NeuroImage 264 (2022) 119724

Contents lists available at ScienceDirect

NeuroImage

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

Speech listening entails neural encoding of invisible articulatory features

A. Pastore[a,b,*], A. Tomassini[a], I. Delis[c], E. Dolfini[a,b], L. Fadiga[a,b,1], A. D'Ausilio[a,b,1,*]

[a] Center for Translational Neurophysiology of Speech and Communication, Istituto Italiano di Tecnologia, Ferrara, Italy
[b] Department of Neuroscience and Rehabilitation, Università di Ferrara, Ferrara, Italy
[c] School of Biomedical Sciences, University of Leeds, Leeds, UK

ARTICLE INFO

Keywords:
Speech entrainment
Speech articulation
EEG
Partial information decomposition
Mutual Information
Audio-motor maps
Articulatory Synergies

ABSTRACT

Speech processing entails a complex interplay between bottom-up and top-down computations. The former is reflected in the neural entrainment to the quasi-rhythmic properties of speech acoustics while the latter is supposed to guide the selection of the most relevant input subspace. Top-down signals are believed to originate mainly from motor regions, yet similar activities have been shown to tune attentional cycles also for simpler, non-speech stimuli. Here we examined whether, during speech listening, the brain reconstructs articulatory patterns associated to speech production. We measured electroencephalographic (EEG) data while participants listened to sentences during the production of which articulatory kinematics of lips, jaws and tongue were also recorded (via Electro-Magnetic Articulography, EMA). We captured the patterns of articulatory coordination through Principal Component Analysis (PCA) and used Partial Information Decomposition (PID) to identify whether the speech envelope and each of the kinematic components provided unique, synergistic and/or redundant information regarding the EEG signals. Interestingly, tongue movements contain both unique as well as synergistic information with the envelope that are encoded in the listener's brain activity. This demonstrates that during speech listening the brain retrieves highly specific and unique motor information that is never accessible through vision, thus leveraging audio-motor maps that arise most likely from the acquisition of speech production during development.

1. Introduction

Verbal interaction is an essential part of human behavior and our brains are tuned to decode speech. Neural oscillations in the delta and theta range are believed to play a key role in shaping speech perception (Giraud and Poeppel, 2012; Meyer, 2018). Indeed, coupling of brain oscillatory activity to the quasi-rhythmic properties of speech, or speech neural entrainment (Obleser and Kayser, 2019), positively scales with speech intelligibility (Ghitza, 2012; Peelle et. al., 2013; Ding and Simon 2014; Kayser et al., 2015; Riecke et al., 2018) and is tightly related to speech comprehension performance (Ahissar et al., 2001; Luo and Poeppel, 2007; Peelle et. al., 2013; Gross et al., 2013; Ding and Simon, 2014). Importantly, rhythmic neural electrical stimulation causally modulates speech comprehension performance (Zoefel et al., 2018a; Riecke et al., 2018; Kösem et al., 2020) providing compelling evidence that the brain oscillatory activity is functionally relevant for speech processing (but see Zoefel et al. 2018b for a discussion on alternative interpretations of the neural entrainment phenomenon).

Brain entrainment to speech in the delta and theta band is suggested to increase comprehension via a facilitation of task-relevant information (Obleser and Kayser, 2019). The cocktail party effect (Cherry, 1953; Ding and Simon, 2012) is an example of this effect where selective attention translates into increased neural entrainment to the attended acoustic stream (Golumbic et al., 2012, 2013a; Kerlin et al., 2010; O'Sullivan et al., 2015; Vander Ghinst et al., 2016). Furthermore, if other speech-related cues are available, neural activity can also entrain to these signals. For instance, when acoustic intelligibility is compromised, oscillatory occipital activity couples to the periodicity of lip/face movements (Giordano et al., 2017; O'Sullivan et al., 2021; Park et al., 2016; Peelle and Sommers, 2015; Giordano et al., 2017).

Unsurprisingly, speech comprehension mostly benefits from visual cues in suboptimal listening conditions (Sumby and Pollack, 1954; Schroeder et al., 2008; Golumbic et al., 2013a). Neural entrainment to speech thus reflects top-down influences (Kösem et al., 2018; Di Liberto et al., 2018; Cope et al., 2017) which are driven by prior knowledge and/or context to predict the temporal structure of the heard stimuli (Calderone et al., 2014; Poeppel, 2003; Keitel, Gross and Kayser, 2018; Poeppel and Assaneo, 2020). One source of top-down modulation is located in the frontal lobe, as supported by the finding that oscillatory activity in the left inferior frontal cortex (between 1 and 3 Hz)

* Corresponding authors at: Center for Translational Neurophysiology of Speech and Communication (CTNSC@Unife), Italian Institute of Technology (IIT), Via Fossato di Mortara, 17-19, Ferrara 44121, Italy.
E-mail addresses: aldo.pastoretp@gmail.com (A. Pastore), dsllsn@unife.it (A. D'Ausilio).
[1] These authors equally contributed to this work.

https://doi.org/10.1016/j.neuroimage.2022.119724.
Received 10 June 2022; Received in revised form 28 September 2022; Accepted 30 October 2022
Available online 31 October 2022.

and motor cortex (between 4 and 8 Hz) modulates the phase of low-frequency activity in auditory areas (Park et al., 2015). This modulation may reflect a domain-general mechanism extending beyond speech processing with the motor system orchestrating sensory processing in time (Morillon and Baillet, 2017). Whether the motor system provides domain-general temporal predictions or richer domain-specific information about articulatory features, is however still unclear. Indeed, top-down motor influences may exploit action circuits to implement an internal 'simulation' of movements (Morillon et al., 2019; Arnal and Giraud 2012; Schubotz 2007).

To investigate this relevant question, we designed an EEG experiment where participants listened to auditorily presented sentences. The sentences were obtained from a publicly available dataset (Canevari et al., 2015) in which acoustic data is synchronized with articulatory data recorded via electromagnetic articulography (EMA). EMA uses miniaturized sensor coils placed on articulators (lips, jaws, tongue) to measure accurate position data with a high sampling frequency during speech production. Of key relevance to the current research is that the EMA provides the accurate description of speech articulators that is essential to uncover whether motor information contributes to the representation of speech in the listener's brain. To this end, we used the Partial Information Decomposition (PID) method (Williams and Beer 2010; Ince, 2017) that is designed to separate unique, redundant (shared), or synergistic (complementary) information provided by two source signals (here speech envelope and kinematic data) about a third target signal (here brain activity). We thus tested whether articulatory kinematics is encoded during listening and conveys information about speech that cannot be obtained from the speech envelope alone, i.e. unique neural information about kinematics or synergistic neural representation of speech envelope and kinematics (a better prediction of the neural response from both modalities simultaneously). Our hypothesis was that, if speech-related neural entrainment entails also a domain-specific motor process, entrainment to speech kinematics will be observed.

2. Methods

2.1. Participants

A total of 23 healthy naive volunteers were recruited for this study and were paid 30€ for their participation. All participants were native speakers of Italian, right-handed (by self-report) and had a normal or corrected-to-normal vision. One participant was excluded because of technical problems during data acquisition. Analysis was performed on data from the remaining 22 participants (13 females; age: 23.04 ± 3.44; MEAN±SD). Participants were informed about the experimental procedure and gave their written consent before participation. The experiment was approved by the local ethical committee "Comitato Etico Unico della Provincia di Ferrara" (approval N. 170592).

2.2. Stimuli

The stimuli were selected from the Multi-SPeaKing-style Articulatory corpus (MSPKA; Canevari et al., 2015) which comprises simultaneous recordings of audio and articulatory (lips, jaws and tongue) data of three mother-tongue speakers pronouncing sentences in Italian. Audio was recorded at a sampling rate of 22.05 kHz. Articulators were tracked at a sampling frequency of 400 Hz by means of an electromagnetic articulography system (EMA; NDI Wave, Northern Digital Instruments, Canada; Berry, 2011). The EMA data provides a very accurate characterization of mouth kinematics and it is commonly used in speech technology research (Savariaux et al., 2017). In the present study, we used data corresponding to x, y, and z positions of 7 sensor coils glued on the upper lip (UL), lower lip (LL), upper jaw (UJ), lower jaw (LJ), tongue tip (TT), tongue middle (TM) and tongue back (TB) (see Fig. 1 for a schematic illustration).

