NeuroImage 235 (2021) 118051

Contents lists available at ScienceDirect

NeuroImage

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

Statistical learning and prosodic bootstrapping differentially affect neural synchronization during speech segmentation

Stefan Elmer[a,f,§,*], Seyed Abolfazl Valizadeh[a,b,c,§], Toni Cunillera[d], Antoni Rodriguez-Fornells[e,f,g,*]

[a] Auditory Research Group Zurich (ARGZ), Division Neuropsychology, Institute of Psychology, University of Zurich, Binzmühlestrasse 14/25, Zurich 8050, Switzerland
[b] Department of Internal Medicine, University Hospital, University of Zurich, Zurich 8091, Switzerland
[c] University Research Priority Program, "Dynamics of Healthy Aging", University of Zurich, Zurich 8050, Switzerland
[d] Department of Cognition, Development and Educational Psychology, Barcelona 08035, University of Barcelona, Spain
[e] Department of Cognition, Development and Educational Psychology, Campus Bellvitge, University of Barcelona, 5L'Hospitalet de Llobregat, Barcelona 08097, Spain
[f] Cognition and Brain Plasticity Group, Bellvitge Biomedical Research Institute, L'Hospitalet de Llobregat, Barcelona 08097, Spain
[g] Institució Catalana de Recerca i Estudis Avançats, ICREA, Barcelona 08010, Spain

ARTICLE INFO

Keywords:
Inter-trial coherence
Event-related potentials
Flat speech
Prosody
Word learning

ABSTRACT

Neural oscillations constitute an intrinsic property of functional brain organization that facilitates the tracking of linguistic units at multiple time scales through brain-to-stimulus alignment. This ubiquitous neural principle has been shown to facilitate speech segmentation and word learning based on statistical regularities. However, there is no common agreement yet on whether speech segmentation is mediated by a transition of neural synchronization from syllable to word rate, or whether the two time scales are concurrently tracked. Furthermore, it is currently unknown whether syllable transition probability contributes to speech segmentation when lexical stress cues can be directly used to extract word forms. Using Inter-Trial Coherence (ITC) analyses in combinations with Event-Related Potentials (ERPs), we showed that speech segmentation based on both statistical regularities and lexical stress cues was accompanied by concurrent neural synchronization to syllables and words. In particular, ITC at the word rate was generally higher in structured compared to random sequences, and this effect was particularly pronounced in the flat condition. Furthermore, ITC at the syllable rate dynamically increased across the blocks of the flat condition, whereas a similar modulation was not observed in the stressed condition. Notably, in the flat condition ITC at both time scales correlated with each other, and changes in neural synchronization were accompanied by a rapid reconfiguration of the P200 and N400 components with a close relationship between ITC and ERPs. These results highlight distinct computational principles governing neural synchronization to pertinent linguistic units while segmenting speech under different listening conditions.

1. Introduction

Speech is a hierarchically organized acoustic signal composed of linguistic units at different time scales, such as phonemes, syllables and words (Ding et al., 2016). However, unlike literary language, speech constitutes a continuous signal without reliable gaps between single words (Lehiste, 1960). Hence, one of the main challenges of learning new words is to segment speech, recognize word boundaries and extract word forms, especially when no lexicon is available for word recognition (Assaneoet al., 2019; Batterink and Paller, 2019; Kuhl, 2004; Rodriguez-Fornells et al., 2009). Currently, at least two processes have been proposed to facilitate speech segmentation, namely statistical learning and prosodic bootstrapping (Brent, 1999; Christophe et al., 1994; Jusczyk et al., 1999; Mattys and Jusczyk, 2001; Saffran et al., 1996b). Statistical learning refers to the ability to extract statistical regularities from the speech signal, and relies on the fact that transitional probabilities between adjacent syllables are higher within words than at the word boundaries (Saffran et al., 1996a, 1996b). Otherwise, prosodic bootstrapping consists of using prosodic cues like rhythm, intonation and lexical stress to infer speech structure and to detect word boundaries (Cutler and Norris, 1988; Jusczyk et al., 1999; Myers et al., 2019; Norris et al., 2000). Previous behavioral studies have demonstrated that both statistical learning (Mattys et al., 2005) and prosodic cues (Johnson and Jusczyk, 2001; Thiessen and Saffran, 2003) can be used to segment speech and extract word forms from continuous acoustic signals. Furthermore, it is noteworthy to mention that EEG (Batterink and

* Corresponding authors.
E-mail addresses: s.elmer@psychologie.uzh.ch (S. Elmer), seyed.valizadeh2@uzh.ch (S.A. Valizadeh), tcunillera@ub.edu (T. Cunillera), arfornells@gmail.com (A. Rodriguez-Fornells).
§ Shared first authorship

https://doi.org/10.1016/j.neuroimage.2021.118051.
Received 13 October 2020; Received in revised form 12 March 2021; Accepted 5 April 2021
Available online 10 April 2021.

Paller, 2017; Cunillera et al., 2009, 2006) and MRI (Cunillera et al., 2009; Lopez-Barroso et al., 2013) studies on speech segmentation have convincingly shown that statistical learning and prosodic bootstrapping partially rely on distinct neural mechanisms. In fact, statistical learning is reflected by an N400-like event-related potential (ERP) with main generators in the posterior supratemporal plane and the ventral premotor cortex (Cunillera et al., 2009), whereas speech segmentation through pitch-based stress differences between syllables (prosodic bootstrapping) has been associated with the P200 ERP component which could be localized in the auditory cortex (Cunillera et al., 2006).

The similarity between speech and neural oscillations is that both signals fluctuate in a rhythmic fashion over time (Giraud and Poeppel, 2012). Drawing on this compliance, it has been proposed that the temporal alignment of neural oscillations with the speech signal at multiple time scales constitutes a fundamental principle governing linguistic structure building, speech segmentation and word form recognition (Ding et al., 2016; Ghitza, 2011; Meyer et al., 2017; Panzeri et al., 2010). In this context, it is important to emphasize that low-frequency oscillations in the delta and theta frequency range have repeatedly been shown to be crucial for tracking syllables and words (Ding et al., 2016; Elmer et al., 2018; Giraud and Poeppel, 2012). On this background, Buiatti and colleagues (Buiatti et al., 2009) used a frequency-tagging approach and computed Fast Fourier Transforms (FFT) to quantify mean neural synchronization to syllable and word rates while participants were exposed to either structured or random streams of flat speech. FFT analyses across multiple trials revealed neural synchronization to the word rate in the structured condition that positively correlated with the percentage of correctly recognized words. However, neural synchronization at the syllabic rate was only discernible in the random condition, suggesting that during word learning adjacent syllables are bound together to recognize single word units. In a more recent EEG study, Batterink and colleagues (Batterink and Paller, 2017) compared Inter-Trial Coherence (ITC) ratio of word rate to syllable rate between structured and random sequences of flat speech. Results showed that in structured streams ITC ratio was generally higher and increased across blocks. Furthermore, this effect was mainly driven by a linear increase in ITC at the word rate, and accompanied by a decrease in ITC at the syllable rate as a function of exposure. However, surprisingly, in both conditions ITC ratio predicted task performance.

As mentioned above, both ERP and ITC metrics can be used as suitable markers for speech segmentation based on transitional probabilities between adjacent syllables or prosodic cues. Increased frontocentral ERP negativities in the time range of 350–550 ms (here referred as N400 component) are thought to reflect the building up of lexical representations or the extraction of word forms as a function of learning (Cunillera et al., 2006; Rodriguez-Fornells et al., 2009). In contrast, frontocentral distributions of the P200 ERP component have been observed when multiple cues (prosodic and statistical information) are used in combination for the isolation of new words during speech segmentation (Cunillera et al., 2009, 2006; De Diego Balaguer et al., 2007). The P200 component has been associated with neural sources in primary and secondary auditory regions (Bosnyak et al., 2004; Liegeois-Chauvel et al., 1994; Picton et al., 1999; Scherg and Von Cramon, 1986), and is particularly sensitive to changes in the acoustic environment that predict and facilitate learning (Shahin et al., 2003a; Tremblay et al., 2014). Furthermore, previous studies have shown increased amplitudes of the P200 component in response to salient stimuli that triggered the selection of relevant information, highlighting the relationship between P200 modulations and attention during learning (Fritz et al., 2007; Luck and Hillyard, 1994; Rentzsch et al., 2008). Such a relationship between attention and word learning was initially proposed by Gleitman & Wanner (Gleitman and Wanner, 1982) who considered that infants might exploit certain perceptual or attentional cues allowing them to extract salient elements from acoustic language streams. Importantly, given that P200 modulations in response to prosodic cues are reduced or disappear when learning is not possible (de Diego-Balaguer et al., 2015), increased P200 amplitudes during speech segmentation tasks are thought to reflect the detection of relevant prosodic cues that might direct attention toward word boundaries and facilitate the extraction of word forms (de Diego-Balaguer et al., 2015; De Diego Balaguer et al., 2007; Rodriguez-Fornells et al., 2009). Finally, ITC measures are sensitive indices to quantify the degree of phase synchronization of neural oscillations. Based on the concept of neural entrainment (Obleser and Kayser, 2019b), one might infer that neural oscillations align with rhythmic fluctuations in the auditory environment. Accordingly, an increase in ITC at the syllable and word rates is thought to reflect the tracking of regular exogenous stimulus attributes or the attentive selection of task-relevant information (Obleser and Kayser, 2019b).

Several previous EEG studies on speech segmentation focused on statistical learning and evaluated neural synchronization across epochs consisting of multiple word units (Batterink and Paller, 2017; Buiatti et al., 2009). Although these studies validated the suitability of frequency-tagging approaches to tackle the neural principles underlying speech segmentation and word learning, some fundamental questions have not yet been systematically addressed and clarified. In fact, it is currently unknown (1) whether neural synchronization to syllables and words likewise operates if additional prosodic cues can be used to segment speech and extract word forms (Meyer et al., 2017). There is also no common agreement on (2) whether speech segmentation and word form recognition are generally mediated by a neural transition from syllabic rate to word rate, or whether the two time scales are concurrently tracked (Batterink and Paller, 2017; Giraud and Poeppel, 2012; Henin et al., 2019). Furthermore, (3) notwithstanding that ERP studies have shown that P200 and N400 responses constitute valid indices of statistical learning and prosodic bootstrapping (Batterink and Paller, 2017, 2019; Cunillera et al., 2009, 2006; De Diego Balaguer et al., 2007), it is unclear whether these two ERP components share a common neural basis with neural synchronization to the syllable and word rates. In this context, it is noteworthy to mention that the relationship between EEG signals averaged in the time domain and neural oscillations is not easy to establish. On the one hand, both measures could in principle reflect the same underlying process, especially regarding single-trial evoked neural activity and phase-alignment of endogenous oscillations. On the other hand, a substantial portion of oscillatory components could also reflect non-phased locked activity that contains information which is lost in the computation of canonical ERPs (Sauseng and Klimesch, 2008). Since the paradigms normally used for statistical learning have a rhythmic structure, single-trial evoked activity could indeed induce an apparent increase in certain frequency bands that match the expected rhythm frequencies. To address the open questions mentioned above, we used EEG and evaluated ITC at the syllable and word rates in structured and random sequences of flat and stressed speech. In addition, we evaluated ERPs in time windows overlapping with the P200 and N400 components (Cunillera et al., 2009, 2006), and assessed the functional compliance between these two ERP manifestations and ITC metrics.

2. Materials and methods

2.1. Participants

Thirty students of the University of Barcelona took part in the flat speech condition (age range = 19–44 years, mean age = 23.69, SD = 5.68, 17 females), and 23 of them were re-invited to perform the stressed condition (age range = 19–38 years, mean age = 23.18, SD = 4.58, 15 females). All participants were right-handed native Spanish-Catalan speakers with normal hearing and no neurological impairments. The experiment was approved by the local ethical committee of the University of Barcelona, and the participants were paid for their participation. All participants provided written informed consent to take part in the study.

2.2. Materials and procedure

The experimental design was the same as the one previously used by Cunillera and colleagues (Cunillera et al., 2009, 2006). Therefore, in the following paragraphs we literally reiterate the description of the stimulus material used in the previous study. "Five words streams (languages) were created for the stressed and flat conditions. The word streams had the same structure as the ones used by Saffran and colleagues. (Saffran et al., 1996a). In particular, each stream consisted of 4 different trisyllabic nonsense words (pseudowords, word duration = 696 ms, syllable duration = 232 ms) and each word was repeated 192 times, resulting in a total of 3840 items per condition (5 languages x 4 words/non-words x 192 repetitions). The words were concatenated to form a text stream and transformed into an acoustic stream using the speech synthesizer MBROLA which relies on concatenation of diphones (Dutoit et al., 1996). The streams did not contain acoustic pauses between single items. Afterwards, the Cooledit software was used to equate the length of the different streams with millisecond precision, resulting in a duration of 8 min 54 s and 528 ms for each stream. Since only 59 syllables could be used for the construction of the five streams, one syllable was repeated in 2 streams. In all streams, the transitional probability across syllables forming a word was 1.0, whereas syllables spanning word boundaries had a transitional probability of 0.33. The same pool of syllables was used for the construction of the languages in both the stressed and flat conditions, however, the syllables were concatenated in a different order. The resulting word streams of the flat condition did not contain pauses or other acoustic cues indicating word onset. In contrast, in the stressed condition, the pitch of the first syllable of each word was increased by 20 Hz to create an artificial stress at the beginning of each possible word (Johnson and Jusczyk, 2001). To use a prosodic cue that does not operate in the native languages of the participants (Catalan / Spanish) bears the advantage of simulating the extraction of word forms from continuous speech and the learning of new words in a foreign language. Furthermore, pitch manipulation of the first instead of the last syllable enables to capture the neural indices of prosodic bootstrapping (P200 component and ITCs) within the same word containing the acoustic manipulation. Although stressed syllables are normally characterized by an increase in length, we maintained the duration of syllables within a word constant (syllable duration = 232 ms) in order to avoid segmentation based on syllable lengths rather than pitch. The fact that all syllables across streams were matched in length enables a direct comparison between the conditions.

