $N_{EEG} = 25$) (left side in Fig. 2) within the frequency range 0 - 10 Hz. As expected, the CTS spectrum shows two peaks in the delta (<4 Hz) and theta (4 - 8 Hz) frequency bands, regardless of the neuroimaging technique and the envelope extraction method. Overall, the analysis revealed that the best performance (i.e., the highest CTS value) was achieved with *gammatone filterbanks*, followed by the *heuristic approach*, the *absHilbert*, the *Bark scale*, and finally, *vocalic energy*. For the following analyses, we computed the mean CTS values within temporal regions and across delta and theta bands for each neuroimaging technique (right side in Fig. 2 and Supplementary Table 1).

The ANOVA (Envelope method x Neuroimage technique) of the mean CTS values showed a main effect of the Envelope method ($F(4164)=37.39$, $p < 0.001$, $\eta 2=0.09$). Post hoc tests showed that the *gammatone filterbanks* method exhibited significantly stronger CTS

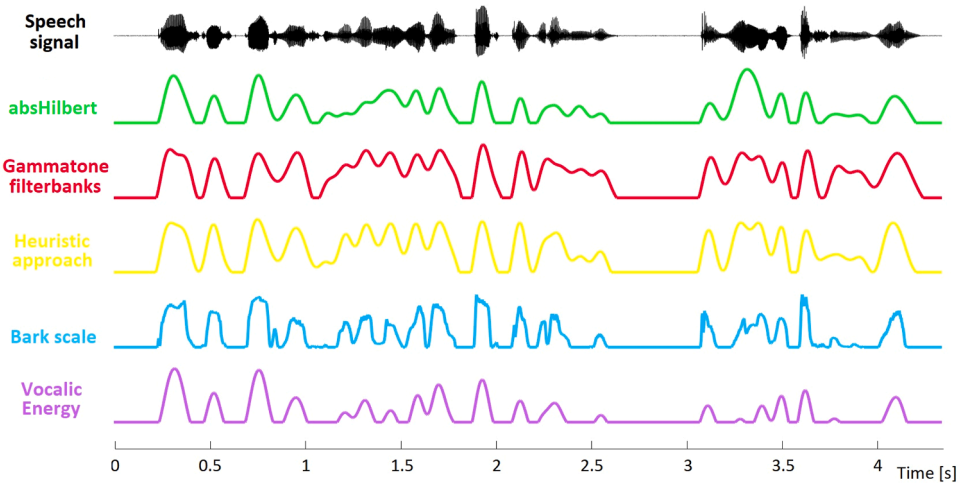


Fig. 1. Speech envelope extraction methods. As an example, we use a 4.5-second sentence. Speech signal (black) and envelopes (i.e., *absHilbert* - green line; *gammatone filterbanks* - red line; *heuristic approach* - yellow line; *bark scale* - blue line; *vocalic energy* - violet line) obtained through the various methods.