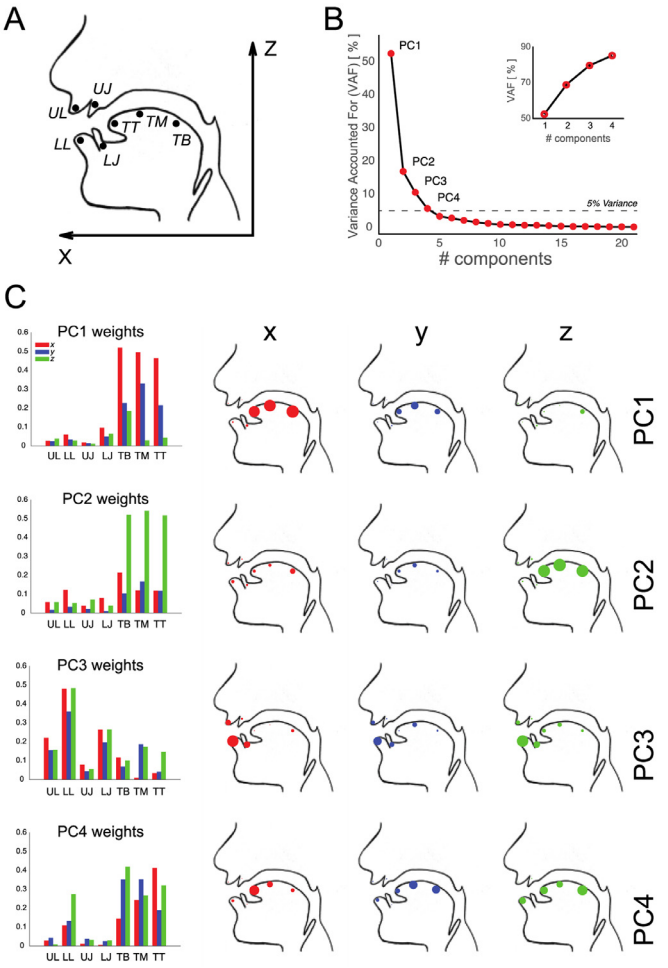

Fig. 1. Kinematic principal components. A. Schematic of the positions of the electromagnetic sensors: upper lip (UL), lower lip (LL), upper jaw (UJ), lower jaw (LJ), tongue tip (TT), tongue middle (TM), tongue back (TB). B. Variance Accounted For (VAF %) for all 21 principal components. Only components explaining at least 5% of variance were retained for further analyes (PC1, 2, 3, 4; cumulative variance in the inset). C. Bar plots represent the weights (absolute values) of each kinematic variable (x-, y- and z-axis for each sensor) for the PC1, 2, 3, 4. Dot size in the three vocal tract schematics show the relative contribution of each sensor across the movement axis (x, y and z respectively in red, blue and green).

For this study, we used 50 sentences (duration ranging from 6.2 to 9.4 s) pronounced by the same female speaker (referred to as "lls" within the dataset). The acoustic stimuli were manually checked and processed to remove any silent and/or noisy part at the beginning and at the end of the sentences. All acoustic stimuli were then normalized to the same average intensity (71 dB). Data corresponding to one sentence (out of 50) were discarded from the analysis because the corresponding EMA data turned out to be corrupted. During the experiment, participants were provided only with the acoustic stimuli. The corresponding EMA data were only used for data analysis (see below).

2.3. Experimental setup and procedure

Participants sat at a ~80 cm distance in front of an LCD monitor (VIEWPixx/EEG; 24", 120 Hz) with their right hand resting on a button-box (Cedrus RB-840 response Box). On each trial, participants were presented with a black fixation cross at the center of a uniformly gray screen; after a variable time (ranging between 0.1 and 1.1 s), a randomly selected sentence was presented acoustically via two loudspeak-

ers placed at ~20 cm from both sides of the screen. The fixation cross was removed after a variable time (between 0.1 and 1.1 s) from the end of the acoustic stimulus, and one word appeared at the center of the screen. The presented word rhymed 50% of the time with one of the words contained in the previously heard sentence. The word was selected to rhyme with any word in the sentence excluding the first, the last and all monosyllabic ones. Participants had to indicate whether the word rhymed or not by pressing one of two buttons, located few centimeters apart, using always the same finger (the right index). The rhyming task was included to encourage participants to listen attentively to the whole sentences. To avoid possible biases in the participants' responses, we ensured that rhyming and non-rhyming words were matched for number of syllables and their frequency of use in the Italian language by means of an online software tool (http://linguistica.sns.it/esploracolfis/home.htm). Different words were presented for each repetition of the same sentence (amounting to 4 words for each sentence, 200 words in total).

Every trial ended when participants provided their response in the rhyming task; trials were automatically ended if no response was provided within 10 s. Participants were asked to reduce blinks as much as possible and maintain their eyes on the fixation cross for the whole duration of the sentence.

The experiment consisted of four separate blocks of 50 trials each (200 trials in total), with short in-between breaks. The whole experiment lasted about 2 h, including the EEG cap mounting and preparation. Stimulus presentation and button-press acquisition were controlled via Matlab (The Math Works, Inc.; https://www.mathworks.com; RRID:SCR_001622) and the PsychToolbox-3 extensions (http://psychtoolbox.org; RRID:SCR_002881). All relevant events in the trial (e.g., trial start, stimulus onset, button press) were converted in a TTL by the VIEWPixx/EEG system to accurately synchronize them with the EEG data.

3. EEG recording and analyzes

EEG data were recorded continuously during the experiment with a 64-channel active electrode system (BrainAmp MR Plus, Brain Products GmbH, Gilching, Germany). Electrooculograms (EOGs) were recorded using 4 electrodes from the cap (FT9, FT10, PO9, and PO10) that were removed from their original scalp sites and placed bilaterally at the outer canthi and below and above the right eye to record horizontal and vertical eye movements, respectively. All electrodes were online referenced to the left mastoid. The impedance of the electrodes was kept below 10 kΩ. EEG signals were acquired at 1000 Hz.

Analyzes were performed within the Matlab and Python computing environments, using open-source toolboxes and libraries such as Fieldtrip (http://www.fieldtriptoolbox.org; RRID:SCR_004849) (Oostenveld et al., 2011), MNE (Gramfort et al., 2013) and PID library (https://github.com/robince/partial-info-decomp) as well as custom-made code. Analyzes were performed only on trials in which participants gave correct responses in the rhyming task (76.3 ± 7.5%; MEAN±SD; range 64.5–89.5%; See supplementary Fig. 1).

3.1. Speech envelope extraction

The amplitude envelope of the acoustic speech signals was calculated by adapting a previously described method (Smith et. al., 2002, Park et al., 2018). As in the Chimera toolbox (Smith et al., 2002), we defined 6 frequency bands in the range 80–8820 Hz that are equally spaced on the cochlear map. The speech signal was first filtered within those six frequency bands (MNE filter_data function, two-pass Butterworth filter, 4th order). Then, we computed the absolute value of the Hilbert transform for each bandpass-filtered signal. Finally, the speech envelope was obtained by summing up the result across all the frequency bands. The envelope was down-sampled to 400 Hz to match the sampling frequency of the EMA data.

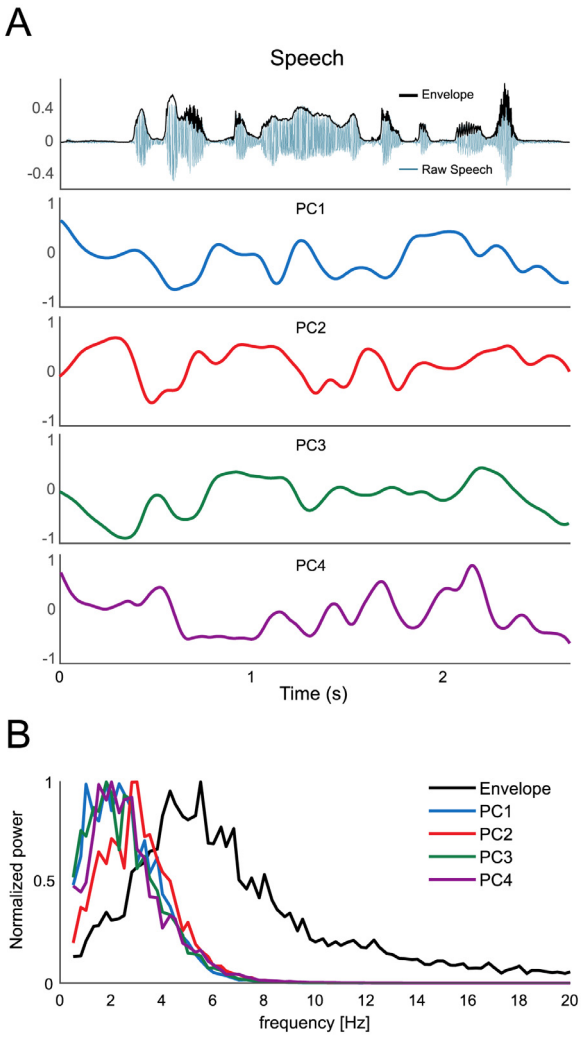

Fig. 2. Acoustic and kinematic stimuli features. A. Example time series of the raw speech signal (blue), its envelope (black) and the kinematic PCs corresponding to the same stimulus. B. Normalized (1/f) power spectra for all features (envelope, PC1, PC2, PC3 and PC4).

3.2. Kinematic features extraction

To capture meaningful speech coordination patterns in the high-dimensional EMA data (i.e., 7 sensors X 3 dimensions = 21 time series of position data) we used a dimensionality reduction technique. We applied Principal Component Analysis (PCA) as implemented in the Fieldtrip Toolbox (function: ft_componentanalysis; method: pca). PCA outputs feature activations over time (principal components [PCs], see Fig. 2A) that explain part of the variance in the EMA measurements and are orthogonal to each other. Furthermore, PCA provides information about the relative contribution of each kinematic feature (PC weights, see Fig. 1C) to the reconstruction of the EMA recordings in single trials. By visually inspecting (the absolute values of) the PC weights it is thus possible to assess the physiological validity of the articulatory coordination pattern identified by each PC.