As a control condition, 10 different streams (five for each condition) were created using the same syllables presented in each structured stream. However, the syllables were concatenated in random order and each syllable in the streams could be followed by any of the other eleven syllables composing the streams, resulting in a transitional probability across syllables of 0.09. Such a low transitional probability should create a condition where the extraction of words is not possible. In the flat condition, the random streams did not contain acoustic information. In contrast, in the stressed condition, the first syllable of each triplet was stressed regardless of which syllable felled in the stressed positions. All stimuli were presented via headphones at a comfortable volume. During the presentation of the auditory streams and the collection of EEG data, a fixation cross was presented at the center of a black screen to help participants to fix their gaze and minimize eye movements.

After the exposure to each stream, a two Alternative Forced-Choice (2AFC) behavioral test was administered to determine whether the participants were able to identify the words previously heard. EEG data were not recorded during this phase. The test comprised eight pairs of randomly presented auditory test items (i.e., a word and a part-word). Part-words were constructed by the concatenation of the third syllable of a word and the first two syllables of another word (3–1–2 part-words; e.g., rutaba, dapiru, bopiru, litoku), or the last two syllables of a word and the first syllable of another word (2–3–1 part-words; e.g., tabago, golito, kudagu, kibopi). Thus, for each stream, the test comprised 8 part-words randomly selected from a pool of 24 possible part-words. While part-words were not repeated during the test in a particular stream, each word was repeated twice but paired with different part-words. For the random streams, the test items were composed of 16 different trisyllabic groupings. After the presentation of each test pairing, participants had to press a response button and to indicate whether it was the first or the second word in the pair that belonged to the stream they just heard. After the two items were presented, a fixation cross was displayed on the screen and the next test pair was presented after response selection. The order of presentation of words and part-words in the test pairs was balanced, and brief rest periods were allowed after each stream. Furthermore, in all test items pitch modulations were removed to match words and part-words.

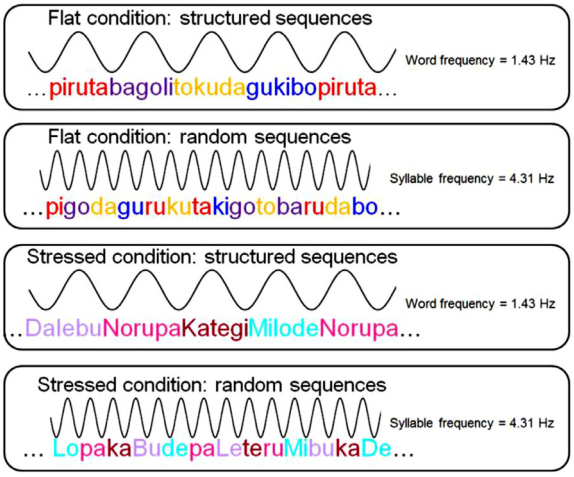

Fig. 1. Schematic representation of the experimental design and theoretical framework of brain-to-stimulus alignment mechanisms.

2.3. EEG data acquisition, pre-processing and ERP analyses

The scalp EEG was recorded from 29 electrodes located at standard positions using Electro-Cap (International). Vertical eye movements were monitored with an electrode at the infraorbital ridge of the right eye, and electrode impedances were kept below 3 kΩ. The electrophysiological signal was filtered on-line with a bandpass of 0.01–50 Hz (half-amplitude cutoffs) and digitized at a rate of 250 Hz. All preprocessing steps were performed with the Brain Vision Analyzer software package (version 2.01; Brain Products). In particular, the EEG signal was re-referenced off-line to the mean of the activity at the two mastoid electrodes, data were filtered with a low-pass filter of 30 Hz (including a Notch filter of 50 Hz), and artifacts (eye movements and blinks) were corrected using an independent component analysis (Jung et al., 2000). Furthermore, an automatic raw data inspection was used to remove remaining artefacts if a voltage gradient criterion of 50 μV/ms or an amplitude criterion of ±100 μV (200 ms before and after the event) was exceeded. Afterwards, each block of the different "languages" was segmented into single epochs of 796 ms (i.e., this epoch covers the entire duration of each word plus a 100 ms baseline), and baseline correction was performed in the time range from −100 to 0 ms. The length of the baseline was selected in accordance with the procedure normally used at our laboratories. The single epochs were subjected to two different types of analyses where we evaluated ERPs and ITCs. ERP analyses focused on two specific components that have previously been shown to be sensitive to speech segmentation based on statistical learning and prosodic bootstrapping, namely the P200 and N400 waveforms (Batterink and Paller, 2017, 2019; Cunillera et al., 2009, 2006). For the ERP analyses, the single baseline-corrected epochs were averaged separately for the structured and random sequences of flat and stressed speech and for

the four blocks. Afterwards, based on the fact that the P200 and N400 components elicited maximal voltage strength at central and anterior electrodes, and for reasons of comparability between ERP and ITC values, we averaged N200 and N400 responses across 6 channels, namely F3, Fz, F4, C3, Cz and C4. Finally, in accordance with previous studies using exactly the same paradigm and stimuli, mean amplitudes were extracted in two time windows overlapping with the P200 (170–250 ms) and N400 (350–550 ms) components (Cunillera et al., 2009, 2006). Otherwise, for ITC analyses with homemade scripts, the pre-stimulus period was removed and the single baseline-corrected epochs were exported to MATLAB.

2.4. ITC and wavelet analysis

ITC across the whole frequency spectrum was computed for each electrode, structured and random sequences of flat and stressed speech and the four blocks using the following Morlet wavelet transform:

$$Morlet(x) = e^{\frac{-x^2}{2}} \cos(5x)$$

Before transferring the signal to the wavelet domain, we added zero-padding of the same length as the single EEG segments (200 sample points) at the beginning as well as at the end of the single epochs to increase resolution at low frequencies. All ITC analyses were computed using the command "cwt" and homemade MATLAB scripts. Phase information was extracted from the wavelet transfer function, and ITC values corresponding to the word (1.43 Hz) and syllable (4.31 Hz) rates were calculated by summation of phase angle (van Diepen and Mazaheri, 2018) of all epochs according to the following formulas:

$$ITC\left(f_{word}, t\right) = \frac{1}{N}\sum_{K=1}^{N} e^{i\phi^k\left(f_{word},t\right)},\ f_{word} = 1.43Hz$$

$$ITC\left(f_{syllable}, t\right) = \frac{1}{N}\sum_{K=1}^{N} e^{i\phi^k\left(f_{syllable},t\right)}, f_{syllable} = 4.31Hz$$

In these equations, N corresponds to the number of trials, whereas ϕ^k depicts the local phase angle of the signal. Since we were interested in absolute phase-shifting, we computed the absolute value of ITC at the word rate and at the syllabic rate. Furthermore, based on a previous work of Batterink and colleagues (Batterink and Paller, 2017), we calculated the ratio of ITC at the word rate to ITC at the syllable rate (word learning index, WLI) as well as the ratio of ITC at the syllable rate to ITC at the word rate (syllable learning index, SLI). Based on the fact that in the structured sequences ITC at the word rate (1.43 Hz) and in the random sequences ITC at the syllabic rate (4.31 Hz) showed the strongest values at frontal and central electrodes (Fig. 2), and in order to increase signal-to-noise ratio, ITC values, WLI and SLI were averaged across six frontal and central electrodes (F3, Fz, F4, C3, Cz and C4) and subjected to statistical analyses. A similar pooling procedure has previously been used by Batterink and colleagues (Batterink and Paller, 2017).

2.5. Statistical analyses

All analyses were performed using parametric statistics implemented in the IBM SPSS Statistics 22 software package (SPSS, an IBM company, Armonk, New York, USA). Significant effects were further inspected by means of post-hoc t-tests or ANOVAs (corrected for multiple comparisons using the Bonferroni procedure), whereas correlative analyses were computed according to Pearson's r (corrected for multiple comparisons using the Bonferroni procedure). Based on the statistical results and the visualization of the means of the contrasts of interest, all post-hoc t-tests and correlations were computed in a one-tailed manner.

2.5.1. Behavioral data

In order to testify that the participants were able to segment speech and learn the words, we performed separate one-sample t-tests for the flat and stressed conditions, and tested the percentage of correct responses against chance level (50%, the behavioral data of 3 participants of the stressed condition are missing). Furthermore, we compared the percentage of correct responses between the flat and stressed condition using a *t*-test for dependent samples.

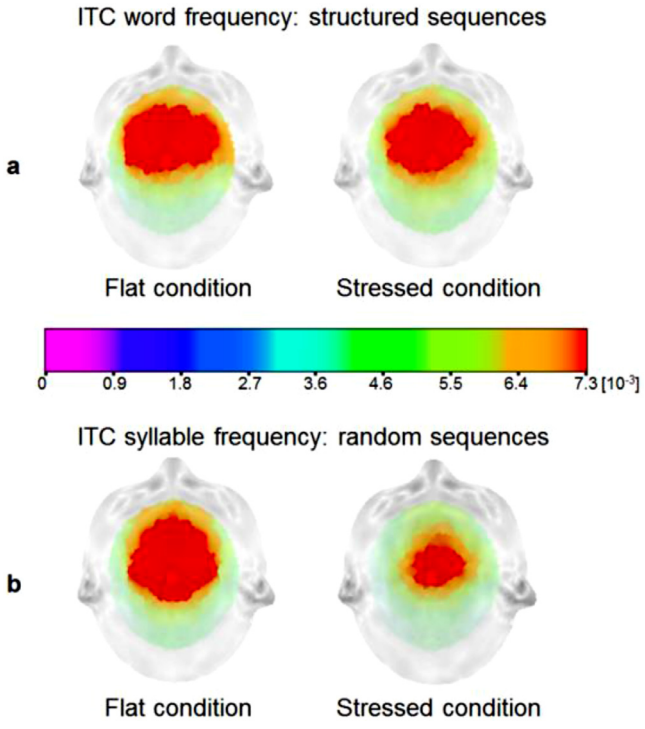

Fig. 2. Topographic distribution (eLORETA software) of mean ITC values at the (a) word rate for structured sequences and at the (b) syllable rate for random sequences.

2.5.2. Event-related potential analyses

In a first EEG analysis, we evaluated mean P200 and N400 amplitudes to provide comparability with previous studies showing the sensitivity of the P200 component to prosodic bootstrapping and of the N400 component to statistical learning (Batterink and Paller, 2017, 2019; Cunillera et al., 2009, 2006; De Diego Balaguer et al., 2007). With this purpose in mind, the ERP data were evaluated by means of separate 2 × 4 ANOVAs (2 structured/random sequences x 4 blocks) for the P200 and N400 components and the flat and stressed conditions.

2.5.3. Main ITC analyses

The ITC data where evaluated to test (1) whether general neural synchronization to syllables and words is comparable in statistical learning and prosodic bootstrapping conditions and (2) whether there are dynamic changes in neural synchronization across the blocks. Given that only a part of the participants performed the stressed condition and missing values in an ANOVA will result in list-wise deletion of individuals, the main ITC analyses were performed on a reduced sample of 23 participants who completed both the flat and stressed conditions. In particular, we computed separate 2 × 2 × 4 ANOVAs for ITC at the syllable and word rates, and directly compared the two conditions (flat and stressed), the two sequences (structured and random) and the four blocks.

2.5.4. Complementary ITC and correlation analyses

To exploit the full range of measured participants (flat condition $n = 30$, stressed condition $n = 23$), we conducted further complementary ITC analyses. Furthermore, this larger sample of participants was used for correlation analyses between ITC and ERP metrics. In a first analysis, we tested (1) whether neural synchronization to pertinent speech units

likewise operates in statistical learning and prosodic bootstrapping conditions. With this purpose in mind, we compared mean ITC (averaged across the four blocks) at the syllable rate, word rate, WLI and SLI between structured and random sequences of flat and stressed speech using separate univariate ANOVAs with the within-subject factor "sequences" (structured and random). In a second analysis, we then focused on neural dynamics as a function of exposure. Accordingly, we evaluated ITC at the syllabic rate and ITC at the word rate across the four blocks, separately for the flat and stressed conditions as well as for structured and random sequences using univariate ANOVAs with the within-subject factor "block". This additional analysis allowed us to test (2) whether in flat and stressed conditions speech segmentation is generally mediated by a neural transition from syllable to word rate or whether the two time scales are concurrently tracked. In the case of neural transitions, ITC at the word rate should increase across the blocks, whereas ITC at the syllable rate is expected to decrease. In contrast, if syllabic rate and word rate are concurrently tracked, one would expect a general linear increase in ITC at both time scales over the blocks.

Finally, we focused on possible relationships between ITC at the syllable rate and mean P200 amplitudes, ITC at the word rate and mean N400 amplitudes as well as on the possible relatedness between ITC at the syllable and word rates. The rational for these correlation analyses was based on the statistical results as well as on previous literature proposing that the N400 component and ITC at the word rate can be used as suitable markers for statistical learning (Batterink and Paller, 2017; Cunillera et al., 2009), the P200 component is sensitive to prosodic bootstrapping (Cunillera et al., 2009), and on the idea that pertinent speech units are possibly concurrently tracked (Ding et al., 2017; Giraud and Poeppel, 2012).

3. Results

3.1. Behavioral data

To test whether the participants were able to learn the new words, we performed separate one-sample t-tests against chance level (50%) for the flat (mean hit rate = 68.33%, SD = 9.27) and stressed (mean hit rate = 62.31%, SD = 11.64) conditions. The evaluation of the percentage of correct responses yielded significance for both the flat ($t_{(29)} = 10.833$, $p < .001$) and stressed conditions ($t_{(19)} = 4.927$, $p < .001$). Accordingly, these behavioral results testify that the participants were able to segment speech and learn the new words based on statistical learning and prosodic bootstrapping. Furthermore, a *t*-test for paired samples revealed that the percentage of correct responses did not differ between the flat and stressed conditions ($t_{(26)} = 1.695$, $p = .102$).

