J Neurophysiol 131: 480–491, 2024.
First published February 7, 2024; doi:10.1152/jn.00388.2023

american physiological society

JNP JOURNAL OF NEUROPHYSIOLOGY

RESEARCH ARTICLE

Higher Neural Functions and Behavior

Speech perception difficulty modulates theta-band encoding of articulatory synergies

Alessandro Corsini,[1,2] Alice Tomassini,[1,2] Aldo Pastore,[3] Ioannis Delis,[4] Luciano Fadiga,[1,2] and Alessandro D'Ausilio[1,2]

[1]*Center for Translational Neurophysiology of Speech and Communication, Istituto Italiano di Tecnologia, Ferrara, Italy;* [2]*Department of Neuroscience and Rehabilitation, Università di Ferrara, Ferrara, Italy;* [3]*Laboratorio NEST, Scuola Normale Superiore, Pisa, Italy; and* [4]*School of Biomedical Sciences, University of Leeds, Leeds, United Kingdom*

Abstract

The human brain tracks available speech acoustics and extrapolates missing information such as the speaker's articulatory patterns. However, the extent to which articulatory reconstruction supports speech perception remains unclear. This study explores the relationship between articulatory reconstruction and task difficulty. Participants listened to sentences and performed a speech-rhyming task. Real kinematic data of the speaker's vocal tract were recorded via electromagnetic articulography (EMA) and aligned to corresponding acoustic outputs. We extracted articulatory synergies from the EMA data with principal component analysis (PCA) and employed partial information decomposition (PID) to separate the electroencephalographic (EEG) encoding of acoustic and articulatory features into unique, redundant, and synergistic atoms of information. We median-split sentences into easy (ES) and hard (HS) based on participants' performance and found that greater task difficulty involved greater encoding of unique articulatory information in the theta band. We conclude that fine-grained articulatory reconstruction plays a complementary role in the encoding of speech acoustics, lending further support to the claim that motor processes support speech perception.

NEW & NOTEWORTHY Top-down processes originating from the motor system contribute to speech perception through the reconstruction of the speaker's articulatory movement. This study investigates the role of such articulatory simulation under variable task difficulty. We show that more challenging listening tasks lead to increased encoding of articulatory kinematics in the theta band and suggest that, in such situations, fine-grained articulatory reconstruction complements acoustic encoding.

articulatory synergies; EEG; mutual information; partial information decomposition; speech entrainment

INTRODUCTION

During speech production, speakers engrave their message in the air by employing vocal cord vibrations and vocal tract resonance to hierarchically organize different units of information, such as phonemes, syllables, words, and sentences. As a result, the acoustic output is shaped around a mixture of quasi-periodic signals, reflecting the intrinsic rhythms of speech-motor commands (1, 2). During speech perception, endogenous neural oscillations are phase-aligned to the physical regularities of the incoming speech stimulus ("speech entrainment") (3) to track different units of information (4, 5) and to support the speech comprehension process (3, 6–10). Neural tracking of slow temporal fluctuations of the speech amplitude envelope has been consistently reported in delta and theta bands in both electroencephalography (EEG) and magnetoencephalography (MEG) studies (3, 6, 11–13).

Speech tracking occurs even in the absence of physical cues in the spectrum of the acoustic stimulus, highlighting the relevance of top-down reconstructive influences on incoming speech signals (14). Phonemic restoration (15, 16) is one such example in which perceptual experience does not coincide with the pure bottom-up encoding of physical acoustic regularities. The cocktail party effect (17) further nourishes these considerations: in multiple-talker scenarios, the cortical representation of speech reflects the encoding of

Correspondence: A. Corsini (crslsn1@unife.it).
Submitted 23 October 2023 / Revised 4 January 2024 / Accepted 25 January 2024

features related to the attended acoustic stream rather than the true content of the acoustic scenario (18–20). Multimodal speech perception (i.e., audiovisual) is another key example of a top-down process (21). In fact, visual cortices track visual speech-related signals to support speech perception under acoustically degraded conditions (22, 23).

Growing experimental evidence suggests that an important source of top-down compensatory modulation may originate from the motor system, underlining the importance of speech production areas during speech listening (11, 24). Indeed, signals coming from the motor cortex causally modulate the phase of delta- and theta-band oscillations in the left auditory cortex only when listening to intelligible speech, and the more motor cortex transfers information to auditory cortex, the stronger low-frequency auditory oscillations entrain to the speech envelope (25). This result is best explained by assuming an internal replay of articulatory movements constrained or guided online by the perception of visible movements. Recently, Pastore et al. (26) found that the brain reconstructs articulatory movements (i.e., the speaker's tongue motion) that are never accessible through vision, either during the experiment or during development (see also Ref. 27). This result was interpreted as demonstrating that acoustic-motor mappings originate from the acquisition of speech production competence (for a recent review see Ref. 28), rather than from exposure to the regularities of a multisensory environment. However, it is unclear to what degree the reconstruction of motor gestures depends on the difficulty of the listening task and whether the neural mechanisms subserving acoustic-motor mappings in such situations rely on delta- or theta-band dynamics.

To investigate this question, we asked participants to listen to sentences followed by a rhyming task to assess their ability to match speech sounds phonetically from memory. We used sentences from the Multi-SPeaKing-style Articulatory Corpus (MSPKA) (29), which provided us with an accurate description of the articulatory movements of the speaker's vocal tract aligned with the produced audio. Considering that the motor system does not control individual articulators (30) but rather orchestrates multiple articulators to achieve an acoustic target, we used principal component analysis (PCA) to extract meaningful patterns of articulatory coordination over time. We then used partial information decomposition (PID) (31) to quantify the unique encoding of articulatory and acoustic features as well as synergistic and redundant information in the neural response (EEG data). We evaluated the difficulty of the presented sentences according to the participants' performance and divided them into easy and hard sentences. Our hypothesis is that the presentation of more difficult stimuli would prompt a stronger engagement of articulatory processes and, as a consequence, enhance the encoding of information uniquely attributable to the articulatory components. In line with our hypothesis, we observed a dissociation between delta and theta bands, with theta-band oscillations playing a complementary role with respect to acoustic encoding when presented with more difficult material. This finding contributes to a number of current areas such as *1*) recent empirical evidence showing the functional dissociation of neural tracking in the two bands and *2*) the idea of the motor system supporting speech perception, especially in more challenging listening conditions.

MATERIALS AND METHODS

Participants

Twenty-three healthy and naive subjects (Italian native speakers, right-handed and with normal vision or corrected-to-normal vision) were recruited for the study and were paid 30€ for their participation. Written informed consent was obtained from all participants. The experiment was approved by the local ethical committee, "Comitato Etico Unico della Provincia di Ferrara" (approval no. 170592). One participant was excluded because of technical problems during data acquisition. The analysis was performed on the remaining 22 subjects (13 females; age: 23.04 ± 3.44 yr, mean ± SD).

Stimuli

This study used stimuli selected from the Multi-SPeaKing-style Articulatory Corpus (MSPKA) (29). This dataset provides simultaneous recordings of audio and articulatory movements of the vocal tracts of three Italian mother-tongue speakers. The audio was recorded at a sampling rate of 22.05 kHz, whereas the kinematics of the articulators (lips, jaws, and tongue) were tracked at a frequency of 400 Hz. An accurate description of the vocal tract motion during audio production was obtained with an electromagnetic articulography (EMA) system (NDI Wave, Northern Digital Instruments, Canada) (32). Research on speech technology commonly utilizes EMA data to accurately describe mouth kinematics owing to its excellent spatial and temporal resolution (33). Seven sensors were placed 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*A* for a schematic illustration) to record their x, y, and z positions (head movement corrected). We used 50 sentences pronounced by the same female speaker ("lls" in the dataset). The duration of the sentences ranged from 6.2 to 9.4 s. Segments corresponding to any silent part at the beginning and end of the acoustic stimuli were removed from all signals (audio and EMA). All the acoustic stimuli were 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 signals were partially corrupted. The acoustic stimuli were presented to the subjects, whereas EMA data were exclusively used during the data analysis.