3.3. EEG pre-processing

The continuous EEG data were band-pass filtered between 0.5 and 100 Hz (two-pass Butterworth filter, 4th order), and down-sampled to 400 Hz to match the sampling frequency of the EMA data. Data were then re-referenced to the common average and time-aligned to the acoustic stimulus onset (from -1 s to the duration of the longest

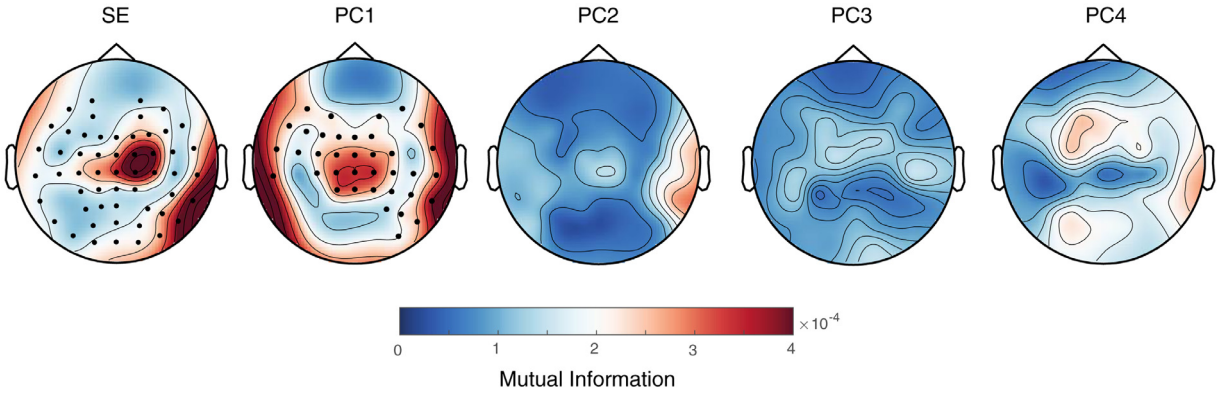

Fig. 3. MI results. Topographical distribution of across-subjects average information values computed via Gaussian Copula Mutual Information performed on the broad-band filtered data (0.5–10 Hz). SE: Speech Envelope; PCi: Principal Components 1 to 4. Black dots highlight the electrodes belonging to the cluster(s) that survived one-tailed cluster-based statistics against circularly shifted data (alpha level = 0.05; see also Methods).

sentence plus one second). Data were visually inspected, and noisy trials were removed. Independent component analysis (ICA) was then applied to identify and remove artifacts related to eye movements and heartbeat. Noisy channels (T8 for one subject) were excluded from the ICA analysis and substituted by linear interpolation of neighboring channels (after ICA-based artifact rejection). The total amount of trials retained for further analysis was 142 ± 21.7 (MEAN±SD).

3.4. Neural coupling to speech envelope and kinematic features

To quantify neural coupling to speech production, we used mutual information (MI), a measure of statistical dependence that captures any type of relationship (even non-linear and non-monotonic) between two signals (Shannon, 1948). Our aim here was to uncover the neural representations of the different kinematic components in the brain of the listener and quantify the contribution of these representations to the neural encoding of speech. To this end, we computed the MI between each recorded EEG signal and a) the speech envelope I(EEG;SE) and b) each one of the i=1,…,4 extracted PCs (kinematic components) I(EEG;PCi). Before computing MI, we first removed 1.5 s after sound onset for each trial to exclude stimulus-locked evoked potentials. This interval was selected to avoid any contamination from late event-related modulations and their slow return to the isoelectric. We then shifted the EEG signals forward in time by 0.2 s relative to the SE and PCs based on preliminary analyzes showing that such a delay maximized information between EEG and all the source signals (See supplementary Fig. 2). The time-shifting is consistent with the assumption that stimulus encoding necessarily follows stimulus presentation. Furthermore, an extensive literature has consistently shown that speech-brain coupling (entrainment) is maximal at comparable lags of about 0.2 s (i.e., for brain activities following auditory/visual speech by 0.2 s; e.g., Gross et al. 2013; O'Sullivan et al. 2015; Keitel et al. 2017; Di Liberto et al. 2018). More specifically, we cut the SE and PCs signals from +1.5 s relative to stimulus onset up to stimulus offset (variable length depending on stimulus duration) and the EEG signals from +1.7 s relative to stimulus onset up to +0.2 s after stimulus offset. Finally, all signals (EEG, SE, PCs) were padded (mirror padding) and then band-pass filtered between 0.5 and 10 Hz (two pass Butterworth, 2nd order). This relatively broad frequency range was set based on prior inspection of the power spectra of both acoustic (SE) and kinematic (PCs) signals, as it encompasses virtually all of their spectral content (see Fig. 3). The choice of the high cut-off frequency (10 Hz) is also consistent with evidence that coupling between brain activities and speech envelope is mostly confined to frequencies below the alpha range (Bröhl and Kayser, 2021). MI was then calculated using a recent implementation of the Gaussian Copula Mutual Information method which provides a lower bound of the actual MI and is robust to high-dimensional signals (Ince et al., 2017).

3.5. Partial information decomposition (PID)

We then focused on a) the contribution of each kinematic component and b) the interactions between speech and kinematic components to the neural encoding of speech. We thus employed Partial Information Decomposition (PID), a recent multivariate mathematical framework, originally proposed in Williams and Beer (2010), to quantify and characterize representational interactions in the human brain.

PID was performed using a recent modification of the original algorithms, which is based on common change in surprisal (Ince, 2017).

PID decomposes the mutual information between a target variable and a multivariate set of predictor variables, called sources (Timme et al., 2014). Indeed, if the sources are not statistically independent from the target, they will provide non-zero joint mutual information about the target which, in other terms, indexes the degree of dependence. PID allows then to disentangle this information, parcelling it out into information that is uniquely carried by each of the sources ('unique'), information that is shared by the sources ('redundant') and information that is accessible only when considering the two sources together ('synergistic').

Here, we considered the EEG measurement at each channel as the target signal. We run 4 different PID models, every time including as sources: (1) the speech envelope (SE; derived directly from the acoustic stimuli), and (2) one of the 4 kinematic features (a different one for each of the 4 models: PC1, 2, 3, 4; obtained from the EMA data through PCA; see above). The decomposition yielded 4 outcome terms:

- U_{SE} (EEG; SE): The unique information that the speech envelope carries about the EEG signal and cannot be obtained from the kinematic PCi.
- U_{PCi} (EEG; PCi): The unique information that the kinematic PCi carries about the EEG signal and cannot be obtained from the speech envelope.
- SYN_i (EEG; SE, PCi): The information that the joint observation of the two predictors {SE; PCi} provides about the EEG signal that cannot be obtained by observing each predictor separately.
- RED_i (EEG; SE, PCi): The information about EEG that is shared by the two sources, SE and PCi, thus reflecting a common neural representation of speech and kinematic component.

If the interaction between SE and PCi is redundant (RED_i), the information (about the EEG) that is carried by PCi can be obtained also from SE and vice versa. In other words, there will be no information loss if either the SE or the PCi is not available. In contrast, if the interaction is synergistic (SYN_i), neural information is encoded by the relationship between SE and PCi. In other words, we would obtain a better estimate of the EEG signal by considering SE and PCi together rather than independently. Finally, unique information (U) is carried by only one of

the two predictors. For example, a significant U_{PCi} would suggest that the corresponding brain response can only be predicted by that specific kinematic signal (PCi) and not by the speech envelope.

In a first PID analysis, all the signals were pre-processed in the same way as described above for MI (i.e., epoching, relative time-shifting of EEG data), including band-pass filtering between 0.5 and 10 Hz. After these preprocessing steps, signals for all trials were concatenated and copula normalized (Ince et al., 2017).

To further increase the granularity of our description, we explored the temporal dimension by running the PID analyzes across different lags (form -0.2 to 0.4 s, in 50-ms steps). We also performed a frequency-resolved PID analysis to evaluate whether the acoustic and kinematic features carry information at different spectral ranges. To this end, all signals were band-pass filtered by applying a sliding window along the frequency axis in the range between 0.5 and 10 Hz in steps of 0.5 Hz and with a frequency window length of 1 Hz. A separate PID was then applied for each band-pass filtered set of signals.