3.2. Event-related potential analyses

For the sake of completeness and to provide comparability with previous EEG studies showing that P200 and N400 responses faithfully mimic speech segmentation based on statistical regularities and prosodic cues (Batterink and Paller, 2017, 2019; Cunillera et al., 2009, 2006), we analyzed mean amplitudes of these two ERPs (Fig. 3b and 3d) in the time window of 170–250 (P200) and 350–550 (N400) ms. With this purpose in mind, we computed separate 2 × 4 ANOVAs for the flat and stressed conditions, and compared structured and random sequences across the four blocks. Moreover, we computed correlation analyses using Pearson's r to determine possible relationships between P200 and N400 manifestations, ITC at the syllabic rate and ITC at the word rate.

3.2.1. P200 component

The statistical analysis of the flat speech condition ($n = 30$) by means of a 2 × 4 ANOVA revealed a main effect of "block" ($F_{(3, 87)} = 27.165$, $p < .001$), whereas the main effect of "sequence" ($F_{(1,29)} = 0.058$, $p = .811$) and the "sequence x block" interaction ($F_{(3, 87)} = 0.638$, $p = .593$) did not reach significance. Post-hoc t-tests (one-tailed, Bonferroni-corrected p value for 6 tests, $p < .008$) indicated an overall increase in mean P200 amplitudes across the four blocks, irrespective of sequence type (Fig. 3b and Table 1).

The evaluation of the stressed condition ($n = 23$) yielded main effects of "sequence" ($F_{(1, 22)} = 126.774$, $p < .001$) and "block" ($F_{(3, 66)} = 5.747$, $p = .001$). The "sequence x block" interaction did not reach significance ($F_{(3, 66)} = 2.458$, $p = .071$). As visible in Fig. 3d, the main effect of "sequence" was related to increased P200 amplitudes in structured compared to the random streams of stressed speech. Furthermore, post-hoc *t*-test (one-tailed, Bonferroni-corrected p value for 6 tests, $p < .008$) revealed that the main effect of "block" was associated with decreased P200 amplitudes from block 1 to block 3 and 4, regardless of sequence (Fig. 3d and Table 1).

3.2.2. N400 component

The statistical analysis of the flat condition ($n = 30$) yielded main effects of "sequence" ($F_{(1, 29)} = 52.268$, $p < .001$) and "block" ($F_{(3, 87)} = 15.010$, $p < .001$) as well as a significant quadratic "sequence x block" interaction ($F_{(1, 29)} = 4.594$, $p = .041$). As visible in Fig. 3b, the quadratic interaction between "sequence" and "block" originated from a U-shaped devolution of the N400 component over time in structured sequences of flat speech, whereas the main effect of "sequence" was related to larger N400 responses in structured compared to random streams (Fig. 3b). Post-hoc t-tests (one-tailed, Bonferroni-corrected p value for 6 tests, $p < .008$) computed to disentangle the "sequence x block" interaction showed a significant reduction in N400 amplitudes from block 2 to 4 ($t_{(29)} = -4.595$, $p < .001$) and from block 3 to 4 ($t_{(29)} = -3.819$, $p < .001$). Furthermore, post-hoc t-tests (one-tailed, Bonferroni-corrected p value for 6 tests, $p < .008$) revealed that the main effect of "block" originated from overall reduced N400 amplitudes in block 4 compared to the first three blocks (block 1_4: $t_{(29)} = -5.277$, $p < .001$; block 2_4: $t_{(29)} = -5.493$, $p < .001$; block 3_4: $t_{(29)} = -5.664$, $p < .001$, Fig. 3b), irrespective of sequence type. All other comparisons did not reach significance (Table 2).

The analysis of the stressed condition ($n = 23$) yielded a main effect of "block" ($F_{(3, 66)} = 3.543$, $p = .019$), whereas the main effect of "sequence" ($F_{(1, 22)} = 1.007$, $p = .326$) and the "sequence x block" interaction ($F_{(3, 66)} = 1.313$, $p = .278$) did not reach significance. According to post-hoc t-tests (one-tailed, Bonferroni-corrected p value for 6 tests, $p < .008$), the main effect of "block" was associated with increased N400 responses in block 2 compared to block 1, regardless of sequences (block 1_2: $t_{(22)} = 2.675$, $p = .007$, Fig. 3d). All others comparisons did not reach significance (Table 2).

3.3. ITC analyses

Fig. 4 shows the mean ITC values for the whole sample of participants, the two conditions (flat and stressed), the two sequences (structured and random) and the four blocks. From this Figure, one can see the expected increase in ITC at the word rate in structured compared to random sequences of flat and stressed speech. Furthermore, ITC values at the syllable rate appear to be higher for random compared to structured sequences.

3.3.1. Main ITC analyses (n = 23)

To statistically evaluate the effects that are shown in Fig. 4, we first carried out an overall repeated measures ANOVA with a reduced sample of participants ($n = 23$) who performed both the flat and stressed conditions. With this purpose in mind, we directly compared the two conditions (flat and stressed), the two sequences (structured and random) and the four blocks by means of separate ANOVAs for ITC at the word rate and ITC at the syllable rate. The results of these analyses are summarized in Table 3.

3.3.1.1. ITC at the word rate. The statistical analysis of ITC at the word rate revealed a significant main effect of "sequence" as well as "condition x sequence" and "condition x block" interaction effects (Table 3). As

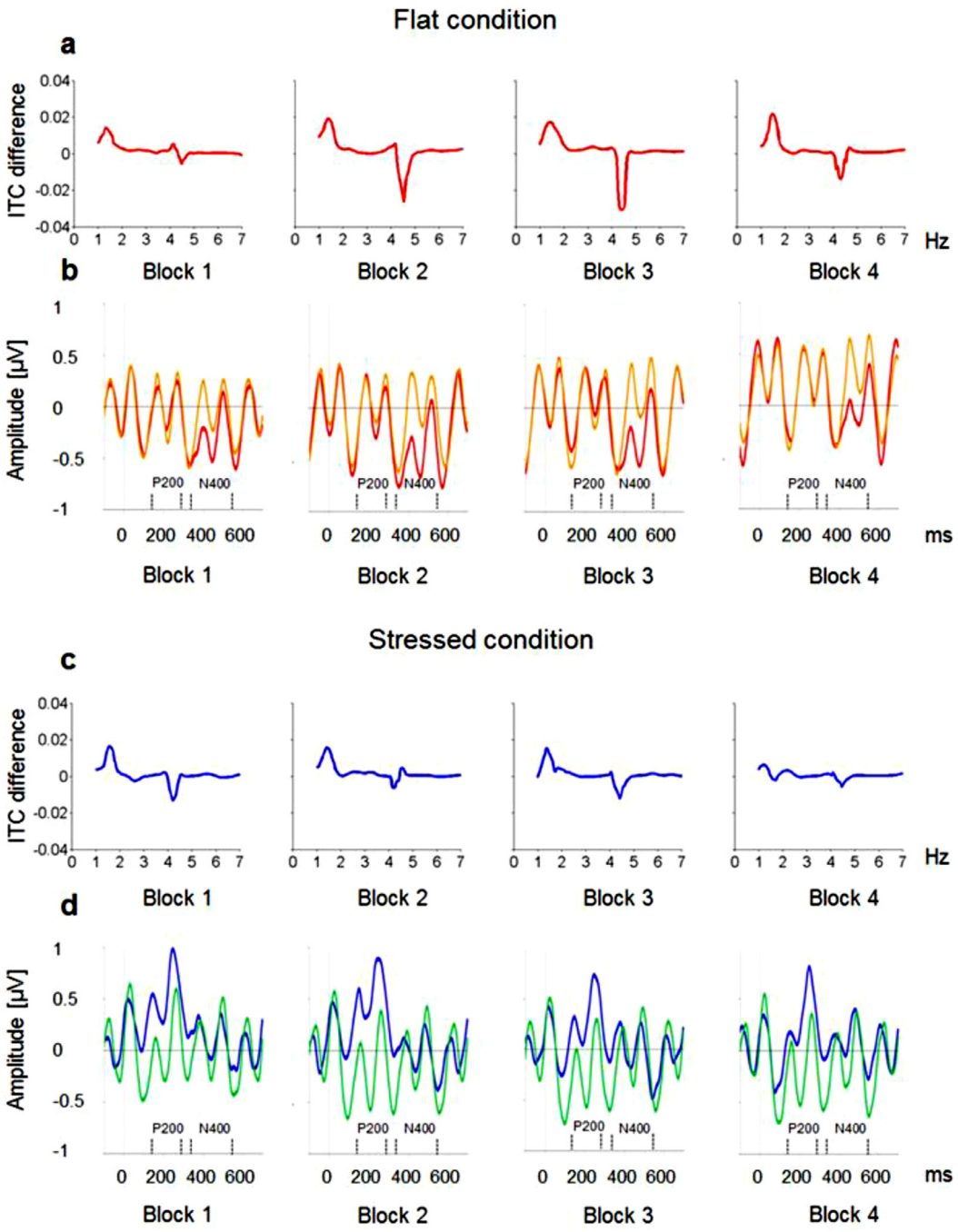

Fig. 3. Grand average event-related potentials of the structured and random sequences of flat (b, $n = 30$) and stressed (d, $n = 23$) speech. In the flat condition (b) the red line depicts brain responses to structured sequences, whereas the orange line represents the random sequences. In the stressed condition (d) brain responses to structured sequences are shown in blue, whereas random sequences are represented in green. Figs. 3a and 3c show ITC differences at the syllabic rate and at the word rate between structured and random sequences of flat (a, $n = 30$) and stressed (c, $n = 23$) speech. All waveforms are shown at a frontocentral pool of electrodes.

visible from Fig. 5a, the main effect of "sequence" was associated with increased ITC values at the word rate in structured compared to random sequences. However, the magnitude of this effect was lower in the stressed compared to the flat condition, as revealed by the significant "condition x sequence" interaction (Fig. 5b). In fact, separate univariate post-hoc ANOVAs for the two conditions (Bonferroni corrected p value for 2 tests, $p < .025$) showed that the effect of "sequence" (structured vs. random) was significant for the flat ($F_{(1, 22)} = 34.647$, $p < .001$) but not for the stressed condition ($F_{(1, 22)} = 2.321$, $p = .142$). Finally, the significant "condition x block" interaction effect was further decomposed by means of separate univariate ANOVAs for the flat and stressed conditions (Bonferroni corrected p value for 2 tests, $p < .025$). According to this procedure, the effect of "block" was only significant in the flat condition (flat: $F_{(3, 66)} = 6.756$, $p < .001$; stressed: $F_{(3, 66)} = 0.535$, $p = .660$). Additional post-hoc t-tests for the flat condition (Bonferroni corrected p value for 6 tests, $p < .008$) revealed increased ITC at the word rate in block 3 ($M = 0.008$) compared to block 1 ($M = 0.005$, $t_{(22)} = -4.787$, $p < .001$, Fig. 5c).

3.3.1.2. ITC at the syllable rate. The analysis of ITC at the syllable rate yielded main effects of "condition" and "block" as well as "condition x block" and "sequence x block" interaction effects (Table 3). As shown in

Table 1
Post-hoc comparisons of the significant main effects of "block" in the omnibus ANOVAs for P200 amplitudes. * Depicts significance after correction for multiple comparisons.

Condition	Sequences	Contrast	Degrees of freedom	t-value	p-value
Flat	All sequences	Block 1 vs. Block 2	29	−2.570	0.008*
		Block 1 vs. Block 3	29	−6.055	< 0.001*
		Block 1 vs. Block 4	29	−6.829	< 0.001*
		Block 2 vs. Block 3	29	−3.261	0.0015*
		Block 2 vs. Block 4	29	−5.496	< 0.001*
		Block 3 vs. Block 4	29	−3.524	< 0.001*
Stressed	All sequences	Block 1 vs. Block 2	22	1.688	0.053
		Block 1 vs. Block 3	22	3.666	< 0.001*
		Block 1 vs. Block 4	22	2.821	0.005*
		Block 2 vs. Block 3	22	2.442	0.011
		Block 2 vs. Block 4	22	1.501	0.074
		Block 3 vs. Block 4	22	−0.434	0.334

Table 2
Post-hoc comparisons of the significant main effects of "block" in the omnibus ANOVAs for N400 amplitudes. * Depicts significance after correction for multiple comparisons.

