


## RESEARCH ARTICLE

# The neural oscillatory markers of phonetic convergence during verbal interaction

Sankar Mukherjee<sup>1</sup>  | Leonardo Badino<sup>1</sup> | Pauline M. Hilt<sup>1</sup> | Alice Tomassini<sup>1</sup> | Alberto Inuggi<sup>4</sup> | Luciano Fadiga<sup>1,3</sup> | Noël Nguyen<sup>2</sup> | Alessandro D'Ausilio<sup>1,3</sup>

<sup>1</sup>Center for Translational Neurophysiology of Speech and Communication, Istituto Italiano di Tecnologia, Ferrara, Italy

<sup>2</sup>CNRS, LPL, Aix Marseille University, Aix-en-Provence, France

<sup>3</sup>Section of Human Physiology, University of Ferrara, Ferrara, Italy

<sup>4</sup>Center for Human Technologies, Istituto Italiano di Tecnologia, Genoa, Italy

## Correspondence

Sankar Mukherjee, Istituto Italiano di Tecnologia, Center for Translational Neuroscience of Speech and Communication, Via Fossato Di Mortara, 19 Ferrara, FE, 44121, Italy.

Email: sankar1535@gmail.com

## Funding information

Brain & Language Research Institute (BLRI), Grant/Award Number: ANR-11-LABX-0036; Excellence Initiative of Aix-Marseille University (A\*MIDEX); Institute of Language Communication and the Brain (ILCB), Grant/Award Number: ANR-16-CONV-0002

## Abstract

During a conversation, the neural processes supporting speech production and perception overlap in time and, based on context, expectations and the dynamics of interaction, they are also continuously modulated in real time. Recently, the growing interest in the neural dynamics underlying interactive tasks, in particular in the language domain, has mainly tackled the temporal aspects of turn-taking in dialogs. Besides temporal coordination, an under-investigated phenomenon is the implicit convergence of the speakers toward a shared phonetic space. Here, we used dual electroencephalography (dual-EEG) to record brain signals from subjects involved in a relatively constrained interactive task where they were asked to take turns in chaining words according to a phonetic rhyming rule. We quantified participants' initial phonetic fingerprints and tracked their phonetic convergence during the interaction via a robust and automatic speaker verification technique. Results show that phonetic convergence is associated to left frontal alpha/low-beta desynchronization during speech preparation and by high-beta suppression before and during listening to speech in right centro-parietal and left frontal sectors, respectively. By this work, we provide evidence that mutual adaptation of speech phonetic targets, correlates with specific alpha and beta oscillatory dynamics. Alpha and beta oscillatory dynamics may index the coordination of the “when” as well as the “how” speech interaction takes place, reinforcing the suggestion that perception and production processes are highly interdependent and co-constructed during a conversation.

## 1 | INTRODUCTION

As two individuals engage in social interaction, they become part of a complex system whose information flow is mediated by visible behavior, prior knowledge, motivations, inferences about the partner's mental states, and history of prior interactions (Schilbach et al., 2013). Linguistic communication is indeed, among other characteristics, a mind-reading exercise requiring the formulation of hypotheses about the mental states of the speaker (Sperber & Wilson, 1986). Although dialog is undeniably the primary form of language use (Levinson, 2016), previous research has mostly investigated the neural processes subtending either speech production or speech perception, separately and almost ignoring the dynamics of the interaction (Price, 2012). This approach has provided fundamental insight on the neural substrates of the speech units for perception and production but the principles

and mechanisms that regulate their coordination during natural interaction are still unclear (Schoot et al., 2016).

The dynamical process of mutual adaptation which occurs at multiple levels is a key component of natural linguistic interaction that is crucially missing in classical laboratory tasks. One interesting phenomenon during linguistic interaction is that of alignment. Alignment refers to the interlocutors' tendency to converge toward similar mental states as the conversation progresses. Conversational success is indeed characterized by the shared understanding of the spoken content, speakers' mutual likability, background environment, and so forth. (Pickering & Garrod, 2004; Garnier et al. 2013). More interestingly, people involved in a dialog automatically and implicitly converge at multiple linguistic levels (Bilous & Krauss, 1988; Pardo, Jay, & Krauss, 2010) as well as with coverbal bodily gestures (West & Turner, 2010). For instance, agreeing interlocutors tend to copy each other's choices of sounds, words, grammatical constructions as well as the

temporal characteristics of speech. Nevertheless, this form of implicit behavioral alignment is still poorly understood, especially regarding its effects on communication efficacy, social and contextual determinants, and neural underpinnings (Stolk et al., 2016).

Although the investigations of these phenomena at the brain level are still sparse, few interesting studies ignited the exploration of inter-brains neural synchrony during conversation by following a hyperscanning approach (Hasson, Ghazanfar, Galantucci, Garrod, & Keysers, 2012). These studies showed that the speaker's brain activity is spatially and temporally coupled with that of the listener (Silbert, Honey, Simony, Poeppel, & Hasson, 2014), and that the degree of coupling and anticipation from the listener's side predicts comprehension (Stephens et al., 2010; Kuhlen et al., 2012; Dikker, Silbert, Hasson, & Zevin, 2014; Liu et al., 2017). Other studies instead explored how social factors affect the neurobehavioral pattern of communication. For example, it has been shown that neural activity in the left-inferior frontal cortex synchronizes between participants during face-to-face interaction (Jiang et al. 2012) and social role does play an important part in such interpersonal neural synchronization (Jiang et al., 2015).

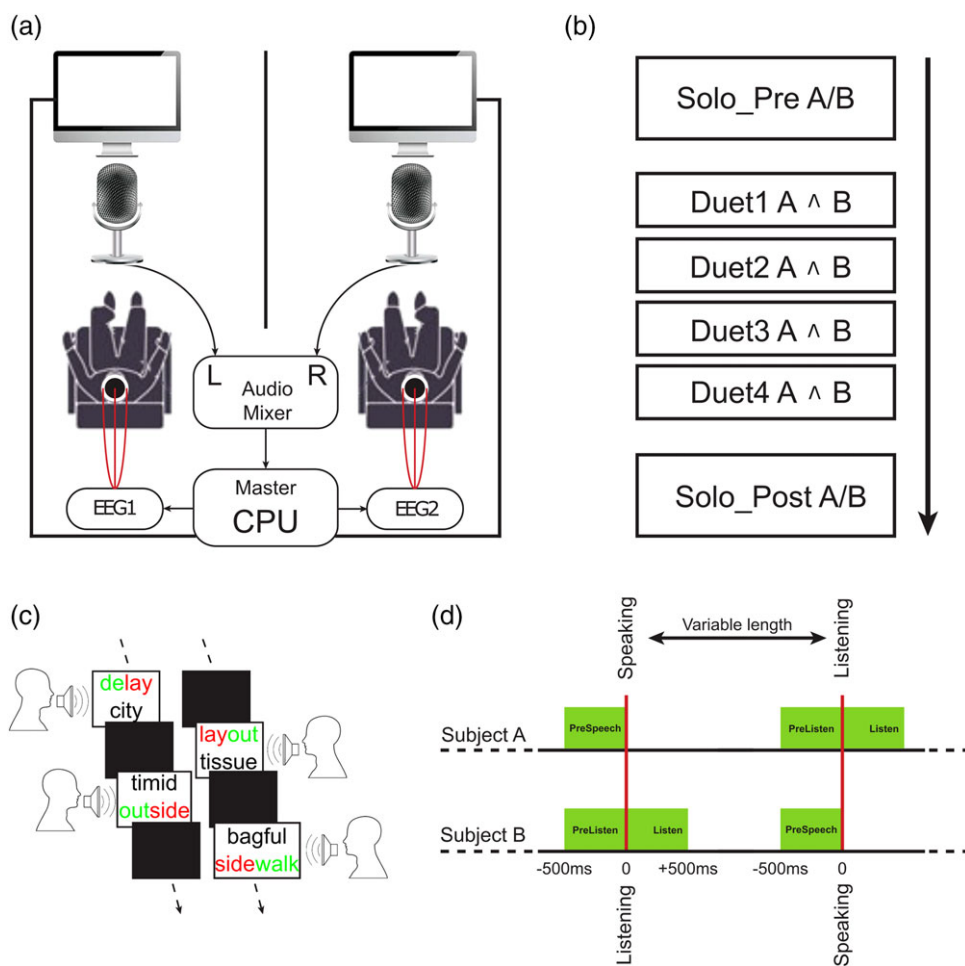
While techniques such as fMRI and fNIRS, due to their poor temporal resolution, are most suited to investigate alignment at the motivational, emotional or semantic level, the higher temporal resolution of EEG or magnetoencephalography (MEG) makes these techniques more suitable to investigate the intrinsically, faster, dynamics intervening during speech coordination. One aspect that has recently received attention regards the online negotiation of each-other turns during dialogs (turn-taking) which relies on an complex between-speaker neurobehavioral coordination (Levinson, 2016). For instance, in a task where subjects alternate in pronouncing letters of the alphabet, interbrain EEG oscillations in the theta/alpha (6–12 Hz) bands were synchronized in temporo-parietal regions as well as linked to behavioral speech synchronization indexes (Kawasaki, Yamada, Ushiku, Miyauchi, & Yamaguchi, 2013). More recently, a dual MEG/EEG study employing a similar number-counting task, reported alpha suppression in left temporal and right centro-parietal electrodes during speech interaction as opposed to a condition of speaking alone (Ahn et al., 2018). Using a dual-MEG set-up during natural conversations allowed to show that rolandic oscillations in the alpha (~10 Hz) and beta (~20 Hz) bands depended on the speaker's versus listener's role (Mandel, Bourguignon, Parkkonen, & Hari, 2016). In the left hemisphere, both bands were attenuated during production compared with listening. Before the speaker and listener swapped roles, power in the alpha band was briefly enhanced in the listener. These studies have thus begun to explore the intrabrain and interbrain oscillatory dynamics underlying one key behavioral coordination aspect during speech interaction, which is our ability to accommodate to the temporal properties of our partners' speech (Jungers & Hupp, 2009).

Beside temporal characteristics, speakers can shape how they speak and listen to speech. In fact, interlocutors adjust both their speech production and perception to their audience. For example, important adjustments are introduced while speaking to infants (Cooper & Aslin, 1990), to foreigners (Uther et al., 2007), or under adverse conditions (i.e., hearing loss or environmental noise; Payton et al., 1994). Adjustments in speech production consider listeners' perceptual salience and effort in listening (Lindblom, 1990). Listeners

adapt to talker characteristics (Nygaard et al., 1994; Bradlow & Bent, 2008), suggesting also great flexibility in speech processing mechanisms (Samuel & Kraljic, 2009). When engaged in a conversation instead, interlocutors align (or converge) their phonetic realizations to each other. Phonetic convergence then, amounts to the gradual and mutual adaptation of our speech targets toward a phonetic space shared by our interlocutor.