3.6. Statistical analysis

The output values obtained both in the MI and PID analysis were statistically evaluated against surrogate data. The original relationship between the two signals (for the MI) or between the target (the EEG activity) and the sources (the SE and the PCi) (for the PID) was destroyed without affecting the statistical properties of each signal, including its autocorrelation structure (Montemurro et al., 2007). More specifically, the EEG activity at each electrode and for each trial (epoched and bandpass filtered as for the original analysis) was circularly shifted by a number of samples that was randomly selected between N/4 and N-N/4, where N represent the number of samples of the shortest trial (i.e., 1892 samples). The time-shifted data were then submitted to the same processing steps as described above for the original data (i.e., trial concatenation, copula normalization) before applying the MI/PID algorithms. As for the original analysis (see above), MI was computed between the EEG and each SE/PC feature (I(EEG;SE) and I(EEG;PCi) with i=1,…,4). The same applies for the PID analysis whereby 4 separate PIDs were run by including as sources the SE and one of the 4 kinematic PCs. This procedure was repeated 1000 times yielding a surrogate distribution for each participant and each information component (I_{SE}, I_{PCi}; U_{SE}, U_{PCi}, $SYN_{SE\text{-}PCi}$, $RED_{SE\text{-}PCi}$). We then applied one-tailed cluster-based permutation statistics (Maris and Oostenveld, 2007) to test at the group level whether the original information values were larger than the mean of the surrogate distribution, i.e., the mean of the information values obtained for the circularly shifted data. In practice, all samples exceeding an a priori decided threshold (here $p < 0.05$, one-tailed) for univariate statistical testing (dependent-samples t-test) were selected and subsequently clustered on the basis of their contiguity along the spatial dimension. Cluster-level statistics was computed by taking the sum of t-values in each cluster. This sum was then used as test statistic and evaluated against the distribution of maximum cluster t-values obtained after permuting the original and circularly shifted data (at the level of participant-specific averages; 1000 permutations). The p-value was finally calculated as the proportion of random permutations yielding a larger test statistic compared with that computed for the non-permuted data. PID results were also statistically evaluated at the single-subject level by computing (separately for each subject and each electrode) the probability that the original information values exceeded the 95% of the distribution for the circularly shifted data. Resulting p-values were corrected for multiple comparisons across electrodes by controlling the False Discovery Rate (FDR; as described in Benjamini and Yekutieli (2001)).

4. Results

Neural entrainment to the speech envelope (Meyer, 2018; Giraud and Poeppel 2012; Keitel et al., 2018) – as well as the lips motion (Park et al., 2016; Giordano et al. 2017; Ozker et al., 2018) – are well-documented phenomena. However, only a fraction of speech articulation is available to vision while most speech-relevant information is in principle contained in hidden articulators (e.g. tongue movement). We here set out to investigate whether articulatory kinematics that is not visually available to the listener still conveys information about the produced speech that goes above and beyond that contained in the speech envelope. We recorded the EEG brain-wide activity while native-language participants were listening to acoustic speech stimuli taken from the Multi-SPeaKing-style Articulatory corpus (MSPKA; Canevari et al. 2015). This corpus contains simultaneous recordings of audio and kinematic data of the articulatory tract (measured via electromagnetic articulography; see methods) while speakers were pronouncing Italian sentences. The dimensionality of the articulatory data was reduced by means of PCA and the first 4 PCs accounting for most of the variance were selected for further analyzes to examine the relationship between the kinematics associated to speech production (in the speaker) and the listener's ongoing brain activity.

4.1. Articulatory features

The first 4 components derived from PCA explained most of the total variance of the kinematic data (85%; Fig. 1). Inspection of the components weights shows that the first 2 components (PC1 and PC2) represented almost entirely movements of the tongue on the antero-posterior (x-) and vertical (z-) axis, respectively (Fig. 1). Two of the movements that contribute significantly to articulation (Perrier et al., 2007), such as that of the tongue towards (and away from) the lips (PC1) or the palate (PC2), were thus automatically identified by PCA. Lower lip and jaw (again along the antero-posterior as well as vertical axes) as well as the tongue mainly contributed to PC3 which, despite explaining a smaller amount of variance (10%) compared to the tongue movements (PC1: 52%; PC2: 17%), appeared to capture another meaningful articulatory component, reflecting most likely mouth opening/closing, lip protrusion and lip-tongue coordination. Finally, a more composite mixture of articulators moving along multiple directions contribute to PC4 (6%), possibly reflecting complex tongue-lip movement synergies. The remaining components explained negligible amounts of variance (<5% each) and their articulatory interpretation was less straightforward; these components were thus excluded from further analysis.

Examples of the reconstructed time series for the 4 retained kinematic components along with the corresponding speech envelope are shown in Fig. 2. The analysis of their spectral content reveals that all the kinematic components show spectral concentration over a low frequency range between 1 and 4 Hz (delta band); the speech envelope instead, in line with previous evidence (Bröhl and Kayser 2021; Doellin et al., 2014; Gross et al., 2013; Luo and Poeppel, 2007; Peelle and Davis,2012; Bosker and Ghitza 2018), is marked by relatively higher frequencies, with a broad spectral peak between 4 and 8 Hz (theta band).

4.2. Neural encoding of speech envelope and articulatory features

Firstly, we evaluated whether the brain encodes the information contained in the speech envelope as well as in the hidden speech kinematics by computing the mutual information (MI; Shannon 1948; Ince et al. 2017) between the respective signal pairs (i.e., I(EEG;SE); I(EEG;PCi) with i=1,…,4). As expected, MI for the speech envelope is statistically significant (higher than that obtained for surrogate data; $p < 0.0001$, one-tailed cluster-based statistics; see Fig. 3 and Methods for details) and maximal in two foci overlaying central electrodes and, more laterally, right temporo-parietal electrodes. Such a topography is indeed very similar to what reported in previous works when quantifying neural speech entrainment with linear (Molinaro and Lizarazu 2018; Boucher et al., 2019) as well as non-linear (Kayser et al., 2015) coupling

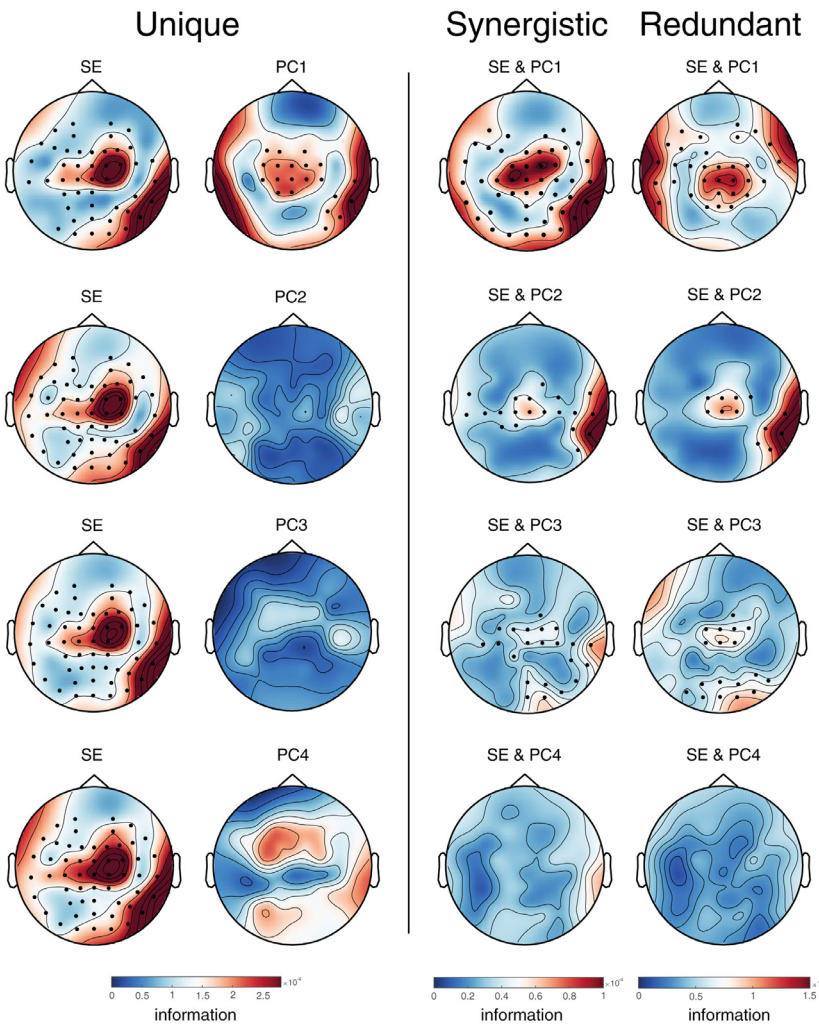

Fig. 4. PID results. Topographical distribution of across-subjects average information values obtained by PID analyzes performed on the broad-band filtered data (0.5–10 Hz). SE: Speech Envelope; PCi: Principal Components 1–4. Black dots highlight the electrodes belonging to the cluster(s) that survived one-tailed cluster-based statistics against circularly shifted data (alpha level= 0.05; see also Methods).

metrics. Remarkably, we found consistent MI also for one of the analyzed kinematic features, i.e., PC1, which mainly relates to the antero-posterior movement of the tongue ($p < 0.0001$). MI for PC1 – similar to what obtained for the speech envelope – largely increases over central electrodes and over right temporo-parietal electrodes (Fig. 3).

The speech envelope and the tongue movements (PC1) could, however, provide (brain-relevant) information that is either exclusive to each feature, fundamentally shared, or complementary between the two features. To disentangle among these different possibilities, we employed a computational approach known as Partial Information Decomposition (PID) (Williams and Beer, 2010). The speech envelope carries unique information in the 4 different PID models (Fig. 4, first column; $p < 0.0001$; one-tailed cluster-based statistics; see Methods for details). The topographic distribution of such activity is very similar across all PID models and closely resembles that already observed for the MI.