Condition	Sequences	Contrast	Degrees of freedom	t-value	p-value
Flat	Structured	Block 1 vs. Block 2	29	2.318	0.014
		Block 1 vs. Block 3	29	1.254	0.110
		Block 1 vs. Block 4	29	−2.540	0.009
		Block 2 vs. Block 3	29	−1.537	0.068
		Block 2 vs. Block 4	29	−4.595	< 0.001*
		Block 3 vs. Block 4	29	−3.819	< 0.001*
Flat	All sequences	Block 1 vs. Block 2	29	1.260	0.109
		Block 1 vs. Block 3	29	0.333	0.370
		Block 1 vs. Block 4	29	−5.277	< 0.001*
		Block 2 vs. Block 3	29	−1.072	0.146
		Block 2 vs. Block 4	29	−5.493	< 0.001*
		Block 3 vs. Block 4	29	−5.664	< 0.001*
Stressed	All sequences	Block 1 vs. Block 2	22	2.675	0.007*
		Block 1 vs. Block 3	22	2.512	0.010
		Block 1 vs. Block 4	22	1.736	0.048
		Block 2 vs. Block 3	22	−0.555	0.292
		Block 2 vs. Block 4	22	−1.296	0.104
		Block 3 vs. Block 4	22	−0.690	0.249

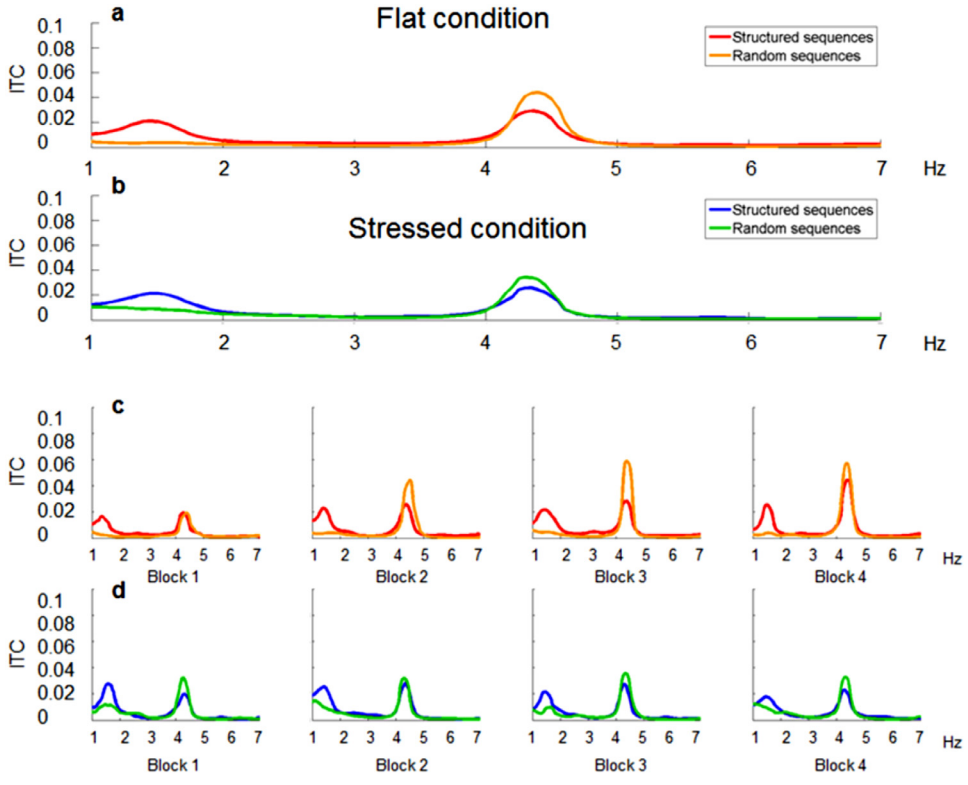

Fig. 4. *a* = mean ITC values in the flat condition (n = 30) for structured (red) and random (orange) sequences. *b* = mean ITC values in the stressed condition (n = 23) for structured (blue) and random (green) sequences. *c* = ITC values across the four blocks of the flat condition (n = 30) for structured (red) and random (orange) sequences. *d* = ITC values across the four blocks of the stressed condition (n = 23) for structured (blue) and random (green) sequences.

Table 3
Overall repeated measures ANOVA results (n = 23) for word rate and syllable rate with the factors "condition" (flat and stressed), "sequence" (structured and random) and "block" (4 levels). * Depicts significance.

	Factor	df	F-value	p-value
Word Rate	Condition	1, 22	.289	0.596
	Sequence	1, 22	12.32	< 0.001*
	Block	3, 66	1.028	0.386
	Condition x Sequence	1, 22	8.13	0.009*
	Condition x Block	3, 66	3.72	0.015*
	Sequence x Block	3, 66	1.611	0.195
	Condition x Sequence x Block	3, 66	1.699	0.176
Syllable Rate	Condition	1, 22	7.41	0.012*
	Sequence	1, 22	2.846	0.106
	Block	3, 66	11.12	< 0.001*
	Condition x Sequence	1, 22	.793	0.383
	Condition x Block	3, 66	8.60	< 0.001*
	Sequence x Block	3, 66	2.77	0.048*
	Condition x Sequence x Block	3, 66	1.714	0.173

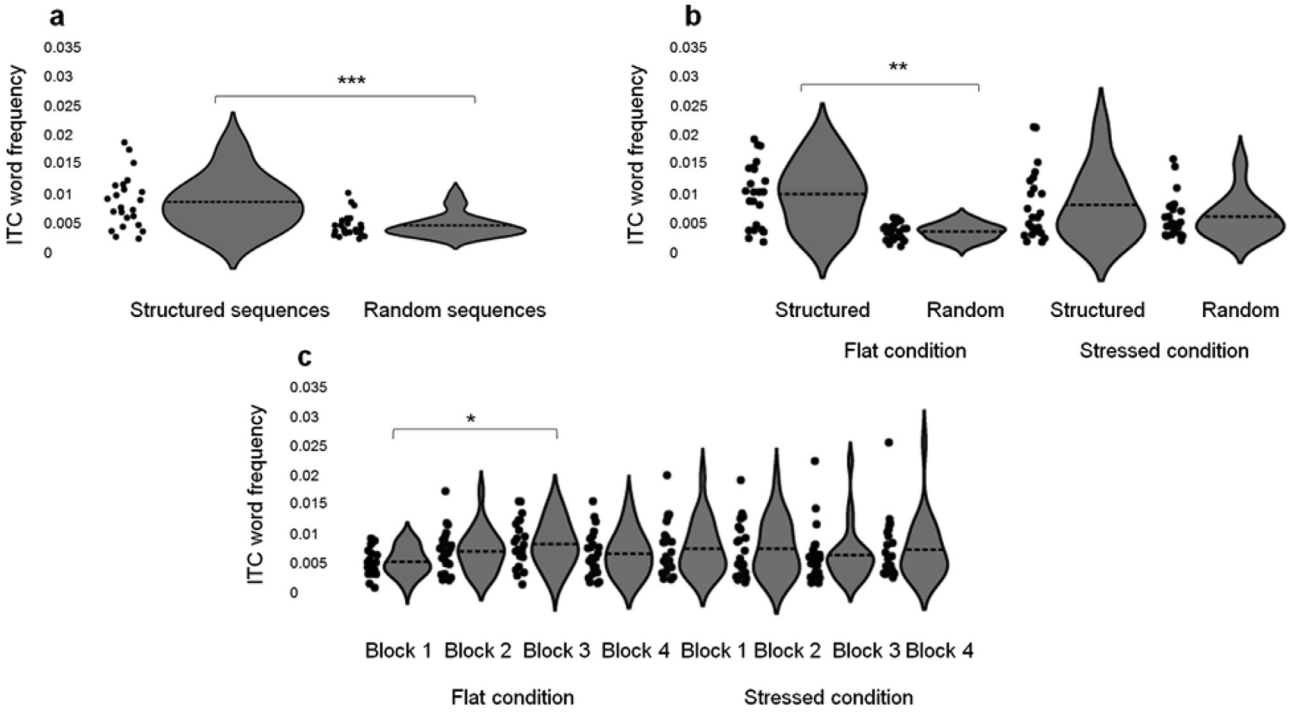

Fig. 5. Single-subject data (n = 23) and violin plots with density distribution and mean ITC values corresponding to the word rate are shown separately for the main effect of "sequence" (a), the "condition x sequence" interaction (b) and the "condition x block" interaction (c). * $p < .05$, ** $p < .01$, *** $p < .001$.

Fig. 6a, ITC values at the syllable rate were higher in the flat compared to the stressed condition. The main effect of "block" was further inspected using post-hoc t-tests (one-tailed, Bonferroni corrected p values for 6 tests, $p < .008$). This procedure revealed overall increased ITC values in block 2 ($t_{(22)} = -4.104$, $p < .001$), 3 ($t_{(22)} = -5.544$, $p < .001$) and 4 ($t_{(22)} = -4.223$, $p < .001$) compared to block 1 (Fig. 6b), irrespective of condition and sequences. However, the main effect of "block" was more pronounced in the flat compared to the stressed condition, as revealed by the "condition x block" interaction effect (Fig. 6c). In fact, separate univariate post-hoc ANOVAs for the two conditions (Bonferroni corrected p value for 2 tests, $p < .025$) yielded a significant effect of "block" for the flat ($F_{(3, 66)} = 12.656$, $p < .001$) but not for the stressed condition ($F_{(3, 66)} = 0.633$, $p = .774$). Additional post-hoc t-tests (one-tailed, Bonferroni corrected p value for 6 tests, $p < .008$) revealed increased ITC values in block 2 ($t_{(22)} = -5.394$, $p < .001$), 3 ($t_{(22)} = -5.657$, $p < .001$) and 4 ($t_{(22)} = -4.381$, $p < .001$) compared to block 1 of the flat condition. Finally, post-hoc t-tests between the four blocks of structured and random sequences (one-tailed, Bonferroni corrected p value for 4 tests, $p < .012$) were used to disentangle the marginally significant "sequence x block" interaction. This strategy revealed increased ITC values in block 3 of random compared to structured sequences ($t_{(22)} = -2.705$, $p = .006$, Fig. 6d).

3.4. Complementary ITC and correlation analyses

To exploit the full range of measured participants (flat condition n = 30, stressed condition n = 23), we conducted additional complementary ITC analyses and assessed relationships between ERPs and ITC metrics. The complementary ITC analyses aimed at re-evaluating (1) overall differences in neural synchronization to pertinent speech units between structured and random sequences of flat and stressed speech as well as (2) dynamic changes in neural synchronization at the syllable and word rates across the blocks.

3.4.1. General neural synchronization to pertinent speech units

In line with previous work (Batterink and Paller, 2017, 2019), ITC analyses yielded clear maxima at the word and syllable rates (Fig. 4). To first testify the overall neural synchronization to pertinent speech units,

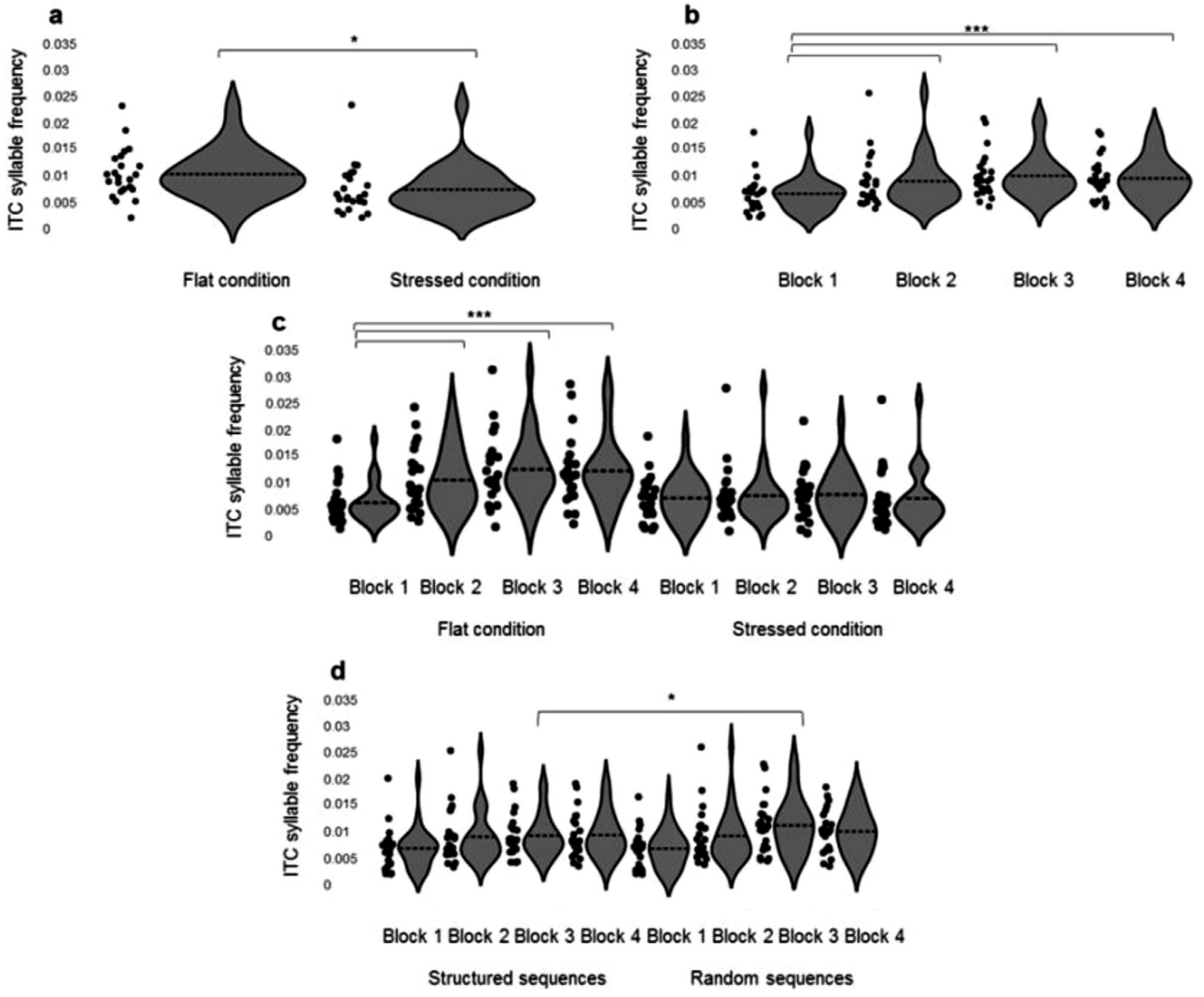

Fig. 6. Single-subject data (n = 23) and violin plots with density distribution and mean ITC values corresponding to the syllable rate are shown separately for the main effect of "condition" (a), the main effect of "block" (b), the "condition x block" interaction (c) and the "sequence x block" interaction (d). * $p < .05$, *** $p < .001$.

we performed separate univariate ANOVAs for the flat and stressed conditions with the within-subject factor "sequence" (structured and random). In this context, we separately evaluated ITC at the word rate, word learning index (WLI), ITC at the syllabic rate and syllable learning index (SLI, see methods). These analyses aimed at testing whether general neural synchronization to syllables and words likewise operates in statistical learning and prosodic bootstrapping conditions.