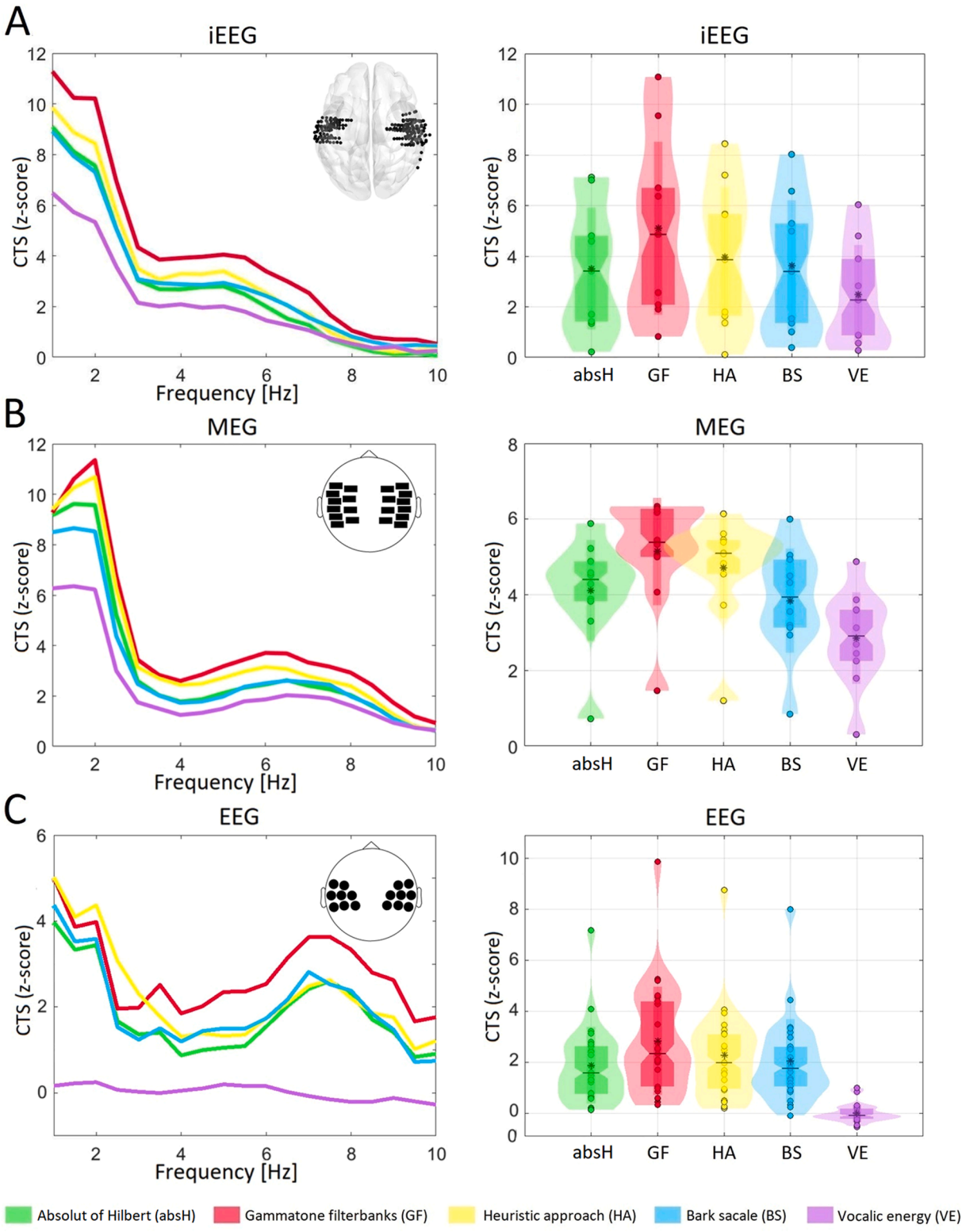


Fig. 2. CTS values for clear speech. On the left side of Figure A, B and C, we present the spectrogram (1 to 10 Hz frequency range) of the cortical tracking of speech (CTS) values (z-scored) for each technique (iEEG, MEG and EEG) and envelope extraction methods (i.e., *absHilbert* - green line; *gammatone filterbanks* - red line; *heuristic approach* - yellow line; *Bark scale* - blue line; *vocalic energy* - violet line) at sensors covering temporal brain regions. The spatial distribution of the sensors of interest is depicted for each technique. On the right side of Figure A, B, and C, boxplots were used to show the average of the CTS values (z-scored) in temporal regions within delta and theta bands. Boxplots are overlaid with individual data points. Each dot represents data from one participant. Boxes cover the 25 th to 75 th percentile (inter-quartile range; IQR). The middle of the box represents the median. Whiskers extend from the 25 th percentile and 75 th percentile to cover all data points lying within 1.5 times the IQR from the 25 th and 75 th percentile respectively.

compared to the *absHilbert* (t = 5.61, $p$bonf<0.001, *Cohen's* d = 0.54), *heuristic approach* (t = 3.39, $p$bonf<0.001, d = 0.33), *Bark scale* (t = 5.27, $p$bonf<0.001, d = 1.15), and *vocalic energy* methods (t = 11.86, $p$bonf<0.001, d = 0.51). Likewise, the *absHilbert* (t = 6.24, $p$bonf<0.001, d = 0.61), the *heuristic approach* (t = 8.46, $p$bonf<0.001, d = 0.82) and the *Bark scale* (t = 6.59, $p$bonf<0.001, d = 0.64) demonstrated significantly stronger CTS compared to *vocalic energy*. We also observed a main effect of Neuroimaging technique (F(2,41)=4.14, p < 0.02, η2=0.13). Post hoc tests showed that CTS values were higher for the iEEG compared to the EEG technique (t = 2.86, $p$bonf=0.02, d = 1.03).

The results also reveal a significant interaction between the Envelope method and the Neuroimaging technique (F(8164)=4.18, p < 0.001, η2=0.02). Fig. 3 shows the statistical results (t-values) comparing CTS values for all possible pairs of methods within each neuroimaging technique. In iEEG, the *gammatone filterbanks* method showed notably stronger CTS compared to the *absHilbert* (t = 4.43, $p$bonf<0.001, d = 0.86), *Bark scale* (t = 4.14, $p$bonf<0.001, d = 0.80), and *vocalic energy* (t = 7.28, $p$bonf<0.001, d = 1.41). Additionally, the *heuristic approach* displayed higher CTS than the *vocal energy* (t = 4.12, $p$bonf<0.001, d = 0.80). Noteworthy differences were observed in EEG, with the *gammatone filterbanks* method showing significantly stronger CTS compared to the *absHilbert* (t = 4.46, $p$bonf<0.001, d = 0.52), *Bark scale* (t = 3.68, $p$bonf=0.03, d = 0.43), and *vocalic energy* (t = 13.22, $p$bonf<0.001, d = 1.53) methods. Similarly, the *absHilbert* (t = 8.75, $p$bonf<0.001, d = 1.02), *heuristic approach* (t = 10.60, $p$bonf<0.001, d = 1.23), and *Bark scale* (t = 9.54, $p$bonf<0.001, d = 1.11) demonstrated stronger CTS compared to *vocalic energy*. In MEG, no significant differences were detected in CTS values among different envelope extraction methods (all *ts* < 2.76, $p$bonf>0.47, d < 0.48).

Speech in noise

The mean intelligibility and mean listening effort scores are shown in Fig. 4. A one-way repeated measure ANOVA was conducted to analyze potential differences across SNRs in the scores for speech intelligibility and listening effort. The findings revealed a significant main effect of SNR on both intelligibility (F(3,72)=193.07, p < 0.001, η2=0.89) and listening effort (F(3,72)=94.54, p < 0.001, η2=0.80) scores. For the intelligibility scores, with the exception of the comparison between 0 dB and 3 dB (t = 1.21, $p$bonf = 1, d = 0.30), all other comparisons revealed significant differences (all *ts* > 5.33, all $p$bonf*s* < 0.01, all *ds* < 1.31), indicating higher intelligibility scores for higher SNR levels. For the listening effort scores, all comparisons showed statistically significant differences (all *ts* < 4.98, all $p$bonf*s* < 0.01, all *ds*<−0.99), indicating lower listening effort scores for higher SNR levels.

The CTS was obtained for each SNR level and envelope extraction method across all participants (N_{EEG} = 25) within the frequency range 0 - 10 Hz (left side in Fig. 5). Similar to clear speech, the CTS spectrum reveals two prominent peaks within the delta and theta frequency bands, irrespective of the SNR level or the method of envelope extraction. Subsequently, mean CTS values were computed within temporal sensors and across delta and theta bands for each SNR level for further analysis (right side in Fig. 5 and Supplementary Table 2).

The ANOVA analysis (Envelope method x SNR level) of the mean CTS values revealed a significant main effect of Envelope method (F(4,96)= 37.34, p < 0.001, η2=0.33). Post hoc comparisons demonstrated that the *gammatone filterbanks* method exhibited notably higher CTS scores compared to the *absHilbert* (t = 3.08, $p$bonf=0.03, d = 0.35), *heuristic approach* (t = 4.56, $p$bonf<0.001, d = 0.52), Bark scale (t = 9.44, $p$bonf<0.001, d = 1.08), and *vocalic energy* methods (t = 10.21, $p$bonf<0.001, d = 1.17). Additionally, the *absHilbert* method showed significantly greater CTS scores compared to both the *Bark scale* (t = 6.37, $p$bonf<0.001, d = 0.73) and the *vocalic energy* (t = 7.13, $p$bonf<0.001, d = 0.82) methods. Similarly, the *heuristic approach* also yielded significantly higher CTS scores compared to both the *Bark scale* (t = 4.88, $p$bonf<0.001, d = 0.56) and the *vocalic energy*(t = 5.64, $p$bonf<0.001, d = 0.65) methods. No other main effects or interactions were observed on the CTS values. Furthermore, no significant correlations were found between the behavioral scores (i.e., intelligibility and listening effort scores) are the CTS values for each SNR level individually, nor when combining all SNR conditions (all *rs* < 0.2, all *ps* > 0.32) (Fig. 6).