In the present study, we aimed at investigating the neural signature of such dynamic phonetic alignment. We asked pairs of participants to engage in an interactive speech game while dual-EEG was recorded. By this manner, we aimed at investigating interpersonal action-perception loops where one person's action (speech articulation) transforms into the sensory input (speech sound) for the other participant, and vice-versa. To this purpose, we used the Verbal Domino Task (VDT) (Bailey & Lelong, 2010), a fast-paced and engaging speech game allowing a relatively well controlled interaction and involving a turn-based phonetic exchange. At each turn, speakers are presented with a pre-selection of two written words and have to choose and read out the one that begins with the same syllable as the final syllable of the word previously uttered by their interlocutor (see Figure 1c). These constraints make the task different from a natural conversation, but contain the fundamental phonetic interactive component we needed for the present investigation. Before the interactive task, each participant's initial phonetic fingerprint was statistically modeled. The phonetic fingerprint was extracted from the individual acoustic spectral properties. We then computed how well the model of one speaker can identify the speech data of her/his interlocutor, at the single word level. In this sense, identification performance of these models allows to estimate how similar the two speakers are during the interactive task. We adopted a conservative criterion for convergence, which as defined as all instances where both participants adapted their speech properties to get closer to each other. All other cases were treated as nonconvergence.

The EEG analysis focused on the oscillatory power modulations during phonetic convergence as compared to nonconvergence, in three different epochs of interest. The first epoch, locked to speech production onset, was selected to investigate the preparatory activities in the speaking brain. The second and third epochs targeted the listening brain's activities: the ongoing brain activities prior to listening and those induced by speech listening. We expect that phonetic convergence could be associated with modulations of oscillatory activity in the alpha and beta range, two prevalent rhythmic modes of the brain, that have been already involved in natural conversation and, in particular, in speech turn-taking (Mandel et al., 2016). In addition, based on previous studies, we can hypothesize a dissociation between effects in the alpha and beta ranges, depending on the role each participant is playing at any given time point. The speaker role would elicit greater suppression in the alpha range, as often reported in behavioral synchronization tasks (Tognoli & Kelso, 2015). By contrast, the listener role would produce beta desynchronization, which has been associated to sensorimotor transformations during speech listening tasks (Bartoli, Maffongelli, Campus, & D'Ausilio, 2016) and top-down predictive coding of upcoming sensory signals (Cope et al., 2017).



**FIGURE 1** (a) Graphical depiction of the experimental setup. (b) Schematic illustration of the experimental timeline including two solo recordings before and after four duet sessions. Here, a/b indicates recording the speech of a and b separately, whereas a ^ b means recording of both together. (c) An example of the sequences of words produced during the verbal domino task by the two speakers. (d) Schematic illustration of the triggers for the EEG signals (red) and the temporal windows considered for the analyses (green) [Color figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

## 2 | MATERIALS AND METHODS

### 2.1 | Participants

A total of 16 healthy participants took part in the task (8 females, age,  $26 \pm 2.3$  years; mean  $\pm$  SD). All participants were right-handed with the exception of one male. They were all native Italian speakers unaware of the purpose of the experiment. We asked dyads of participants to perform the task in English with a view to augmenting the likelihood for these participants to converge toward each other at the phonetic level. Previous studies have shown that the use of a nonnative language induces greater phonetic convergence (Gambi & Pickering, 2013; Trofimovich & Kennedy, 2014) and greater sensorimotor compensatory activities while listening, compared with the native language (Schmitz et al., 2018). We asked participants to self-rate their English reading ( $7.87 \pm 1.08$ ), writing ( $7.31 \pm 0.95$ ), understanding ( $7.56 \pm 1.03$ ), and fluency ( $7.19 \pm 1.17$ , mean  $\pm$  SD) capabilities on a 1–10 scale. Self-reported proficiency in a nonnative language is routinely used in experimental psycholinguistic studies, and has been repeatedly shown to be closely related to a large variety of objective measures of proficiency (Gollan, Weissberger, Runnqvist, Montoya, &

Cera, 2012; Marian, Blumenfeld, & Kaushanskaya, 2007). Participants were grouped into 8 pairs (pair 1 to 8) before the start of the experiment. Groups consisted in 4 male–male and 4 female–female pairs. Within each pair, speakers will be referred to as A and B in what follows. The participants did not know each other nor interacted before the experiment. The study was approved by the local Ethics Committee and a written informed consent was obtained from the subjects according to the Declaration of Helsinki.

### 2.2 | Task and stimuli

We used the Verbal Domino Task (VDT, Mukherjee, D'Ausilio, Nguyen, Fadiga, & Badino, 2017; Mukherjee et al., 2018; adapted version from Bailly & Lelong, 2010) with English words. The VDT is an interactive, collaborative task jointly performed by two participants who pronounce a sequence of disyllabic words with an alternation between participants across words. Along the sequence, there is a match between the last syllable of each word and the first syllable of the following one, in its pronounced form, written form, or both. The task shows some resemblance with verbal games that are popular with children throughout the world, and that are referred to as

"Grab on Behind," "Last and First," or "Alpha and Omega," in English, and "Shiritori" in Japanese (Bailly & Martin, 2014). It shares some characteristics of conversational interactions, though fundamentally focusing on the phonetic aspect. Interestingly, participants speak one at time in alternating turns, while allowing us to both control the linguistic material used by the participants, and to avoid overlaps between the participants' turns. We are aware that the VDT shows important differences with respect to a natural conversation, but the no-overlap constraint is essential to temporally isolate brain activities associated to speaking and listening (see section 4).

During the task, the two speakers are alternatively presented with two words on a screen and must read aloud the word whose initial syllable coincides with the final syllable of the word previously produced by the other speaker. For example, on having heard Speaker A pronouncing the word "delay," Speaker B is offered to choose between "layoff" and "tissue" and is expected to pronounce "layoff" (see Figure 1c). The task is collaborative in that participants have to make the right choices for them to reach a common goal, that is, to jointly go through the domino chain up to the end. The choice of presenting the words visually, rather than allowing self-generation, was required to avoid participants using (a variable amount of) time "searching" for possible candidate words. In fact, this would have added a fundamental memory retrieval component that could not be controlled. Word self-generation can also introduce frequent stops in the chain. Moreover, the presentation of two alternatives, instead of a single option, forced participant to actively listen to their partner to select the correct word.

To build the word chain, we first selected disyllabic words from the WebCelex English lexical database (<http://celex.mpi.nl/>). Then, we sorted these words depending on spoken frequency (Collins Birmingham University International Language Database - COBUILD). The chain was built using a custom-made iterative algorithm in Matlab. The algorithm started from the highest frequency word and then looked for the next highest frequency item that both fulfilled the rhyming criterion and did not already occur in the chain. This frequency criterion was introduced to avoid low frequency, and thus, very specialized terminology that would have taxed participant second language skills. By this manner, we generated sequences of at least 200 items that were manually checked to exclude those with crude or offensive words. We finally selected one sequence of 200 unique disyllabic words. This sequence is freely available online at <https://www.gipsa-lab.grenoble-inp.fr/~gerard.bailly/Resources/DOMINOS.xlsx>.

## 2.3 | Procedure

For each pair, we recorded the speech and EEG signals of the two participants simultaneously. They were sitting in a quiet room with the experimenter monitoring the whole experiment.

The experiment was divided into three main sections (Figure 1b). Solo recordings were performed before (Solo\_Pre) and after (Solo\_Post) the Duet session. Solo data were needed to establish a participant-wise baseline. During the Solo task, the other participant wore noise insulating headphones. The Solo task required participants to pronounce 40 words randomly selected from the 200 word set of

the domino chain. Words were presented one after the other on a black screen and subjects had to read them out. Voice onset for each word triggered the appearance of the following word's written form with a random delay (0.6–1.5 s). This random delay was introduced to avoid anticipation and entrainment to an external rhythm of presentation. Each Solo session lasted about 2 min. The resulting dataset was made up of a total of 1,280 words (including all subjects).

In the Duet session, the task started with one word visually presented on the screen of one of the two participants (e.g., Participant A), while the other participant's screen was blank. When Participant A read the word aloud, her/his screen went immediately blank and two words appeared on Participant's B screen. Participant B chose that of the two words which best fulfilled the rhyming criterion and, as soon as she/he read that word aloud, her/his screen went blank, and two other words appeared on Subject A's screen. This cycle was repeated until the end of the list.

The 200 word Verbal Domino chain was divided into two lists of 100 words, both repeated twice so that the Duet part was composed by four separate blocks. During Duet blocks 1 and 2, Participant A started the Verbal Domino, whereas for Duet blocks 3 and 4, Participant B was the first. In each of the four blocks, the two participants spoke 50 words each, summing up to 200 words per participant for the Duet part and resulting in a total of 3,200 words. The duet session lasted about 25 min.

After each of the four duet blocks, participants could rest for about 2 min. After the resting period, we recorded 1 min of baseline EEG activity by asking participants to fixate a cross at the center of the black screen. The whole experiment was monitored by one experimenter who checked that participants correctly performed the tasks. Whenever one participant chose the wrong word in a Duet session, the task was halted. The experimenter would then tell the other participant which word he/she was to continue with.

## 2.4 | Data acquisition

Neural activities were recorded by a dual-EEG recording setup consisting of two Biosemi Active Two systems (Amsterdam, The Netherlands), each with 64 channels mounted onto an elastic cap according to the 10–20 international system. The left mastoid was used as online reference. EEG data were digitized at 1,024 Hz. Electrodes' sensitivity, as expressed by the Biosemi hardware, was kept below 20  $\mu$ V.

Three Central Processing Units (CPU) were used for the whole experiment: one master CPU for the control of experimental events, and two other CPUs to acquire EEG data from the two Biosemi systems. The master CPU controlled the presentation of the stimuli and the detection of the voice onset, and sent triggers to the two other CPUs via a parallel port. All of the operations of the master CPU were controlled with Psychtoolbox-3 running in the Matlab environment.

Speech data were also recorded by the master CPU (16 bits, stereo, 44,100 Hz sampling frequency) using two high-quality microphones (AKG C1000S) and an external dedicated amplifier (M-Audio Fast Track USB II Audio Interface). Voice onset was detected using an adaptive energy-based speech detector (Reynolds, Quatieri, & Dunn, 2000). This detector tracks the noise energy floor of the input signal

and labels that signal as speech if any feature vector (computed over a 10 ms window) energy exceeds the current noise floor by a fixed energy threshold. For each participant, and before the experiment, we set up this threshold by controlling the microphone channel amplifier gain (M-Audio Fast Track USB II Audio Interface).

### 3 | AUTOMATIC ANALYSIS OF CONVERGENCE

#### 3.1 | Acoustic data preprocessing

Trials in which automatic voice trigger was incorrect (e.g., premature triggering by noise or failure of the triggering system because verbal response was not loud enough, wrong choice of words) were excluded from the analysis. This resulted in the removal of 33 out of 1,280 Solo words, and 98 out of 3,200 Duet words. As a result, we collected an average of  $387.75 \pm 16.36$  (mean  $\pm$  SD) words for each dyad in the Duet session and an average of  $68.19 \pm 18.76$  (mean  $\pm$  SD) words per participant in the Solo session. The number of incorrect word was small  $6.375$  ( $SD = 5.677$ ) and there was no correlation between number of incorrect word choice and number of convergence points ( $R = 0.25$ ;  $p = .54$ ). The low number of incorrect word choices indicates that the participants had little or no difficulty in performing the task in English.

Acoustic feature extraction was performed as follows. Periods of silence were discarded using an energy-based Speech Activity Detector. We then extracted MFCCs (Mel Frequency Cepstral Coefficients), which are a short-term power spectrum representation of sounds, based on a linear cosine transform of log power spectrum on a non-linear mel-scale of frequency, widely used in speech technology applications (Kinnunen & Li, 2010). This choice allowed us not to make a priori assumptions about which subset of acoustic features is associated with between-speaker phonetic convergence, and instead to exploit the entire informational content of the acoustic spectrum.

