NeuroImage 44 (2009) 509–519

Contents lists available at ScienceDirect

NeuroImage

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

Investigating the neural correlates of continuous speech computation with frequency-tagged neuroelectric responses

Marco Buiatti [a,b,c,d,*,1], Marcela Peña [e,f,g,1], Ghislaine Dehaene-Lambertz [a,b,c,h]

[a] INSERM, U562, Cognitive Neuroimaging Unit, F-91191 Gif/Yvette, France
[b] CEA, DSV/I2BM, NeuroSpin Center, F-91191 Gif/Yvette, France
[c] Univ Paris-Sud, IFR49, F-91191 Gif/Yvette, France
[d] Functional NeuroImaging Laboratory, Center for Mind/Brain Sciences, University of Trento, Via delle Regole, 101, 38060, Mattarello (TN) 38060, Italy
[e] International School for Advanced Studies, Trieste, Italy
[f] Escuela de Psicología, Pontificia Universidad Católica de Chile, Chile
[g] Centro de Investigación Avanzada en Educación, Universidad de Chile, Chile
[h] AP-HP, Service de Neuropédiatrie, Hôpital Bicêtre, F- 94275 Kremlin Bicêtre, France

ARTICLE INFO

Article history:
Received 8 May 2008
Revised 19 August 2008
Accepted 11 September 2008
Available online 27 September 2008

Keywords:
EEG
Explicit learning
Prosody
Speech segmentation
Steady-state response

ABSTRACT

In order to learn an oral language, humans have to discover words from a continuous signal. Streams of artificial monotonous speech can be readily segmented based on the statistical analysis of the syllables' distribution. This parsing is considerably improved when acoustic cues, such as subliminal pauses, are added suggesting that a different mechanism is involved. Here we used a frequency-tagging approach to explore the neural mechanisms underlying word learning while listening to continuous speech. High-density EEG was recorded in adults listening to a concatenation of either random syllables or tri-syllabic artificial words, with or without subliminal pauses added every three syllables. Peaks in the EEG power spectrum at the frequencies of one and three syllables occurrence were used to tag the perception of a monosyllabic or tri-syllabic structure, respectively. Word streams elicited the suppression of a one-syllable frequency peak, steadily present during random streams, suggesting that syllables are no more perceived as isolated segments but bounded to adjacent syllables. Crucially, three-syllable frequency peaks were only observed during word streams with pauses, and were positively correlated to the explicit recall of the detected words. This result shows that pauses facilitate a fast, explicit and successful extraction of words from continuous speech, and that the frequency-tagging approach is a powerful tool to track brain responses to different hierarchical units of the speech structure.

Introduction

While adults easily find the words in a sentence from their native language, the task becomes impossible when they listen to an unknown language (Pilon, 1981). This difficulty is due to the absence of robust physical cues at the boundaries of words inside utterances (Echols, 1993; Hayes and Clark, 1970; Pisoni and Luce, 1986), as spaces observed in written language. Computation of transitional probabilities between adjacent syllables has been proposed as one solution to extract words from the continuous speech stream. Since transitional probabilities are higher between two syllables within words than between two syllables encompassing different words, a word boundary would be placed when transitional probabilities systematically drop. Indeed, infants and adults have been shown able to exploit this cue in order to extract words from an artificial language (Saffran et al., 1996a,b). It was also shown that this computation is not restricted to contiguous syllables but can be realized on non-adjacent segments, either on the consonantal tier and the vowel tier of the words (Bonatti et al., 2005; Newport and Aslin, 2004) or between non adjacent syllables (De Diego Balaguer et al., 2007; Endress and Bonatti, 2007; Pena et al., 2002).

Natural speech is not a monotonous signal and is organized into cohesive prosodic units, such as intonational and phonological phrases, that reflect the morpho-syntactic organization of the sentences. These units encompass one or several words and their boundaries are signaled by acoustic cues, such as pitch decrease, final lengthening or pauses. These cues are spontaneously used by listeners to facilitate word access and limit word recognition to the segments within the prosodic unit. For example, responses to detect a word target (e.g. "chat" in French) are slower when neighboring words present in the lexicon have to be discarded (e.g. "chagrin", "chateau", etc.). This happens when there is a local lexical ambiguity within a

[*] Corresponding author. Functional NeuroImaging Laboratory, Center for Mind/Brain Sciences, University of Trento, Via delle Regole, 101, 38060, Mattarello (TN) 38060, Italy. Fax: +39 461 88 3066.
E-mail address: marco.buiatti@unitn.it (M. Buiatti).
[1] MB and MP equally contributed to this study.