4. Discussion

Our study investigates the relationship between speech envelope extraction methods and their impact on the cortical tracking of speech (CTS). We employed iEEG, MEG, and EEG to systematically evaluate five standard envelope extraction methods (*absHilbert, gammatone filterbanks, heuristic approach, Bark scale* and *vocalic energy*) on the CTS, calculated by the weighted pairwise phase consistency (WPPC, Vinck et al., 2010). Overall, we observed a strong correlation in the temporal characteristics of the speech envelopes across methods (all *rs* > 0.69, all *ps* < 0.01). Interestingly, results from different experiments reveal that the selection of the envelope extraction method significantly affects the CTS, unveiling a notable preference for the *gammatone filterbanks*, followed by the *heuristic approach*, the *absHilbert*, the *Bark scale*, and finally, *vocalic energy*.

In the analysis of clear speech, the superior performance of the *gammatone filterbanks* method is evident in both iEEG and EEG datasets. In both techniques, the *gammatone filterbanks* method exhibited significantly higher CTS values compared to the *absHilbert, Bark scale*, and *vocalic energy* methods. Conversely, in the MEG dataset, while a similar trend was observed, no statistically significant differences in CTS values were detected among the methods. This discrepancy in outcomes between MEG and other neuroimaging techniques may be attributed to the smaller sample size (N = 10) in the MEG dataset compared to EEG (N = 25), as well as the inherently lower signal-to-noise ratio in MEG recordings compared to iEEG. For the speech in noise experiment, CTS

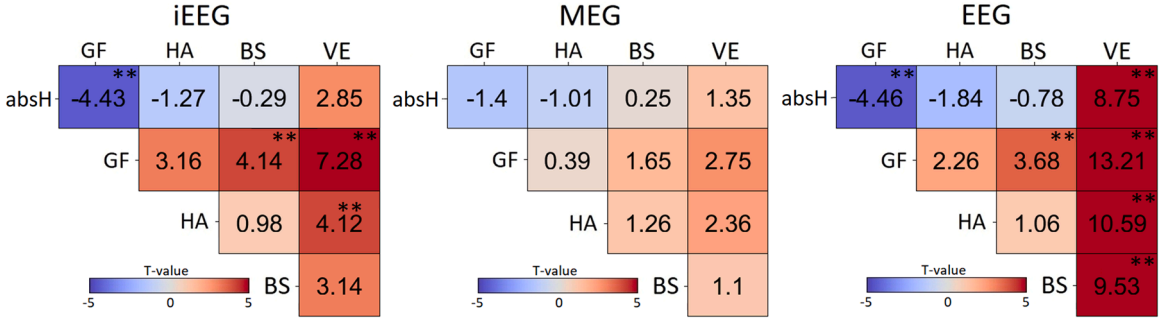


Fig. 3. Statistical comparison of CTS for clear speech. For each neuroimaging technique (iEEG, EEG and MEG), CTS values were compared between amplitude extraction methods (*absH, absolute of Hilbert transform; GF, gammatone filterbanks; HA, heuristic approach; BS, Bark scale; VE, vocalica energy*) using T-tests.For each comparison, the T-values and significant ("**" indicates p-value of <0.001, Bonferroni corrected) differences are indicated using an asterisk.

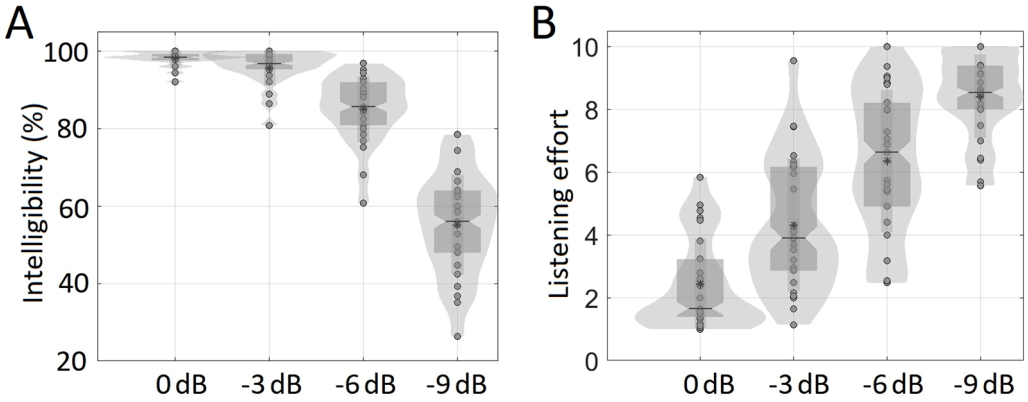


Fig. 4. Behavioral responses for speech in noise. (A) The intelligibility of scores increases with increasing SNR. (B) The listening effort score decreases with increasing SNR. Boxplots are overlaid with individual data points. Each dot represents data from one participant. Boxes cover the 25 th to 75 th percentile (interquartile range; IQR). The middle of the box represents the median. Whiskers extend from the 25 th percentile and 75 th percentile to cover all data points lying within 1.5 times the IQR from the 25 th and 75 th percentile respectively.