Most remarkably, passive listening also entails neural encoding of kinematic information that is not accounted for by the speech amplitude fluctuations, i.e., the SE. Specifically, the PC1 provides unique information (not carried by the SE; U_{PC1}) that is consistently represented in the listener's brain ($p = 0.002$; cluster-based statistics; see Fig. 4, first row). PC1 not only carries unique informational content but also interacts significantly with the acoustic information in a synergistic fashion; in other words, its combination with the SE leads to a net increase in information encoded in the listener's brain ($SYN_{SE\text{-}PC1}$; $p < 0.0001$; Fig. 4, third column). We also observe significant synergistic information between the SE and both PC2 ($p < 0.0001$) and PC3 ($p < 0.0001$). Indeed, PID enables to uncover representational interactions in the listener's brain that cannot be directly observable in pairwise measures of dependence (MI for PC2 and PC3), suggesting that the listener's brain may integrate articulatory information with SE in a super-additive way. Redundancy is then observed between SE and the first 3 PCs (PC1: $p < 0.0001$; PC2: $p = 0.029$ and $p = 0.045$ for the first and second significant cluster; PC3: $p = 0.007$ and $p = 0.012$, for the first and second significant cluster), highlighting the informational overlap that exists between articulation and speech output.

Overall, the same pattern of results is observed also when statistical evaluation is performed at the single-subject level, with a large proportion of participants showing significant U_{SE} and U_{PC1}. All PID systems resulted in 14 subjects having at least one significant channel for the SE unique information (U_{SE}) while PC1 unique information (U_{PC1}) is significant in 8 participants (supplementary Fig. 3). Very few subjects show significant unique information for the other kinematic components (0, 1 and 3 subjects for U_{PC2}, U_{PC3} and U_{PC4}, respectively). Synergistic information between SE and PC1 is significant in 14 subjects and fewer for the other components (10, 8 and 3 subjects for PC2, PC3 and PC4, respectively). A non-negligible number of participants also report significant redundant information between SE and the first three PCs (9, 8, 5 subjects, respectively), which is in agreement with corresponding group-level statistics.

The results reported above indicate that the brain encodes a certain amount of information carried by articulatory kinematics that cannot be equivalently extracted from the speech envelope. To explore the temporal dynamics of information encoding in the brain we repeated the PID analyzes by systematically varying the EEG lag with respect to all

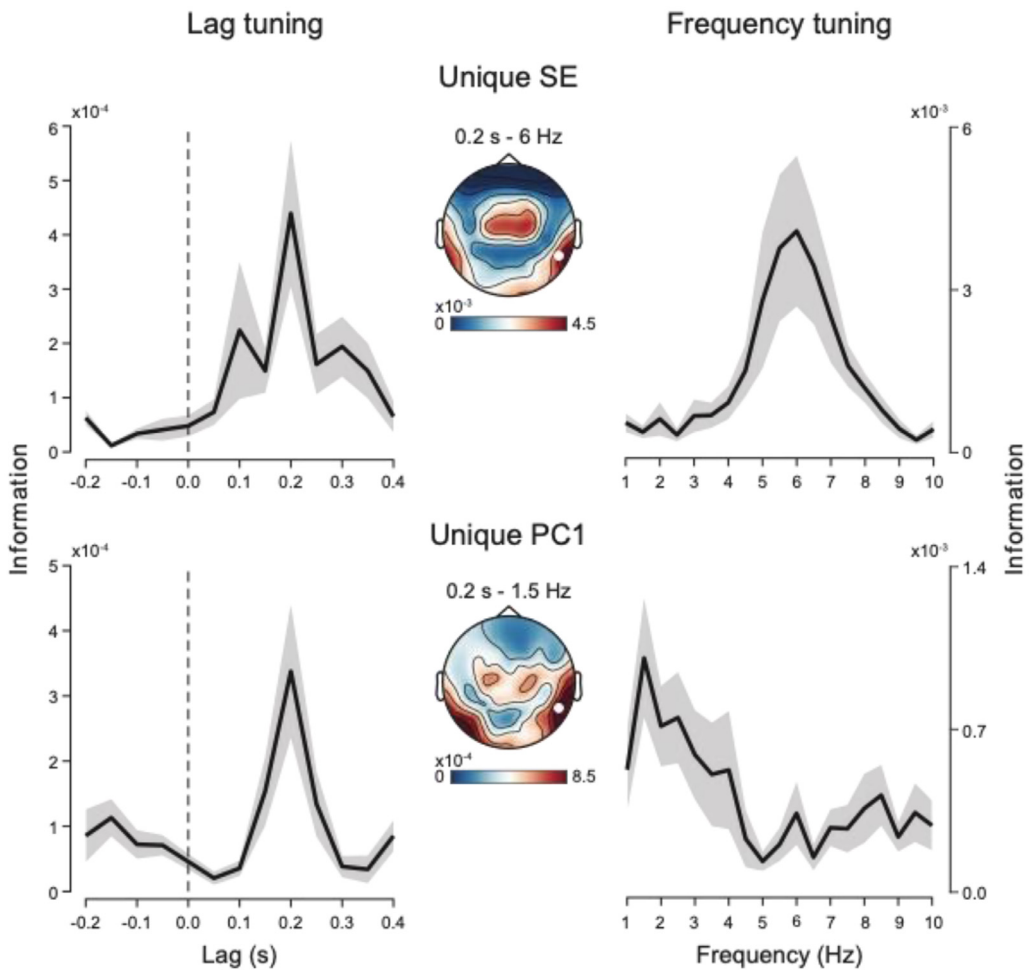

Fig. 5. Lag and frequency tuning for SE and PC1 Unique information. Group average of the unique information for speech envelope (top) and PC1 (bottom) as a function of lag (left) and frequency (right) for the electrode (TP8; marked in white in the topographies) showing maximal information in the broad-band analysis (see Fig. 4). The shaded areas represent ± standard errors of the mean (SEM). Topographies are shown for the frequency and lag where information is maximal.

sources. In line with results on MI (supplementary Fig. 2), maximal information for all the most relevant components (see Fig. 5 left column and supplementary Fig. 4) is obtained at ~0.2-s lag. To further gain insight into whether information conveyed by the kinematic and acoustic signals is band-limited and possibly concentrated within distinct frequency ranges, we repeated the PID analysis in a frequency-resolved way (all results are presented as supplementary Fig. 5). Fig. 5 (right column) shows the outcome of PID as a function of frequency for SE and PC1 for which we show significant unique information in the previous (broadband) analysis (see above). The information that is uniquely carried by the SE is clearly spectrally selective with maximal values being observed at ~5-6 Hz, i.e., in the theta band. A different spectral fingerprint, however, marks the information that is uniquely contained in the kinematics (PC1) which is enhanced within a lower frequency range between 1 and 4 Hz, i.e., in the delta band.

5. Discussion

Neural entrainment to speech originates from the integration of bottom-up processing and top-down projections from higher-order functional nodes in the brain to select and isolate particular signals of interest (Rimmele et al., 2018). Top-down projections, based on context and *a priori* knowledge, bias bottom-up sensory sampling via predictive models operating at multiple levels (e.g. Meyer 2018; Keitel et al. 2018). For instance, the neural computations run in the motor system may provide key top-down signals to isolate segmental or suprasegmental cues (Giraud and Poeppel, 2012). In fact, the sensory ambiguity characterizing the acoustic stream may partially be solved by unambiguous (or less ambiguous) endogenous signals (Meyer et al., 2020) arising from the inherent rhythmicity in speech articulation (Poeppel and Assaneo, 2020). Yet, for such a claim to be tenable, one should be able to find traces of articulatory signals in brain activities of speech-listening participants. Most importantly, neural activities should encode articulatory information in a way that is not trivially explained by the encoding of other tightly-coupled speech acoustic features (i.e. mouth opening and speech envelope; Chandrasekaran et al. 2009). Here we show that vocal tract configurations are encoded in the EEG signal and that they contain information that goes above and beyond the one carried by the speech envelope.

This result was obtained through the combination of a series of targeted novel approaches. First and foremost, participants listened to a set of acoustic speech stimuli for which synchronized articulatory data was available (Canevari et al., 2015). EMA data is characterized by a spatial and temporal resolution (Rebernik et al., 2021) that is not otherwise achievable by other technologies for speech kinematic analysis (i.e. Ultrasound or MRI). Moreover, as it is customary in the field of motor control (Ting, 2007), we reduced the dimensionality of the data (i.e. PCA) to derive a tractable number of signals explaining most of the variance (Lambert-Shirzad and Van del Loos, 2017). Yet, this choice has also key theoretical implications since data reduction techniques may be used to extract physiologically meaningful data-driven patterns of coordination across articulators (i.e. vocal tract synergies), which have far

greater functional relevance than the raw time-varying spatial positions of isolated articulators (Story, 2005; Sorensen et al., 2019). Secondly, we used a mathematical framework (Partial Information Decomposition – PID; Williams and Beer 2010; Ince 2017) that captures complex nonlinear relations among variables and decomposes these relations into atoms of information between the target (i.e. EEG data) and the sources (i.e. speech envelope and speech kinematics principal components). The PID can indeed extract unique, synergistic or redundant information contained in the sources (Park et al., 2018; Daube et al., 2019; Delis et al., 2022).