Analyses of the flat condition (Fig. 7) revealed that ITC at the word rate ($F_{(1, 29)} = 39.84$, $p < .001$) and the WLI ($F_{(1, 29)} = 7.92$, $p = .009$) were increased in structured compared to random sequences. In contrast, the SLI was higher in random compared to the structured sequences ($F_{(1, 29)} = 5.01$, $p = .033$), whereas ITC analyses at the syllabic rate did not reveal significant differences between structured and random streams ($F_{(1, 29)} = 3.43$, $p = .074$). These results suggest that statistical learning is mediated by neural synchronization to word units, whereas word rate and syllabic rate are possibly concurrently tracked. The latter assumption is consistent with the significant correlation we revealed between ITC at the syllabic rate and ITC at the word rate (Pearson's r, one-tailed, $r = 0.315$, $p = .045$, Fig. 8a) in structured sequences of flat speech. Furthermore, based on the ITC and ERP results, we correlated mean ITC at the word rate and the WLI with mean N400 amplitudes. Correlation analyses (Pearson's r, one-tailed, Bonferroni-corrected p-value for 2 tests, $p < .025$) yielded a significant negative relationship between ITC at the word rate and mean N400 amplitudes ($r = -0.604$, $p < .001$, Fig. 8b), whereas the correlation between WLI and mean N400 responses did not reach significance ($r = -0.050$, $p = .396$). Interestingly, the statistical analyses of the stressed condition (Fig. 7) did not reveal significant differences between structured and random sequences with respect to ITC at the word rate ($F_{(1, 22)} = 2.32$, $p = .142$), WLI ($F_{(1, 22)} = 1.42$, $p = .246$), ITC at the syllabic rate ($F_{(1, 22)} = 0.332$, $p = .57$) or SLI ($F_{(1, 22)} = 1.065$, $p = .313$).

3.4.2. *Neural synchronization as a function of exposure across blocks*

In a next complementary analysis, we examined whether there are dynamic changes in neural synchronization at the syllable and word rates across the blocks, and tested whether such neural modulations were differentially influenced by conditions and sequences. Accordingly, dynamic changes in neural synchronization were quantified by analyzing ITC values at the word and syllable rates across the four blocks using separate univariate ANOVAs with the within-subject factor "blocks" for structured and random sequences of flat and stressed speech. Analyses of the structured sequences of flat speech (Fig. 9) revealed main effects of "block" for both ITC at the word ($F_{(3, 87)} = 4.656$, $p = .005$) and syllable rates ($F_{(3, 87)} = 6.795$, $p < .001$), whereas in the random condition statistical analyses yielded significance only for ITC at the syllable rate (syllable rate: $F_{(3, 87)} = 16.958$, $p < .001$; word rate: $F_{(3, 87)} = 2.156$, $p = .099$).

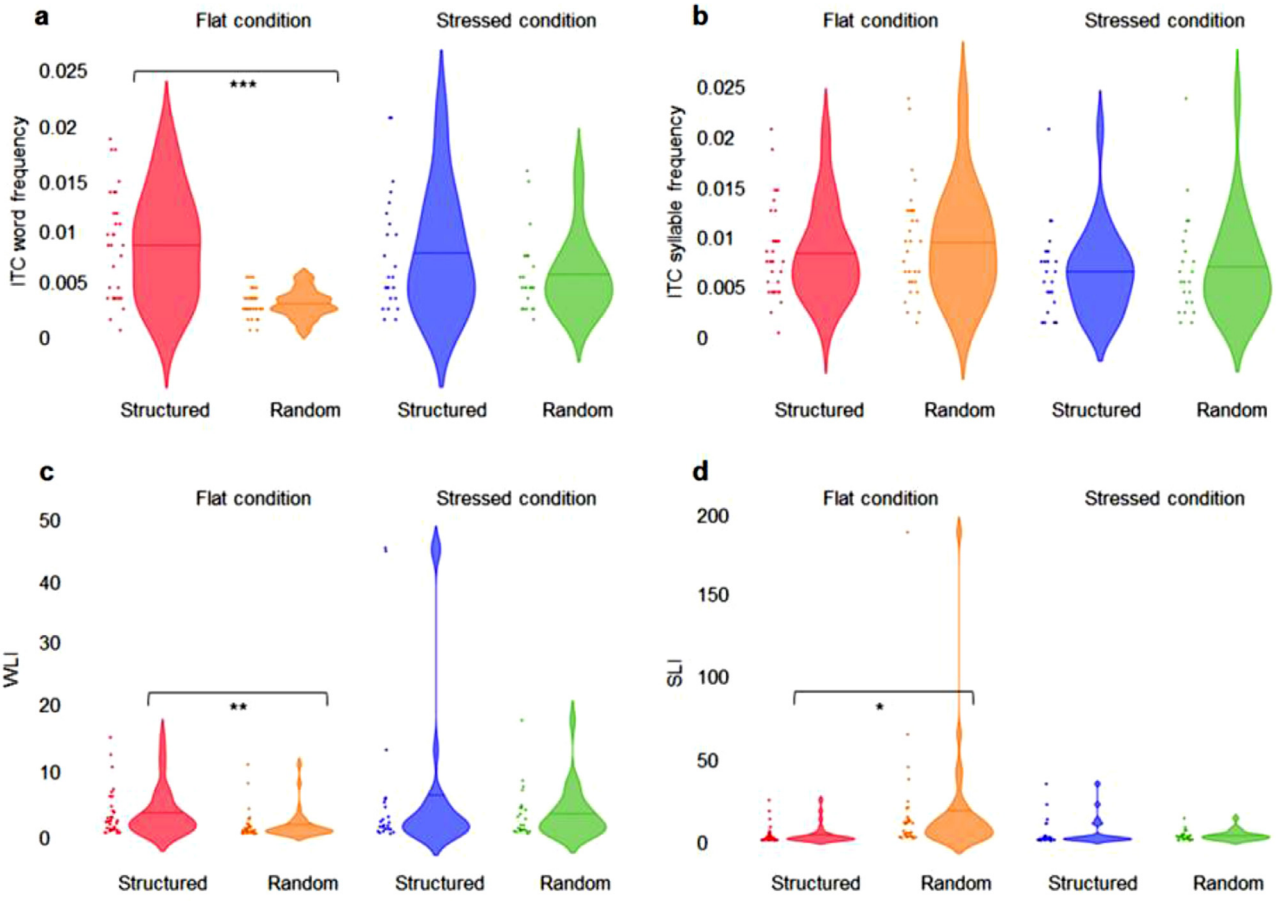

Fig. 7. Single-subject data and violin plots with density distribution and mean for the flat (left, $n = 30$) and stressed (right, $n = 23$) conditions and structured (red and blue) and random (orange and green) sequences. a = ITC at the word rate, b = ITC at the syllable rate, c = word learning index (WLI), d = syllable learning index (SLI). $^{*} = p < .05$, $^{**} = p < .01$, $^{***} = p < .001$.

Post-hoc comparisons (one-tailed, Bonferroni-corrected p value for 6 tests, $p < .008$) of the structured sequences of flat speech revealed increased neural synchronization to the word rate from the first to the third and fourth blocks (block 1_3: $t_{(29)} = -3.416$, $p = .001$; block 1_4: $t_{(29)} = -2.754$, $p = .005$). Furthermore, neural synchronization to the syllabic rate increased from the first to the second, third and fourth block (block 1_2: $t_{(29)} = -2.654$, $p = .006$; block 1_3: $t_{(29)} = -3.312$, $p = .001$; block 1_4: $t_{(29)} = -3.62$, $p < .001$). In a similar way, post-hoc analyses of the random sequences of flat speech revealed increased neural synchronization to the syllabic rate in the second, third and fourth block compared to the first one (block 1_2: $t_{(29)} = -6.113$, $p < .001$; block 1_3: $t_{(29)} = -5.823$, $p < .001$; block 1_4: $t_{(29)} = -5.177$, $p < .001$). All other comparisons did not reach significance (Table 4). Taken together, these results suggest that during flat speech neural oscillations dynamically synchronized to the syllabic rate, whereas speech segmentation based on statistical regularities was mediated by increased neural synchronization to both syllables and words across the four blocks.

Based on the parallel dynamic changes we observed in terms of ITC at the word rate and N400 amplitudes across the blocks of structured sequences of flat speech, we performed additional correlation analyses and assessed possible relationships between these two electrophysiological parameters separately for each block. Correlation analyses (Pearson's r, one-tailed, Bonferroni-corrected p value for 4 tests, $p < .012$, Fig. 8c-f) consistently yielded significant negative relationships between ITC at the word rate and N400 amplitudes in the first ($r = -0.612$, $p < .001$), second ($r = -0.634$, $p < .001$) and fourth ($r = -0.482$, $p = .003$) but not in the third block ($r = -0.359$, $p = .026$). In a similar way, drawing on the main effects of "block" we revealed for ITC at the syllable rate and P200 responses in both structured and random sequences of flat speech, we correlated these neural markers across the four blocks (Pearson's r, one-tailed, Bonferroni-corrected p value for 4 tests, $p < .012$). In structured sequences of flat speech, correlation analyses revealed a significant positive relationship between ITC at the syllabic rate and mean P200 amplitude in the fourth block (block 4: $r = 0.677$, $p < .001$; block 1: $r = 0.232$, $p = .109$; block 2: $r = 0.266$, $p = .078$; block 3: $r = 0.075$, $p = .347$, Fig. 10a). Analyses of the random sequences of flat speech highlighted consistent positive relationships between ITC at the syllable rate and mean P200 amplitude in the second, third and fourth block (block 2: $r = 0.522$, $p = .002$; block 3: $r = 0.652$, $p < .001$; block 4: $r = 0.534$, $p = .001$; block 1: $r = 0.034$, $p = .429$, Fig. 10b-d).

Statistical analysis of the stressed condition (Fig. 9) did not reveal significant effects of "block" neither for structured (word rate: $F_{(3, 66)} = 1.144$, $p = .338$; syllable rate: $F_{(3, 66)} = 1.266$, $p = .293$) nor for random (word rate: $F_{(3, 66)} = 0.298$, $p = .827$; syllable rate: $F_{(3, 66)} = 0.512$, $p = .675$) sequences. Accordingly, these results might suggest that in the presence of lexical stress cues ITC metrics were not sensitive enough to uncover dynamic changes in neural synchronization to syllables, whereas neural synchronization to the word rate possibly reflected a phase-resetting mechanism induced by the lexical stress cues.

4. Discussion

In the present EEG study, we examined the neural computations governing speech segmentation based on statistical learning and prosodic bootstrapping while participants learned new words embedded in continuous speech streams. In the main ITC analyses ($n = 23$), we directly

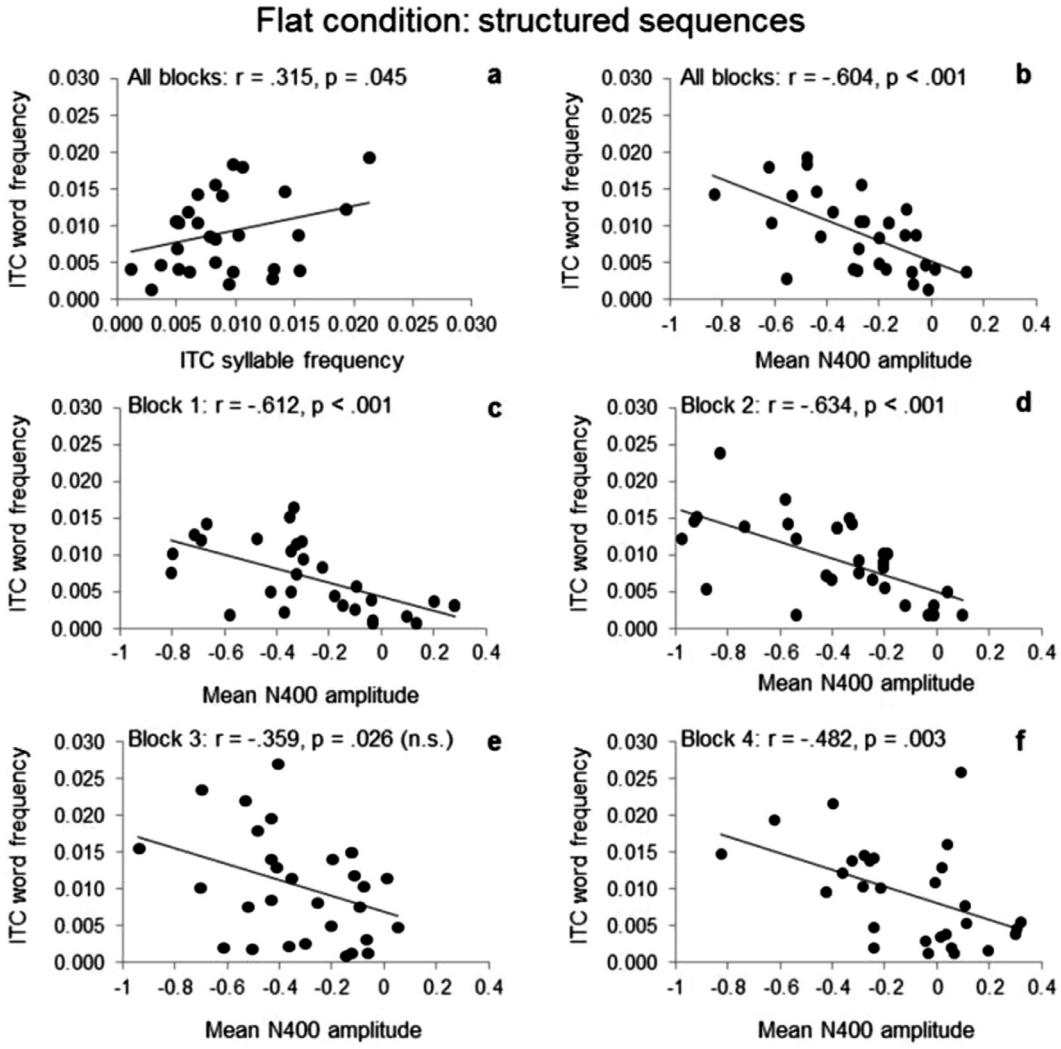

Fig. 8. Correlation analyses for structured sequences of flat speech (n = 30). a = Correlation between mean ITC at the syllable rate and word rate. b = Correlation between mean ITC at the word rate and mean N400 amplitude. c, d, e, f = Correlations between mean ITC at the word rate and mean N400 amplitude in the first (c), second (d), third (e) and fourth (f) block. n.s. = not significant after correction for multiple comparisons.

Table 4
Post-hoc comparisons of the significant main effects of "block" in the complementary ITC analyses. * Depicts significance after correction for multiple comparisons.