Experimental Setup and Procedure

During the experiment, the participants sat in front of an LCD monitor (VIEWPixx/EEG; 24 in., 120 Hz) at a distance of ~80 cm and were asked to put their right hand on a button box (Cedrus RB-840 response box). Two loudspeakers were placed ~20 cm from each side of the screen. The session consisted of 200 trials divided into four blocks (50 trials each), with short in-between breaks. In each trial, the following sequence of events occurred: *1*) participants were presented with a black fixation cross at the center of a uniformly colored gray screen; *2*) after a variable time (between 0.1 and 1.1 s), a randomly selected sentence was presented acoustically;

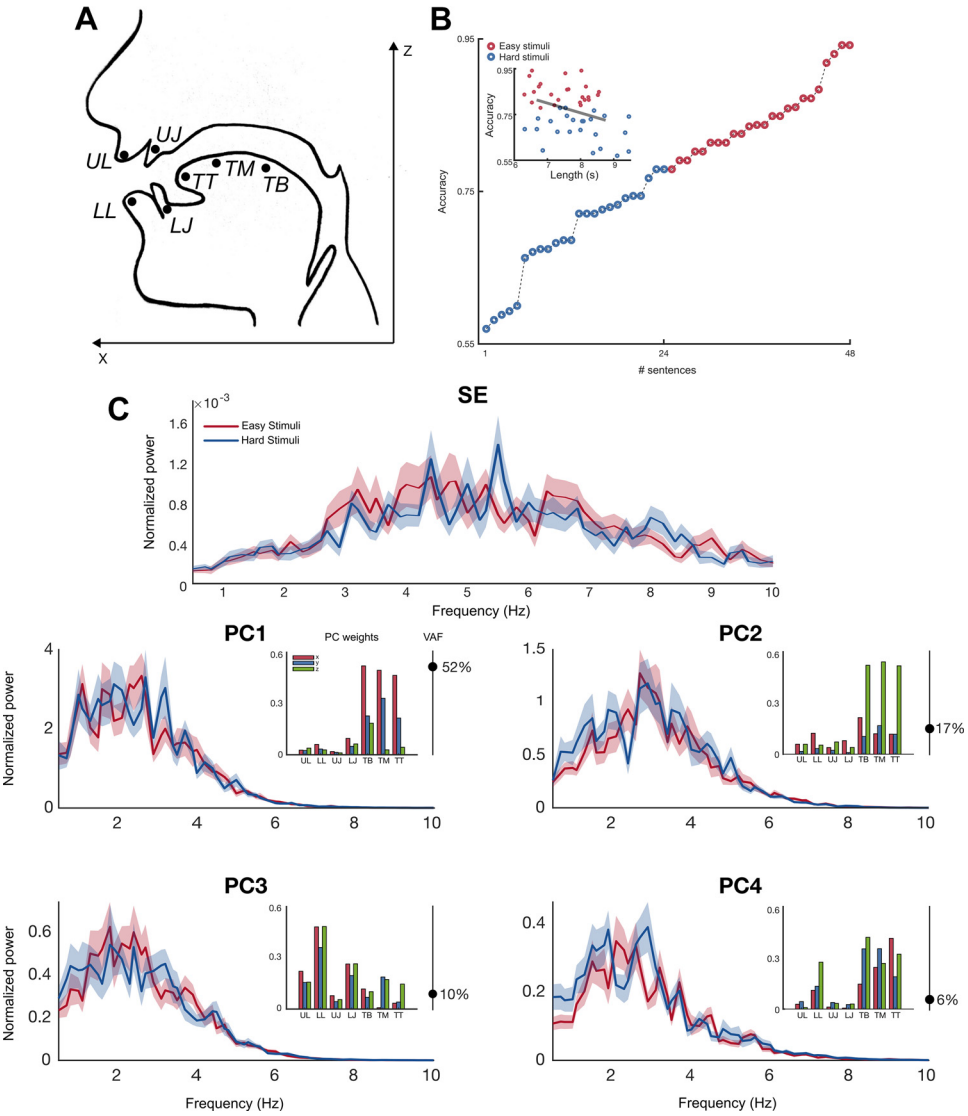

Figure 1. Stimuli split and feature description. *A*: schematic illustration of the positions of the electromagnetic articulography (EMA) coils: upper lip (UL), lower lip (LL), upper jaw (UJ), lower jaw (LJ), tongue tip (TT), tongue middle (TM), and tongue back (TB). *B*: split of the 48 sentences between easy stimuli (ES) and hard stimuli (HS). Accuracy was obtained from the pooled answers of all subjects. The scatterplot shows the negative correlation between accuracy and stimulus length ($r = -0.37$, $P = 0.01$). *C*: mean and SEM of the normalized (1/f) power spectra for all features [speech envelope (SE), principal component (PC)1, PC2, PC3, PC4] in both classes of stimuli (ES and HS). Only PCs explaining at least 5% of the variance [variance accounted for (VAF)] were selected for the analyses. Bar plots show the contribution of all sensors to each of the selected kinematic components.

3) after a variable time (between 0.1 and 1.1 s) from the end of the acoustic stimulus, a word appeared at the center of the screen in place of the fixation cross; *4*) participants had to indicate whether or not the word presented rhymed with one of the words in the previously heard sentence by pressing one of two buttons, always using the right index finger; and *5*) the trial ended when the participant gave a response or after a time-out of 10 s if no response was given. In each block, all 50 sentences were presented to the subject (1 for each trial) in random order. Rhyming words were selected to exclude any rhyme with the first and last words of the presented sentence and any monosyllabic word. To avoid possible biases in the participants' responses, we ensured that rhyming and nonrhyming words were matched for the number of syllables and their frequency of use in Italian, using an online software tool (http://linguistica.sns.it/esploracolfis/home.htm). Different words were selected for each repetition of the same sentence, resulting in four words per sentence (2 rhyming and 2 nonrhyming), for a total of 200 words. Therefore, words presented in the session rhymed 50% of the time. Participants were asked to reduce their blinks as much as possible and to maintain their eyes on the fixation cross for the entire duration of the sentence. The entire experiment lasted ~2 h, including the EEG cap mounting and preparation. Stimulus presentation and button-press acquisition were controlled via MATLAB (The MathWorks, 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, and button press) were converted into a TTL (transistor-transistor logic) by the VIEWPixx/EEG system to accurately synchronize them with the EEG data.

EEG Recording

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 with four electrodes from the cap (FT9, FT10, PO9, and PO10) that were removed from the 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 the electrodes were online referenced to the left mastoid. The impedance of the electrodes was kept below 10 kΩ. EEG signals were acquired at 1,000 Hz.

Data Analysis

The present study involved reanalyzing a dataset previously recorded by our group (26) to answer a different research question. Analyzes were performed within the MATLAB and Python computing environments with open-source toolboxes and libraries: FieldTrip (http://www.fieldtriptoolbox.org; RRID:SCR_004849) (34), MNE (https://mne.tools/stable/index.html; RRID:SCR_005972) (35), GCMI (https://github.com/robince/gcmi) (36), and PID (https://github.com/robince/partial-info-decomp) (37) libraries as well as custom-made code (https://doi.org/10.5281/zenodo.10580338).

Behavioral Data Analysis and Dataset Split

To classify the selected stimuli (i.e., sentences) into two classes of difficulty, we examined the participants' responses to the rhyming task. In practice, we used the 22 participants as the a posteriori jury. Before proceeding, we excluded one randomly selected sentence to achieve class balance (i.e., an equal number of items), thus retaining 48 sentences in total. We then pooled the responses of all participants and calculated, separately for each sentence (4 presentations × 22 participants = 88 answers per sentence), the accuracy (the ratio of the total number of correct responses to the number of pooled trials). Sentences with higher accuracy were considered easier than those with lower accuracy values, which was assumed to pose a greater challenge in the speech-rhyming task. We then performed a median split on the accuracy of all sentences, yielding two classes containing 24 stimuli each: easy sentences (ES) and hard sentences (HS) (Fig. 1*B*).

Speech Envelope Extraction

Speech signals were processed to obtain the amplitude envelope of each sentence, using an adapted version of a previously described method (38, 39). As in the Chimera toolbox (39), we defined six frequency bands in the range of 80–8,820 Hz equally spaced on the cochlear map. We then filtered the speech signals within these six frequency bands (MNE filter_data function, 2-pass Butterworth filter, 4th order). We then computed the absolute value of the Hilbert transform for each band-pass filtered signal. Finally, we obtained the speech envelope (SE) by summing the results over all frequency bands. The envelope was then downsampled to 400 Hz to match the sampling rate of the EMA data.

Kinematic Feature Extraction

The high-dimensional EMA data (i.e., 7 sensors × 3 dimensions = 21 time series of position data) were reduced in dimensionality by applying principal component analysis (PCA; FieldTrip function: ft_componentanalysis; method: pca) to extract meaningful synergies between the individual articulators. Precisely, each stimulus i is represented as a matrix of shape (21,$Nsamples_i$): recording dimensions by number of samples in the ith stimulus recording. We concatenated all the stimuli matrices along the last dimension to obtain a single matrix D of shape $\left(21, \sum_{i=1}^{N} Nsamples_i\right)$, where N is the number of stimuli, and applied the algorithm to the matrix of concatenated data D. PCA rotates the original (sensor) space to maximize the amount of information stored in the projection of the data along the first principal components (PCs), thus allowing for dimensionality reduction. This result is obtained by applying eigendecomposition to the data covariance matrix Σ_D. Each PC is a vector obtained as a linear combination of the coordinates in the original sensor space. Therefore, each PC attributes a weight to each of the original coordinates, which can easily be used to interpret the composition of each feature dimension. In practice, we visually inspected the absolute value of the PC weights to assess the physiological validity of the identified articulatory pattern (Fig. 1*C*). By selecting a subset of the PCs (the first 4), we reduced the dimensionality of the data while retaining most of the information. Importantly, performing a single PCA on concatenated data, rather than one PCA for each stimulus, maximized the extraction of articulatory synergies consistent across sentences. This can better approximate prototypical patterns of articulation in the spoken material and therefore, possibly, the most salient signal for the listener's brain.