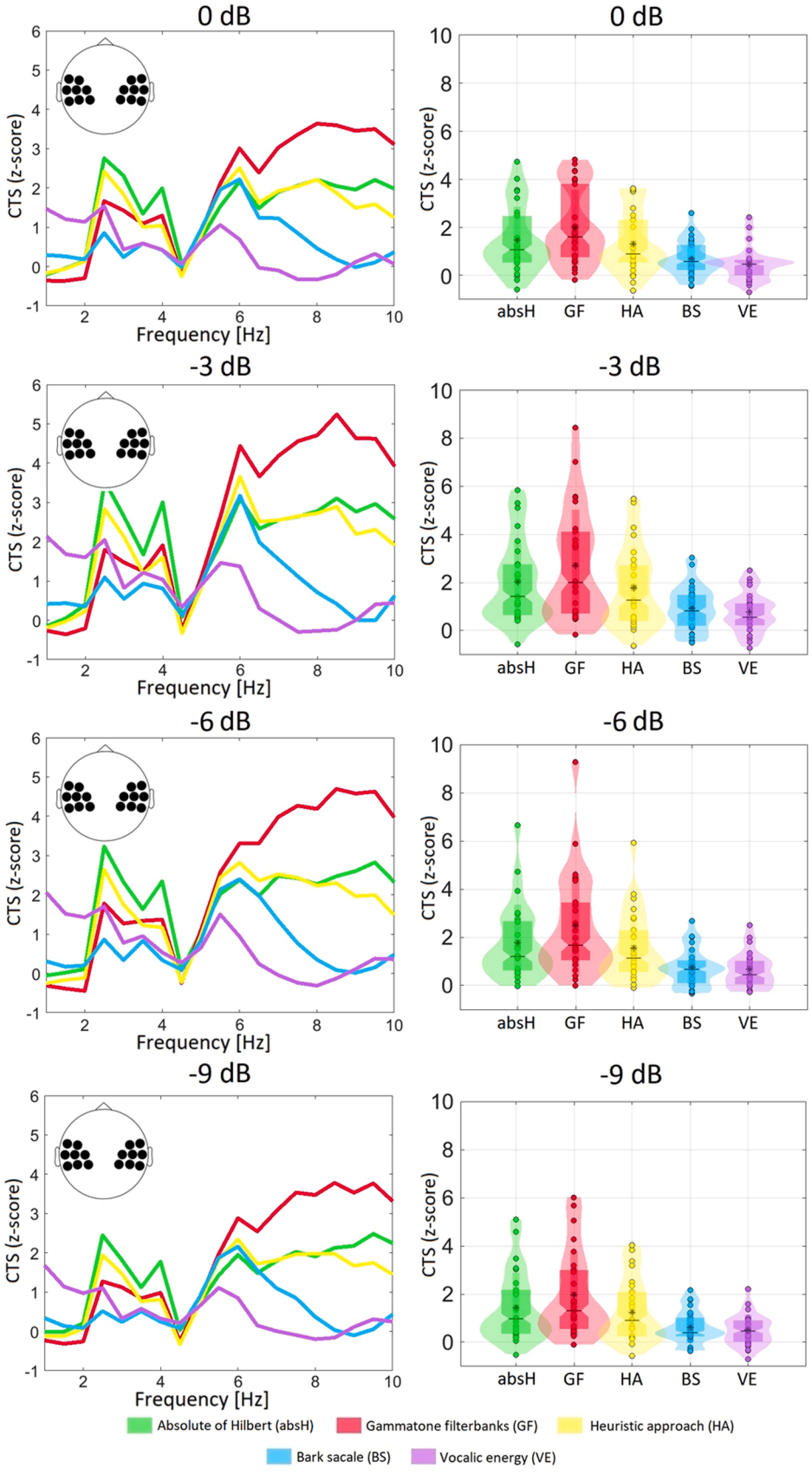
values consistently favored the *gammatone filterbanks* method over other envelope extraction methods, regardless of SNR level (e.g., 0, −3, −6 and −9 dB). Furthermore, another noteworthy point is that we did not observe a relationship between the levels of intelligibility and listening effort associated with different SNR conditions and the corresponding CTS values. The relationship between CTS strength and intelligibility remains a subject of debate. Some studies have indicated higher CTS for intelligible speech compared to unintelligible speech, while others have found no relationship or even an opposite effect (Kösem and Van Wassenhove, 2017). Further research is needed to shed light on this matter, and we believe that our findings will contribute to better characterizing this relationship.

Our findings align with those of Biesmans et al., 2017, who conducted a comparative analysis of envelope extraction methods in the context of auditory attention decoding using EEG. Their study revealed that the utilization of "auditory-inspired modifications," such as *gammatone filterbanks*, led to an improvement in classification performance when compared to more straightforward methods like half-wave rectification (e.g., Dellwo, Leemann and Kolly, 2015; Kolly and Dellwo, 2014) or the *absHilbert* (e.g., Gervain and Geffen, 2019). The utilization of the *gammatone filterbanks* in speech envelope extraction offers several notable benefits grounded in a biological-based approach. Inspired by cochlear mechanics, this filterbank demonstrates efficacy in capturing the intricate details of speech signals, mirroring the frequency selectivity observed in the auditory system (Irino and Patterson, 1997; Lyon, 2017). Its physiological relevance enhances the representation of speech envelope information in the brain, providing a more realistic account of auditory processing (Huang and Avendano, 2002). The *heuristic approach* (Caetano and Rodet, 2011; Jarne, 2018) is the method that most closely resembles the performance achieved with *gammatone filterbanks*, although, in general, the values of CTS remain smaller. The temporal characteristics of the envelope obtained through the *heuristic approach* and the *gammatone filterbanks* exhibit a substantial overlap, as evidenced by a strong correlation ($r = 0.95$) between these two methods. While it is true that computing the envelope through the *gammatone filterbanks* method is more complex compared to the *heuristic approach*, this is justifiable considering the improvement in CTS. The *absHilbert* method for speech envelope extraction is probably the most widely employed approach in CTS studies (e.g., O'Sullivan et al., 2015; Assaneoet al., 2019; Braiman et al., 2018; Molinaro and Lizarazu, 2018), performed more inconsistently compared to the *gammatone filterbanks*. The utilization of the *absHilbert* for speech envelope extraction presents certain challenges and limitations. One notable issue is its sensitivity to noise, which can lead to inaccuracies in the extracted envelope (Schimmel, 1992). Additionally, the *absHilbert* may struggle with non-stationary signals, impacting its performance in scenarios where the speech characteristics vary over time (Rangayyan, 2001). Another concern is the potential introduction of phase distortions, especially in the presence of abrupt changes in the signal (Boashash, 1992). The *Bark scale*, a widely utilized tool in auditory signal processing, is subject to criticism when applied to speech envelope extraction. A notable limitation of the *Bark scale* is its potential oversimplification of critical bandwidths, raising concerns about its ability to accurately represent the frequency content in speech signals (Moore and Glasberg, 1983; Zwicker and Terhardt, 1980). The *Bark scale*'s design, which is based on psychoacoustic principles, may not provide the necessary granularity to accurately capture the intricate frequency characteristics present in speech (Greenwood, 1961). This limitation becomes particularly apparent in scenarios involving complex acoustic stimuli, such as speech, where precise frequency resolution is crucial for comprehensive analysis. Finally, we observe that the *vocalic energy* method yields the lowest CTS values. When comparing these values, for instance, with those obtained using the *gammatone filterbanks* method, they decrease by half. In fact, the temporal correlation between the envelope obtained with the vocalic energy method and the *gammatone filterbank* ($r = 0.72$) or the *heuristic approach* ($r = 0.69$) is the lowest.