As expected from the abundant literature on neural entrainment to speech, the PID analysis highlights that the speech envelope contains unique information encoded in the EEG data. The topographical distribution of this effect matches the one normally observed with other analytical methods, with the involvements of central and right temporo-parietal electrodes (Ding et al., 2017). At the same time, we found that the first kinematic principal component, reflecting the antero-posterior movement of the tongue, also carries unique information about the neural signals. The topography of this effect is confined to central and right temporo-parietal electrodes. Interestingly, the synergistic interaction between the speech envelope and the first, second and third kinematic components conveys additional information about the EEG data (PC2 represents vertical tongue movements while PC3 captures the coordination between mouth opening/closing and tongue motion). In these three cases, the topography similarly shows a distribution covering both central and right temporo-parietal electrodes. The PID analyzes also reported redundant information between the first three kinematic components and speech envelope, stemming from the obvious coupling existing between articulators motion and speech sound output.

Importantly, we found a spectral dissociation that emerged in the frequency-resolved PID analysis. Lower frequencies (from ~0.5 to 4 Hz) appear to be relevant for kinematics-specific information (unique information provided by the PC1), in agreement with previous evidence that entrainment in the delta-range originates from higher-level processes in frontal (Molinaro et al., 2016) and motor cortices in particular (Park et al., 2015; Morillon et al., 2019; Biau et al., 2022). Instead, higher frequencies (between 4 and 8 Hz) contain unique information carried by the speech envelope. Such a delta/theta dissociation is compatible with the idea that neural entrainment in the theta-band is associated to the phonetic features that are critical for speech recognition, while the delta range entrainment is more closely related to the perceived acoustic rhythm of speech (Ding and Simon, 2014; Meyer et al., 2020). Overall, our results offer new insight regarding the functional origin of the delta/theta dissociation observed in speech neural entrainment, especially regarding the contribution provided by domain-specific motor processes.

In fact, the goal of our study was to investigate if speech listening does entail neural coupling to highly granular speech kinematic information that is not readily available to the participants. In our study, participants listened to auditory speech signals and were never – explicitly nor implicitly – exposed to the articulatory side of speech. Recent studies showed that brain signals encode missing information such as acoustic features when only silent lip-reading is allowed (Hauswald et al., 2018; Bourguignon et al., 2020). In this case, the tight audio-visual contingencies experienced during early childhood (as well as throughout life; Chandrasekaran et al. 2009) offer a solid ground to explain these phenomena according to a Bayesian perspective and as the result of multimodal associative learning (van Wassenhove, 2013). In our case, kinematic data contain information that is neither available during the experiment nor ever visually accessible throughout life (i.e. tongue kinematics). It follows that neural coupling to unavailable information cannot be explained by the life-long learning of audio-visual associations (i.e. as is the case for lip motion). Rather, we advance the proposal that speech production learning must play a key ontogenetic role in explaining our results.

Neural coupling to tongue kinematics could still imply that (at least part of the) articulatory information is retrieved from speech acoustics. Recent investigations have looked for example at the temporal fine structure (TFS) of speech. The recent study by Teng et al. (2019) showed that TFS and SE are both coupled to brain signals in the same frequency band (3–6 Hz) and suggested that TFS can be exploited by the brain to reconstruct relevant temporal information when speech is distorted. Importantly though, SE and TFS appear to share the same informational content when it comes to temporal cues useful for speech segmentation (Teng et al., 2019). Notwithstanding the obvious fact that speech acoustics contains more information than that conveyed by the envelope, the mapping between such an acoustic information and the articulatory space is tremendously complex (Atal et al., 1978). During speech production, different phono-articulatory tract configurations produce the same acoustic target depending on phonetic context (i.e., coarticulation; Grimme et al. 2011). Known as the “many to one” mapping problem, it would force the brain to solve an ill-posed inverse problem when listening to speech (known as acoustic-to-articulatory inversion). As a consequence, how is it possible that information related to the coordination of an invisible articulator is retrieved (and potentially partially reconstructed) from speech acoustics?

An answer to this conundrum could be that, during development, the brain approximates a solution for this inverse problem, mapping intended acoustic targets back to vocal tract articulatory parameters to allow intelligible speech production (Guenther, 1995; Tourville and Guenther, 2011). Indeed, infants explore how sounds are produced by experimenting the full range of their vocal tract configurations (Bruderer et al., 2015; Kuhl et al., 2014). In support of this idea, automatic speech recognition models trained with both acoustic and articulatory data achieve better classification performance with far fewer examples than acoustic-only training regimes (King et al., 2007; Ghosh and Narayanan, 2011). These models recapitulate some key properties of speech production development (Canevari et al., 2013; Badino et al., 2014) and demonstrate that learning auditory-motor mappings grants more compact and efficient representations of speech acoustics (Badino et al., 2016). As a consequence of learning audio-motor contingencies, speech auditory processing should in principle be tuned to capture those cues that allow triggering of endogenously guided reconstruction of missing articulatory signals (Meyer et al., 2020). To date, regardless of how fine-grained articulatory information is mapped onto the acoustic space, there was no evidence that the brain encodes articulatory configurations whose relevance is functionally dependent on the acquisition of speech production - and thus reflecting an intrinsically domain specific process. However, a conclusive demonstration of this point will only come from future developmental research (i.e. by investigating the developmental trajectory of the auditory-based neural encoding of invisible articulators) or exploiting a multi-language approach (i.e. by studying similar phenomena in L2).

Yet, the idea of a tight functional relationship between speech production and speech perception is not a new concept (Pulvermuller and Fadiga, 2010; Fadiga et al., 2002; Watkins et al., 2003, Wilson et al., 2004, Pulvermüller et al., 2006; D’Ausilio et al., 2014; but see also the pioneering insight provided by other scientists such as A. M. Liberman, L. A. Chistovich or A. N. Leontiev). Indeed, transcranial magnetic stimulation of the motor system produce specific (Meister et al., 2007; Möttönen et al., 2009; Sato et al., 2009) and somatotopic effects on speech discrimination performance (D’Ausilio et al., 2009; Bartoli et al., 2015). A recent series of studies proposed a more detailed, oscillation-based mechanism through which the motor system could have an impact on speech perception. Assaneo and Poeppel (2018) found synchronized brain activity between motor and auditory areas during a syllable listening task and successfully modelled the speech motor cortex as an oscillator coupled to the auditory system. According to this model, neuronal oscillations observed in auditory and motor cortices indeed synchronize in a frequency range corresponding to the mean syllable rate across languages (~4.5 Hz). Endogenous signals from the motor system

would phase-reset neuronal oscillations in the auditory cortex to align the most excitable states to the occurrence of expected events and/or changes in the speech stream (Rimmele et al., 2018) with benefits on perceptual/comprehension performance (Assaneo et al., 2021).

Perception is an inherently noisy process and in order to cope with speech-intrinsic (talker-specific) and speech-extrinsic (environment-specific) noise (Ru et al., 2003) our brain needs to integrate and weigh multiple sources of information depending on their reliability (Golumbic et al., 2013a,b; Schroeder et al., 2008). In this regard, when the acoustic signal is corrupted, the increased importance of visual cues translates into stronger entrainment to lip movements (Giordano et al., 2017; O'Sullivan et al., 2021; Park et al., 2016, 2018; Peelle and Sommers, 2015). Here, we provide a demonstration that neural speech processing can draw inferences based on highly granular endogenous domain-specific motor signals whose relevance for perception necessarily derives from the acquisition of speech production capabilities. Substantial progress in our understanding could come by providing a clearer picture on: (1) the detailed acoustic encoding of the different articulatory synergies, (2) whether those acoustic cues overlap fully or partly with articulatory information, and (3) to which extent information is then truly synthesized via endogenous processes.

Funding

This work was supported by the BIAL Foundation – Grant for Scientific Research 2020 (No. 246/20) to A.T., Ministero della Salute, Ricerca Finalizzata 2016 – Giovani Ricercatori (GR-2016-02361008) and Ministero della Salute, Ricerca Finalizzata 2018 – Giovani Ricercatori (GR-2018-12366027) to A.D., Ministero della Ricerca (20208RB4N9) - PRIN 2020 - and the European Union H2020 – EnTimeMent (FETPROACT-824160) to L.F.

Data code availability statement

The data and code that support the findings of this study are openly available in Mendeley Data as DOI: 10.17632/svy9m6987n.1.

Declaration of Competing Interest

Authors declare no competing interest.

Credit authorship contribution statement

A. Pastore: Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft. **A. Tomassini:** Methodology, Writing – original draft, Writing – review & editing, Supervision. **I. Delis:** Methodology, Writing – review & editing, Supervision. **E. Dolfini:** Investigation. **L. Fadiga:** Conceptualization, Writing – review & editing, Funding acquisition. **A. D'Ausilio:** Conceptualization, Writing – review & editing, Visualization, Supervision, Project administration.

Acknowledgments

The authors would like to thank Leonardo Badino for the many fruitful discussions that motivated this study prior to its completion.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2022.119724.

References

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

Arnal, L.H., Giraud, A.L., 2012. Cortical oscillations and sensory predictions. Trends Cogn. Sci. 16 (7), 390–398.