EEG Preprocessing

The continuous EEG data were band-pass filtered between 0.5 and 100 Hz (2-pass Butterworth filter, 4th order) and downsampled to 400 Hz to match the sampling rate of the EMA data. Data were then rereferenced to the common average and epoched around the acoustic stimulus onset (from −1 s to the duration of the longest sentence plus 1 s). The noisy trials were removed after visual inspection. Artifacts related to eye movements and heartbeat were identified and removed by independent component analysis (ICA). Noisy channels (T8 for 1 subject) were excluded from the ICA analysis and substituted by the linear interpolation of neighboring channels. The total number of trials retained for further analysis was 179.2 ± 21.1 (mean ± SD).

Neural Coupling to Speech Envelope and Kinematic Features

We quantified the information encoding of delta- and theta-band features of the speech envelope and the articulatory features in the two classes of sentences (ES and HS) employing the mutual information measure (MI) (40). Mutual information is a measure of the reduction of uncertainty in the output of a random variable X given the observation of a second variable Y and can be viewed as a test against the null hypothesis that the two variables are statistically independent, thus considering also nonlinear and nonmonotonic relationships. In other words, MI measures the extent to which the variability of X can be predicted by looking only at Y. We used the Gaussian Copula Mutual Information (GCMI) estimator, which provides a lower bound to the true MI value and is robust to a limited quantity of collected data (36). During speech listening, the brain encodes information about the speech envelope as well as the movement of the invisible articulators of the speaker's vocal tract (26). As in Ref. 26, we

only considered the first four PCs (accounting for most of the variance in the kinematic data) and separately computed the MI between the EEG and each of the selected features (speech envelope: SE; principal components: PCi, $i = 1,\ldots,4$). Precisely, we first cut 0.5 s at the beginning of every trial to remove the initial transient components evoked by the onset of speech and shifted the EEG signals 0.2 s backward in time with respect to the speech-related stimulus to account for a natural lag between stimulus presentation and brain response. This means that the EEG timestamp corresponding to 0.2 s of each epoch was associated to the start of the trial (0.0 s) and all the other EEG timestamps shifted accordingly, while the stimuli signals remained unchanged. The selection of this time interval is justified by the large literature on brain coupling to audiovisual speech cues (41–43) and past analyses performed on this dataset showing maximal values of MI at 0.2-s time lag (26). An inspection of the power spectra of the stimuli (SE and PCs) showed that most of the spectral information was confined to the delta and theta bands and that ES and HS stimuli showed comparable spectral power distribution in all five selected features (Fig. 1*C*). Consequently, all signals (EEG, SE, PCs) were mirror-padded and band-pass filtered between 0.5 and 4 Hz (delta band) and between 5 and 7 Hz (theta band) with a two-pass Butterworth second-order filter in both cases. We then concatenated and copula-normalized the trials belonging to the same class. Specifically, a copula is a statistical structure that represents the dependence of two or more variables independently of their marginal distributions. It is a useful concept since mutual information between X and Y [$I(X,Y)$] is directly linked to it. Precisely,

$$I(X,Y) = H(X) + H(Y) - H(X,Y)$$

$$H(X,Y) = H(X) + H(Y) + H(C)$$

where H is the entropy function and C the statistical copula. Consequently, $I(X,Y)$ is equal to the opposite of the entropy of the statistical copula:

$$I(X,Y) = -H(C)$$

Therefore, if we preserve the copula linking the variables X and Y, mutual information does not change. In this context, we transform the marginal distributions of X and Y to be univariate Gaussian while still preserving their statistical copula (copula normalization). Finally, we computed the mutual information $I(X,Y)$, using the analytical expression for Gaussian variables (function: gcmi_cc):

$$I(X,Y) = \frac{1}{2\ln 2}\ln\left[\frac{|\Sigma_X||\Sigma_Y|}{|\Sigma_{XY}|}\right]$$

[where Σ_X, Σ_Y, and Σ_{XY} are the covariance matrix of X, Y, and the joint variable (X, Y), respectively] between each recorded EEG channel and *1*) the speech envelope I(EEG;SE), and *2*) each of the first four extracted PCs (kinematic components) I(EEG;PCi), $i = 1,\ldots,4$.

Partial Information Decomposition

Acoustic and kinematic features are strongly related to each other, as speech acoustic outputs are directly constrained by phonoarticulatory movements. Therefore, the neural encoding of the two could produce overlapping information. Shannon's mutual information only measures the relationship between two variables at a time (SE and EEG or PC and EEG). For systems modeled by $n > 2$ variables, the mutual information between one of these variables and the joint distribution of the others can be computed. However, this procedure does not reveal the internal structure of multivariate information. Interaction information (44) (or coinformation; Ref. 45) partially solves this problem: it identifies components of unique information provided by the predictors on a target value and one of interaction information. However, this last variable merges the redundant and synergistic effects of predictors (38). Instead, partial information decomposition (PID) is a mathematical framework proposed by Williams and Beer (31) that is capable of decomposing the total information provided by a set of variables, called sources, about a target variable into clearly interpretable atoms of partial information (46, 47). In the bivariate case (2 sources and 1 target), PID outputs four atoms of partial information: two atoms of information exclusively provided by each of the two sources ("unique"), one accounting for the information only available when the two sources are considered together and never accessible when looking at one source at a time ("synergistic"), and the last one for information shared between the two sources ("redundant"). Thus, we employed PID to account for the simultaneous encoding of the speech envelope and first kinematic component (the only one showing significant results in the MI values) during listening. We used a recent modification of the algorithm based on the common change in surprisal, which can qualitatively capture redundant information between variables (37) (MATLAB function: calc_pi_mvn; see also https://github.com/AlessandroCorsini/PyPID_mvn for a precise implementation in Python). We ran the PID separately for each class of sentences (ES and HS), considering the speech envelope (SE) and the first kinematic component (PC1) as sources and each single EEG channel as the target variable. All signals were preprocessed as described for MI (exclusion of onset-locked evoked potentials, EEG backward shifting by 0.2 s, and band-pass filtering in delta and theta bands separately). The choice of the lag is again justified by previous analyses on the same dataset, which showed how maximal information encoding of both SE and PC1 is reached with a 0.2-s lag (26). We then concatenated and copula-normalized trials belonging to the same class of stimuli and computed the PID using the redundancy measure proposed by Ince (37) (function: Iccs_mvn). The algorithm provides four terms (for each channel):

- Unq_1(EEG; SE, PC1): the information encoded in the EEG signal conveyed by the speech envelope and not by the kinematic principal component. This piece of information is feature specific, meaning that it would be lost if the SE was not available to the brain.
- Unq_2(EEG; SE, PC1): the information encoded in the EEG signal conveyed by the kinematic principal component and not by the speech envelope. This piece of information is specific to the reconstruction of PC1 by the listener's brain, meaning that it would be lost if the brain did not encode the kinematic feature.
- Syn(EEG; SE, PC1): the information conveyed by the simultaneous presence of SE and PC1. Instead, this information is specific to the set of sources {SE, PC1} and

cannot be accessed by the brain if one of the two sources is unavailable. Therefore, it is related to how the two features are integrated into the neural response.

- Rdn(EEG; SE, PC1): the information shared between the two sources, which can be equally retrieved from SE or PC1. This information is specific neither to the SE nor to PC1 and is lost if and only if both features are not available to the brain.

Hereafter, we refer to these quantities as Unq(SE), Unq (PC1), Syn(SE,PC1), and Rdn(SE,PC1).

Notably, the concatenation of the trials belonging to the ES class was shorter than that of the HS trials in all the subjects (ES: 618.04 ± 74.15 s; HS: 663.14 ± 77.86 s; mean ± SD). To exclude the possibility that a possible bias in the PID computation dependent on the length of the signals affected our results, we matched the lengths of the ES and HS concatenations 1,000 times for each subject by cutting a segment of adjacent samples (every time at a randomly selected point) from the longer one and computed the PID at each repetition. Eventually, we ran the statistical analysis (see below) on the average (across 1,000 repetitions) PID output.

Statistical Analysis

To test whether the neural encoding of acoustic and kinematic stimulus information differed in the two classes of sentences at the group level, the MI/PID outputs were statistically evaluated for one condition against the other by applying two-tailed cluster-based permutation statistics, as described by Maris and Oostenveld (48). In practice, we tested the null hypothesis that the two (ES, HS) multisensor MI/PID values belong to the same probability distribution. We did this using a nonparametric statistic, thus without making any assumption on the probability distributions of the information values. All MI/PID output samples (1 per condition for each subject) were collected in a single set and then randomly partitioned into two sets, and the test statistic was calculated. Specifically, a univariate statistical test was first employed to compute a channel-specific t value (using the formula for dependent samples). All channels whose absolute t values exceeded an a priori decided threshold were clustered based on spatial adjacency (separately for positive and negative t values). We fixed the threshold to the 97.5th quantile of the T distribution, which, in a two-sided test, sets the probability of rejecting the null hypothesis to a critical alpha level of 0.05. At this point, the cluster-level statistics were calculated by summing all t values of the channels inside a cluster. This procedure was repeated for 5,000 random permutations to construct a histogram of the distribution of the largest (in absolute value) of the cluster-level statistics found at each repetition. Finally, the test statistic was performed on the nonpermuted data (original ES and HS conditions), clusters of channels were identified, and their P value was calculated as the proportion of random permutations yielding a larger (in absolute value) test statistic compared with that computed for the nonpermuted data (Monte Carlo estimate). Crucially, this method inherently controls the FA (false alarm) rate setting the probability of type 1 errors equal to the critical alpha level (0.05 here) and solves the MCP (multiple comparison problem) by testing a hypothesis only once at the cluster level.