MFCCs were derived using the speech signal, segmented into 10 ms frames (5 ms overlap) and a Hamming window. The short-time magnitude spectrum, obtained by applying fast Fourier transform (FFT), was passed to a bank of 32 Mel-spaced triangular bandpass filters, spanning the frequency region from 0 Hz to 3,800 Hz. The outputs of all 32 filters were transformed into 12 static, 12 velocity, and 12 acceleration MFCCs with the 0th coefficients resulting in 39 MFCC dimensions in total. Velocity and acceleration features were included to incorporate information about the way the 12 static vectors varied over time. Finally, the distribution of these cepstral features was wrapped (Pelecanos & Sridharan, 2001) per word to the standard normal distribution to mitigate the effects of mismatch between microphones and recording environments.

#### 3.2 | GMM-UBM

To extract unbiased measures of convergence, we used a data-driven, text-independent, automatic speaker identification technique, based on Gaussian Markov Modeling (GMM) Universal Background Model (UBM). The Gaussian components model the underlying broad

phonetic features (i.e., MFCCs) that characterize a speaker's voice and are based on a well-understood statistical model (Reynolds et al., 2000). In previous work (Bailly & Martin, 2014), a similar method was used to extract phonetic convergence, with some important differences. In Bailly and Martin (2014), the model was trained and tested on phonemes, whereas we applied it at the whole word level.

We used the MSR Identity Toolbox (Sadjadi et al., 2013) for GMM-UBM modeling. A 32-component UBM was trained with the pooled Solo\_Pre speech data of all the participants (a total of 124,068 speech frames). Then, individual speaker-dependent models were obtained via maximum a posteriori (MAP) adaptation of the UBMs to the Solo\_Pre speech data of each speaker separately. The GMM-UBM has multiple hyperparameters and different settings of these hyperparameters can affect the performance of speaker-dependent models.

A cross-validation technique was used to choose the optimum hyper-parameter settings. Solo\_Post speech data were used as a validation set, and each speaker-dependent model's performance was verified against the UBM model (Figure 3a). Furthermore, to verify if there had been any prepost change in the acoustic properties of speech, due to the duet interaction, we further grouped the data within participant, within dyad, and across dyads. In principle, if the interaction has been able to affect the phonetic fingerprint of the participants, the cross-validation performance should be better within than across dyads. Differences were verified using Bonferroni corrected paired *t* tests.

#### 3.3 | Phonetic convergence computation

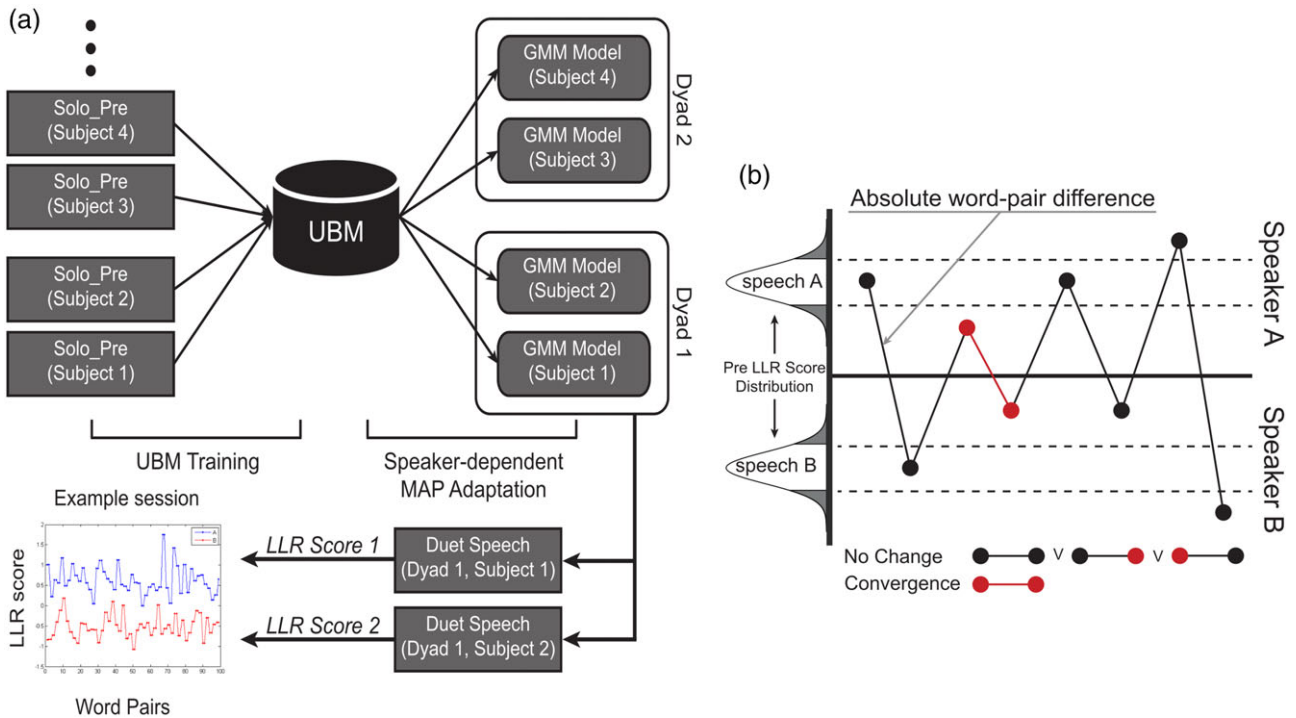
Phonetic convergence is computed on word pairs. For a word pair to be a convergent one, the acoustic properties of the words for the two speakers must become more similar to each other.

First of all, speaker-dependent models are grouped together according to their dyads, that is, if speaker A and speaker B interacted in the experiment, speaker-dependent model A and speaker-dependent model B are used for the following analysis (Figure 2a). During the duet, each word (i.e., the MFCCs of the speech) is tested with its corresponding grouped dyad models (Figure 2a). The test is performed by using the log-likelihood ratio score (LLR) which allows us to compare how well two statistical models can predict test samples. LLR of samples  $y_x$  ( $y$  is the MFCCs and  $x$  is the speaker identity) during the duet is computed using Equation (1):

$$LLR_{DUET}(y_x) = \log\left(\frac{p(y_x|H_A)}{p(y_x|H_B)}\right), \quad (1)$$

where  $H_A$  and  $H_B$  are the speaker-dependent models of speaker A and B, respectively. Now, when  $x$  is speaker A, LLR scores are positive (numerator greater than the denominator), whereas if  $x$  is speaker B, we get a negative score. The same computation, run on Solo\_Pre data, is then used to obtain the distribution of  $LLR_{PRE}$  scores which represents the baseline for each subject.

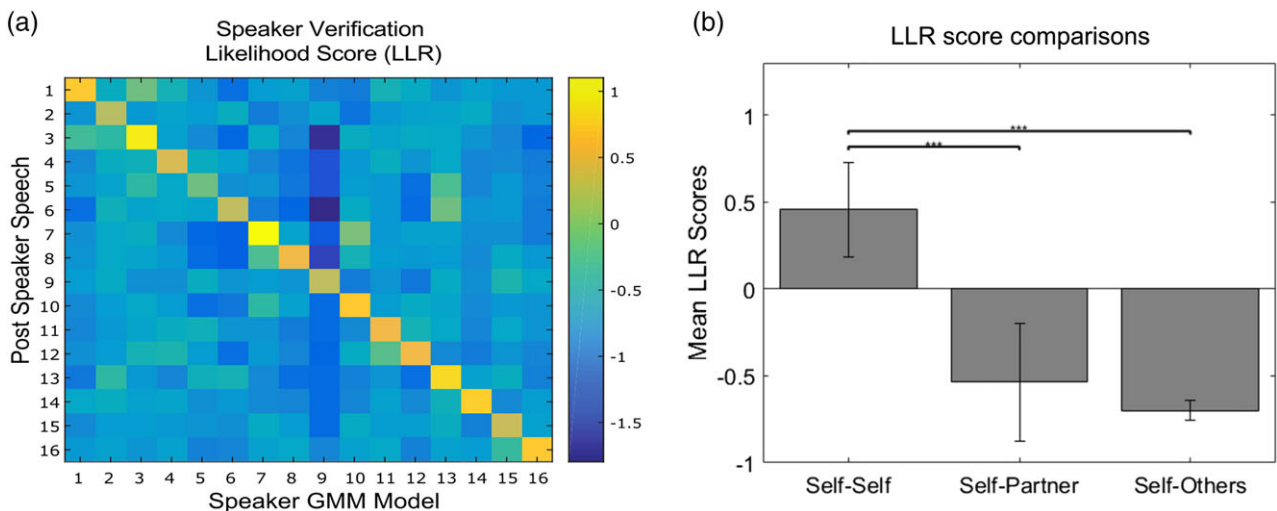
Then, to define convergence, we set two criteria that must be fulfilled at the same time. In the first one, we evaluate if the speech of both participants in the dyad becomes more similar in two consecutive words. A threshold on the  $LLR_{PRE}$  scores distribution allows us to consider only events for which  $LLR_{DUET}(y_x)$  becomes close to zero. In this case, as  $\log_{x-1}(x) = 0$ , the value of zero means that both speaker-



**FIGURE 2** (a) Schematic diagram of GMM-UBM modeling and how LLR score of test speech is predicted. (b) Graphical depiction of how we selected word pair points, based on how they relate to the distribution of PreSpeech LLR scores and how close they are to each other in the Duet condition [Color figure can be viewed at wileyonlinelibrary.com]

dependent models contribute equally to the prediction. In other words, the test speech is similar to both speakers. When at least two consecutive  $LLR_{DUET}(y_x)$  words fulfill this criterion, we consider them as Convergent (see Figure 2b). All other words are considered instead as NoChange. The second criterion controls that the convergent words in the Duet are not random phenomena (Ramseyer & Tschacher, 2010; Ward & Litman, 2007). We built 48 surrogate pairs (combining participants of the same gender only) from participants who never interacted

with each other in the task. We then ran the same computation as before, on the newly built surrogate pairs. This provides the distribution of surrogate consecutive  $LLR_{DUET}(y_x)$  difference scores, which we use to threshold the real consecutive  $LLR_{DUET}(y_x)$  difference scores and thus define true convergence. The threshold for both criteria was set at 1.5 SD. This threshold was set so that the words considered as Convergent were at the same time (1) extreme values along the continuum and (2) enough represented to be analyzed separately.



**FIGURE 3** (a) Speaker verification confusion matrix of all the speaker-dependent models against background UBM in the Solo\_Post. Here the diagonal positive score line indicates a good MAP adaptation. (b) The bar plot represents the same data in plot a, only by grouping the data within participant, within dyads, and across dyads. This plot shows the obvious better fit of the model when it is tested on the same speaker (self-self) and far lower performance when testing on another one [Color figure can be viewed at wileyonlinelibrary.com]

## 4 | EEG DATA ANALYSIS

### 4.1 | Preprocessing

EEG data were analyzed using the EEGLAB software (Delorme & Makeig, 2004), the Fieldtrip toolbox (Oostenveld, Fries, Maris, & Schoffelen, 2011) and custom-made MATLAB code. EEG data were first bandpass-filtered (two-pass Butterworth filter, fourth-order) between 0.1 and 40 Hz and then down-sampled to 256 Hz. Data recorded during speech production were discarded from the analysis because of strong speech-related artifacts. The remaining EEG data were first visually inspected for bad channels and/or artifacts in the time domain. Noisy channels were interpolated using a distance-weighted nearest-neighbor approach. To identify and remove artifacts related to participants' eye movements, eye blinks, and muscle activity, we used Independent Component Analysis (ICA) according to a consolidated approach (Delorme & Makeig, 2004). Finally, data were rereferenced using a common average reference over all electrodes.