Assaneo, M.F., Poeppel, D., 2018. The coupling between auditory and motor cortices is rate-restricted: evidence for an intrinsic speech-motor rhythm. Sci. Adv. 4 (2), eaao3842.

Assaneo, M.F., Rimmele, J.M., Perl, Y.S., Poeppel, D., 2021. Speaking rhythmically can shape hearing. Nat. Hum. Behav. 5 (1), 71–82.

Atal, B.S., Chang, J.J., Mathews, M.V., Tukey, J.W., 1978. Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. J. Acoust. Soc. Am. 63 (5), 1535–1555.

Badino, L., Canevari, C., Fadiga, L., Metta, G., 2016. Integrating articulatory data in deep neural network-based acoustic modeling. Comput. Speech Lang. 36, 173–195.

Badino, L., D'Ausilio, A., Fadiga, L., Metta, G., 2014. Computational validation of the motor contribution to speech perception. Top. Cogn. Sci. 6 (3), 461–475.

Bartoli, E., D'Ausilio, A., Berry, J., Badino, L., Bever, T., Fadiga, L., 2015. Listener–speaker perceived distance predicts the degree of motor contribution to speech perception. Cereb. Cortex 25 (2), 281–288.

Benjamini, Y., Yekutieli, D., 2001. The control of the false discovery rate in multiple testing under dependency. Ann. Stat. 1165–1188.

Biau, E., Schultz, B.G., Gunter, T.C., Kotz, S.A., 2022. Left motor delta oscillations reflect asynchrony detection in multisensory speech perception. J. Neurosci. 42 (11), 2313–2326.

Bosker, H.R., Ghitza, O., 2018. Entrained theta oscillations guide perception of subsequent speech: behavioural evidence from rate normalisation. Lang. Cogn. Neurosci. 33 (8), 955–967.

Boucher, V.J., Gilbert, A.C., Jemel, B., 2019. The role of low-frequency neural oscillations in speech processing: revisiting delta entrainment. J. Cogn. Neurosci. 31 (8), 1205–1215.

Bourguignon, M., Baart, M., Kapnoula, E.C., Molinaro, N., 2020. Lip-reading enables the brain to synthesize auditory features of unknown silent speech. J. Neurosci. 40 (5), 1053–1065.

Bröhl, F., Kayser, C., 2021. Delta/theta band EEG differentially tracks low and high frequency speech-derived envelopes. NeuroImage 233, 117958.

Bruderer, A.G., Danielson, D.K., Kandhadai, P., Werker, J.F., 2015. Sensorimotor influences on speech perception in infancy. Proc. Natl. Acad. Sci. 112 (44), 13531–13536.

Calderone, D.J., Lakatos, P., Butler, P.D., Castellanos, F.X., 2014. Entrainment of neural oscillations as a modifiable substrate of attention. Trends Cogn. Sci. 18 (6), 300–309.

Canevari, C., Badino, L., Fadiga, L., 2015. A new Italian dataset of parallel acoustic and articulatory data. In: Proceedings of the 16th Annual Conference of the International Speech Communication Association.

Canevari, C., Badino, L., D'Ausilio, A., Fadiga, L., Metta, G., 2013. Modeling speech imitation and ecological learning of auditory-motor maps. Front. Psychol. 4, 364.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., Ghazanfar, A.A., 2009. The natural statistics of audiovisual speech. PLoS Comput. Biol. 5 (7), e1000436.

Cherry, E.C., 1953. Some experiments on the recognition of speech, with one and with two ears. J. Acoust. Soc. Am. 25 (5), 975–979.

Cope, T.E., Sohoglu, E., Sedley, W., Patterson, K., Jones, P.S., Wiggins, J., Rowe, J.B., 2017. Evidence for causal top-down frontal contributions to predictive processes in speech perception. Nat. Commun. 8 (1), 1–16.

Daube, C., Ince, R.A., Gross, J., 2019. Simple acoustic features can explain phoneme-based predictions of cortical responses to speech. Curr. Biol. 29 (12), 1924–1937.

D'Ausilio, A, Bartoli, E, Maffongelli, L, Berry, J.J., Fadiga, L, 2014. Vision of tongue movements bias auditory speech perception. Neuropsychologia 63, 85–91.

D'Ausilio, A., Pulvermüller, F., Salmas, P., Bufalari, I., Begliomini, C., Fadiga, L., 2009. The motor somatotopy of speech perception. Curr. Biol. 19 (5), 381–385.

Delis, I., Ince, R.A., Sajda, P., Wang, Q., 2022. Neural encoding of active multi-sensing enhances perceptual decision-making via a synergistic cross-modal interaction. J. Neurosci. doi:10.1523/JNEUROSCI.0861-21.2022.

Di Liberto, G.M., Lalor, E.C., Millman, R.E., 2018. Causal cortical dynamics of a predictive enhancement of speech intelligibility. NeuroImage 166, 247–258.

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

Ding, N., Simon, J.Z., 2014. Cortical entrainment to continuous speech: functional roles and interpretations. Front. Hum. Neurosci. 8, 311.

Ding, N., Melloni, L., Yang, A., Wang, Y., Zhang, W., Poeppel, D., 2017. Characterizing neural entrainment to hierarchical linguistic units using electroencephalography (EEG). Front. Hum. Neurosci. 11, 481.

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.

Fadiga, L., Craighero, L., Buccino, G., Rizzolatti, G., 2002. Speech listening specifically modulates the excitability of tongue muscles: a TMS study. Eur. J. Neurosci. 15 (2), 399–402.

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

Giordano, B.L., Ince, R.A., Gross, J., Schyns, P.G., Panzeri, S., Kayser, C., 2017. Contributions of local speech encoding and functional connectivity to audio-visual speech perception. eLife 6, e24763.

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

Ghosh, P.K., Narayanan, S.S., 2011. Automatic speech recognition using articulatory features from subject-independent acoustic-to-articulatory inversion. J. Acoust. Soc. Am. 130 (4), EL251–EL257.

Golumbic, E.M.Z., Ding, N., Bickel, S., Lakatos, P., Schevon, C.A., McKhann, G.M., Schroeder, C.E., 2013a. Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party. Neuron 77 (5), 980–991.