RESULTS

Behavioral Performance and Articulatory Feature Extraction

We first analyzed behavioral data to split the dataset into easy and hard stimuli. The pooled-subjects accuracies related to easy sentences (ES) varied from 78% to 94%, whereas those for the hard sentences (HS) ranged from 57% to 78% (Fig. 1*B*). Accuracy was also negatively correlated with sentence length ($r = -0.37$; $P = 0.01$; Fig. 1*B*). We then evaluated the consistency of the extracted stimuli features (speech envelopes and kinematic components). In both classes of stimuli, the peaks of the power spectra of speech envelopes were mostly confined between 4 and 8 Hz, which is consistent with consolidated evidence (3, 49–52). The first four principal components (PCs) together accounted for 85% of the total variance of the EMA data (Fig. 1*C*), whereas each of the remaining components explained a negligible amount of variance [variance accounted for (VAF): <5% each]. During speech production, the motor system regulates the activation of several muscles to reach a vocal tract configuration (2, 30). PCA extracted physiologically meaningful kinematic synergies: PC1 (VAF = 52%) and PC2 (VAF = 17%) described tongue movement toward (and away from) the lips (PC1) and the palate (PC2), PC3 (VAF = 10%) instead accounted for mouth opening and closing, whereas PC4 (VAF = 6%) represented more complex synergies between tongue and lip movement. The spectral peaks of all four PCs fell between 0.5 and 4 Hz (delta band) under both conditions (ES and HS).

Neural Encoding of Acoustic and Kinematic Features

We quantified neural encoding of the speech envelope (SE) and of each of the selected kinematic components (PCi, $i = 1,\ldots,4$) separately for the two types of stimuli by mutual information (MI) (40) and then statistically evaluated their difference (ES vs. HS). Only the neural tracking of the speech envelope [I(EEG,SE)] and the first kinematic component [I(EEG,PC1)] showed significant results (2-tailed cluster-based statistics). Importantly, the only frequency band showing consistent differences in the encoded information was the delta band, whereas tracking of the theta band dynamics was not affected by the class of stimuli (Fig. 2; 2-tailed cluster-based statistics). One significant cluster covering parieto-occipital electrodes (P4, CP4, P2, PO7, PO4; $P = 0.02$) was obtained for the speech envelope, indicating greater information encoding for HS than for ES (negative cluster). On the other hand, two significant positive clusters, indicating greater information encoding for ES than HS, were found for PC1 (Fig. 2): the first covers parieto-occipital electrodes (CP1, CP2, Pz, C2, CPz, CP4, P1, P2, PO3, POz; $P = 0.003$), whereas the other is distributed over right frontal electrodes (Fp2, F4, F8, AF4, AF8, F6; $P = 0.012$). All in all, this initial analysis speaks in favor of a functional dissociation between the encoding of acoustic and articulatory data in relation to task difficulty, which is, however, limited to the delta band.

Delta-Theta Dissociation

The MI metric, however, may conflate more refined unique and synergistic modulatory effects that can instead

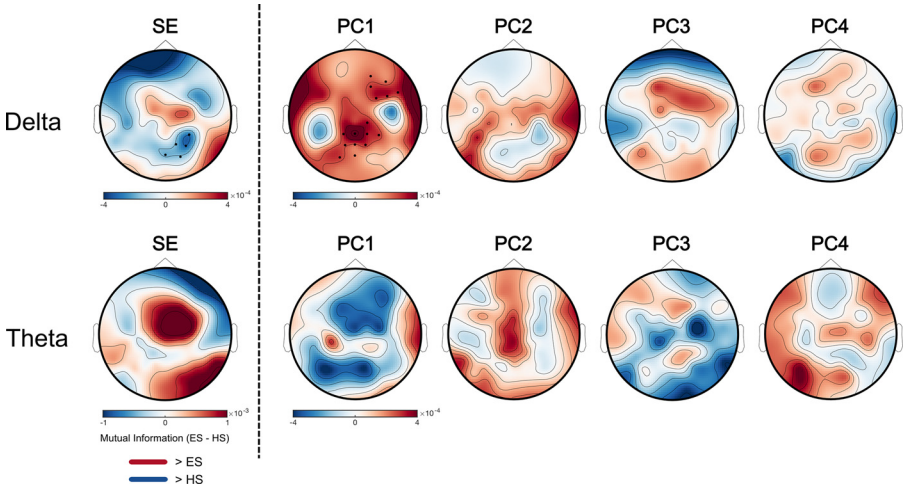

Figure 2. Topographical distributions show the mean mutual information (MI) difference across subjects [easy stimuli (ES) − hard stimuli (HS)] computed for the speech envelope [*I*(EEG,SE)] and the 4 principal components [*I*(EEG, PC*i*), $i = 1,\ldots, 4$] for band-pass filtered data (EEG, SE, PCs) in the delta (0.5–4 Hz) and theta (5–7 Hz) bands. Black dots highlight the electrodes belonging to the clusters that survived 2-tailed cluster-based statistics (ES vs. HS; alpha level = 0.05).

be distinguished by the PID method. Interestingly, PID analysis confirmed the pattern of results obtained with MI in the delta band while providing novel insights into the nature of speech processing in the theta band (Fig. 3). First, a statistically significant difference in the theta band also emerged for PC1. In fact, the encoding of unique articulatory information [Unq(PC1)] is greater for HS than ES over two distinct clusters of electrodes, showing right parieto-occipital (CP6, P4, O2, P6, PO4, PO8; $P = 0.011$) and fronto-central (F3, Fz, FC1, FCz, AF3, F1, F2; $P = 0.008$) distributions. Second, ES was associated with significantly greater encoding of redundant information (shared between SE and PC1) than HS in the theta band over right parieto-occipital (CP6, P8, TP8, P6, PO8; $P = 0.008$) and fronto-central (F4, AF4, F2, FC4; $P = 0.014$) electrodes (Fig. 3*A*). In the delta band, greater encoding of unique acoustic information [Unq(SE); $P = 0.008$] was observed for HS than for ES, whereas the opposite pattern (ES > HS) holds for unique articulatory information [Unq(PC1); parietal cluster, $P = 0.004$; right frontal cluster, $P = 0.023$], closely resembling the results reported above for MI.

Overall, these results highlight a delta-theta dissociation for Unq(PC1) and an opposite trend for Unq(SE) in the delta

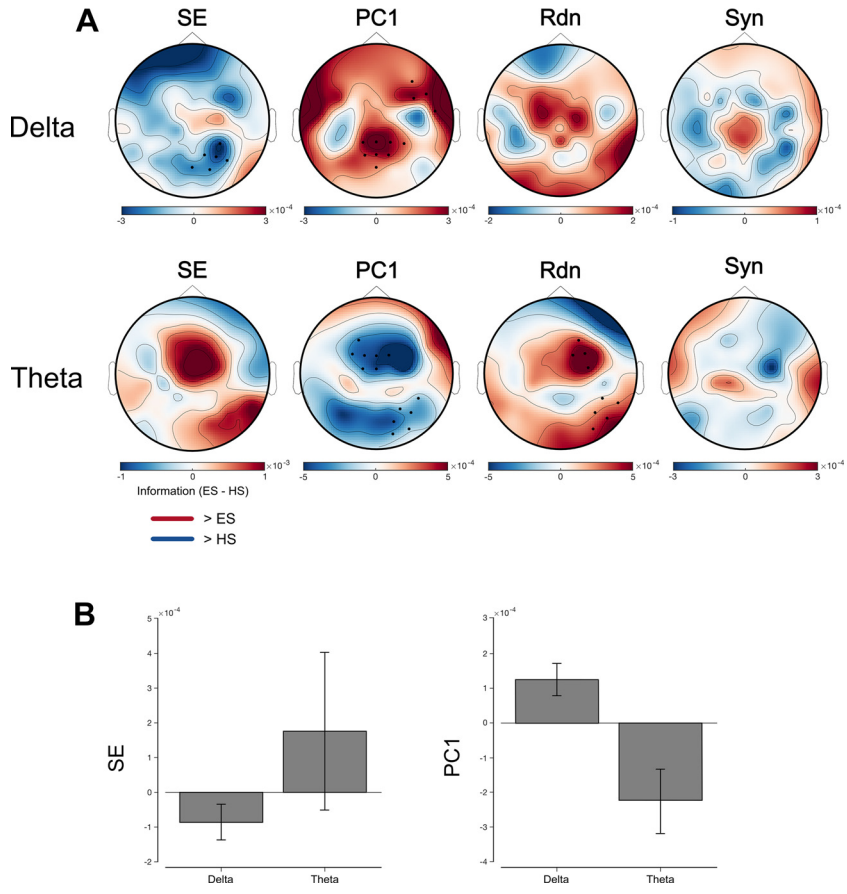

Figure 3. Partial information decomposition (PID) delta-theta analysis. *A*: topographical distributions show the mean of the information difference across subjects (ES − HS) for the unique [Unq(SE), Unq(PC1)], redundant [Rdn(SE,PC1)] and synergistic [Syn(SE, PC1)] atoms of information obtained for band-pass filtered data in the delta and theta bands. SE, speech envelope; PC, principal component. Black dots highlight the electrodes belonging to the clusters that survived 2-tailed cluster-based statistics [easy stimuli (ES) vs. hard stimuli (HS); alpha level = 0.05]. *B*: group average and SEM of the difference between the average information across all channels in the 2 conditions (ES − HS) for Unq(SE) (*left*) and Unq(PC1) (*right*) in delta and theta.

band, whereas theta-band modulations for the acoustic features did not reach statistical significance (Fig. 3*B*). In other words, listening to easier stimuli resulted in greater encoding of articulatory information in the delta band, whereas preferential encoding of the same articulatory features for harder stimuli occurred in the theta band.