EEG analyses of the Duet condition were constrained by the self-paced structure of the task which, by definition, made the timing of the events of interest (i.e., listening and speaking) not under experimental control. Given that speaking and listening phases alternated at a fast and variable rate, we restricted our analyses to short epochs of 500 ms that allowed us to avoid (1) artifacts due to speech production, and (2) temporal superposition of speech-related and listening-related neural processes.

We defined three 500 ms epochs of interest (Figure 1d):

1. Before speech production (PreSpeech): from  $-500$  to  $0$  ms relative to (one's own) voice onset.
2. Before speech listening (PreListen): from  $-500$  to  $0$  ms relative to (the partner's) voice onset.
3. During speech listening (Listen): from  $0$  to  $+500$  ms relative to the partner's voice onset.

### 4.2 | Time-frequency analysis

Time-frequency representations (TFRs) for the three different epochs (PreSpeech, PreListen, Listen) were calculated using a Fourier transform approach applied to short sliding time windows. All the epochs were zero-padded to avoid edge artifacts and spectral bleeding from contiguous EEG signal possibly contaminated by speech-related artifacts. The power values were calculated for frequencies between 8 and 40 Hz (in steps of 2 Hz) using a Hanning-tapered adaptive time window of 4 cycles ( $\Delta t = 4/f$ ) that was advanced in steps of 50 ms. This procedure results in a frequency-dependent spectral smoothing of  $\Delta f = 1/\Delta t$ . As a consequence of analyzing 500 ms epochs (see above) using 4 cycle time windows, the lowest frequency for which we could derive a power estimate (based on the entire epoch) was 8 Hz. In other terms, the relatively fast and self-paced nature of our task did not permit a reliable estimation of the power of slow oscillations (in the delta and theta frequency range).

### 4.3 | Statistical analysis

Statistical analysis was performed on the whole-brain oscillatory power (between 8 and 40 Hz). To evaluate statistically whether Convergence and NoChange data (as defined in the "Automatic analysis of convergence" section) showed a difference in oscillatory power, we performed a group-level nonparametric cluster-based permutation test (Maris & Oostenveld, 2007), separately for each epoch of interest (PreSpeech, PreListen, Listen). For every sample (here defined as [channel, frequency, time] triplet), a dependent-sample  $t$  value was computed. All samples for which this  $t$  value exceeded an a priori decided threshold (uncorrected  $p < .05$ ) were selected and subsequently clustered on the basis of temporal, spatial, and spectral contiguity. Then, cluster-level statistics was computed by taking the sum of  $t$  values in each cluster. The cluster yielding the maximum sum was subsequently used for evaluating the difference between the two data sets (with the maximum sum used as test statistic). We randomized the data across the two data sets, and for each random permutation (10,000 iterations), we calculated again the test statistics in the same way as previously described for the original data. This procedure generates a surrogate distribution of maximum cluster  $t$  values against which we can evaluate the actual data. The  $p$  value of this test is given by the proportion of random permutations that yields a larger test statistic compared to that computed for the original data.

## 5 | RESULTS

### 5.1 | Behavioral results and GMM-UBM performance

Turn-taking reaction time (RT) during the Duet sessions, measured as the time elapsed between visual presentation of words and voice onset, did not differ between NoChange ( $427 \pm 262$  ms, mean  $\pm$  SD) and Convergence ( $426 \pm 298$  ms, mean  $\pm$  SD) trials (Wilcoxon rank sum test:  $z = -0.46$ ,  $p = .64$ ). Turn-taking was self-paced and thus RTs are also a direct measure of the turn-taking pace and the rhythm established by the dyad. This analysis suggests that Convergence and NoChange trials share similar temporal turn-taking dynamics. Furthermore, the Pearson correlation between turn-taking RT and the number of convergence points did not show any significant relationship ( $R = .57$ ;  $p = .14$ ).

As far as the GMM-UBM modeling was concerned, we verified each speaker-dependent model's performance against the UBM model (Reynolds et al., 2000). The confusion matrix for the Post speech showed that modeling performance was good. This is measured using the equal error rate (EER) which indicates that the proportion of false acceptances is equal to the proportion of false rejections. The lower the EER value, the higher the accuracy of the classifiers. EER for the training is 2.26% and validation is 10.55% as shown in Figure 3a. As visible in Figure 3b, the speaker-dependent model's performance was far better when tested on the same subject (self-self). Instead, testing on the Solo\_Post of another speaker, led to critically lower performances (self-self vs. self-partner,  $t_{(30)} = 9.1$ ;  $p < .00001$ ; self-self vs. self-other,  $t_{(30)} = 16.63$ ;  $p < .00001$ ; self-partner

vs. self–other,  $t_{(30)} = 1.89$ ;  $p = .07$ ). This result suggests that the VDT did not affect the phonetic fingerprint of the participants.

The proportion of convergence points that fulfilled both convergence criteria in each dyad (see section on “automatic analysis of convergence”) was on average  $12.62 \pm 9.02\%$  (mean  $\pm$  SD). The large interindividual variability is consistent with previous reports showing that convergence may not equally distribute across dyads (Pardo, Urmanche, Wilman, & Wiener, 2017). It has also been reported that female dyads tend to converge more than male dyads (Bailly & Martin, 2014; Pardo, 2006). However, gender differences may potentially be affected by task and psycho-social factors and indeed in our data, female (114 words in total;  $22 \pm 15.56$ , mean  $\pm$  SD) and male dyads (88 in total;  $28.5 \pm 22.13$ , mean  $\pm$  SD), did not show any statistical difference in convergence (Wilcoxon rank sum test,  $p = .68$ ).

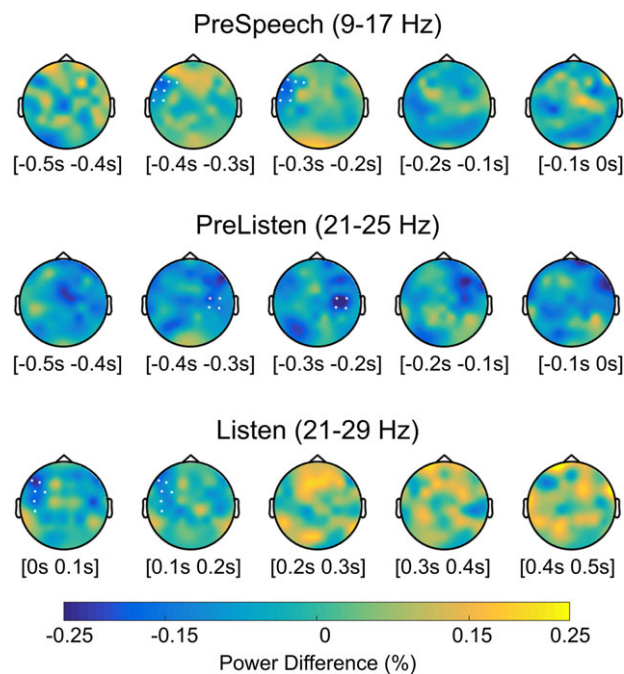
Finally, some studies in this area have modeled convergence as a linear process, that is, it grows as the conversation proceeds (Natale, 1975; Suzuki & Katagiri, 2007). However, subjects do not remain involved to the same degree over the whole course of a conversation, suggesting that convergence can be a dynamic phenomenon (Edlund et al., 2009; De Looze, Scherer, Vaughan, & Campbell, 2014). The one-way repeated-measures ANOVA with session (Duet1, Duet2, Duet3, Duet4) as within-subject factor did not reveal any significant effect ( $F_{(3,21)} = 2.63$ ,  $p = .08$ ), offering no conclusive evidence for a change over the experimental blocks.

## 5.2 | EEG results

The comparison between the oscillatory power in the Convergent and NoChange data sets for the three epochs of interest showed significant results that are summarized in Figure 4. Specifically, in the epoch preceding speech onset (PreSpeech), oscillatory power in the alpha/low beta band (9–17 Hz) was attenuated for Convergence compared to NoChange trials ( $p = .035$ ; see Figure 5). This alpha/low beta power suppression was more pronounced over left anterior scalp sites (F3, F5, F7, FT7, FC5, T7) and during early stages of speech preparation (from  $-400$  to  $-150$  ms relative to speech onset).

The observed power modulation did not depend on the reaction time (possibly indexing task difficulty), as no difference in reaction times was found between the two data sets ( $p = .64$ , see behavioral results). Moreover, we ensured that trial-by-trial fluctuations in reactions times were not associated with corresponding alpha/low beta-band power fluctuations. In fact, in all turn-taking behaviors, a confounding factor could be related to the temporal aspects of behavioral synchronization to the rhythm of the task (Fujioka, Trainor, Large, & Ross, 2012).

To this end, we calculated the Pearson correlation between single-trial reaction times and oscillatory power averaged across the time points (from  $-400$  to  $-150$  ms), frequencies (from 9 to 17 Hz) and electrodes (F3, F5, F7, FT7, FC5, T7) where we found the strongest power modulations between Convergence and NoChange (see above). Correlation was not significant for both data sets (Convergence,  $r = -.05$ ,  $p = .77$ ; NoChange,  $r = -.02$ ,  $p = .74$ ), confirming that the oscillatory activity that is modulated by phonetic convergence is not related to the (within-data set) variability in reaction times.