Rate	Condition	Contrast	Degrees of freedom	t-value	p-value
Word	Structured flat speech	Block 1 vs. Block 2	29	−2.440	0.010
		Block 1 vs. Block 3	29	−3.416	0.001*
		Block 1 vs. Block 4	29	−2.754	0.005*
		Block 2 vs. Block 3	29	−0.748	0.230
		Block 2 vs. Block 4	29	0.189	0.425
		Block 3 vs. Block 4	29	1.003	0.162
Syllable	Structured flat speech	Block 1 vs. Block 2	29	−2.654	0.006*
		Block 1 vs. Block 3	29	−3.312	0.001*
		Block 1 vs. Block 4	29	−3.620	< 0.001*
		Block 2 vs. Block 3	29	−0.965	0.172
		Block 2 vs. Block 4	29	−1.625	0.058
		Block 3 vs. Block 4	29	−1.290	0.104
Syllable	Random flat speech	Block 1 vs. Block 2	29	−6.113	< 0.001*
		Block 1 vs. Block 3	29	−5.823	< 0.001*
		Block 1 vs. Block 4	29	−5.177	< 0.001*
		Block 2 vs. Block 3	29	−1.749	0.045
		Block 2 vs. Block 4	29	−0.493	0.313
		Block 3 vs. Block 4	29	1.206	0.119

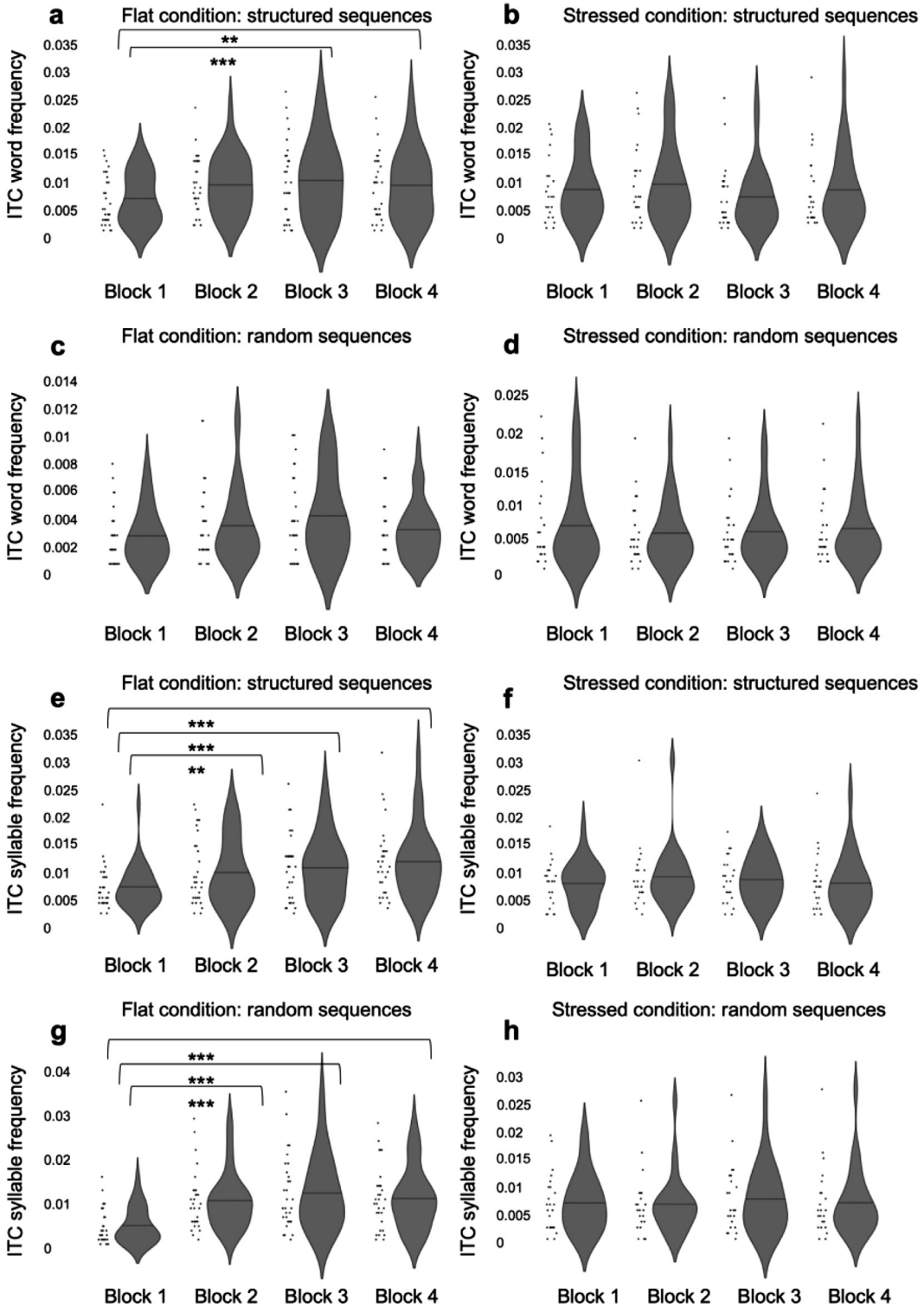

Fig. 9. Single-subject data and violin plots with density distribution and mean for the four blocks. ITC at the word rate (a, b, c, d) and ITC at the syllable rate (e, f, g, h) are shown for the flat (a, c, e, g; $n = 30$) and stressed conditions (b, d, f, h; $n = 23$) as well as for structured (a, b, e, f) and random sequences (c, d, g, h). ** = $p < .01$, *** = $p < .001$.

compared the flat and stressed conditions using separate 2 × 2 × 4 (2 conditions, 2 sequences and 4 blocks) ANOVAs, and analyzed ITC at the syllable rate and ITC at the word rate. This approach aimed at testing (1) whether neural synchronization to words and syllables likewise operates under statistical learning and prosodic bootstrapping conditions. Moreover, we examined (2) whether syllable transition probability contributes to speech segmentation when lexical stress cues can be used to directly extract word forms. Finally, we were also interested in (3) whether speech segmentation and word learning are generally mediated by a neural transition from syllabic rate to word rate, or whether the two time scales are concurrently tracked. For reasons of completeness and for the sake of comparability with previous studies (Batterink and

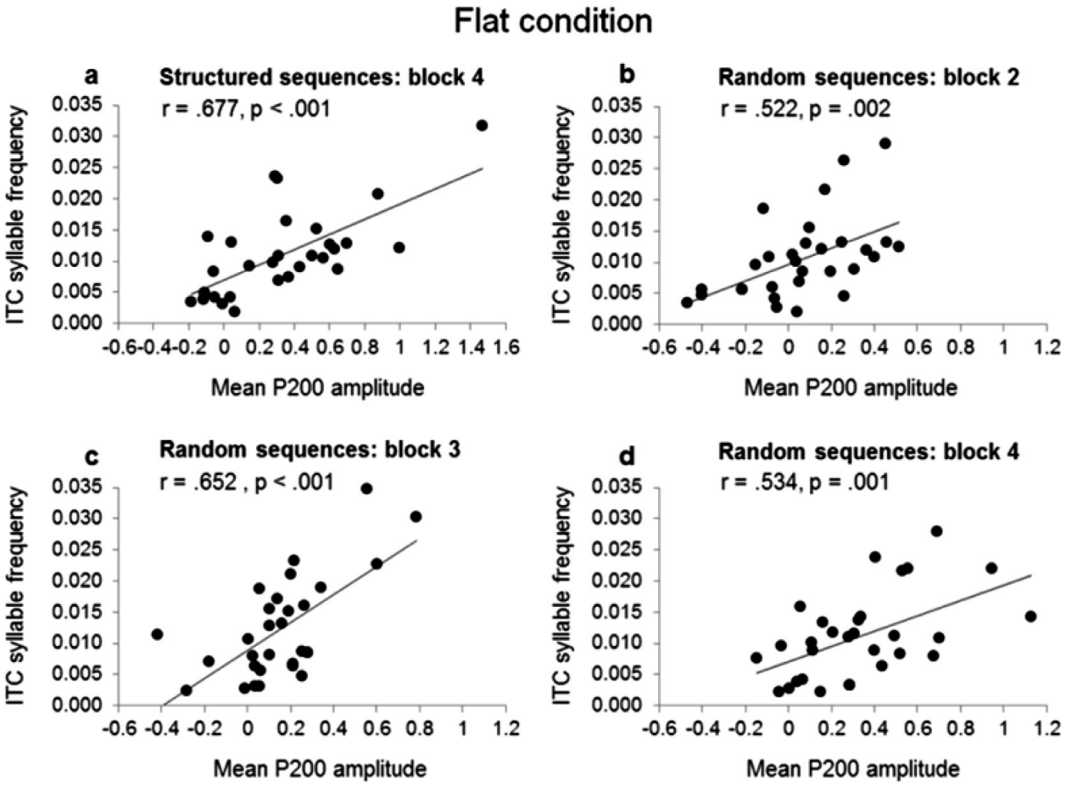

Fig. 10. Significant correlations between mean P200 amplitudes and ITC at the syllable rate in the flat condition (n = 30). *a* = block 4 of structured sequences, *b* = block 2 of random sequences, *c* = block 3 of random sequences, *d* = block 4 of random sequences.

Paller, 2017, 2019; Cunillera et al., 2009, 2006), we also evaluated mean P200 and N400 amplitudes. Furthermore, to exploit the full range of measured participants (flat condition $n = 30$, stressed condition $n = 23$), we conducted complementary ITC analyses and assessed relationships between ERPs and ITC metrics.

The results of the main analyses showed that ITC at the word rate was (1) generally higher in structured compared to random sequences (main effect of sequence), and that (2) this effect was more pronounced in the flat condition (condition x sequence interaction). Furthermore, somewhat surprisingly, in the flat condition we revealed (3) a marginally significant modulation of ITC at the word rate across the blocks (condition x block interaction), with higher ITC values in block 3 compared to block 1. Otherwise, the analysis of ITC at the syllable rate revealed (1) generally higher values in the flat compared to the stressed condition (main effect of condition). Moreover, (2) in both the flat and stressed conditions ITC at the syllable rate increased across the blocks (main effect of block). However, as also confirmed by the complementary ITC analyses, (3) this effect was mainly restricted to the flat condition (condition x block interaction). Interestingly, (4) we also noticed a marginal "sequence x block" interaction effect that was related to increased ITC values at the syllabic rate in the third block of random compared to structured sequences, irrespective of condition.

The evaluation of the P200 and N400 ERP components yielded results that were broadly compatible with the ITC metrics. In fact, in the flat condition, (1) P200 responses did not differ between structured and random sequences but generally increased across the blocks. Furthermore, in the flat condition, (2) N400 responses were larger in structured compared to random sequences and dynamically changed across the blocks. Notably, in the flat condition we also revealed several relationships between EEG metrics. In particular, (1) in structured sequences ITC at the word rate correlated with N400 amplitudes, whereas (2) in random sequences ITC at the syllable rate correlated with P200 responses. Furthermore, in structured sequences of flat speech, we revealed (3) a positive relationship between ITC at the syllabic rate and ITC at the word rate.

4.1. Main ITC analyses: neural synchronization to pertinent speech units

The results of the main ITC analyses generally confirmed previous findings showing increased neural synchronization to the word rate in structured compared to random sequences (Batterink and Paller, 2017, 2019; Buiatti et al., 2009). Such a synchronization of neural oscillations to words may constitute the neural basis of speech segmentation based on statistical regularities and lexical stress cues. However, both the main and the complementary ITC results suggested that the neural tracking of words was particularly pronounced in the flat condition where prosodic cues could not be used to segment speech and extract word forms. In fact, the main analysis of ITC at the word rate also yielded a significant "condition x sequence" interaction effect, and post-hoc comparisons between structured and random sequences only reached significance in the flat condition. However, this does not mean that in the stressed condition there was no neural synchronization to words. In fact, as visible in Fig. 4 and 5, structured sequences of stressed speech were associated with a clear peak corresponding to the word rate. In this context, it is important to mention that the main difference between structured sequences of flat and stressed speech is that in the latter case statistical learning and prosodic bootstrapping interact. This implies that in the stressed condition both transitional probabilities between adjacent syllables and prosodic cues can be used to recognize word boundaries. Nevertheless, since in the stressed condition ITC at the word rate did not markedly differ between structured and random sequences, results might suggest that neural synchronization at the word rate was induced by the lexical stress cues, possibly through evoked activity or a phase-resetting mechanism (Zoefel et al., 2018). Otherwise, based on a previous EEG study showing that pitch and final lengthening have to occur in combination to trigger speech segmentation (Holzgrefe-Lang et al.,

2016), it is possible that the pitch manipulation we used was insufficient to trigger participants' prosodic bootstrapping.

The main ITC analysis performed on word rate metrics also revealed a "condition x block" interaction effect that was related to increased ITC values in block 3 compared to block 1 of the flat condition. Since this effect was not differentially affected by structured and random sequences, and in random sequences word extraction was not possible, such a result is difficult to explain. Nevertheless, we might speculate that the increased ITC at the word rate we observed in the third compared to the first block of the flat condition was possibly influenced by superimposed theta oscillations reflecting progressively increased selective attention to the auditory streams in order to optimize learning in the flat compared to the stressed condition (Clayton et al., 2015).

The computation of ITC metrics at the syllable rate yielded clear peaks in both structured and random sequences of flat and stressed speech (Fig. 4). Furthermore, ITC at the syllable rate was higher in the flat compared to the stressed condition (Fig. 6a), and in the flat condition, ITC values increased from block 1 to blocks 2, 3 and 4. Since such a modulation across the blocks of the stressed condition did not reach significance in the post-hoc analyses, our results suggest that in the presence of lexical stress cues syllabic information is constantly tracked without the need to recruit additional neural resources over time.