While the application of the *absHilbert*, the *heuristic approach*, and the *vocalic energy* is straightforward, *gammatone filterbanks* and the *Bark scale* are more complex and depend on numerous parameters (e.g., number of filters, spacing, order, weighting assigned to each filter), which can potentially influence a study. In this case, we strictly followed the methodology employed by Biesman et al. for the design of the *gammatone filterbanks* and by Oganian and Chang, 2019, for the *Bark scale* filters. Investigating whether alternative filter parameters modify the classification of methods based on performance in CTS evaluation would be highly interesting.

Our analysis revealed that the CTS is significantly affected by the speech envelope extraction method. Among all the methods investigated, *gammatone filterbanks* consistently exhibit superior performance in CTS estimation. The choice of a specific method becomes particularly significant when dealing with noisy neurophysiological data, as is often the case in studies involving children and the elderly, or when the data length for CTS calculation is limited.

Our study strategically employed varying stimuli lengths, types, and languages across different recording methods (iEEG versus MEG/EEG) to robustly test the method. This diversity in experimental conditions is a deliberate design choice that highlights the strength of our research. It allows us to consistently demonstrate that the effects on cortical tracking are stable across various auditory contexts and diverse subject groups, underscoring the reliability and wide applicability of our findings.



(caption on next page)

Fig. 5. CTS values for speech in noise. On the left side, we present the spectrogram of the CTS values for each SNL level (0 dB, −3 dB, −6 dB and −9 dB) and envelope extraction methods. The spatial distribution of the sensors of interest is also included. On the right side, boxplots were used to show the average of the CTS values (z-scored) in temporal regions within delta and theta bands. Boxplots are overlaid with individual data points. Each dot represents data from one participant. Boxes cover the 25 th to 75 th percentile (inter-quartile range; IQR). The middle of the box represents the median. Whiskers extend from the 25 th percentile and 75 th percentile to cover all data points lying within 1.5 times the IQR from the 25 th and 75 th percentile respectively.

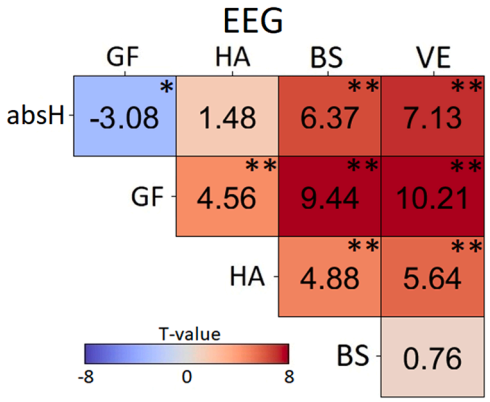


Fig. 6. Statistical comparison for speech in noise. CTS values were compared between amplitude extraction methods (*absHilbert, absolute of Hilbert transform; GF, gammatone filterbanks; HA, heuristic approach; BS, Bark scale; VE, vocalica energy*) across all SNR conditions using T-tests. For each comparison, T-values are reported, and significant differences are indicated using an asterisk ("*" indicates p-value of <0.05, "**" indicates p-value of <0.001, Bonferroni corrected).

Furthermore, our analysis incorporated a comprehensive range of datasets, including EEG, MEG, and intracranial EEG, as well as datasets with clear speech and EEG data at different background noise levels. These varied datasets reinforce the generalizability of our conclusions, showcasing the robustness of our approach in diverse experimental settings.

Among the notable limitations of our study, the constrained number of participants, particularly in the iEEG and MEG techniques, stands out prominently,and the lack of audiometric data for iEEG subjects. Future studies could benefit from recruiting a larger, more diverse group and including audiometric testing for all participants. Expanding the sample size holds promise for enhancing alignment among the results derived from the various techniques employed. Another significant consideration is the choice of measure utilized to evaluate Cortical Tracking of Speech (CTS); we selected weighted pairwise phase consistency (WPPC) due to its optimization of CTS, as demonstrated by Gross et al. (2013). Each connectivity method exhibits sensitivity to distinct properties of EEG and speech signals, thus warranting investigation into whether a consistent hierarchy in envelope extraction methods persists across alternative approaches, such as phase-locking value, Gaussian-Copula mutual information, Rayleigh test, magnitude squared coherence, and entropy. Furthermore, our study primarily concentrated on analyzing CTS within temporal regions, a focus supported by previous research (e.g., Gross et al., 2013; Lizarazu et al., 2021c). However, evidence suggests CTS extends to frontal regions like the inferior frontal gyrus (Molinaro et al., 2016), raising questions about generalizability to other brain regions. Lastly, exploring if the observed effects extend to other linguistic stimuli or languages would be intriguing. While our investigation centered on CTS in continuous speech (with or without noise), it remains uncertain if these effects replicate across stimuli such as words, syllables, or phonemes. Additionally, although our study encompassed Spanish, Danish, English, and French, examining the generalizability of effects to other languages would be advantageous.

To ensure transparency and reproducibility, our analysis scripts and the iEEG, MEG, and EEG data are openly accessible on the Open Science Framework (https://osf.io/gtsa5/). This availability enables the testing of novel envelope extraction methods or modifications to those presented here, allowing for an in-depth exploration of their performance in CTS analysis.

Author contribution

All stated authors have made a significant, direct, and intellectual contribution to the work and have given their approval for it to be published.

CRediT authorship contribution statement

Mohamed F. Issa: Writing – original draft, Software, Methodology, Formal analysis. **Izhar Khan:** Methodology, Formal analysis. **Manuela Ruzzoli:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Funding acquisition, Conceptualization. **Nicola Molinaro:** Writing – review & editing, Supervision, Resources, Project administration, Investigation, Funding acquisition, Conceptualization. **Mikel Lizarazu:** Writing – review & editing, Writing – original draft, Visualization, Validation, Resources, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

This research is supported by the Basque Government through the BERC 2022-2025 program and funded by the Spanish State Research Agency (AEI) through BCBL Severo Ochoa excellence accreditation CEX2020-001010/AEI/10.13039/501100011033. Additionally, it is supported by project PDC2022-133917-I00 funded by the Spanish Ministry of Science, Innovation, and Universities (MICIU) and the AEI. M.L. acknowledges funding support from the Ramón y Cajal program of the Spanish MICIU (grant RYC2022-035497-I) and the PIBA-2022-1-0015 from the Basque Government. MR is supported by MICIU and the AEI under the Ramón y Cajal program (RYC2019-027538-I/10.13039/501100011033), as well as the Basque Foundation for Science (Ikerbasque). NM was supported by the Spanish MICIU (grants RTI2018-096311-B-I00, PCI2022-135031-2, PID2022-136991NB-I00), the AEI, and the European Regional Development Fund (FEDER). The authors express their gratitude to Prof. Mathieu Bourguignon for sharing the MEG and EEG data at the CUB Hôpital Erasme.