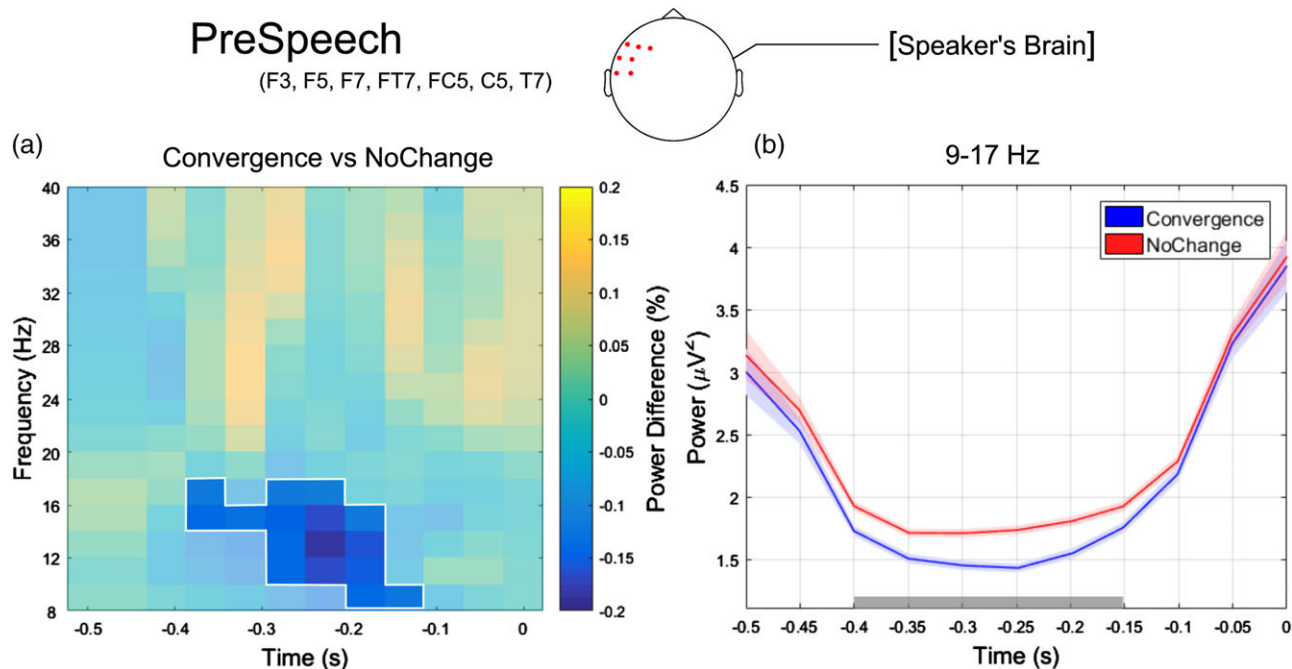
**FIGURE 4** Topographical plots of the relative power changes between convergence and NoChange ( $\frac{\text{Convergence} - \text{NoChange}}{\text{NoChange}}$ ) are shown for the frequency ranges for which the cluster-based permutation test yielded a significant difference. PreSpeech epoch refers to the preparation to speak. PreListen and Listen epochs instead refers to listener's brain activities, respectively, while the partner is speaking and before he/she speaks. Each topographic plot shows the change in power across the two data sets in 100 ms time windows, covering the entire 500 ms epoch of interest. The white dots mark the channels for which significant differences were found [Color figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

We investigated the effect of convergence in the listener's brain, before the interlocutor started speaking. With this analysis, we sought to establish whether a specific neural state, as indexed by the ongoing oscillatory power, preceded convergent words. We found a statistically significant difference in oscillatory power across data sets ( $p = .02$ ; Figure 6). Again, this difference consisted in a reduction of power for the Convergence compared to the NoChange data set, which was most consistent in the beta band (21–25 Hz), over right centro-parietal electrodes (C4, C6, CP6, CP4) and between  $-290$  and  $-190$  ms relative to the partner's voice onset.

The Pearson correlation showed no significant relationship between beta-band power (averaged across significant frequencies, electrodes, and time points) and reaction times at the single-trial level, indicating that beta power before listening does not covary with reaction time (Convergence:  $r = .08$ ,  $p = .18$ ; NoChange:  $r = .05$ ,  $p = .2$ ).

Finally, we also found a significant difference in oscillatory power between Convergence and NoChange in the listener ( $p = .03$ ; Figure 7). In particular, Convergence trials showed a reduction in power compared to NoChange trials that was most pronounced in the beta band (21–29 Hz), over left frontal electrodes (F5, F7, FC5, FC3, C5, CP5) and just after the partner's voice onset (50–120 ms; i.e., at the very beginning of the listening phase).





**FIGURE 5** (a) Relative power changes between convergence and NoChange ( $\frac{\text{Convergence} - \text{NoChange}}{\text{NoChange}}$ ) for the PreSpeech epoch, averaged over left anterior channels (F3, F5, F7, FT7, FC5, T7; channels showing statistically significant differences according to the cluster-based permutation test) and plotted as a function of time (−0.5–0 s) and frequency (8–40 Hz). The unshaded spectrotemporal region where converge trials show strongest and statistically significant power decrease compared to NoChange trials (i.e., 9–17 Hz frequency range and −0.4–0.15 s time window). (b) Temporal evolution of the oscillatory power averaged across frequencies ranging from 9 to 17 Hz and left anterior channels F3, F5, F7, FT7, FC5, T7 for convergence (blue line) and NoChange (red line). The gray horizontal line indicates the time points where the cluster-based permutation test revealed a significant difference between data sets. Colored shaded areas indicate standard error of the mean [Color figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

Again, we checked whether oscillatory power was related to reaction times in the two data sets. The Pearson correlation (calculated in the same way as already described for the PreSpeech epoch) yielded no significant relationship between trial-by-trial fluctuations in the beta-band power (in the 21–29 Hz range) and reaction times for both data sets (Convergence:  $r = -.17$ ,  $p = .97$ ; NoChange:  $r = .006$ ,  $p = .35$ ). As for the other epochs, our results are not driven by the rhythmic nature of the task and thus by behavioral synchronization in the temporal domain.

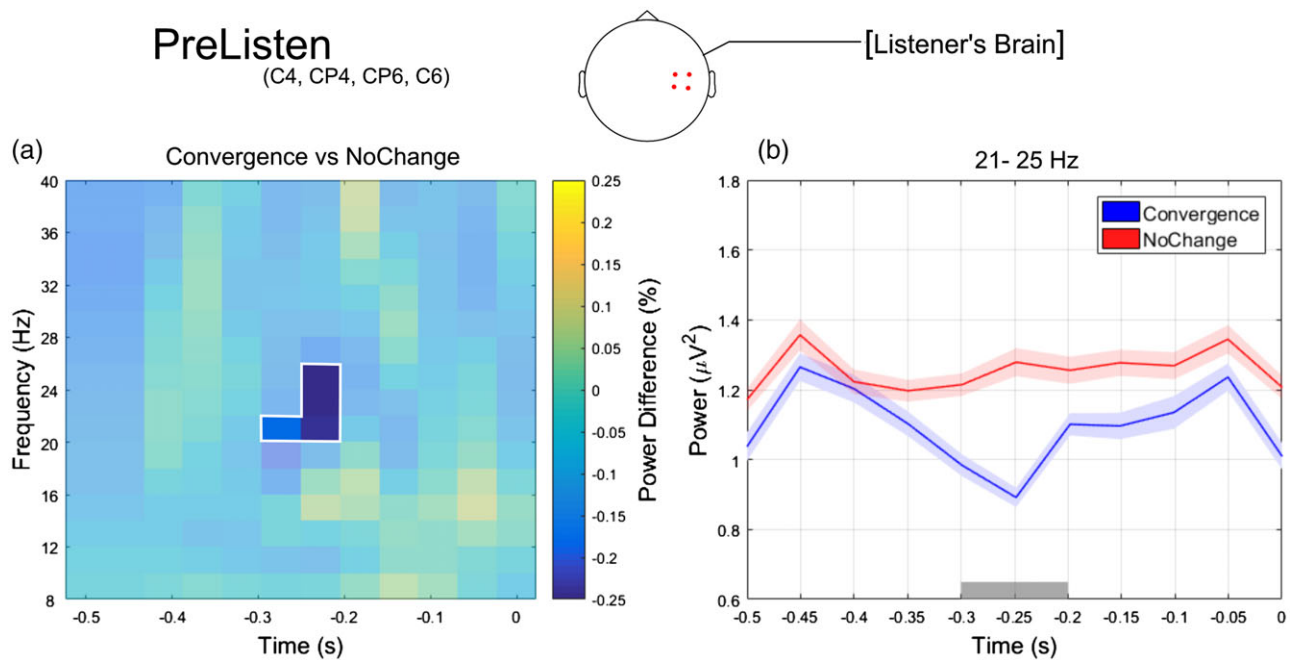
## 6 | DISCUSSION

Phonetic convergence is the phenomenon by which participants in a dialog tend to naturally align with each other in their phonetic characteristics (Pardo, 2013). Although convergence (phonetic alignment) is a well-known phenomenon, its quantitative assessment is still an open area of research. Several studies have focused on subjective evaluations (Pardo, 2013), whereas others have used a variety of objective acoustic measures (Goldinger, 1998). Therefore, a great deal of inconsistency and variability still exists among studies (Pardo, Urmache, Wilman, & Wiener, 2017). One key novelty of our study is that we implemented a quantitative method to extract phonetic convergence from a game-like task, allowing an engaging, yet relatively constrained, phonetic interaction. Phonetic convergence was computed using a robust and automatic speaker identification technique applied to the

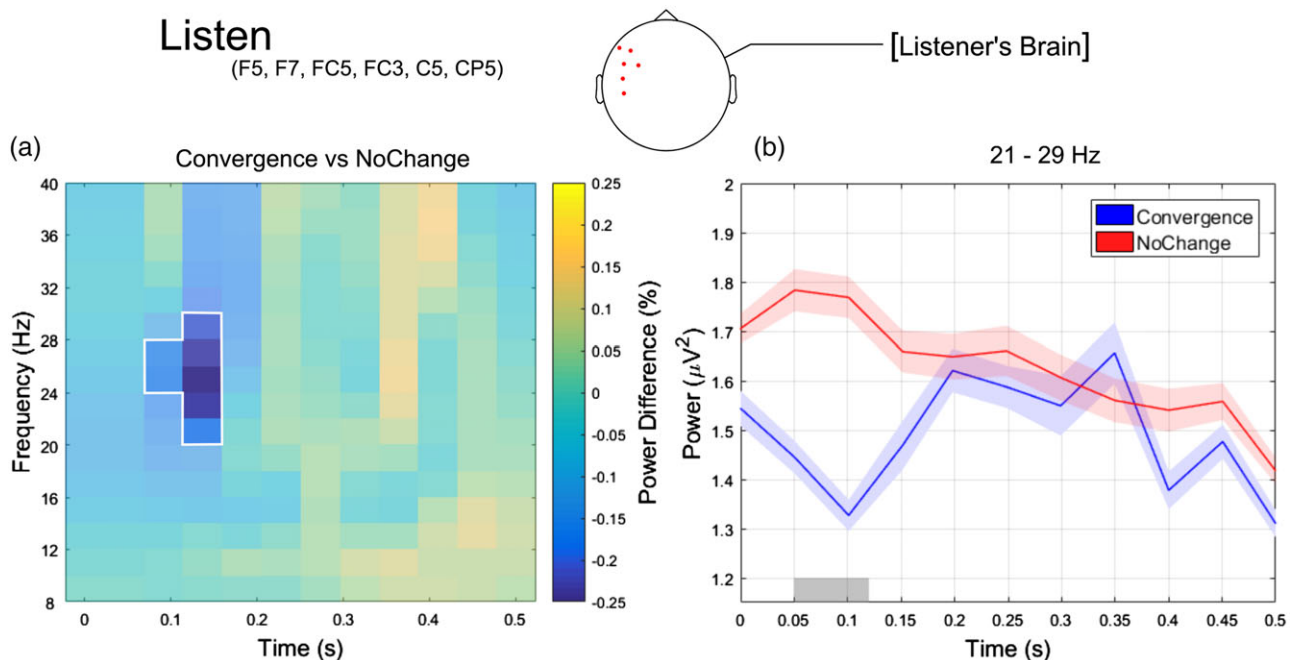
full acoustic spectrum, thus reducing the number of a priori hypotheses about which acoustic feature shows alignment (Mukherjee et al., 2017). This method was designed to specifically evaluate cooperative speech behavior. Indeed, phonetic convergence is not extracted from individual speech characteristics, but is rather computed out of the combination of both speakers' speech production. Therefore, we quantified participants' joint efforts to imitate each other's acoustic targets.