In contrast to the results of Buiatti and colleagues (Buiatti et al., 2009) as well as of Batterink and co-workers (Batterink and Paller, 2017), our data did not support the idea of a neural suppression or a linear decrease in ITC at the syllabic rate in structured sequences of flat speech. Given that we noticed an increase in ITC at the syllable rate across the blocks, our results rather suggest that during statistical learning syllables and words are concurrently tracked. Such a parallel synchronization to basic speech elements and learned higher-order word units would be in agreement with the assumption that single words can only be recognized based on transitional probabilities between adjacent syllables (Saffran et al., 1996a, 1996b) or chunk of syllables (Perruchet et al., 2014; Perruchet and Vinter, 1998). Furthermore, it is noteworthy to mention that several studies have shown that regularities in the envelope of the acoustic signal correlate with syllabic information (Giraud and Poeppel, 2012; Myers et al., 2019; Poeppel and Assaneo, 2020), and that syllables can even be tracked if speech is unattended or unintelligible (Howard and Poeppel, 2010). Drawing on this background, the linear increase in neural synchronization to syllables we noticed in the flat condition across the blocks might suggest an additional recruitment of top-down resources to optimize learning. In particular, based on previous work showing that selective attention leads to more robust neural synchronization to acoustic features (Obleser and Kayser, 2019a), the increased ITC at the syllabic rate we revealed across the blocks of the flat condition leads to suggest that the participants focused more attention on the syllables to further ameliorate word learning performance.

The argument that syllables and words are concurrently tracked has previously already been proposed by other authors in the context of sentence processing (Ding et al., 2017; Giraud and Poeppel, 2012), and would also explain the significant positive relationship we revealed between ITC at the syllabic rate and ITC at the word rate in structured sequences of flat speech (Fig. 8a). Such a perspective would also be in line with previous work showing that speech segmentation is facilitated when multiple cues are available in one or even more sensory modalities (Altmann, 2002; Cunillera et al., 2010). Interestingly, in structured sequences of flat speech we also revealed negative correlations between ITC at the word rate and mean N400 amplitudes (Fig. 8b-f). Since the N400 component oscillates in the range of the delta (1–2 Hz) frequency (Cunillera et al., 2006) which roughly corresponds to the word rate (1.43 Hz), we may speculate that single-trial N400 responses were reflected in the ITC spectra. Nevertheless, further methodological studies are needed to clarify the direction of interaction between ITCs and ERPs (Van Diepen et al., 2019; Zoefel et al., 2018).

Finally, it is important to mention that a dynamic increase in neural synchronization to the syllabic rate was also observed in random sequences of flat speech. This perspective is compatible with previous work showing that syllabic information correlates with regularities in the envelope of the acoustic signal (Giraud and Poeppel, 2012; Myers et al., 2019; Poeppel and Assaneo, 2020), and substantiates the conclusion that neural synchronization to syllabic units constitutes an intrinsic neural principle underlying speech processing (Giraud and Poeppel, 2012; Hyafil et al., 2015; Makov et al., 2017; Pefkou et al., 2017). Somewhat surprisingly, the main analysis of ITC at the syllable rate also led to a "sequence x block" interaction effect that originated from increased ITC values in block 3 of random compared to structured sequences. Even though this result was unexpected and is difficult to explain, it might possibly reflect a reorientation of attention (Obleser and Kayser, 2019b) toward syllables to try to capture statistical regularities and extract word forms, even if learning was not possible in random sequences.

4.2. Additional insights from the complementary ITC analyses: neural synchronization as a function of exposure across blocks

Since the results of the main and complementary ITC analyses were broadly similar, in the present section we will only discuss some divergent findings that have not yet been addressed. A main result of the complementary ITC analyses of structured and random sequences across the four blocks was that we only revealed dynamic changes in neural synchronization in the flat speech condition. In particular, in structured streams ITC at the word and syllabic rate conjointly increased across blocks, whereas in random sequences ITC only increased at the syllable rate. In contrast, in the stressed condition ITC at the syllable and word rate did not linearly increase across the four blocks neither in structured nor in random sequences.

The increased neural synchronization to word units we revealed in the third and fourth blocks of structured sequences of flat speech is in line with previous results of Batterink and colleagues (Batterink and Paller, 2017) showing a linear increase in ITC across three blocks of approximately four minutes each. Recently, also Henin and co-workers (Henin et al., 2019) who collected intracranial recordings in 23 patients found that phase coherence at the word rate emerged after only about four minutes of exposure. Furthermore, in accordance with previous studies showing a contribution of the left dorsal stream to speech segmentation based on statistical regularities (Cunillera et al., 2009; Lopez-Barroso et al., 2013), Henin and colleagues demonstrated that neural synchronization to word units was particularly pronounced in two cluster of electrodes located over the supratemporal plane and the inferior frontal gyrus (Henin et al., 2019). The rapid neural synchronization we observed at the word rate is also compatible with a previous ERP study of Cunillera and colleagues (Cunillera et al., 2009), who compared structured and random sequences and found N400 manifestations as early as in the second block.

Most importantly, no previous studies have examined neural synchronization mechanisms at the intersection between statistical learning and prosodic bootstrapping in comparison to a pure statistical learning condition. Therefore, our study provides important insights into the influence of stress cues on ITC sensitivities and dynamics. Thereby, it is noteworthy to mention that we did not find any evidence for a dynamic increase in neural synchronization at the word rate across the four blocks in structured sequences of stressed speech. Therefore, our data suggest that lexical stress cues induced evoked activity or even a trial-by-trial phase-resetting of neural oscillations at the word rate without further spectrum for dynamic changes over time (Zoefel et al., 2018).

4.3. Statistical learning vs. prosodic bootstrapping, what makes the difference? Some additional theoretical considerations

Our results revealed distinct neural synchronization to syllables and words while the participants segmented speech and learned new words

based on statistical learning and prosodic bootstrapping. Although we already discussed the mechanisms that might be at the basis of the observed effects, some other theoretical considerations have to be discussed. First, it is important to remark that the data were collected in Barcelona and that in the Spanish and Catalan languages not the first but rather the penultimate and the final syllable are stressed. Therefore, it is possible that the lexical stress cues did not influence ITC at the word rate as expected because the prosodic manipulation was unusual for the participants. This argument would also be in line with the behavioral data showing that the mean hit rate in the 2AFC task was lower in the stressed compared to the flat condition. The view that native prosody influences the segmentation of artificial speech has, for example, been shown in American (Saffran et al., 1996b) and Swiss French (Bagou et al., 2002) adults.

Several year ago, Shukla and colleagues (Shukla et al., 2007) performed a series of behavioral experiments to investigate the interaction between statistical and prosodic cues to extract words from speech streams. In this context, the authors proposed that statistical learning and prosodic bootstrapping are processed independently, and that prosodic cues can block the computations of statistical regularities that span prosodic boundaries by filtering them. Even though in our experiment we did not analyze statistical regularities across prosodic boundaries, we may speculate whether the filtering mechanisms proposed by Shukla and colleagues might also explain that we did not find strong arguments for dynamic changes in ITC at the syllable rate across blocks in the stressed condition. Nevertheless, further empirical data are needed to validate this theoretical model under different listening conditions.

4.4. P200 component

For reasons of consistency and comparability with other studies, ERP analyses were restricted to two specific components that have been shown to be sensitive to speech segmentation based on statistical regularities and prosodic cues (Batterink and Paller, 2017, 2019; Cunillera et al., 2006). With this purpose in mind, we focused on the P200 and N400 components, and separately evaluated mean amplitudes across the four blocks of the flat and stressed conditions. In line with previous work (Cunillera et al., 2006), structured sequences of stressed speech elicited larger P200 responses than random sequences, whereas a comparable P200 modulation was not observed in the flat condition. However, P200 analyses also documented new findings that might be particularly interesting for a better understanding of the relationships between ERPs and ITCs. In particular, we noticed that P200 amplitudes conjointly increased with ITC at the syllabic rate over time in structured and random sequences of flat speech. In contrast, in the stressed condition this was not the case and P200 amplitudes generally decreased in the last two blocks compared to the first one.

In the context of an EEG study carried out on a subgroup of participants also included in the present work, Cunillera and colleagues (Cunillera et al., 2006) identified the P200 component as a distinctive electrophysiological marker of speech segmentation in structured stressed sequences. As revised in the introduction, P200 modulations in the context of speech segmentation tasks are usually associated with neural sources in primary and secondary auditory regions. Hence, increased P200 amplitudes are thought to reflect the detection of relevant prosodic cues that might direct attention toward word boundaries and facilitate the extraction of word forms during learning (Cunillera et al., 2006; de Diego-Balaguer et al., 2015; De Diego Balaguer et al., 2007; Francois and Schon, 2011; Rodriguez-Fornells et al., 2009). This perspective is also compatible with previous EEG studies showing that the P200 component is modulated by auditory attention (Rif et al., 1991; Rosburg et al., 2009) and sensitive to pitch (Shahin et al., 2003b; Trainor et al., 2003) and prosody (Paulmann and Kotz, 2008; Pinheiro et al., 2015). Nevertheless, it is important to mention that in the flat condition P200 responses and ITC at the syllabic rate conjointly increased across the four blocks (Fig. 3 and 6), and ITC at the syllable rate correlated with mean P200 amplitudes (Fig. 10) in both structured (block 4) and random sequences (blocks 2, 3 and 4). Such a compliance of ITCs and ERPs is particularly interesting for two reasons. First, because it introduces the idea that both EEG parameters are anchored on a common neural mechanism, and second, because it confirms previous ideas that the attentive tracking of syllabic units over time is related to P200 manifestations. However, since in the stressed condition a similar progression was not visible and P200 amplitudes generally decreased in the last two blocks compared to the first one, we might speculate that the presence of additional prosodic cues induced neural adaptation in the auditory cortex as reflected by lower P200 amplitudes (Grill-Spector et al., 2006; Hyde et al., 2008).

4.5. N400 component

The results of the flat condition replicated previous EEG findings showing increased N400 amplitudes in structured compared to random sequences (Batterink and Paller, 2017, 2019; Cunillera et al., 2009, 2006). Furthermore, as previously already reported by Cunillera and colleagues (Cunillera et al., 2009), the N400 component was characterized by a U-shaped response pattern with smallest amplitudes in the last block. The fast emergence and configuration of the N400 component in the first three blocks of structured sequences is interpreted as a marker of speech segmentation reflecting the codification and strengthening of episodic memory traces for linguistic representations or novel words (Batterink and Paller, 2017, 2019; Cunillera et al., 2009, 2006). This argument is reinforced by the significant correlations we observed between mean N400 amplitudes and ITC at the word rate (Fig. 8). Otherwise, the intrinsic meaning of decreased N400 amplitudes in the fourth block is somewhat unclear. In fact, reduced N400 amplitudes have previously been associated with an optimized access to verbal memory as a function of learning (Kutas and Hillyard, 1980; Vanpetten and Kutas, 1990). Nevertheless, based on our ERP and ITC data it is necessary to envisage an alternative interpretation. Notably, as visible in Fig. 4c, in structured sequences neural synchronization to the syllable rate significantly increased after the first block, and in the fourth block it reached almost the same level as that of random sequences. Furthermore, by subtracting neural synchronization to the syllable rate in random sequences from structured sequences (Fig. 3a), we obtained a U-shaped function that roughly coincides with the time course of the N400 component (Fig. 3b). However, this was not the case in the stressed condition (Fig. 3c). This observation is rooted in the notion that the reduced N400 amplitude we revealed in the fourth block might be somehow related to increased syllabic tracking. This argument is not only supported by the positive correlation we revealed between ITC at the syllable rate and P200 amplitudes in the fourth block of structured sentences, but also by the steady increase of both parameters over time. A possible explanation might be that learning based on statistical principles follows a logarithmic function (Mirman et al., 2008), where the increment of learning is the strongest at the beginning and saturates with deliberate practice. Hence, we speculate that a change in attentional focus from words to syllables in the last block might constitute a strategy to further improve learning.

Conclusions

In the present EEG study, we examined neural synchronization to syllables and words during speech segmentation based on statistical information and lexical stress cues. Results demonstrated concurrent neural synchronization to pertinent speech units in both experimental conditions. However, neural synchronization to words in structured compared to random sequences was more pronounced in the flat condition. Otherwise, the tracking of syllabic information was increased in the flat compared to the stressed condition, and a dynamic increase in ITC at the syllable rate was only observed across the blocks of the flat condition. Importantly, we also revealed robust correlations between ITC indexes

and ERP components (P200/N400) that have previously been associated with speech segmentation. Taken together, our results corroborate the existence of different computational principles governing neural synchronization to pertinent linguistic units during statistical learning with and without concurrent prosodic cues.

Data and code availability statement

The codes used for computing ITC analyses are available from SAV, whereas the EEG data of this study are available from TC upon request. Due to ethical considerations the data cannot be made openly available. However, the data will be shared upon request without any restrictions.

Authors contribution

ARF, SE and TC planned the study, TC performed the EEG measurements and SAV wrote the MATLAB scripts and computed ITC analyses. SE and SAV analyzed the EEG data and performed the statistical analyses. SE, SAV, TC and ARF contributed to the interpretation of the data and wrote the manuscript.

Code availability

The codes used for computing ITC analyses are available from SAV upon reasonable request.

Data availability

The EEG data of this study are available from TC upon request.
Fig. 1

Declaration of Competing Interest

The authors declare no competing interests.

Credit authorship contribution statement

Stefan Elmer: Conceptualization, Formal analysis, Writing – original draft, Writing – review & editing, Visualization, Project administration, Supervision. **Seyed Abolfazl Valizadeh:** Conceptualization, Formal analysis, Writing – review & editing. **Toni Cunillera:** Conceptualization, Investigation, Data curtion, Writing – review & editing. **Antoni Rodriguez-Fornells:** Conceptualization, Writing – review & editing, Project administration, Supervision.

Acknowledgments

This research was supported by research grants from the Spanish Government (MCYT) to A.R.F. (with EC Funds FEDER SEJ2005-06067/PSIC).