Data availability

Our analytical scripts, along with the iEEG, EEG, and MEG datasets, are publicly available on the Open Science Framework at https://osf.io/gtsa5/.

Acknowledgments

This research is supported by the Basque Government through the BERC 2022-2025 program and funded by the Spanish State Research Agency (AEI) through BCBL Severo Ochoa excellence accreditation CEX2020-001010/AEI/10.13039/501100011033. Additionally, it is supported by project PDC2022-133917-I00 funded by the Spanish Ministry of Science, Innovation, and Universities (MICIU) and the AEI. M.L. acknowledges funding support from the Ramón y Cajal program of

the Spanish MICIU (grant RYC2022-035497-I) and the PIBA-2022-1-0015 from the Basque Government. MR is supported by MICIU and the AEI under the Ramón y Cajal program (RYC2019-027538-I/10.13039/501100011033), as well as the Basque Foundation for Science (Ikerbasque). NM was supported by the Spanish MICIU (grants RTI2018-096311-B-I00, PCI2022-135031-2, PID2022-136991NB-I00), the AEI, and the European Regional Development Fund (FEDER). The authors express their gratitude to Prof. Mathieu Bourguignon for sharing the MEG and EEG data at the CUB Hôpital Erasme.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2024.120675.

References

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

Andrzejak, R.G., Kraskov, A., Stögbauer, H., Mormann, F., Kreuz, T., 2003. Bivariate surrogate techniques: necessity, strengths, and caveats. Phys. Rev. E 68 (6), 066202.

Assaneo, M.F., Rimmele, J.M., Orpella, J., Ripollés, P., de Diego-Balaguer, R., Poeppel, D., 2019. The lateralization of speech-brain coupling is differentially modulated by intrinsic auditory and top-down mechanisms. Front. Integr. Neurosci. 13, 28.

Bastos, A.M., Schoffelen, J.M., 2016. A tutorial review of functional connectivity analysis methods and their interpretational pitfalls. Front. Syst. Neurosci. 9, 175.

Bertels, J., Niesen, M., Destoky, F., Coolen, T., Vander Ghinst, M., Wens, V., Bourguignon, M., 2023. Neurodevelopmental oscillatory basis of speech processing in noise. Dev. Cogn. Neurosci. 59, 101181.

Biesmans, W., Das, N., Francart, T., Bertrand, A., 2017. Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection in a cocktail party scenario. IEEE Transact. Neur. Syst. Rehabil. Eng. 25 (5), 402–412.

Boashash, B., 1992. Estimating and interpreting the instantaneous frequency of a signal. IEE Proceed. F (Rad. Signal Process.) 139 (4), 362–370.

Boersma, P. 2007. Praat: doing phonetics by computer. Retrieved from http://www.praat.org/.

Bourguignon, M., De Tiege, X., De Beeck, M.O., Ligot, N., Paquier, P., Van Bogaert, P., Jousmäki, V., 2013. The pace of prosodic phrasing couples the listener's cortex to the reader's voice. Hum. Brain. Mapp. 34 (2), 314–326.

Bourguignon, M., Molinaro, N., Lizarazu, M., Taulu, S., Jousmäki, V., Lallier, M., De Tiege, X., 2020. Neocortical activity tracks the hierarchical linguistic structures of self-produced speech during reading aloud. Neuroimage 216, 116788.

Braiman, C., Fridman, E.A., Conte, M.M., Voss, H.U., Reichenbach, C.S., Reichenbach, T., Schiff, N.D., 2018. Cortical response to the natural speech envelope correlates with neuroimaging evidence of cognition in severe brain injury. Curr. Biol. 28 (23), 3833–3839.

Caetano, M., Rodet, X., 2011. Improved estimation of the amplitude envelope of time-domain signals using true envelope cepstral smoothing. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4244–4247.

Darling, A.M., 1991. Properties and implementation of the gammatone filter: a tutorial. Speech Hearing and Language, Work in Progress. University College London, Department of Phonetics and Linguistics, pp. 43–61.

David, O., Cosmelli, D., Friston, K.J., 2004. Evaluation of different measures of functional connectivity using a neural mass model. Neuroimage 21 (2), 659–673.

Davis, T.S., Caston, R.M., Philip, B., Charlebois, C.M., Anderson, D.N., Weaver, K.E., Rolston, J.D., 2021. LeGUI: a fast and accurate graphical user interface for automated detection and anatomical localization of intracranial electrodes. Front. Neurosci. 15, 769872.

Decruy, L., Vanthornhout, J., Francart, T., 2020. Hearing impairment is associated with enhanced neural tracking of the speech envelope. Hear. Res. 393, 107961.

Dellwo, V., Leemann, A., Kolly, M.J., 2015. Rhythmic variability between speakers: articulatory, prosodic, and linguistic factors. J. Acoust. Soc. Am. 137 (3), 1513–1528.

... & Destoky, F., Philippe, M., Bertels, J., Verhasselt, M., Coquelet, N., Vander Ghinst, M., Bourguignon, M., 2019. Comparing the potential of MEG and EEG to uncover brain tracking of speech temporal envelope. Neuroimage 184, 201–213.

Dial, H.R., Gnanateja, G.N., Tessmer, R.S., Gorno-Tempini, M.L., Chandrasekaran, B., Henry, M.L., 2021. Cortical tracking of the speech envelope in logopenic variant primary progressive aphasia. Front. Hum. Neurosci. 14, 597694.

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

Doelling, K.B., Arnal, L.H., Ghitza, O., Poeppel, D., 2014. Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing. Neuroimage 85, 761–768.

Ershaid, H., Lizarazu, M., McLaughlin, D., Cooke, M., Simantiraki, O., Koutsogiannaki, M., Lallier, M., 2024. Contributions of listening effort and intelligibility to cortical tracking of speech in adverse listening conditions. Cortex 172, 54–71.

Fosler-Lussier, E., Dilley, L., Na'im, R., Pitt, M.A., 2007. The buckeye corpus of speech: updates and enhancements. In: Interspeech, pp. 934–937.