The quantification of phonetic convergence as the result of a joint-action behavior was the prerequisite to investigate its neural markers. Here, convergence was associated to specific oscillatory modulations in the alpha and beta bands. Convergent speech preparation in the speaker's brain was characterized by alpha/low beta power suppression which was most prominent over left fronto-central electrodes and early before speech onset (from −400 ms to −150 ms). Convergence in the listener's brain, instead, showed significant beta suppression peaking over left fronto-central sites just after the partner's speech onset (from 50 to 120 ms). At the same time, phonetic convergence is also characterized by lower power of the ongoing beta rhythm over right centro-parietal electrodes before listening. Overall, these findings suggest that alpha and beta oscillatory dynamics are associated with phonetic convergence.

These results are in line with previous studies reporting modulation of alpha (Kawasaki et al., 2013; Mandel et al., 2016; Pérez, Carreiras, & Duñabeitia, 2017; Ahn, et al., 2018) and beta rhythms (Mandel et al., 2016; Pérez et al., 2017) during speech-based



**FIGURE 6** (a) Relative power changes between convergence and NoChange ( $\frac{\text{Convergence} - \text{NoChange}}{\text{NoChange}}$ ) for the PreListen epoch, averaged over left anterior channels (C4 C6 CP6 CP4; channels showing statistically significant differences according to the cluster-based permutation test) and plotted as a function of time (–0.5–0 s) and frequency (8–40 HZ). The unshaded spectrotemporal region where converge trials show strongest and statistically significant power decrease compared to NoChange trials (i.e., 21–25 Hz frequency range and – 0.3–0.2 s time window). (b) Temporal evolution of the oscillatory power averaged across frequencies ranging from 21 to 25 Hz and right centro-parietal channels C4 C6 CP6 CP4 for Convergence (blue line) and NoChange (red line). The gray horizontal line indicates the time points where the cluster-based permutation test revealed a significant difference between data sets. Shaded areas indicate SE of the mean [Color figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]



**FIGURE 7** (a) Relative power changes between convergence and NoChange ( $\frac{\text{Convergence} - \text{NoChange}}{\text{NoChange}}$ ) for the Listen epoch, averaged over left anterior channels (F5, F7, FC5, FC3, C5, CP5; channels showing statistically significant differences according to the cluster-based permutation test) and plotted as a function of time (0–0.5 s) and frequency (8–40 HZ). The unshaded spectrotemporal region where converge trials show strongest and statistically significant power decrease compared to NoChange trials (i.e., 21–29 Hz frequency range and + 0.05–0.15 s time window). (b) Temporal evolution of the oscillatory power averaged across frequencies ranging from 21 to 29 Hz and left anterior channels F5, F7, FC5, FC3, C5, CP5 for Convergence (blue line) and NoChange (red line). The gray horizontal line indicates the time points where the cluster-based permutation test revealed a significant difference between data sets. Shaded areas indicate SE of the mean [Color figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

interaction tasks. However, one key aspect differentiates our study with respect to prior hyperscanning investigation of speech interaction. We used a joint-action behavioral feature as a searchlight for the neural underpinnings of speech coordination. In fact, our EEG analysis was driven by a behavioral index that cannot directly or independently be controlled by any of the partners during the interaction, that is, phonetic alignment.

## 6.1 | The sensorimotor nature of phonetic convergence

As far as the alpha/low beta effect in the speaker is concerned, we first observe that fronto-central desynchronization in the upper alpha and lower beta bands, always precedes voluntary movements (Leocani, Toro, Manganotti, Zhuang, & Hallett, 1997; Pfurtscheller & Aranibar, 1979; Pfurtscheller & Berghold, 1989). Interestingly, similar results can be observed across hyperscanning studies. Tognoli and Kelso (2015) used an interactive finger movement task and manipulated the subjects' view of each other's hands. The results showed that neural oscillations in the alpha range (the phi complex) were modulated by the control of participants' own behavior in relation to that of the partners. Following this pioneering dual-EEG study, a few others have confirmed the role of alpha oscillations, overlaying sensorimotor regions, in behavioral coordination (Dumas, Nadel, Sousignan, Martinerie, & Garnero, 2010; Konvalinka et al., 2014). In general, the comparison between interactive and noninteractive behaviors has consistently shown the suppression of alpha range oscillations (Tognoli & Kelso, 2015). However, task differences can produce slightly different topographical maps of alpha/low beta suppression. For instance, a centro-parietal topography in a joint attention task (Lachat & George, 2012), a frontal one in a finger-tapping task (Konvalinka et al., 2014), while a central effect was present in a nonverbal hand movement task (Ménoret et al., 2014). All in all, our fronto-central effect matches similar hyper-scanning results, while its left topography may be explained by the lateralization of the speech production function.

To discuss about the functional meaning of our results, we refer to the fact that a rolandic alpha desynchronization is usually found during execution, observation, or mental imagery of movements, possibly reflecting the activation or release from inhibition of the sensorimotor cortex (Caetano, Jousmäki, & Hari, 2007; Cochin, Barthelemy, Roux, & Martineau, 1999; Pfurtscheller & Da Silva, 1999). In fact, multi agent action coordination requires that participants produce their own actions, while simultaneously perceiving the actions of their partners. Similarly, a speech conversation creates the need for a tight action-perception coupling (Hari & Kujala, 2009). In fact, the central alpha band suppression has been proposed to be an index of action-perception coupling (De Lange et al., 2008; Hari, 2006), and thus sensorimotor information transfer during behavioral coordination. Within this context, our study provides evidence that alpha suppression, extending to the low beta range, is present also during speech interaction, in a task that critically requires coordination of articulatory gestures. More importantly, these EEG features were modulated by the efficacy with which participants jointly (as opposed to independently)

managed to coordinate each other while converging toward a shared phonetic space.

Moving to the listener's brain activities, phonetic convergence leads to the suppression of beta oscillations. In general, as for the rolandic alpha, fronto-central beta-band desynchronization has been related to the activation of the sensorimotor cortices (Parkes, Bastiaansen, & Norris, 2006; Salmelin, Hämäläinen, Kajola, & Hari, 1995). However, using electrocorticography (ECoG) it was shown that beta event-related desynchronization (ERD) is more focused and somatotopically specific than alpha ERD (Crone et al., 1998). In this sense, it has been proposed that the rolandic alpha ERD reflects the unspecific activation of sensorimotor areas, while the beta ERD signals a relatively more focal motor recruitment (Pfurtscheller & Da Silva, 1999; Pfurtscheller, Pregenzer, & Neuper, 1994). More specifically and in line with our findings, somatotopic beta attenuation has also been shown for speech listening (Jenson et al., 2014; Bartoli et al., 2016). In fact, specific sensorimotor regions recruited during speech production are also activated during speech listening (Fadiga, Craighero, Buccino, & Rizzolatti, 2002; Watkins et al., 2003; D'Ausilio et al., 2014) and the perturbation of these sensorimotor centers affects speech discrimination performance (Meister, Wilson, Deblieck, Wu, & Iacoboni, 2007; D'Ausilio et al., 2009; D'Ausilio, Bufalari, Salmas, & Fadiga, 2012; Bartoli et al., 2015; Möttönen & Watkins, 2009). The beta ERD we observe after speech presentation may thus be interpreted as supporting the perceptual discrimination processes. The effect is localized in a left fronto-central cluster of electrodes, supporting the claim that top-down sensorimotor predictions can exert a functional contribution to the hierarchical generative models underlying speech perception (Cope et al., 2017).

## 6.2 | Phonetic convergence and predictive coding

Beta ERD was also shown to precede speech listening, though with a right centro-parietal topography. This pattern of lateralization is consistent with a possible attentional role (Petit et al., 2007; Gao et al., 2017). In agreement with this possibility, it has been proposed that the functional role of the prestimulus beta rhythm is to convey motor information (efferent copies) to suppress self-generated sensory stimulations, freeing up resources to respond to external sensory stimuli (Engel & Fries, 2010). The mechanism of action might be that beta rhythms interact with other modality specific rhythms to anticipate sensory events by boosting neural excitability at specific moments in time when salient stimuli are expected to happen (Arnal, Wyart, & Giraud, 2011). Such a predictive top-down influence is supposed to play a key role in attentional selection (Bastos et al., 2015; Lee, Whittington, & Kopell, 2013; Lewis, Wang, & Bastiaansen, 2015; Morillon & Baillet, 2017). Interestingly, the pre-stimulus beta suppression has shown to be involved in predictions about the precision of a specific processing channel, thus, establishing the attentional context for perceptual processing (Bauer, Stenner, Friston, & Dolan, 2014).

In this framework, the brain acts like a predictive engine (Friston, Mattout, & Kilner, 2011), aiming at reducing the cost of analyzing the full set of incoming information, by formulating specific perceptual hypotheses that are tested against the temporal flow of sensory evidences (Donnarumma, Costantini, Ambrosini, Friston, & Pezzulo,

2017). In our turn-taking task, the organization of one's own speech output could bias the subsequent active listening processes by allowing faster or more efficient discrimination of similar acoustic targets. On the other hand, motor activations elicited by speech perception could in turn prime the organization of the immediately following speech planning required in the task. Based on general principles of neural reuse (Anderson, 2010) of action-perception circuits for speech communication (Pulvermüller, 2018) phonetic convergence may depend on the amount of sensorimotor detail extracted, while discriminating the speech produced by the partner. In this sense, the degree of neurofunctional sensorimotor overlap between speech perception and production may translate into larger likelihood of motor contagion (Bisio et al., 2014; Bisio, Stucchi, Jacono, Fadiga, & Pozzo, 2010; D'Ausilio et al., 2015).

### 6.3 | Predicting the “how” rather than the “when” of speech interaction

*Effective prediction however requires task predictability.* In our interactive task, the listeners have critical prior information to constrain perceptual analysis. From the listener's point of view, the word spoken by the partner shares one out of the two syllables of the word just produced by herself. The other syllable, the novel one, is contained in one of the two words that the participant can now read on the screen, and that she will have to pronounce. These task dependencies offer strong anchoring points to predict the dynamics of the ongoing interaction. Importantly, the listeners are forced into a predictive mode of operation regarding the phonetic content (i.e., what syllables I'm going to hear) rather than the timing characteristics of the turn-taking action. This aspect is of particular interest if we consider that previous studies investigated the neural dynamics subtending the estimation of “when” a partner is going to speak in a conversation (Mandel et al., 2016). In fact, the estimation of this temporal information is fundamental in establishing effective turn taking, as well as supporting word segmentation and parsing sentence-level syntax. However, temporal prediction may not be the only anticipatory mechanism at play during speech interaction. In the present study, we took a different direction by mixing a set of task constraints together with specific computational methods, to investigate how people engage in highly predictive behaviors regarding the (phonetic) “how” component of speech interaction. In this regard, phonological convergence has been here considered as the tendency to align phono-articulatory tract gestures during the interaction (Mukherjee et al., 2018).