Golumbic, E.M.Z., Poeppel, D., Schroeder, C.E., 2012. Temporal context in speech processing and attentional stream selection: a behavioral and neural perspective. Brain Lang. 122 (3), 151–161.
Golumbic, E.Z., Cogan, G.B., Schroeder, C.E., Poeppel, D., 2013b. Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party. J. Neurosci. 33 (4), 1417–1426.
Gramfort, A., et al., 2013. MEG and EEG data analysis with MNE-Python. Front. Neurosci. 267.
Grimme, B., Fuchs, S., Perrier, P., Schöner, G., 2011. Limb versus speech motor control: a conceptual review. Motor Control 15 (1), 5–33.
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.
Guenther, F.H., 1995. Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. Psychol. Rev. 102 (3), 594.
Hauswald, A., Lithari, C., Collignon, O., Leonardelli, E., Weisz, N., 2018. A visual cortical network for deriving phonological information from intelligible lip movements. Curr. Biol. 28 (9), 1453–1459.
Ince, R.A. (2017). The partial entropy decomposition: decomposing multivariate entropy and mutual information via pointwise common surprisal. arXiv preprint arXiv:1702.01591v2.
Ince, R.A., Giordano, B.L., Kayser, C., Rousselet, G.A., Gross, J., Schyns, P.G., 2017. A statistical framework for neuroimaging data analysis based on mutual information estimated via a gaussian copula. Hum. Brain Mapp. 38 (3), 1541–1573.
Kayser, S.J., Ince, R.A., Gross, J., Kayser, C., 2015. Irregular speech rate dissociates auditory cortical entrainment, evoked responses, and frontal alpha. J. Neurosci. 35 (44), 14691–14701.
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.
Keitel, A., Ince, R.A., Gross, J., Kayser, C., 2017. Auditory cortical delta-entrainment interacts with oscillatory power in multiple fronto-parietal networks. NeuroImage 147, 32–42.
Kerlin, J.R., Shahin, A.J., Miller, L.M., 2010. Attentional gain control of ongoing cortical speech representations in a "cocktail party. J. Neurosci. 30 (2), 620–628.
King, S., Frankel, J., Livescu, K., McDermott, E., Richmond, K., Wester, M., 2007. Speech production knowledge in automatic speech recognition. J. Acoust. Soc. Am. 121 (2), 723–742.
Kösem, A., Bosker, H.R., Jensen, O., Hagoort, P., Riecke, L., 2020. Biasing the perception of spoken words with transcranial alternating current stimulation. J. Cogn. Neurosci. 32 (8), 1428–1437.
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.
Kuhl, P.K., Ramírez, R.R., Bosseler, A., Lin, J.F.L., Imada, T., 2014. Infants' brain responses to speech suggest analysis by synthesis. Proc. Natl. Acad. Sci. 111 (31), 11238–11245.
Lambert-Shirzad, N., Van der Loos, H.M., 2017. On identifying kinematic and muscle synergies: a comparison of matrix factorization methods using experimental data from the healthy population. J. Neurophysiol. 117 (1), 290–302.
Luo, H., Poeppel, D., 2007. Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. Neuron 54 (6), 1001–1010.
Maris, E., Oostenveld, R., 2007. Nonparametric statistical testing of EEG- and MEG-data. J. Neurosci. Methods 164 (1), 177–190.
Meister, I.G., Wilson, S.M., Deblieck, C., Wu, A.D., Iacoboni, M., 2007. The essential role of premotor cortex in speech perception. Curr. Biol. 17 (19), 1692–1696.
Meyer, L., 2018. The neural oscillations of speech processing and language comprehension: state of the art and emerging mechanisms. Europ. J. Neurosci. 48 (7), 2609–2621.
Meyer, L., Sun, Y., Martin, A.E., 2020. Synchronous, but not entrained: exogenous and endogenous cortical rhythms of speech and language processing. Lang. Cogn. Neurosci. 35 (9), 1089–1099.
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., Lallier, M., Bourguignon, M., Carreiras, M., 2016. Out-of-synchrony speech entrainment in developmental dyslexia. Hum. Brain Mapp. 37 (8), 2767–2783.
Montemurro, M.A., Senatore, R., Panzeri, S., 2007. Tight data-robust bounds to mutual information combining shuffling and model selection techniques. Neural Comput. 19 (11), 2913–2957.
Morillon, B., Baillet, S., 2017. Motor origin of temporal predictions in auditory attention. Proc. Natl. Acad. Sci. 114 (42), E8913–E8921.
Morillon, B., Arnal, L.H., Schroeder, C.E., Keitel, A., 2019. Prominence of delta oscillatory rhythms in the motor cortex and their relevance for auditory and speech perception. Neurosci. Biobehav.Rev. 107, 136–142.
Möttönen, R., Watkins, K.E., 2009. Motor representations of articulators contribute to categorical perception of speech sounds. J. Neurosci. 29 (31), 9819–9825.
O'Sullivan, A.E., Crosse, M.J., Di Liberto, G.M., de Cheveigné, A., Lalor, E.C., 2021. Neurophysiological indices of audiovisual speech processing reveal a hierarchy of multisensory integration effects. J. Neurosci. 41 (23), 4991–5003.
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. Cortex 25 (7), 1697–1706.
Obleser, J., Kayser, C., 2019. Neural entrainment and attentional selection in the listening brain. Trends Cogn. Sci. 23 (11), 913–926.
Oostenveld, R., Fries, P., Maris, E., Schoffelen, J.M., 2011. FieldTrip: open source software for advanced analysisof MEG, EEG, and invasive electrophysiological data. Comput. Intell. Neurosc. 2011, 156869. doi:10.1155/2011/156869, Epub 2011/01/22PMID: 21253357.
Ozker, M., Yoshor, D., Beauchamp, M.S., 2018. Frontal cortex selects representations of the talker's mouth to aid in speech perception. eLife 7, e30387.
Park, H., Ince, R.A., Schyns, P.G., Thut, G., Gross, J., 2015. Frontal top-down signals increase coupling of auditory low-frequency oscillations to continuous speech in human listeners. Curr. Biol. 25 (12), 1649–1653.
Park, H., Ince, R.A., Schyns, P.G., Thut, G., Gross, J., 2018. Representational interactions during audiovisual speech entrainment: Redundancy in left posterior superior temporal gyrus and synergy in left motor cortex. PLoS Biol. 16 (8), e2006558.
Park, H., Kayser, C., Thut, G., Gross, J., 2016. Lip movements entrain the observers' low-frequency brain oscillations to facilitate speech intelligibility. eLife 5, e14521.
Peelle, J.E., Davis, M.H., 2012. Neural oscillations carry speech rhythm through to comprehension. Front. Psychol. 3, 320.
Peelle, J.E., Sommers, M.S., 2015. Prediction and constraint in audiovisual speech perception. Cortex 68, 169–181.
Peelle, J.E., Gross, J., Davis, M.H., 2013. Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. Cereb. Cortex 23 (6), 1378–1387.
Perrier, P., Perkell, J., Payan, Y., Zandipour, M., Guenther, F., & Khalighi, A. (2007). Degrees of freedom of tongue movements in speech may be constrained by biomechanics. arXiv:0709.1405v1. https://doi.org/10.48550/arXiv.0709.1405
Poeppel, D., 2003. The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time. Speech Commun. 41 (1), 245–255.
Poeppel, D., Assaneo, M.F., 2020. Speech rhythms and their neural foundations. Nat. Rev. Neurosci. 21 (6), 322–334.
Pulvermüller, F., Fadiga, L., 2010. Active perception: sensorimotor circuits as a cortical basis for language. Nat. Rev. Neurosci. 11 (5), 351–360.
Pulvermüller, F., Huss, M., Kherif, F., del Prado Martin, F.M., Hauk, O., Shtyrov, Y., 2006. Motor cortex maps articulatory features of speech sounds. Proc. Natl. Acad. Sci. 103 (20), 7865–7870.
Rebernik, T., Jacobi, J., Jonkers, R., Noiray, A., Wieling, M., 2021. A review of data collection practices using electromagnetic articulography. Lab. Phonol. 12 (1), 6.
Riecke, L., Formisano, E., Sorger, B., Başkent, D., Gaudrain, E., 2018. Neural entrainment to speech modulates speech intelligibility. Curr. Biol. 28 (2), 161–169.
Rimmele, J.M., Morillon, B., Poeppel, D., Arnal, L.H., 2018. Proactive sensing of periodic and aperiodic auditory patterns. Trends Cogn. Sci. 22 (10), 870–882.
Ru, P., Chi, T., Shamma, S., 2003. The synergy between speech production and perception. J. Acoust. Soc. Am. 113 (1), 498–515.
Sato, M., Tremblay, P., Gracco, V.L., 2009. A mediating role of the premotor cortex in phoneme segmentation. Brain Lang. 111 (1), 1–7.
Savariaux, C., Badin, P., Samson, A., Gerber, S., 2017. A comparative study of the precision of Carstens and Northern digital instruments electromagnetic articulographs. J. Speech Lang. Hear. Res. 60 (2), 322–340.
Schroeder, C.E., Lakatos, P., Kajikawa, Y., Partan, S., Puce, A., 2008. Neuronal oscillations and visual amplification of speech. Trends Cogn. Sci. 12 (3), 106–113.
Schubotz, R.I., 2007. Prediction of external events with our motor system: towards a new framework. Trends Cogn. Sci. 11 (5), 211–218.
Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27 (3), 379–423.
Smith, Z.M., Delgutte, B., Oxenham, A.J., 2002. Chimaeric sounds reveal dichotomies in auditory perception. Nature 416 (6876), 87–90.
Sorensen, T., Toutios, A., Goldstein, L., Narayanan, S., 2019. Task-dependence of articulator synergies. J. Acoust. Soc. Am. 145 (3), 1504–1520.
Story, B.H., 2005. Synergistic modes of vocal tract articulation for American english vowels. J. Acoust. Soc. Am. 118 (6), 3834–3859.
Sumby, W.H., Pollack, I., 1954. Visual contribution to speech intelligibility in noise. J. Acoust. Soc. Am. 26 (2), 212–215.
Timme, N., Alford, W., Flecker, B., Beggs, J.M., 2014. Synergy, redundancy, and multivariate information measures: an experimentalist's perspective. J. Comput. Neurosci. 36 (2), 119–140.
Teng, X., Cogan, B.G., Poeppel, D., 2019. Speech fine structure contains critical temporal cues to support speech segmentation. NeuroImage 202, 116152.
Ting, L.H., 2007. Dimensional reduction in sensorimotor systems: a framework for understanding muscle coordination of posture. Prog. Brain Res. 165, 299–321.
Tourville, J.A., Guenther, F.H., 2011. The DIVA model: a neural theory of speech acquisition and production. Lang. Cogn. Process. 26 (7), 952–981.
van Wassenhove, V., 2013. Speech through ears and eyes: interfacing the senses with the supramodal brain. Front. Psychol. 4, 388.
Ghinst, M.V., Bourguignon, M., de Beeck, M.O., Wens, V., Marty, B., Hassid, S., De Tiege, X., 2016. Left superior temporal gyrus is coupled to attended speech in a cocktail-party auditory scene. J. Neurosci. 36 (5), 1596–1606.
Watkins, K.E., Strafella, A.P., Paus, T., 2003. Seeing and hearing speech excites the motor system involved in speech production. Neuropsychologia 41 (8), 989–994.
Williams, P.L., & Beer, R.D. (2010). Nonnegative decomposition of multivariate information. arXiv:1004.2515v1. https://doi.org/10.48550/arXiv.1004.2515
Wilson, S.M., Saygin, A.P., Sereno, M.I., Iacoboni, M., 2004. Listening to speech activates motor areas involved in speech production. Nat. Neurosci. 7 (7), 701–702.
Zoefel, B., Archer-Boyd, A., Davis, M.H., 2018a. Phase entrainment of brain oscillations causally modulates neural responses to intelligible speech. Curr. Biol. 28 (3), 401–408.
Zoefel, B., ten Oever, S., Sack, A.T., 2018b. The involvement of endogenous neural oscillations in the processing of rhythmic input: more than a regular repetition of evoked neural responses. Front. Neurosci. 12, 95.