However, harder sentences are, at a first-order approximation, longer than easy sentences (see above). Therefore, to further test the robustness of the results and to account for a possible bias in the PID algorithm depending on the length of the dataset, the two sets of data (ES, HS) were matched in length, and PID was computed 1,000 times. We then repeated the statistical analysis of the average PID outputs at the group level. Overall, the analysis completely validated the reported pattern of results in both the delta and theta bands (Supplemental Fig. S1) and excluded potential confounds owing to unbalanced data length. In fact, for features filtered in the delta band, the statistics run on Unq(SE) yielded one significant negative cluster on the same (and extending to others) parieto-occipital electrodes ($P = 0.002$) and confirmed the significant positive cluster on central parieto-occipital channels ($P = 0.004$) for Unq(PC1). Similarly, theta-band analysis confirmed the significant clusters for both Unq(PC1) (negative clusters; both $P = 0.007$) and Rdn (SE,PC1) (positive clusters; $P = 0.006$ and $P = 0.012$). This suggests that the difference between ES and HS is not length per se, but rather it must be found in the complexity of storing into and retrieving from memory the phonetic content of each sentence. Indeed, considering the nature of our task, which forces participants to concentrate on the acoustic/phonetic properties of the words rather than on their semantics, harder listening tasks may have forced subjects to enrich the content of the acoustical objects via integration of alternative streams (kinematic representations).

DISCUSSION

During speech perception, brain rhythmic activity synchronizes with the quasi-periodic properties of speech signals (3). This synchronization is believed to support the segregation of relevant units of information such as syllables and phrases (1). Recent evidence suggests that brain oscillations originating from motor or premotor areas may contribute to speech entrainment via top-down mechanisms (25). However, determining the motor nature of brain-speech coupling phenomena is inherently ambiguous, as the localization of activities in motor areas may not necessarily describe *motor processes*. Consider, for example, the classic increase in firing rate in neurons of the primary motor cortex following passive finger displacement (53) and the recent demonstration in mice that spontaneous activity in the primary visual cortex is largely explained by motor behavior and is not interrupted by visual stimulation (54). Here, we approached this problem by exploiting a dataset containing synchronized acoustic and articulatory information, applying specific information-theoretic tools that led us to quantify information encoding in the brain uniquely attributable to the acoustic cues and articulatory movements as well as to both (redundancy) and to the ensemble of the two (synergy). Finally, we evaluated how such encoding schemes differ based on task difficulty.

The Importance of True Articulatory Data

To achieve this goal it was essential to get access to the articulatory side of speech, and, unlike techniques normally used to track vocal tract productions (i.e., ultrasound, real-time MRI, or electromyography), EMA allows for the refined evaluation of vocal tract coordination dynamics with high temporal and spatial resolution. Prior studies have made use of articulatory features recovered from vocal acoustics (55, 56) or limited the investigation to one external articulator (i.e., lip motion from video recordings, e.g., Refs. 24, 56). However, it must be considered that the motor system coordinates the activation of multiple articulators simultaneously, according to synergistic principles (57), to reach specific acoustic targets (58–60).

Indeed, the same acoustic target can be achieved via multiple context-based (i.e., coarticulation) configurations of the phono-articulatory tract. This many-to-one mapping is a clear example of an ill-posed inverse problem, and, although it has been a matter of extensive investigations for the last 30 years, the speech technology field has yet to find a solution to the acoustic-to-articulatory inversion problem (e.g., Ref. 61). This fact places a hard limit on what can be demonstrated concerning the encoding of motor features in brain signals unless we record the articulatory output. The currently accepted view is that articulation in all its spatial-temporal details is not directly computable from the speech signal (61–67).

Therefore, we deliberately circumvented the acoustic-to-articulatory inversion problem by identifying patterns of articulatory coordination directly from kinematic data, thus avoiding the arbitrariness inherent in adopting a (number of) combination(s) of acoustic features that loosely map to these articulators. In this study, as is often the case with studies on upper limb motor control (68, 69), we used a data reduction technique to identify movement synergies.

Delta/Theta Encoding of Articulatory Synergies

Our analyses based on the PID framework showed that increased task difficulty gave rise to greater encoding of the speech envelope in the delta band (0.5–4 Hz) as opposed to a reduced encoding of articulatory information. The delta band appears to be related to the processing of slower acoustic features, such as the pitch contour (12), and is thus concerned with the grouping of words and phrases (4, 16, 70, 71). Greater entrainment for more difficult sentences may therefore be the consequence of increased listening effort (72), probably associated with information integration across longer temporal windows and thus larger contextual tokens to compensate for the increased difficulty of the task. Speech tracking in the left auditory cortex is, in fact, also modulated by delta oscillations in motor areas (25), whereas left motor/premotor cortex engagement explains speech comprehension (73) and multisensory integration (74). Delta-band entrainment to speech is stronger when listening to meaningful sentences than to lists of randomized words (75, 76), suggesting its involvement in encoding meaningful content above and beyond phonetics (77, 78). In fact, as long as the speech signal remains intelligible, delta entrainment is reduced neither by the injection of increasing levels of noise nor by the spectral impoverishment of acoustic cues (79).

In contrast, theta-band activity has been linked to acoustic-phonetic processing and thus to the encoding of features critical for speech intelligibility (3, 80, 81). Here we show that increased task complexity causes the brain to encode articulatory information in a unique and thus complementary neural pattern. This finding fits well with the principle of inverse effectiveness in multimodal integration (82), according to which perception obtains maximum gain from a second modality when unimodal stimuli evoke weaker individual responses. This idea has been tested in audiovisual speech processing and started from an observation by Sumby and Pollack (83), who showed that the presence of visual cues becomes more important as speech recognition tasks become more difficult (larger vocabulary sizes and/or higher speech-to-noise ratios). This research has led to the conclusion that visual information aids speech processing in two concomitant ways, that is, "correlated" and "complementary" (84). Visual processing assumes a complementary role when it conveys information inaccessible through the auditory stream (the same applies here for articulatory dynamics), thus being particularly useful in adverse listening conditions (85, 86). Instead, it takes on a correlated role when it provides a redundant contribution (87, 88), which is especially the case under optimal listening conditions, and at best helps reduce cognitive demands during speech recognition (23). In line with this hypothesis, our data suggest that the contribution of the motor system to the perception of auditory speech at the theta-band timescale is comparable to that of the visual system in audiovisual speech. Indeed, when listening to ES stimuli, the fine-grained articulatory simulation is highly redundant to the encoding of the speech envelope [see the significantly positive clusters in Rdn(SE,PC1)]: in this sense motor processes assume a "correlated" (redundant) role with respect to acoustic encoding. Instead, HS stimuli forced listeners to acquire novel information from motor processes [see the significantly negative clusters in Unq(PC1)], which is not retrieved in any way from the speech envelope, thus supporting a "complementary" role.

Delta- and theta-band oscillations contribute distinctively to speech encoding, yet both encoded unique kinematic information [see Unq(PC1)]. Delta-band top-down signals in speech listening originate from the frontal and precentral gyri, whereas theta-band sources are located in the left precentral gyrus and posterior temporal area (25). Here, the topographical distribution of information patterns is suggestive of a possible dissociation of delta and theta sources in the unique articulatory encoding that could be explored in the future. However, the uniqueness of the kinematic encoding is bounded to the comparison with the acoustical feature that we utilized. Indeed, the speech envelope is one fundamental feature of speech recordings and particularly salient for the human brain, as its computation highly resembles the physical preprocessing undertaken by the cochlea in the inner ear. Nonetheless, it does not preserve the entirety of information in the acoustic input. This limitation extends to each PC of kinematic data, which merely provides a linear approximation of the full content of the signals. Consequently, addressing these challenges demands further scientific inquiry in the future.