Here, we show that phonetic convergence elicits specific patterns of alpha and beta suppressions that dissociate the speaking, preparing to listen and listening phases. The novelty of the current study arises also from the characterization of phonetic convergence as a dynamic and interactive process. In doing so, these results add to the few recent studies aiming at the investigation of speech and language processes during (quasi)-realistic verbal interactions. In fact, we need to bear in mind that in fast-paced natural dialogs, comprehension and production tend to greatly overlap in time (Levinson & Torreira, 2015). Based on this evidence, it has been suggested that one key issue is to which extent current models of language, developed for isolated individuals (Hickok & Poeppel, 2007), are still valid in interactive contexts

(Pickering & Garrod, 2013; Schoot et al., 2016). On one hand, turn-taking involves multitasking comprehension and production (Levinson, 2016) and indeed the neural network for language production and comprehension may at least partially overlap (Menenti et al., 2012). On the other, the now classical neural entrainment to surface auditory features during attentive listening (Luo & Poeppel, 2007; Schroeder & Lakatos, 2009; Giraud & Poeppel, 2012; Ding & Simon, 2014; Park, Kayser, Thut, & Gross, 2016) does not seem to fully explain interbrain synchronization occurring during conversations (Pérez et al., 2017). Therefore, when we extrapolate results to the complexity of ecological scenarios, the listener, apart from speech comprehension, may adapt the phono-articulatory properties of speech preparation through substantially incomplete understanding and at the same time may influence the speaker's brain processes through back-channeling.

## 7 | CONCLUSIONS

In conclusion, mutual understanding might be the result of a joint process whereby alignment of situation models is facilitated when interlocutors align their behavioral output (Pickering & Garrod, 2004; Schoot et al., 2016). Also, the fast-paced interactive nature of dialogs suggests that speech and language understanding and production form a shared process that is co-constructed by participants (Donnarumma, Dindo, Iodice, & Pezzulo, 2017). Along these lines, an emerging trend suggests that a complete grasp of the neural and cognitive processes involved in speech-based communication cannot be achieved without examining more realistic interactions among individuals (Hasson et al., 2012; Pickering & Garrod, 2013; Schoot et al., 2016). However, it is important to highlight that, to investigate the phonetic aspect of linguistic convergence, the current study implemented a series of task constraints to allow a moderate level of experimental control. Eventually, the investigation of whether more realistic and open-ended scenarios result in similar neurobehavioral phenomena will have to be tackled by future studies.

### ACKNOWLEDGMENT

We thank ANR-16-CONV-0002 (ILCB), ANR-11-LABX-0036 (BLRI), and the Excellence Initiative of Aix-Marseille University (A\*MIDEX) for support.

### ORCID

Sankar Mukherjee  <https://orcid.org/0000-0003-3927-6365>

### REFERENCES

- Anderson, M. L. (2010). Neural reuse: A fundamental organizational principle of the brain. *Behavioral and Brain Sciences*, 33(4), 245–266.
- Arnal, L. H., Wyart, V., & Giraud, A. L. (2011). Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. *Nature Neuroscience*, 14(6), 797–801.
- Ahn, S., Cho, H., Kwon, M., Kim, K., Kwon, H., Kim, B. S., ... Jun, S. C. (2018). Interbrain phase synchronization during turn-taking verbal interaction—a hyperscanning study using simultaneous EEG/MEG. *Human brain mapping*, 39(1), 171–188.

- Bailly, G., Lelong, A. (2010). Speech dominoes and phonetic convergence. In *11th Annual Conference of the International Speech Communication Association (Interspeech)* (pp. 1153–1156).
- Bailly, G., Martin, A. (2014). Assessing objective characterizations of phonetic convergence. In *15th Annual Conference of the International Speech Communication Association (Interspeech)* (pp. P-19).
- 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. *Cerebral Cortex*, *25*(2), 281–288.
- Bartoli, E., Maffongelli, L., Campus, C., & D'Ausilio, A. (2016). Beta rhythm modulation by speech sounds: Somatotopic mapping in somatosensory cortex. *Scientific Reports*, *6*, 31182.
- Bastos, A. M., Vezoli, J., Bosman, C. A., Schoffelen, J. M., Oostenveld, R., Dowdall, J. R., ... Fries, P. (2015). Visual areas exert feedforward and feedback influences through distinct frequency channels. *Neuron*, *85*(2), 390–401.
- Bauer, M., Stenner, M. P., Friston, K. J., & Dolan, R. J. (2014). Attentional modulation of alpha/beta and gamma oscillations reflect functionally distinct processes. *Journal of Neuroscience*, *34*(48), 16117–16125.
- Bilous, F. R., & Krauss, R. M. (1988). Dominance and accommodation in the conversational behaviors of same-and mixed-gender dyads. *Language & Communication*, *8*(3–4), 183–194.
- Bisio, A., Sciutti, A., Nori, F., Metta, G., Fadiga, L., Sandini, G., & Pozzo, T. (2014). Motor contagion during human-human and human-robot interaction. *PLoS One*, *9*(8), e106172.
- Bisio, A., Stucchi, N., Jacono, M., Fadiga, L., & Pozzo, T. (2010). Automatic versus voluntary motor imitation: Effect of visual context and stimulus velocity. *PLoS One*, *5*(10), e13506.
- Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707–729.
- Caetano, G., Jousmäki, V., & Hari, R. (2007). Actor's and observer's primary motor cortices stabilize similarly after seen or heard motor actions. *Proceedings of the National Academy of Sciences*, *104*(21), 9058–9062.
- Cochin, S., Barthelemy, C., Roux, S., & Martineau, J. (1999). Observation and execution of movement: Similarities demonstrated by quantified electroencephalography. *European Journal of Neuroscience*, *11*(5), 1839–1842.
- Cope, T. E., Sohoglu, E., Sedley, W., Patterson, K., Jones, P. S., Wiggins, J., ... Davis, M. H. (2017). Evidence for causal top-down frontal contributions to predictive processes in speech perception. *Nature Communications*, *8*(1), 2154.
- Cooper, R. P., & Aslin, R. N. (1990). Preference for infant-directed speech in the first month after birth. *Child Development*, *61*(5), 1584–1595.
- Crone, N. E., Miglioretti, D. L., Gordon, B., Sieracki, J. M., Wilson, M. T., Uematsu, S., & Lesser, R. P. (1998). Functional mapping of human sensorimotor cortex with electrocorticographic spectral analysis I. Alpha and beta event-related desynchronization. *Brain: A Journal Of Neurology*, *121*(12), 2271–2299.
- D'Ausilio, A., Badino, L., Cipresso, P., Chirico, A., Ferrari, E., Riva, G., & Gaggioli, A. (2015). Automatic imitation of the arm kinematic profile in interacting partners. *Cognitive Processing*, *16*(1), 197–201.
- D'Ausilio, A., Bufalari, I., Salmas, P., & Fadiga, L. (2012). The role of the motor system in discriminating normal and degraded speech sounds. *Cortex*, *48*(7), 882–887.
- D'Ausilio, A., Maffongelli, L., Bartoli, E., Campanella, M., Ferrari, E., Berry, J., & Fadiga, L. (2014). Listening to speech recruits specific tongue motor synergies as revealed by transcranial magnetic stimulation and tissue-Doppler ultrasound imaging. *Philosophical Transactions of the Royal Society B*, *369*(1644), 20130418.
- D'Ausilio, A., Pulvermüller, F., Salmas, P., Bufalari, I., Begliomini, C., & Fadiga, L. (2009). The motor somatotopy of speech perception. *Current Biology*, *19*(5), 381–385.
- De Lange, F. P., Jensen, O., Bauer, M., & Toni, I. (2008). Interactions between posterior gamma and frontal alpha/beta oscillations during imagined actions. *Frontiers in Human Neuroscience*, *2*, 7.
- De Looze, C., Scherer, S., Vaughan, B., & Campbell, N. (2014). Investigating automatic measurements of prosodic accommodation and its dynamics in social interaction. *Speech Communication*, *58*, 11–34.
- Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21.
- Dikker, S., Silbert, L. J., Hasson, U., & Zevin, J. D. (2014). On the same wavelength: Predictable language enhances speaker–listener brain-to-brain synchrony in posterior superior temporal gyrus. *Journal of Neuroscience*, *34*(18), 6267–6272.
- Ding, N., & Simon, J. Z. (2014). Cortical entrainment to continuous speech: Functional roles and interpretations. *Frontiers in Human Neuroscience*, *8*, 311.
- Donnarumma, F., Costantini, M., Ambrosini, E., Friston, K., & Pezzulo, G. (2017). Action perception as hypothesis testing. *Cortex*, *89*, 45–60.
- Donnarumma, F., Dindo, H., Iodice, P., & Pezzulo, G. (2017). You cannot speak and listen at the same time: A probabilistic model of turn-taking. *Biological Cybernetics*, *111*(2), 165–183.
- Dumas, G., Nadel, J., Soussignan, R., Martinerie, J., & Garnero, L. (2010). Inter-brain synchronization during social interaction. *PLoS One*, *5*, e12166.
- Edlund, J., Heldner, M., Hirschberg, J., 2009. Pause and gap length in face-to-face interaction. In *10th Annual Conference of the International Speech Communication Association*, pp. 2779–2782.
- Engel, A. K., & Fries, P. (2010). Beta-band oscillations—Signaling the status quo? *Current Opinion in Neurobiology*, *20*(2), 156–165.
- Fadiga, L., Craighero, L., Buccino, G., & Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles: A TMS study. *European Journal of Neuroscience*, *15*(2), 399–402.
- Friston, K., Mattout, J., & Kilner, J. (2011). Action understanding and active inference. *Biological Cybernetics*, *104*(1–2), 137–160.
- Fujioka, T., Trainor, L. J., Large, E. W., & Ross, B. (2012). Internalized timing of isochronous sounds is represented in neuromagnetic beta oscillations. *Journal of Neuroscience*, *32*(5), 1791–1802.
- Gambi, C., & Pickering, M. J. (2013). Prediction and imitation in speech. *Frontiers in Psychology*, *4*, 340.
- Gao, Y., Wang, Q., Ding, Y., Wang, C., Li, H., Wu, X., ... Li, L. (2017). Selective attention enhances Beta-band cortical oscillation to speech under “cocktail-party” listening conditions. *Frontiers in Human Neuroscience*, *11*, 34.
- Garnier, M., Lamalle, L., & Sato, M. (2013). Neural correlates of phonetic convergence and speech imitation. *Frontiers in psychology*, *4*, 600.
- Giraud, A. L., & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience*, *15*(4), 511.
- Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*, 251–279.
- Gollan, T. H., Weissberger, G. H., Runnqvist, E., Montoya, R. I., & Cera, C. M. (2012). Self-ratings of spoken language dominance: A multilingual naming test (MINT) and preliminary norms for young and aging Spanish–English bilinguals. *Bilingualism: Language and Cognition*, *15*(3), 594–615.
- Hari, R., & Kujala, M. V. (2009). Brain basis of human social interaction: From concepts to brain imaging. *Physiological Reviews*, *89*, 453–479.
- Hari, R. (2006). Action–perception connection and the cortical mu rhythm. *Progress in Brain Research*, *159*, 253–260.
- Hasson, U., Ghazanfar, A. A., Galantucci, B., Garrod, S., & Keysers, C. (2012). Brain-to-brain coupling: A mechanism for creating and sharing a social world. *Trends in Cognitive Sciences*, *16*, 114–121.
- Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature reviews. Neuroscience*, *8*(5), 393.
- Jenson, D., Bowers, A. L., Harkrider, A. W., Thornton, D., Cuellar, M., & Saltuklaroglu, T. (2014). Temporal dynamics of sensorimotor integration in speech perception and production: Independent component analysis of EEG data. *Frontiers in Psychology*, *5*, 656.
- Jiang, J., Dai, B., Peng, D., Zhu, C., Liu, L., & Lu, C. (2012). Neural synchronization during face-to-face communication. *Journal of Neuroscience*, *32*(45), 16064–16069.
- Jiang, J., Chen, C., Dai, B., Shi, G., Ding, G., Liu, L., & Lu, C. (2015). Leader emergence through interpersonal neural synchronization. In *Proceedings of the National Academy of Sciences* (p. 201422930).
- Jungers, M. K., & Hupp, J. M. (2009). Speech priming: Evidence for rate persistence in unscripted speech. *Language and Cognitive Processes*, *24*(4), 611–624.
- Kawasaki, M., Yamada, Y., Ushiku, Y., Miyauchi, E., & Yamaguchi, Y. (2013). Inter-brain synchronization during coordination of speech rhythm in human-to-human social interaction. *Scientific Reports*, *3*, 1692.