References

Altmann, G.T.M., 2002. Statistical learning in infants. In: Proceedings of the National Academy of Sciences of the United States of America, 99, pp. 15250–15251.
Assaneo, M.F., Ripolles, P., Orpella, J., Lin, W.M., de Diego-Balaguer, R., Poeppell, D., 2019. Spontaneous synchronization to speech reveals neural mechanisms facilitating language learning. Nat. Neurosci. 22 627–+.
Bagou, O., Fougeron, C., Frauenfelder, U., 2002. Contribution of prosody to the segmentation and storage of "words" in the acquisition of a new mini-language. In: Proceedings of the Speech Prosody 2002 conference, pp. 59–62.
Batterink, L.J., Paller, K.A., 2017. Online neural monitoring of statistical learning. Cortex 90, 31–45.
Batterink, L.J., Paller, K.A., 2019. Statistical learning of speech regularities can occur outside the focus of attention. Cortex 115, 56–71.
Bosnyak, D.J., Eaton, R.A., Roberts, L.E., 2004. Distributed auditory cortical representations are modified when non-musicians are trained at pitch discrimination with 40Hz amplitude modulated tones. Cereb. Cortex 14, 1088–1099.
Brent, M.R., 1999. Speech segmentation and word discovery: a computational perspective. Trends Cogn. Sci. 3, 294–301.
Buiatti, M., Pena, M., Dehaene-Lambertz, G., 2009. Investigating the neural correlates of continuous speech computation with frequency-tagged neuroelectric responses. Neuroimage 44, 509–519.
Christophe, A., Dupoux, E., Bertoncini, J., Mehler, J., 1994. Do infants perceive word boundaries? An empirical study of the bootstrapping of lexical acquisition. J. Acoust. Soc. Am. 95, 1570–1580.
Clayton, M.S., Yeung, N., Cohen Kadosh, R., 2015. The roles of cortical oscillations in sustained attention. Trends Cogn. Sci. 19, 188–195.
Cunillera, T., Camara, E., Toro, J.M., Marco-Pallares, J., Sebastian-Galles, N., Ortiz, H., Pujol, J., Rodriguez-Fornells, A., 2009. Time course and functional neuroanatomy of speech segmentation in adults. Neuroimage 48, 541–553.
Cunillera, T., Laine, M., Camara, E., Rodriguez-Fornells, A., 2010. Bridging the gap between speech segmentation and word-to-world mappings: evidence from an audiovisual statistical learning task. J. Mem. Lang. 63, 295–305.
Cunillera, T., Toro, J.M., Sebastian-Galles, N., Rodriguez-Fornells, A., 2006. The effects of stress and statistical cues on continuous speech segmentation: an event-related brain potential study. Brain Res. 1123, 168–178.
Cutler, A., Norris, D., 1988. The role of strong syllables in segmentation for lexical access. J. Exp. Psychol. -Hum. Percept. Perform. 14, 113–121.
de Diego-Balaguer, R., Rodriguez-Fornells, A., Bachoud-Levi, A.C., 2015. Prosodic cues enhance rule learning by changing speech segmentation mechanisms. Front. Psychol. 6, 1478.
De Diego Balaguer, R., Toro, J.M., Rodriguez-Fornells, A., Bachoud-Levi, A.C., 2007. Different neurophysiological mechanisms underlying word and rule extraction from speech. PLoS ONE 2, e1175.
Ding, N., Melloni, L., Yang, A., Wang, Y., Zhang, W., Poeppel, D., 2017. Characterizing neural entrainment to hierarchical linguistic units using electroencephalography (EEG). Front. Hum. Neurosci. 11.
Ding, N., Melloni, L., Zhang, H., Tian, X., Poeppel, D., 2016. Cortical tracking of hierarchical linguistic structures in connected speech. Nat. Neurosci. 19 158-+.
Dutoit, T., Pagel, V., Pierret, N., Bataille, F., vanderVrecken, O., 1996. The MBROLA project: towards a set of high quality speech synthesizers free of use for non commercial purposes. In: Icslp 96 - Fourth International Conference on Spoken Language Processing, Proceedings, 1-4, pp. 1393–1396 Vols.
Elmer, S., Albrecht, J., Valizadeh, S.A., Francois, C., Rodriguez-Fornells, A., 2018. Theta coherence asymmetry in the dorsal stream of musicians facilitates word learning. Sci. Rep. 8.
Francois, C., Schon, D., 2011. Musical expertise boosts implicit learning of both musical and linguistic structures. Cereb. Cortex 21, 2357–2365.
Fritz, J.B., Elhilali, M., David, S.V., Shamma, S.A., 2007. Auditory attention–focusing the searchlight on sound. Curr. Opin. Neurobiol. 17, 437–455.
Ghitza, O., 2011. Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input rhythm. Front. Psychol. 2.
Giraud, A.L., Poeppel, D., 2012. Cortical oscillations and speech processing: emerging computational principles and operations. Nat. Neurosci. 15, 511–517.
Gleitman, L.R., Wanner, E., 1982. Language acquisition: the state of the state of the art. In: Wanner, E., Gleitman, L.R. (Eds.), Language Acquisition: The State of the Art. Cambridge University Press, Cambridge, England.
Grill-Spector, K., Henson, R., Martin, A., 2006. Repetition and the brain: neural models of stimulus-specific effects. Trends Cogn. Sci. 10, 14–23.
Henin, S., Turk-Browne, N., Friedman, D., Liu, A., Dugan, P., Flinker, A., Doyle, W., DEvinsky, O., Melloni, L., 2019. Statistical learning shapes neural sequence representations. bioRxiv.
Holzgrefe-Lang, J., Wellmann, C., Petrone, C., Räling, R., Truckenbrodt, H., Höhle, B., Wartenburger, I., 2016. How pitch change and final lengthening cueboundary perception in German: convergingevidence from ERPs and prosodic judgements. Language. Cogn. Neurosci. 31, 904–920.
Howard, M.F., Poeppel, D., 2010. Discrimination of speech stimuli based on neuronal response phase patterns depends on acoustics but not comprehension. J. Neurophysiol. 104, 2500–2511.
Hyafil, A., Fontolan, L., Kabdebon, C., Gutkin, B., Giraud, A.L., 2015. Speech encoding by coupled cortical theta and gamma oscillations. Elife 4.
Hyde, K.L., Peretz, I., Zatorre, R.J., 2008. Evidence for the role of the right auditory cortex in fine pitch resolution. Neuropsychologia 46, 632–639.
Johnson, E.K., Jusczyk, P.W., 2001. Word segmentation by 8-month-olds: when speech cues count more than statistics. J. Mem. Lang. 44, 548–567.
Jung, T.P., Makeig, S., Humphries, C., Lee, T.W., McKeown, M.J., Iragui, V., Sejnowski, T.J., 2000. Removing electroencephalographic artifacts by blind source separation. Psychophysiology 37, 163–178.
Jusczyk, P.W., Houston, D.M., Newsome, M., 1999. The beginnings of word segmentation in English-learning infants. Cognit. Psychol. 39, 159–207.
Kuhl, P.K., 2004. Early language acquisition: cracking the speech code. Nat. Rev. Neurosci. 5, 831–843.
Kutas, M., Hillyard, S.A., 1980. Reading senseless sentences - brain potentials reflect semantic incongruity. Science 207, 203–205.
Lehiste, I., 1960. An acoustic-phonetic study of internal open juncture. Phonetica 5, S3–S54.
Liegeois-Chauvel, C., Musolino, A., Badier, J.M., Marquis, P., Chauvel, P., 1994. Evoked potentials recorded from the auditory cortex in man: evaluation and topography of the middle latency components. Electroencephalogr. Clin. Neurophysiol. 92, 204–214.
Lopez-Barroso, D., Catani, M., Ripolles, P., Dell'Acqua, F., Rodriguez-Fornells, A., de Diego-Balaguer, R., 2013. In: Word Learning is Mediated by the Left Arcuate Fasciculus, 110. Proceedings of the National Academy of Sciences of the United States of America, pp. 13168–13173.
Luck, S.J., Hillyard, S.A., 1994. Electrophysiological correlates of feature analysis during visual search. Psychophysiology 31, 291–308.

Makov, S., Sharon, O., Ding, N., Ben-Shachar, M., Nir, Y., Golumbic, E.Z., 2017. Sleep disrupts high-level speech parsing despite significant basic auditory processing. J. Neurosci. 37, 7772–7781.
Mattys, S.L., Jusczyk, P.W., 2001. Do infants segment words or recurring contiguous patterns? J. Exp. Psychol. Hum. Percept. Perform. 27, 644–655.
Mattys, S.L., White, L., Melhorn, J.F., 2005. Integration of multiple speech segmentation cues: a hierarchical framework. J. Exp. Psychol. -Gen. 134, 477–500.
Meyer, L., Henry, M.J., Gaston, P., Schmuck, N., Friederici, A.D., 2017. Linguistic bias modulates interpretation of speech via neural delta-band oscillations. Cereb. Cortex 27, 4293–4302.
Mirman, D., Magnuson, J.S., Estes, K.G., Dixon, J.A., 2008. The link between statistical segmentation and word learning in adults. Cognition 108, 271–280.
Myers, B.R., Lense, M.D., Gordon, R.L., 2019. Pushing the envelope: developments in neural entrainment to speech and the biological underpinnings of prosody perception. Brain Sci. 9.
Norris, D., McQueen, J.M., Cutler, A., 2000. Merging information in speech recognition: feedback is never necessary. Behav. Brain Sci. 23 299-+.
Obleser, J., Kayser, C., 2019a. Neural entrainment and attentional selection in the listening brain. Trends Cogn. Sci. 23, 913–926.
Obleser, J., Kayser, C., 2019b. Neural entrainment and attentional selection in the listening brain. Trends Cogn. Sci. 23, 913–926.
Panzeri, S., Brunel, N., Logothetis, N.K., Kayser, C., 2010. Sensory neural codes using multiplexed temporal scales. Trends Neurosci. 33, 111–120.
Paulmann, S., Kotz, S.A., 2008. Early emotional prosody perception based on different speaker voices. Neuroreport 19, 209–213.
Pefkou, M., Arnal, L.H., Fontolan, L., Giraud, A.L., 2017. Theta-band and beta-band neural activity reflects independent syllable tracking and comprehension of time-compressed speech. J. Neurosci. 37, 7930–7938.
Perruchet, P., Poulin-Charronnat, B., Tillmann, B., Peereman, R., 2014. New evidence for chunk-based models in word segmentation. Acta Psychol. (Amst) 149, 1–8.
Perruchet, P., Vinter, A., 1998. PARSER: a model for word segmentation. J. Mem. Lang. 39, 246–263.
Picton, T.W., Alain, C., Woods, D.L., John, M.S., Scherg, M., Valdes-Sosa, P., Bosch-Bayard, J., Trujillo, N.J., 1999. Intracerebral sources of human auditory-evoked potentials. Audiol. Neurootol. 4, 64–79.
Pinheiro, A.P., Vasconcelos, M., Dias, M., Arrais, N., Goncalves, O.F., 2015. The music of language: an ERP investigation of the effects of musical training on emotional prosody processing. Brain Lang. 140, 24–34.
Poeppel, D., Assaneo, M.F., 2020. Speech rhythms and their neural foundations. Nat. Rev. Neurosci. 21, 322–334.
Rentzsch, J., Jockers-Scherubl, M.C., Boutros, N.N., Gallinat, J., 2008. Test-retest reliability of P50, N100 and P200 auditory sensory gating in healthy subjects. Int. J. Psychophysiol. 67, 81–90.
Rif, J., Hari, R., Hamalainen, M.S., Sams, M., 1991. Auditory attention affects two different areas in the human supratemporal cortex. Electroencephalogr. Clin. Neurophysiol. 79, 464–472.
Rodriguez-Fornells, A., Cunillera, T., Mestres-Misse, A., de Diego-Balaguer, R., 2009. Neurophysiological mechanisms involved in language learning in adults. Philos. Trans. R. Soc. B-Biol. Sci. 364, 3711–3735.
Rosburg, T., Trautner, P., Elger, C.E., Kurthen, M., 2009. Attention effects on sensory gating - Intracranial and scalp recordings. Neuroimage 48, 554–563.
Saffran, J.R., Aslin, R.N., Newport, E.L., 1996a. Statistical learning by 8-month-old infants. Science 274, 1926–1928.
Saffran, J.R., Newport, E.L., Aslin, R.N., 1996b. Word segmentation: the role of distributional cues. J. Mem. Lang. 35, 606–621.
Sauseng, P., Klimesch, W., 2008. What does phase information of oscillatory brain activity tell us about cognitive processes? Neurosci. Biobehav. Rev. 32, 1001–1013.
Scherg, M., Von Cramon, D., 1986. Evoked dipole source potentials of the human auditory cortex. Electroencephalogr. Clin. Neurophysiol. 65, 344–360.
Shahin, A., Bosnyak, D.J., Trainor, L.J., Roberts, L.E., 2003a. Enhancement of neuroplastic P2 and N1c auditory evoked potentials in musicians. J. Neurosci. 23, 5545–5552.
Shahin, A., Bosnyak, D.J., Trainor, L.J., Roberts, L.E., 2003b. Enhancement of neuroplastic P2 and N1c auditory evoked potentials in musicians. J. Neurosci. 23, 5545–5552.
Shukla, M., Nespor, M., Mehler, J., 2007. An interaction between prosody and statistics in the segmentation of fluent speech. Cognit. Psychol. 54, 1–32.
Thiessen, E.D., Saffran, J.R., 2003. When cues collide: use of stress and statistical cues to word boundaries by 7-to 9-month-old infants. Dev. Psychol. 39, 706–716.
Trainor, L.J., Shahin, A., Roberts, L.E., 2003. Effects of musical training on the auditory cortex in children. Neurosci. Music 999, 506–513.
Tremblay, K.L., Ross, B., Inoue, K., McClannahan, K., Collet, G., 2014. Is the auditory evoked P2 response a biomarker of learning? Front. Syst. Neurosci. 8, 28.
Van Diepen, R.M., Foxe, J.J., Mazaheri, A., 2019. The functional role of alpha-band activity in attentional processing: the current zeitgeist and future outlook. Curr. Opin. Psychol. 29, 229–238.
van Diepen, R.M., Mazaheri, A., 2018. The caveats of observing inter-trial phase-coherence in cognitive neuroscience. Sci. Rep. 8.
Vanpetten, C., Kutas, M., 1990. Interactions between sentence context and word-frequency in event-related brain potentials. Mem. Cognit. 18, 380–393.
Zoefel, B., ten Oever, S., Sack, A.T., 2018. The involvement of endogenous neural oscillations in the processing of rhythmic input: more than a regular repetition of evoked neural responses. Front. Neurosci. 12.