Friston, K., 2003. Learning and inference in the brain. Neural Netw. 16 (9), 1325–1352.

Gervain, J., Geffen, M.N., 2019. Efficient neural coding in auditory and speech perception. Trend. Neurosci. 42 (1), 56–65.

Gillis, M., Decruy, L., Vanthornhout, J., Francart, T., 2022. Hearing loss is associated with delayed neural responses to continuous speech. Eur. J. Neurosci. 55 (6), 1671–1690.

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

Gramfort, A., Luessi, M., Larson, E., Engemann, D.A., Strohmeier, D., Brodbeck, C., Hämäläinen, M., 2013. MEG and EEG data analysis with MNE-Python. Front. Neurosci. 267.

Greenwood, D.D., 1961. Critical bandwidth and the frequency coordinates of the basilar membrane. J. Acoust. Soc. Am. 33 (10), 1344–1356.

Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., Garrod, S., 2013. Speech rhythms and multiplexed oscillatory sensory coding in the human brain. PLoS Biol. 11 (12), e1001752.

Gross, J., Kluger, D.S., Abbasi, O., Chalas, N., Steingräber, N., Daube, C., Schoffelen, J. M., 2021. Comparison of undirected frequency-domain connectivity measures for cerebro-peripheral analysis. Neuroimage 245, 118660.

Hart, S.G., Staveland, L.E., 1988. Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In: Advances in Psychology, 52, pp. 139–183. North-Holland.

Henry, M.J., Herrmann, B., Kunke, D., Obleser, J., 2017. Aging affects the balance of neural entrainment and top-down neural modulation in the listening brain. Nat. Commun. 8 (1), 15801.

Hilbert, D., 1912. Begründung der kinetischen Gastheorie. Mathematische Annalen 72 (4), 562–577.

Horton, C., Srinivasan, R., D'Zmura, M, 2014. Envelope responses in single-trial EEG indicate attended speaker in a 'cocktail party. J. Neural. Eng. 11 (4), 046015.

Huang, J., Avendano, C., 2002. Speech enhancement based on perceptual wavelet packet thresholding. IEEE Transact. Speech Audio Process. 10 (6), 341–350.

Irino, T., Patterson, R.D., 1997. A time-domain, level-dependent auditory filter: the gammachirp. J. Acoust. Soc. Am. 101 (1), 412–419.

Jarne, C.G., 2018. A heuristic approach to obtain signal envelope with a simple software implementation. Asociación Física Argentina Anales AFA 29 (2), 51–57.

Kaganovich, N., Schumaker, J., Leonard, L.B., Gustafson, D., Macias, D., 2014. Children with a history of SLI show reduced sensitivity to audiovisual temporal asynchrony: an ERP study. J. Speech Lang. Hear. Res. 57 (4), 1480–1502.

Keitel, A., Gross, J., Kayser, C., 2018. Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. PLoS Biol. 16 (3), e2004473.

Kolly, M.J., Dellwo, V., 2014. Cues to linguistic origin: the contribution of speech temporal information to foreign accent recognition. J. Phon. 42, 12–23.

Kösem, A., Van Wassenhove, V., 2017. Distinct contributions of low-and high-frequency neural oscillations to speech comprehension. Lang. Cogn. Neurosci. 32 (5), 536–544.

Kösem, A., Bosker, H.R., Takashima, A., Meyer, A., Jensen, O., Hagoort, P., 2018. Neural entrainment determines the words we hear. Curr. Biol. 28 (18), 2867–2875.

Kreuz, T., Mormann, F., Andrzejak, R.G., Kraskov, A., Lehnertz, K., Grassberger, P., 2007. Measuring synchronization in coupled model systems: a comparison of different approaches. Phys. D: Nonlin. Phenom. 225 (1), 29–42.

Kries, J., De Clercq, P., Lemmens, R., Francart, T., Vandermosten, M., 2023. Acoustic and phonemic processing are impaired in individuals with aphasia. Sci. Rep. 13 (1), 11208.

Kurthen, I., Galbier, J., Jagoda, L., Neuschwander, P., Giroud, N., Meyer, M., 2021. Selective attention modulates neural envelope tracking of informationally masked speech in healthy older adults. Hum. Brain. Mapp. 42 (10), 3042–3057.

Lallier, M., Molinaro, N., Lizarazu, M., Bourguignon, M., Carreiras, M., 2017. Amodal atypical neural oscillatory activity in dyslexia: a cross-linguistic perspective. Clin. Psycholog. Sci. 5 (2), 379–401.

Lallier, M., Lizarazu, M., Molinaro, N., Bourguignon, M., Ríos-López, P., Carreiras, M., 2018. From auditory rhythm processing to grapheme-to-phoneme conversion: how neural oscillations can shed light on developmental dyslexia. Read. Dyslex.: From Bas. Funct. High. Order Cognit. 147–163.

Lancaster, G., Iatsenko, D., Pidde, A., Ticcinelli, V., Stefanovska, A., 2018. Surrogate data for hypothesis testing of physical systems. Phys Rep 748, 1–60.

Lizarazu, M., Lallier, M., Molinaro, N., Bourguignon, M., Paz-Alonso, P.M., Lerma-Usabiaga, G., Carreiras, M., 2015. Developmental evaluation of atypical auditory sampling in dyslexia: functional and structural evidence. Hum. Brain. Mapp. 36 (12), 4986–5002.

Lizarazu, M., Lallier, M., Bourguignon, M., Carreiras, M., Molinaro, N., 2021a. Impaired neural response to speech edges in dyslexia. Cortex 135, 207–218.

Lizarazu, M., di Covella, L.S., van Wassenhove, V., Rivière, D., Mizzi, R., Lehongre, K., Ramus, F., 2021b. Neural entrainment to speech and nonspeech in dyslexia: conceptual replication and extension of previous investigations. Cortex 137, 160–178.

Lizarazu, M., Carreiras, M., Bourguignon, M., Zarraga, A., Molinaro, N., 2021c. Language proficiency entails tuning cortical activity to second language speech. Cereb. Cort. 31 (8), 3820–3831.

Lizarazu, M., Carreiras, M., Molinaro, N., 2023. Theta-gamma phase-amplitude coupling in auditory cortex is modulated by language proficiency. Hum. Brain. Mapp. 44 (7), 2862–2872.

Lyon, R.F., 2017. Human and Machine hearing: Extracting Meaning from Sound. Cambridge University Press.