- Kinnunen, T., & Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication*, 52, 12–40.
- Konvalinka, I., Bauer, M., Stahlhut, C., Hansen, L. K., Roepstorff, A., & Frith, C. D. (2014). Frontal alpha oscillations distinguish leaders from followers: Multivariate decoding of mutually interacting brains. *NeuroImage*, 94, 79–88.
- Kuhlen, A. K., Allefeld, C., & Haynes, J. D. (2012). Content-specific coordination of listeners' to speakers' EEG during communication. *Frontiers in human neuroscience*, 6, 266.
- Lachat, F., & George, N. (2012). Oscillatory brain correlates of live joint attention: A dual-EEG study. *Frontiers in Human Neuroscience*, 6, 156.
- Lee, J. H., Whittington, M. A., & Kopell, N. J. (2013). Top-down beta rhythms support selective attention via interlaminar interaction: A model. *PLoS Computational Biology*, 9(8), e1003164.
- Leocani, L., Toro, C., Manganotti, P., Zhuang, P., & Hallett, M. (1997). Event-related coherence and event-related desynchronization/synchronization in the 10 Hz and 20 Hz EEG during self-paced movements. *Clinical Neurophysiology*, 104(3), 199–206.
- Levinson, S. C. (2016). Turn-taking in human communication—origins and implications for language processing. *Trends in Cognitive Sciences*, 20(1), 6–14.
- Levinson, S. C., & Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, 6, 731.
- Lewis, A. G., Wang, L., & Bastiaansen, M. (2015). Fast oscillatory dynamics during language comprehension: Unification versus maintenance and prediction? *Brain and Language*, 148, 51–63.
- Liu, Y., Piazza, E. A., Simony, E., Shewokis, P. A., Onaral, B., Hasson, U., & Ayaz, H. (2017). Measuring speaker–listener neural coupling with functional near infrared spectroscopy. *Scientific Reports*, 7, 43293.
- Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In *Speech production and speech modelling* (pp. 403–439). Dordrecht: Springer.
- Luo, H., & Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron*, 54(6), 1001–1010.
- Mandel, A., Bourguignon, M., Parkkonen, L., & Hari, R. (2016). Sensorimotor activation related to speaker vs. listener role during natural conversation. *Neuroscience Letters*, 614, 99–104.
- Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The language experience and proficiency questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research*, 50(4), 940–967.
- Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, 164, 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. *Current Biology*, 17(19), 1692–1696.
- Menenti, L., Garrod, S. C., & Pickering, M. J. (2012). Toward a neural basis of interactive alignment in conversation. *Frontiers in Human Neuroscience*, 6, 185.
- Ménoret, M., Varnet, L., Fargier, R., Cheylus, A., Curie, A., Des Portes, V., ... Paulignan, Y. (2014). Neural correlates of non-verbal social interactions: A dual-EEG study. *Neuropsychologia*, 55, 85–97.
- Morillon, B., & Baillet, S. (2017). Motor origin of temporal predictions in auditory attention. *Proceedings of the National Academy of Sciences*, 114(42), E8913–E8921.
- Möttönen, R., & Watkins, K. E. (2009). Motor representations of articulators contribute to categorical perception of speech sounds. *Journal of Neuroscience*, 29(31), 9819–9825.
- Mukherjee, S., D'Ausilio, A., Nguyen, N., Fadiga, L., & Badino, L. (2017). The relationship between F0 synchrony and speech convergence in dyadic interaction. *Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech*, 2341–2345.
- Mukherjee, S., Legou, T., Lancia, L., Hilt, P., Tomassini, A., Fadiga, L., D'Ausilio, A., Badino, L., Nguyen, N., 2018. Analyzing vocal tract movements during speech accommodation. *Interspeech*.
- Natale, M. (1975). Convergence of mean vocal intensity in dyadic communication as a function of social desirability. *Journal of Personality and Social Psychology*, 32, 790–804.
- Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, 5(1), 42–46.
- Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J. M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, 2011, 1–9.
- Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, 119(4), 2382–2393.
- Pardo, J. S. (2013). Measuring phonetic convergence in speech production. *Frontiers in Psychology*, 4, 559.
- Pardo, J. S., Jay, I. C., & Krauss, R. M. (2010). Conversational role influences speech imitation. *Attention, Perception, & Psychophysics*, 72, 2254–2264.
- Pardo, J. S., Urmache, A., Wilman, S., & Wiener, J. (2017). Phonetic convergence across multiple measures and model talkers. *Attention, Perception, & Psychophysics*, 79(2), 637–659.
- 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.
- Parkes, L. M., Bastiaansen, M. C., & Norris, D. G. (2006). Combining EEG and fMRI to investigate the post-movement beta rebound. *NeuroImage*, 29(3), 685–696.
- Payton, K. L., Uchanski, R. M., & Braida, L. D. (1994). Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing. *The Journal of the Acoustical Society of America*, 95(3), 1581–1592.
- Pelecanos, J., & Sridharan, S. (2001). Feature warping for robust speaker verification. In: *Proceeding of Speaker Odyssey: The Speaker Recognition Workshop, Crete, Greece, June, 2001*, 213–218.
- Pérez, A., Carreiras, M., & Duñabeitia, J. A. (2017). Brain-to-brain entrainment: EEG interbrain synchronization while speaking and listening. *Scientific Reports*, 7, 4190.
- Petit, L., Simon, G., Joliot, M., Andersson, F., Bertin, T., Zago, L., ... Tzourio-Mazoyer, N. (2007). Right hemisphere dominance for auditory attention and its modulation by eye position: An event related fMRI study. *Restorative Neurology and Neuroscience*, 25(3–4), 211–225.
- Pfurtscheller, G., & Aranibar, A. (1979). Evaluation of event-related desynchronization (ERD) preceding and following voluntary self-paced movement. *Clinical Neurophysiology*, 46(2), 138–146.
- Pfurtscheller, G., & Berghold, A. (1989). Patterns of cortical activation during planning of voluntary movement. *Clinical Neurophysiology*, 72(3), 250–258.
- Pfurtscheller, G., & Da Silva, F. L. (1999). Event-related EEG/MEG synchronization and desynchronization: Basic principles. *Clinical Neurophysiology*, 110(11), 1842–1857.
- Pfurtscheller, G., Pregezer, M., & Neuper, C. (1994). Visualization of sensorimotor areas involved in preparation for hand movement based on classification of  $\mu$  and central  $\beta$  rhythms in single EEG trials in man. *Neuroscience Letters*, 181(1–2), 43–46.
- Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2), 169–190.
- Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, 36(4), 329–347.
- Price, C. J. (2012). A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. *NeuroImage*, 62(2), 816–847.
- Pulvermüller, F. (2018). Neural reuse of action perception circuits for language, concepts and communication. *Progress in Neurobiology*, 160, 1–44.
- Ramseyer, F., & Tschacher, W. (2010). Nonverbal synchrony or random coincidence? How to tell the difference. In *Development of multimodal interfaces: Active listening and synchrony* (pp. 182–196). Springer Berlin Heidelberg.
- Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10(1–3), 19–41.
- Sadjadi, S. O., Slaney, M., & Heck, L. (2013). MSR identity toolbox v1.0: A MATLAB toolbox for speaker recognition research. *Speech and Language Processing Technical Committee: Newsletter*.

- Salmelin, R., Hämäläinen, M., Kajola, M., & Hari, R. (1995). Functional segregation of movement-related rhythmic activity in the human brain. *NeuroImage*, 2(4), 237–243.
- Samuel, A. G., & Kraljic, T. (2009). Perceptual learning for speech. *Attention, Perception, & Psychophysics*, 71(6), 1207–1218.
- Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., & Voegeley, K. (2013). A second-person neuroscience in interaction. *Behavioral and Brain Sciences*, 36(4), 441–462.
- Schmitz, J., Bartoli, E., Maffongelli, L., Fadiga, L., Sebastian-Galles, N., & D'Ausilio, A. (2018). Motor cortex compensates for lack of sensory and motor experience during auditory speech perception. *Neuropsychologia*. In Press, DOI: doi.org/10.1016/j.neuropsychologia.2018.01.006
- Schoot, L., Hagoort, P., & Segaert, K. (2016). What can we learn from a two-brain approach to verbal interaction. *Neuroscience & Biobehavioral Reviews*, 68, 454–459.
- Schroeder, C. E., & Lakatos, P. (2009). Low-frequency neuronal oscillations as instruments of sensory selection. *Trends in Neurosciences*, 32(1), 9–18.
- Silbert, L. J., Honey, C. J., Simony, E., Poeppel, D., & Hasson, U. (2014). Coupled neural systems underlie the production and comprehension of naturalistic narrative speech. *Proceedings of the National Academy of Sciences*, 111(43), E4687–E4696.
- Sperber, D., & Wilson, D. (1998). The mapping between the mental and the public lexicon. In *Language and thought: Interdisciplinary themes* (pp. 184–200).
- Stephens, G. J., Silbert, L. J., & Hasson, U. (2010). Speaker–listener neural coupling underlies successful communication. *Proceedings of the National Academy of Sciences*, 107(32), 14425–14430.
- Stolk, A., Verhagen, L., & Toni, I. (2016). Conceptual alignment: How brains achieve mutual understanding. *Trends in Cognitive Sciences*, 20(3), 180–191.
- Suzuki, N., & Katagiri, Y. (2007). Prosodic alignment in human–computer interaction. *Connection Science*, 19(2), 131–141.
- Tognoli, E., & Kelso, J. S. (2015). The coordination dynamics of social neuromarkers. *Frontiers in Human Neuroscience*, 9, 563.
- Trofimovich, P., & Kennedy, S. (2014). Interactive alignment between bilingual interlocutors: Evidence from two information-exchange tasks. *Bilingualism: Language and Cognition*, 17(4), 822–836.
- Uther, M., Knoll, M. A., & Burnham, D. (2007). Do you speak E-NG-LI-SH? A comparison of foreigner-and infant-directed speech. *Speech Communication*, 49(1), 2–7.
- Ward, Diane, Litman, Arthur, 2007. Automatically measuring lexical and acoustic/prosodic convergence in tutorial dialog corpora. In: ISCA Tutorial and Research Workshop, 4.
- 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.
- West, R., & Turner, L. H. (2010). Understanding interpersonal communication: Making choices in changing times. Cengage learning.

**How to cite this article:** Mukherjee S, Badino L, Hilt PM, et al. The neural oscillatory markers of phonetic convergence during verbal interaction. *Hum Brain Mapp*. 2019;40:187–201. <https://doi.org/10.1002/hbm.24364>