Conclusions

Neural signals are known to track speech acoustics in both the delta and theta bands. Pastore et al. (26) found that the brain also encodes the articulatory movements of the speaker. Here we show that heightened task complexity involves a fundamental enhancement of articulatory encoding in the theta band. More importantly, these articulatory features relate to the tongue, which is (almost) never visually accessible to the listener either during conversations or during development. As a consequence, the encoding of articulatory features cannot emerge from passive exposure to environmental statistics, as in the case of audiovisual speech perception, but requires a speech-producing agent to learn the mapping between motor and sensory (acoustic and proprioceptive) feedback (89, 90). At the same time, the kinematic information of the tongue must necessarily be represented (at least part of it) in the acoustic stream for the participant to pick it up and possibly complete it through internal model reconstruction. Crucially, the present study and the previous one (26) demonstrate that (the combination of) acoustic cues reflecting (all or part of) tongue movements are particularly salient for brain speech processing. These results add more stringent evidence that motor processes contribute to speech perception (90–102).

DATA AVAILABILITY

As in the previous work (26), data are openly available in Mendeley Data at https://doi.org/10.17632/svy9m6987n.1. Source code for this study is openly available in GitHub at https://doi.org/10.5281/zenodo.10027877.

SUPPLEMENTAL MATERIAL

Supplemental Fig. S1: https://github.com/AlessandroCorsini/articEncod_difficulty/blob/main/SupplementalMaterial.pdf.

GRANTS

This work was supported by the BIAL Foundation–Grant for Scientific Research 2020 (No. 246/20) to A.T., 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 and 859588) to L.F.

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the authors.

AUTHOR CONTRIBUTIONS

A.T., L.F., and A.D. conceived and designed research; A.C. and A.P. analyzed data; A.C., A.T., and A.D. interpreted results of experiments; A.C. prepared figures; A.C. drafted manuscript; A.C., A.T., I.D., L.F., and A.D. edited and revised manuscript; A.C., A.T., A.P., I.D., L.F., and A.D. approved final version of manuscript.

REFERENCES

1. **Giraud AL**, **Poeppel D.** Cortical oscillations and speech processing: emerging computational principles and operations. *Nat Neurosci* 15: 511–517, 2012. doi:10.1038/nn.3063.

2. **Poeppel D**, **Assaneo MF.** Speech rhythms and their neural foundations. *Nat Rev Neurosci* 21: 322–334, 2020. doi:10.1038/s41583-020-0304-4.
3. **Luo H**, **Poeppel D.** Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron* 54: 1001–1010, 2007. doi:10.1016/j.neuron.2007.06.004.
4. **Meyer L.** The neural oscillations of speech processing and language comprehension: state of the art and emerging mechanisms. *Eur J Neurosci* 48: 2609–2621, 2018. doi:10.1111/ejn.13748.
5. **Obleser J**, **Kayser C.** Neural entrainment and attentional selection in the listening brain. *Trends Cogn Sci* 23: 913–926, 2019. doi:10.1016/j.tics.2019.08.004.
6. **Ahissar E**, **Nagarajan S**, **Ahissar M**, **Protopapas A**, **Mahncke H**, **Merzenich MM.** Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proc Natl Acad Sci USA* 98: 13367–13372, 2001. doi:10.1073/pnas.201400998.
7. **Ghitza O.** Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input rhythm. *Front Psychol* 2: 130, 2011. doi:10.3389/fpsyg.2011.00130.
8. **Kösem A**, **Bosker HR**, **Takashima A**, **Meyer A**, **Jensen O**, **Hagoort P.** Neural entrainment determines the words we hear. *Curr Biol* 28: 2867–2875.e3, 2018.doi:10.1016/j.cub.2018.07.023.
9. **Riecke L**, **Formisano E**, **Sorger B**, **Başkent D**, **Gaudrain E.** Neural entrainment to speech modulates speech intelligibility. *Curr Biol* 28: 161–169.e5, 2018. doi:10.1016/j.cub.2017.11.033.
10. **Zoefel B**, **Archer-Boyd A**, **Davis MH.** Phase entrainment of brain oscillations causally modulates neural responses to intelligible speech. *Curr Biol* 28: 401–408.e5, 2018. doi:10.1016/j.cub.2017.11.071.
11. **Chalas N**, **Daube C**, **Kluger DS**, **Abbasi O**, **Nitsch R**, **Gross J.** Multivariate analysis of speech envelope tracking reveals coupling beyond auditory cortex. *Neuroimage* 258: 119395, 2022. doi:10.1016/j.neuroimage.2022.119395.
12. **Ding N**, **Simon JZ.** Cortical entrainment to continuous speech: functional roles and interpretations. *Front Hum Neurosci* 8: 311, 2014. doi:10.3389/fnhum.2014.00311.
13. **Kayser SJ**, **Ince RA**, **Gross J**, **Kayser C.** Irregular speech rate dissociates auditory cortical entrainment, evoked responses, and frontal alpha. *J Neurosci* 35: 14691–14701, 2015. doi:10.1523/JNEUROSCI.2243-15.2015.
14. **Ding N**, **Melloni L**, **Zhang H**, **Tian X**, **Poeppel D.** Cortical tracking of hierarchical linguistic structures in connected speech. *Nat Neurosci* 19: 158–164, 2016. doi:10.1038/nn.4186.
15. **Warren RM.** Perceptual restoration of missing speech sounds. *Science* 167: 392–393, 1970. doi:10.1126/science.167.3917.392.
16. **Kösem A**, **van Wassenhove V.** Distinct contributions of low- and high-frequency neural oscillations to speech comprehension. *Lang Cogn Neurosci* 32: 536–544, 2017. doi:10.1080/23273798.2016.1238495.
17. **Cherry EC.** Some experiments on the recognition of speech, with one and with two ears. *J Acoust Soc Am* 25: 975–979, 1953. doi:10.1121/1.1907229.
18. **Ding N**, **Simon JZ.** Emergence of neural encoding of auditory objects while listening to competing speakers. *Proc Natl Acad Sci USA* 109: 11854–11859, 2012. doi:10.1073/pnas.1205381109.
19. **Mesgarani N**, **Chang EF.** Selective cortical representation of attended speaker in multi-talker speech perception. *Nature* 485: 233–236, 2012. doi:10.1038/nature11020.
20. **Vander Ghinst M**, **Bourguignon M**, **Op de Beeck M**, **Wens V**, **Marty B**, **Hassid S**, **Choufani G**, **Jousmäki V**, **Hari R**, **Van Bogaert P**, **Van Goldman S**, **De Tiège X.** Left superior temporal gyrus is coupled to attended speech in a cocktail-party auditory scene. *J Neurosci* 36: 1596–1606, 2016. doi:10.1523/JNEUROSCI.1730-15.2016.
21. **McGurk H**, **MacDonald J.** Hearing lips and seeing voices. *Nature* 264: 746–748, 1976. doi:10.1038/264746a0.
22. **Schroeder CE**, **Lakatos P**, **Kajikawa Y**, **Partan S**, **Puce A.** Neuronal oscillations and visual amplification of speech. *Trends Cogn Sci* 12: 106–113, 2008. doi:10.1016/j.tics.2008.01.002.
23. **Peelle JE**, **Sommers MS.** Prediction and constraint in audiovisual speech perception. *Cortex* 68: 169–181, 2015. doi:10.1016/j.cortex.2015.03.006.
24. **Park H**, **Kayser C**, **Thut G**, **Gross J.** Lip movements entrain the observers' low-frequency brain oscillations to facilitate speech intelligibility. *eLife* 5: e14521, 2016. doi:10.7554/eLife.14521.
25. **Park H**, **Ince RA**, **Schyns PG**, **Thut G**, **Gross J.** Frontal top-down signals increase coupling of auditory low-frequency oscillations to continuous speech in human listeners. *Curr Biol* 25: 1649–1653, 2015. doi:10.1016/j.cub.2015.04.049.
26. **Pastore A**, **Tomassini A**, **Delis I**, **Dolfini E**, **Fadiga L**, **D'Ausilio A.** Speech listening entails neural encoding of invisible articulatory features. *Neuroimage* 264: 119724, 2022. doi:10.1016/j.neuroimage.2022.119724.
27. **D'Ausilio A**, **Bartoli E**, **Maffongelli L**, **Berry JJ**, **Fadiga L.** Vision of tongue movements bias auditory speech perception. *Neuropsychologia* 63: 85–91, 2014. doi:10.1016/j.neuropsychologia.2014.08.018.
28. **Choi D**, **Yeung HH**, **Werker JF.** Sensorimotor foundations of speech perception in infancy. *Trends Cogn Sci* 27: 773–784, 2023. doi:10.1016/j.tics.2023.05.007.
29. **Canevari C**, **Badino L**, **Fadiga L.** A new Italian dataset of parallel acoustic and articulatory data. *Sixteenth Annual Conference of the International Speech Communication Association*. Dresden, Germany, September 6–10, 2015.
30. **Perrier P**, **Fuchs S.** Motor equivalence in speech production. In: *The Handbook of Speech Production*, edited by Redford MA. Chichester, UK: Wiley, 2015, p. 223–247.
31. **Williams PL**, **Beer RD.** Nonnegative decomposition of multivariate information. *arXiv* arXiv.1004.2515, 2023. doi:10.48550/arXiv.1004.2515.
32. **Berry JJ.** Accuracy of the NDI wave speech research system. *J Speech Lang Hear Res* 54: 1295–1301, 2011. doi:10.1044/1092-4388(2011/10-0226).
33. **Savariaux C**, **Badin P**, **Samson A**, **Gerber S.** A comparative study of the precision of Carstens and Northern Digital Instruments electromagnetic articulographs. *J Speech Lang Hear Res* 60: 322–340, 2017. doi:10.1044/2016_JSLHR-S-15-0223.
34. **Oostenveld R**, **Fries P**, **Maris E**, **Schoffelen JM.** FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput Intell Neurosci* 2011: 156869, 2011. doi:10.1155/2011/156869.
35. **Gramfort A**, **Luessi M**, **Larson E**, **Engemann DA**, **Strohmeier D**, **Brodbeck C**, **Goj R**, **Jas M**, **Brooks T**, **Parkkonen L**, **Hämäläinen M.** MEG and EEG data analysis with MNE-Python. *Front Neurosci* 7: 267, 2013. doi:10.3389/fnins.2013.00267.
36. **Ince RA**, **Giordano BL**, **Kayser C**, **Rousselet GA**, **Gross J**, **Schyns PG.** A statistical framework for neuroimaging data analysis based on mutual information estimated via a gaussian copula. *Hum Brain Mapp* 38: 1541–1573, 2017. doi:10.1002/hbm.23471.
37. **Ince RA.** Measuring multivariate redundant information with pointwise common change in surprisal. *Entropy* 19: 318, 2017. doi:10.3390/e19070318.
38. **Park H**, **Ince RA**, **Schyns PG**, **Thut G**, **Gross J.** Representational interactions during audiovisual speech entrainment: redundancy in left posterior superior temporal gyrus and synergy in left motor cortex. *PLoS Biol* 16: e2006558, 2018. doi:10.1371/journal.pbio.2006558.
39. **Smith ZM**, **Delgutte B**, **Oxenham AJ.** Chimaeric sounds reveal dichotomies in auditory perception. *Nature* 416: 87–90, 2002. doi:10.1038/416087a.
40. **Shannon CE.** A mathematical theory of communication. *Bell Syst Tech J* 27: 379–423, 1948. doi:10.1002/j.1538-7305.1948.tb01338.x.
41. **Di Liberto GM**, **Lalor EC**, **Millman RE.** Causal cortical dynamics of a predictive enhancement of speech intelligibility. *Neuroimage* 166: 247–258, 2018. doi:10.1016/j.neuroimage.2017.10.066.
42. **Keitel A**, **Ince RA**, **Gross J**, **Kayser C.** Auditory cortical delta-entrainment interacts with oscillatory power in multiple fronto-parietal networks. *Neuroimage* 147: 32–42, 2017. doi:10.1016/j.neuroimage.2016.11.062.
43. **O'Sullivan JA**, **Power AJ**, **Mesgarani N**, **Rajaram S**, **Foxe JJ**, **Shinn-Cunningham BG**, **Slaney M**, **Shamma SA**, **Lalor EC.** Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cereb Cortex* 25: 1697–1706, 2015. doi:10.1093/cercor/bht355.
44. **McGill W.** Multivariate information transmission. *Trans IRE Prof Group Inf Theory* 4: 93–111, 1954. doi:10.1109/TIT.1954.1057469.
45. **Bell A.** The co-information lattice. *ICA 2003*. 2003, p. 5102.
46. **Timme N**, **Alford W**, **Flecker B**, **Beggs JM.** Synergy, redundancy, and multivariate information measures: an experimentalist's