MacIntyre, A.D., Cai, C.Q., Scott, S.K., 2022. Pushing the envelope: evaluating speech rhythm with different envelope extraction techniques. J. Acoust. Soc. Am. 151 (3), 2002–2026.
Mai, A., Riès, S., Ben-Haim, S., Shih, J.J., Gentner, T.Q., 2024. Acoustic and language-specific sources for phonemic abstraction from speech. Nat. Commun. 15 (1), 677.
Menn, K.H., Michel, C., Meyer, L., Hoehl, S., Männel, C., 2022. Natural infant-directed speech facilitates neural tracking of prosody. Neuroimage 251, 118991.
Meyer, L., 2018. The neural oscillations of speech processing and language comprehension: state of the art and emerging mechanisms. Eur. J. Neurosci. 48 (7), 2609–2621.
Meyer, L., Gumbert, M., 2018. Synchronization of electrophysiological responses with speech benefits syntactic information processing. J. Cogn. Neurosci. 30 (8), 1066–1074.
Meyer, L., Sun, Y., Martin, A.E., 2020. Entraining" to speech, generating language? Lang. Cogn. Neurosci. 35 (9), 1138–1148.
Mohammadi, Y., Graversen, C., Østergaard, J., Andersen, O.K., Reichenbach, T., 2023. Phase-locking of neural activity to the envelope of speech in the delta frequency band reflects differences between word lists and sentences. J. Cogn. Neurosci. 35 (8), 1301–1311.
Molinaro, N., Lizarazu, M., Lallier, M., Bourguignon, M., Carreiras, M., 2016. Out-of-synchrony speech entrainment in developmental dyslexia. Hum. Brain. Mapp. 37 (8), 2767–2783.
Molinaro, N., Lizarazu, M., 2018. Delta (but not theta)-band cortical entrainment involves speech-specific processing. Eur. J. Neurosci. 48 (7), 2642–2650.
Molinaro, N., Lizarazu, M., Baldin, V., Pérez-Navarro, J., Lallier, M., Ríos-López, P., 2021. Speech-brain phase coupling is enhanced in low contextual semantic predictability conditions. Neuropsychologia 156, 107830.
Moore, B.C., Glasberg, B.R., 1983. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. J. Acoust. Soc. Am. 74 (3), 750–753.
Obleser, J., Kayser, C., 2019. Neural entrainment and attentional selection in the listening brain. Trend. Cogn. Sci. (Regul. Ed.) 23 (11), 913–926.
Oganian, Y., Chang, E.F., 2019. A speech envelope landmark for syllable encoding in human superior temporal gyrus. Sci. Adv. 5 (11), eaay6279.
Oostenveld, R., Fries, P., Maris, E., Schoffelen, J.M., 2011. FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Comput. Intell. Neurosci. 2011 (1), 156869.
Ortiz-Barajas, M.C., Guevara, R., Gervain, J., 2023. Neural oscillations and speech processing at birth. Iscience 26 (11).
... & O'sullivan, J.A., Power, A.J., Mesgarani, N., Rajaram, S., Foxe, J.J., Shinn-Cunningham, B.G., Lalor, E.C., 2015. Attentional selection in a cocktail party environment can be decoded from single-trial EEG. Cereb. Cort. 25 (7), 1697–1706.
Peelle, J.E., Davis, M.H., 2012. Neural oscillations carry speech rhythm through to comprehension. Front. Psychol. 3, 320.
Peelle, J.E., Gross, J., Davis, M.H., 2013. Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. Cereb. Cort. 23 (6), 1378–1387.
Peirce, C.S., 2009. *Writings of Charles S. Peirce: a chronological edition*, 8. Indiana University Press, pp. 1890–1892.
Quiroga, R.Q., Kraskov, A., Kreuz, T., Grassberger, P., 2002. Performance of different synchronization measures in real data: a case study on electroencephalographic signals. Phys. Rev. E 65 (4), 041903.
Quique, Y.M., Gnanateja, G.N., Dickey, M.W., Evans, W.S., Chandrasekaran, B., 2023. Examining cortical tracking of the speech envelope in post-stroke aphasia. Front. Hum. Neurosci. 17.
Rangayyan, R.M., 2001. Biomedical Signal analysis: a Case-Study Approach. IEEE Press.
Ríos-López, P., Molnar, M.T., Lizarazu, M., Lallier, M., 2017. The role of slow speech amplitude envelope for speech processing and reading development. Front. Psychol. 8, 1497.
Schimmel, H., 1992. Hilbert Transform applications in mechanical vibration. Mech. Syst. Signal. Process. 6 (3), 209–214.
Schreiber, T., Schmitz, A., 2000. Surrogate time series. Phys. D: Nonlin. Phenom. 142 (3–4), 346–382.
Schwarz, J., Lizarazu, M., Lallier, M., Klimovich-Gray, A., 2024. Phonological deficits in dyslexia impede lexical processing of spoken words: linking behavioural and MEG data. Cortex 171, 204–222.
Taulu, S., Simola, J., 2006. Spatiotemporal signal space separation method for rejecting nearby interference in MEG measurements. Phys. Med. Biol. 51 (7), 1759.
Tilsen, S., Arvaniti, A., 2013. Speech rhythm analysis with decomposition of the amplitude envelope: characterizing rhythmic patterns within and across languages. J. Acoust. Soc. Am. 134 (1), 628–639.
Vinck, M., van Wingerden, M., Womelsdorf, T., Fries, P., Pennartz, C.M., 2010. The pairwise phase consistency: a bias-free measure of rhythmic neuronal synchronization. Neuroimage 51 (1), 112–122.
Wagener, K., Josvassen, J.L., Ardenkjær, R., 2003. Design, optimization and evaluation of a Danish sentence test in noise: Diseño, optimización y evaluación de la prueba Danesa de frases en ruido. Int. J. Audiol. 42 (1), 10–17.
Xia, M., Wang, J., He, Y., 2013. BrainNet Viewer: a network visualization tool for human brain connectomics. PLoS ONE 8 (7), e68910.
Zoefel, B., VanRullen, R., 2015. The role of high-level processes for oscillatory phase entrainment to speech sound. Front. Hum. Neurosci. 9, 651.
Zwicker, E., Flottorp, G., Stevens, S.S., 1979. Critical band width in loudness summation. J. Acoust. Soc. Am. 65 (4), 819–821.
Zwicker, E., Terhardt, E., 1980. Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. J. Acoust. Soc. Am. 68 (5), 1523–1525.