perspective. *J Comput Neurosci* 36: 119–140, 2014. doi:10.1007/s10827-013-0458-4.
47. **Timme NM**, **Lapish C.** A tutorial for information theory in neuroscience. *eNeuro* 5: ENEURO.0052-18.2018, 2018. doi:10.1523/ENEURO.0052-18.2018.
48. **Maris E**, **Oostenveld R.** Nonparametric statistical testing of EEG- and MEG-data. *J Neurosci Methods* 164: 177–190, 2007. doi:10.1016/j.jneumeth.2007.03.024.
49. **Bosker HR**, **Ghitza O.** Entrained theta oscillations guide perception of subsequent speech: behavioural evidence from rate normalisation. *Lang Cogn Neurosci* 33: 955–967, 2018. doi:10.1080/23273798.2018.1439179.
50. **Bröhl F**, **Kayser C.** Delta/theta band EEG differentially tracks low and high frequency speech-derived envelopes. *Neuroimage* 233: 117958, 2021. doi:10.1016/j.neuroimage.2021.117958.
51. **Doelling KB**, **Arnal LH**, **Ghitza O**, **Poeppel D.** Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing. *Neuroimage* 85: 761–768, 2014. doi:10.1016/j.neuroimage.2013.06.035.
52. **Peelle JE**, **Davis MH.** Neural oscillations carry speech rhythm through to comprehension. *Front Psychol* 3: 320, 2012. doi:10.3389/fpsyg.2012.00320.
53. **Fetz EE**, **Finocchio DV**, **Baker MA**, **Soso MJ.** Sensory and motor responses of precentral cortex cells during comparable passive and active joint movements. *J Neurophysiol* 43: 1070–1089, 1980. doi:10.1152/jn.1980.43.4.1070.
54. **Stringer C**, **Pachitariu M**, **Steinmetz N**, **Reddy CB**, **Carandini M**, **Harris KD.** Spontaneous behaviors drive multidimensional, brainwide activity. *Science* 364: eaav7893, 2019. doi:10.1126/science.aav7893.
55. **Gwilliams L**, **King JR**, **Marantz A**, **Poeppel D.** Neural dynamics of phoneme sequences: position-invariant code for content and order. *Nat Commun* 13: 6606, 2022. doi:10.1038/s41467-022-34326-1.
56. **Daube C**, **Ince RA**, **Gross J.** Simple acoustic features can explain phoneme-based predictions of cortical responses to speech. *Curr Biol* 29: 1924–1937.e9, 2019.doi:10.1016/j.cub.2019.04.067.
57. **Kelso JA.** Synergies: atoms of brain and behavior. *Adv Exp Med Biol* 639: 83–91, 2009. doi:10.1007/978-0-387-77064-2_5.
58. **Bouchard KE**, **Mesgarani N**, **Johnson K**, **Chang EF.** Functional organization of human sensorimotor cortex for speech articulation. *Nature* 495: 327–332, 2013 [Erratum in *Nature* 498: 526, 2013]. doi:10.1038/nature11911.
59. **Gracco VL**, **Löfqvist A.** Speech motor coordination and control: evidence from lip, jaw, and laryngeal movements. *J Neurosci* 14: 6585–6597, 1994. doi:10.1523/JNEUROSCI.14-11-06585.1994.
60. **Krause PA**, **Kawamoto AH.** On the timing and coordination of articulatory movements: historical perspectives and current theoretical challenges. *Lang Linguist Compass* 14: e12373, 2020. doi:10.1111/lnc3.12373.
61. **Badino L**, **Canevari C**, **Fadiga L**, **Metta G.** Integrating articulatory data in deep neural network-based acoustic modeling. *Comput Speech Lang* 36: 173–195, 2016. doi:10.1016/j.csl.2015.05.005.
62. **Ghosh PK**, **Goldstein LM**, **Narayanan SS.** Processing speech signal using auditory-like filterbank provides least uncertainty about articulatory gestures. *J Acoust Soc Am* 129: 4014–4022, 2011. doi:10.1121/1.3573987.
63. **Ghosh PK**, **Narayanan S.** A generalized smoothness criterion for acoustic-to-articulatory inversion. *J Acoust Soc Am* 128: 2162–2172, 2010. doi:10.1121/1.3455847.
64. **Lammert A**, **Goldstein L**, **Narayanan S**, **Iskarous K.** Statistical methods for estimation of direct and differential kinematics of the vocal tract. *Speech Commun* 55: 147–161, 2013. doi:10.1016/j.specom.2012.08.001.
65. **Li M**, **Kim J**, **Lammert A**, **Ghosh PK**, **Ramanarayanan V**, **Narayanan S.** Speaker verification based on the fusion of speech acoustics and inverted articulatory signals. *Comput Speech Lang* 36: 196–211, 2016. doi:10.1016/j.csl.2015.05.003.
66. **Mitra V**, **Nam H**, **Espy-Wilson CY**, **Saltzman E**, **Goldstein L.** Retrieving tract variables from acoustics: a comparison of different machine learning strategies. *IEEE J Sel Top Signal Process* 4: 1027–1045, 2010. doi:10.1109/JSTSP.2010.2076013.
67. **Ramanarayanan V**, **Van Segbroeck M**, **Narayanan SS.** Directly data-derived articulatory gesture-like representations retain discriminatory information about phone categories. *Comput Speech Lang* 36: 330–346, 2016. doi:10.1016/j.csl.2015.03.004.
68. **Lee S**, **Bresch E**, **Narayanan S.** An exploratory study of emotional speech production using functional data analysis techniques. In: *Proceedings of the International Seminar on Speech Production*, 2006.
69. **Wrench A**, **Richmond K.** Continuous speech recognition using articulatory data. *Sixth International Conference on Spoken Language Processing, ICSLP 2000*. 2000.
70. **Lu C**, **Zong Y**, **Zheng W**, **Li Y**, **Tang C**, **Schuller BW.** Domain invariant feature learning for speaker-independent speech emotion recognition. *IEEE/ACM Trans Audio Speech Lang Process* 30: 2217–2230, 2022. doi:10.1109/TASLP.2022.3178232.
71. **Meyer L**, **Sun Y**, **Martin AE.** Synchronous, but not entrained: exogenous and endogenous cortical rhythms of speech and language processing. *Lang Cogn Neurosci* 35: 1089–1099, 2020. doi:10.1080/23273798.2019.1693050.
72. **Ershaid H**, **Lizarazu M**, **McLaughlin DJ**, **Cooke M**, **Simantiraki O**, **Koutsogiannaki M**, **Lallier M.** Listening effort contributes to cortical tracking of speech in adverse listening conditions (Preprint). *PsyArXiv* osf.io/ym8zb, 2023. doi:10.31234/osf.io/ym8zb.
73. **Keitel A**, **Gross J**, **Kayser C.** Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. *PLoS Biol* 16: e2004473, 2018. doi:10.1371/journal.pbio.2004473.
74. **Biau E**, **Schultz BG**, **Gunter TC**, **Kotz SA.** Left motor δ oscillations reflect asynchrony detection in multisensory speech perception. *J Neurosci* 42: 2313–2326, 2022. doi:10.1523/JNEUROSCI.2965-20.2022.
75. **Mohammadi Y**, **Graversen C**, **Østergaard J**, **Andersen OK**, **Reichenbach T.** 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: 1301–1311, 2023. doi:10.1162/jocn_a_02016.
76. **Slaats S**, **Weissbart H**, **Schoffelen JM**, **Meyer AS**, **Martin AE.** Delta-band neural responses to individual words are modulated by sentence processing. *J Neurosci* 43: 4867–4883, 2023. doi:10.1523/JNEUROSCI.0964-22.2023.
77. **Coopmans CW**, **de Hoop H**, **Hagoort P**, **Martin AE.** Effects of structure and meaning on cortical tracking of linguistic units in naturalistic speech. *Neurobiol Lang (Camb)* 3: 386–412, 2022. doi:10.1162/nol_a_00070.
78. **Kaufeld G**, **Bosker HR**, **Ten Oever S**, **Alday PM**, **Meyer AS**, **Martin AE.** Linguistic structure and meaning organize neural oscillations into a content-specific hierarchy. *J Neurosci* 40: 9467–9475, 2020. doi:10.1523/JNEUROSCI.0302-20.2020.
79. **Etard O**, **Reichenbach T.** Neural speech tracking in the theta and in the delta frequency band differentially encode clarity and comprehension of speech in noise. *J Neurosci* 39: 5750–5759, 2019. doi:10.1523/JNEUROSCI.1828-18.2019.
80. **Ghitza O.** On the role of theta-driven syllabic parsing in decoding speech: intelligibility of speech with a manipulated modulation spectrum. *Front Psychol* 3: 238, 2012. doi:10.3389/fpsyg.2012.00238.
81. **Peelle JE**, **Gross J**, **Davis MH.** Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cereb Cortex* 23: 1378–1387, 2013. doi:10.1093/cercor/bhs118.
82. **Stein BE**, **Meredith MA.** *The Merging of the Senses*. Cambridge, MA: The MIT Press, 1993.
83. **Sumby WH**, **Pollack I.** Visual contribution to speech intelligibility in noise. *J Acoust Soc Am* 26: 212–215, 1954. doi:10.1121/1.1907309.
84. **Campbell R.** The processing of audio-visual speech: empirical and neural bases. *Philos Trans R Soc Lond B Biol Sci* 363: 1001–1010, 2008. doi:10.1098/rstb.2007.2155.
85. **Crosse MJ**, **Di Liberto G**, **Lalor EC.** Eye can hear clearly now: inverse effectiveness in natural audiovisual speech processing relies on long-term crossmodal temporal integration. *J Neurosci* 36: 9888–9895, 2016. doi:10.1523/JNEUROSCI.1396-16.2016.
86. **Ross LA**, **Saint-Amour D**, **Leavitt VM**, **Javitt DC**, **Foxe JJ.** Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cereb Cortex* 17: 1 147–1153, 2007. doi:10.1093/cercor/bhl024.
87. **Crosse MJ**, **Butler JS**, **Lalor EC.** Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions. *J Neurosci* 35: 14195–14204, 2015. doi:10.1523/JNEUROSCI.1829-15.2015.

88. **O'Sullivan AE**, **Crosse MJ**, **Liberto GM**, **de Cheveigné A**, **Lalor EC.** Neurophysiological indices of audiovisual speech processing reveal a hierarchy of multisensory integration effects. *J Neurosci* 41: 4991–5003, 2021. doi:10.1523/JNEUROSCI.0906-20.2021.
89. **Bartoli E**, **Maffongelli L**, **Campus C**, **D'Ausilio A.** Beta rhythm modulation by speech sounds: somatotopic mapping in somatosensory cortex. *Sci Rep* 6: 31182, 2016. doi:10.1038/srep31182.
90. **D'Ausilio A**, **Maffongelli L**, **Bartoli E**, **Campanella M**, **Ferrari E**, **Berry J**, **Fadiga L.** Listening to speech recruits specific tongue motor synergies as revealed by transcranial magnetic stimulation and tissue-Doppler ultrasound imaging. *Philos Trans R Soc Lond B Biol Sci* 369: 20130418, 2014. doi:10.1098/rstb.2013.0418.
91. **Bartoli E**, **D'Ausilio A**, **Berry J**, **Badino L**, **Bever T**, **Fadiga L.** Listener-speaker perceived distance predicts the degree of motor contribution to speech perception. *Cereb Cortex* 25: 281–288, 2015. doi:10.1093/cercor/bht257.
92. **D'Ausilio A**, **Pulvermüller F**, **Salmas P**, **Bufalari I**, **Begliomini C**, **Fadiga L.** The motor somatotopy of speech perception. *Curr Biol* 19: 381–385, 2009. doi:10.1016/j.cub.2009.01.017.
93. **Fadiga L**, **Craighero L**, **Buccino G**, **Rizzolatti G.** Speech listening specifically modulates the excitability of tongue muscles: a TMS study. *Eur J Neurosci* 15: 399–402, 2002. doi:10.1046/j.0953-816x.2001.01874.x.
94. **Meister IG**, **Wilson SM**, **Deblieck C**, **Wu AD**, **Iacoboni M.** The essential role of premotor cortex in speech perception. *Curr Biol* 17: 1692–1696, 2007. doi:10.1016/j.cub.2007.08.064.
95. **Möttönen R**, **Watkins KE.** Motor representations of articulators contribute to categorical perception of speech sounds. *J Neurosci* 29: 9819–9825, 2009. doi:10.1523/JNEUROSCI.6018-08.2009.
96. **Murakami T**, **Restle J**, **Ziemann U.** Effective connectivity hierarchically links temporoparietal and frontal areas of the auditory dorsal stream with the motor cortex lip area during speech perception. *Brain Lang* 122: 135–141, 2012. doi:10.1016/j.bandl.2011.09.005.
97. **Nuttall HE**, **Kennedy-Higgins D**, **Hogan J**, **Devlin JT**, **Adank P.** The effect of speech distortion on the excitability of articulatory motor cortex. *Neuroimage* 128: 218–226, 2016. doi:10.1016/j.neuroimage.2015.12.038.
98. **Sato M**, **Tremblay P**, **Gracco VL.** A mediating role of the premotor cortex in phoneme segmentation. *Brain Lang* 111: 1–7, 2009. doi:10.1016/j.bandl.2009.03.002.
99. **Schmitz J**, **Bartoli E**, **Maffongelli L**, **Fadiga L**, **Sebastian-Galles N**, **D'Ausilio A.** Motor cortex compensates for lack of sensory and motor experience during auditory speech perception. *Neuropsychologia* 128: 290–296, 2019. doi:10.1016/j.neuropsychologia.2018.01.006.
100. **Smalle EH**, **Rogers J**, **Möttönen R.** Dissociating contributions of the motor cortex to speech perception and response bias by using transcranial magnetic stimulation. *Cereb Cortex* 25: 3690–3698, 2015. doi:10.1093/cercor/bhu218.
101. **Watkins KE**, **Strafella AP**, **Paus T.** Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia* 41: 989–994, 2003. doi:10.1016/S0028-3932(02)00316-0.
102. **Mukherjee S**, **Badino L**, **Hilt PM**, **Tomassini A**, **Inuggi A**, **Fadiga L**, **Nguyen N**, **D'Ausilio A.** The neural oscillatory markers of phonetic convergence during verbal interaction. *Hum Brain Mapp* 40: 187–201, 2019. doi:10.1002/hbm.24364.