1053-8119/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2008.09.015

phonological phrase like in [un chat grincheux] where both words "chat" and "chagrin" are activated before resolving the ambiguity. However, this delay is not observed when the local lexical ambiguity straddles a phonological phrase boundary like in [son grand chat] [grimpait…]. In this case, lexical access is limited to the domain of the phonological phrase and "chat" has no competitors (Christophe et al., 2004). Similarly using artificial streams, Shukla et al. (2007) have observed that the computation of transitional probabilities between adjacent syllables was limited within prosodic constituents and that the correct identification of artificial words significantly dropped from 68% to 45% when high transitional probabilities between adjacent syllables straddled over two prosodic constituents. Pena et al. (2002) also reported that word segmentation was facilitated, requiring a shorter exposure when a subliminal pause was inserted at the end of the words embedded in the stream. Thus, acoustical cues, especially those used to mark prosodic boundaries, can modulate and facilitate statistical information computation.

In order to investigate the neural correlates of word learning from continuous speech and the respective role of statistical distribution and of acoustical cues, we recorded high-density EEG in adult participants exposed to four experimental conditions using different continuous speech streams. Two of the streams conveyed statistical information (Word streams) and the other two (Random streams) consisted of a random concatenation of the same syllables used in the word streams. In Word streams, words were defined as structured tri-syllabic items in which the first syllable -A- predicted the occurrence of the third one -C- (see Table 1). Thus, the transitional probability between these syllables was 1, while the ones between adjacent syllables within and between words were considerably lower (0.33 to 0.5). Acoustic information was manipulated by adding a sub-liminal pause (25 ms) every three syllables in one stream of each condition (Random and Word).

Numerous studies on word segmentation in continuous streams share an intrinsic limitation: computations during the learning phase are classically inferred from recognition of words presented afterwards during a test phase (Pena et al., 2002; Saffran et al., 1996b). This precludes the study of on-line segmentation computations and transfers the question from on-line segmentation to recognition of memorized items. Strategic effects related to the statistical properties of the test items themselves may also appear during the test phase. Brain-imaging techniques, such as electro-encephalography, can be proposed to follow learning on-line during continuous speech presentation. However, several difficulties are encountered. First, the absence of on-line information about the success of learning makes it difficult to isolate its brain correlates. Second, learning during several minutes of exposure might be erratic and fickle, some items being segmented as possible words at one time, then others at another time, blurring an average response computed across all words presentation. Third, the computation of event-related potentials (ERPs) is problematic because the continuous nature of the stream complicates the estimation of a proper baseline. Finally, the continuous stream of information provided by the acoustic stream decreases the amplitude of the brain responses, and thus the signal to noise ratio. Nevertheless, different studies have proposed several ERP components that might signal word-segmentation, the N1, the P2 and the N400. Sanders et al. (2002) observed an increase of N1 for the first syllable of previously learned words. This N1 increase for word-onset has not been always recorded in following studies (Cunillera, 2008; Cunillera, 2006; Sanders and Neville, 2003a,b) and seems to be related to an attentional effect, being present when subjects already know the words and are expecting them in the stream. Cunillera et al. (2006, 2008) did not observe a N1 effect but rather report a P2 increase for stressed syllables embedded in non-sense words but not for stressed syllables embedded in a random stream. They did not observe any N1 nor P2 enhancement for monotonous streams without stressed syllables although the percentage of correctly detected words in that unstressed condition was similar to the performance in the stressed condition. De Diego Balaguer et al. (2007) also noted a P2 increase along the 4 min of exposure to an artificial language. This effect was more important in good than in poor learners. Thus, as for N1, P2 increase seems to be present when subjects are expecting a precise word onset once they have identified the words. A N400 increase appears to be a more robust and automatic index of word segmentation. It has been observed consistently when non-sense words are compared to unlearned words or to non-words in the artificial language (Cunillera et al., 2006, 2008; De Diego Balaguer et al., 2007; Sanders et al., 2002). De Diego Balaguer et al. (2007) suggested that this N400 modulation represents the construction of a pre-lexical trace for new words. In their experiment, the N400 appeared before the increase in P2 amplitude.

Table 1
Material used to generate the artificial speech streams

Continuous speech		Test items		
Stream	'Words'	'Words'	'Rule-words'	'Part-words'
1	puloki	puloki	pumiki	lokimi
	pudaki	pudaki	pufoki	lobemi
	punuki	punuki	pubeki	loRapu
	folobe	folobe	fokibe	dabepu
	fodabe	fodabe	fopube	daRapu
	fonube	fonube	fomibe	dakifo
	miloRa	miloRa	mipuRa	nuRafo
	midaRa	midaRa	mifoRa	nukifo
	minuRa	minuRa	mibeRa	nubemi
2	tomudu	tomudu	tolidu	mufeto
	togadu	togadu	tobadu	muvoto
	topidu	topidu	tofedu	muduli
	bamuvo	bamuvo	balivo	gaduba
	bagavo	bagavo	bafevo	gafeba
	bapivo	bapivo	baduvo	gavoto
	limufe	limufe	lidufe	piduli
	ligafe	ligafe	litofe	pivoli
	lipife	lipife	libafe	pifeba

List of 'words', 'rule words' and 'part words' used during learning and test phase. Each stream was generated by concatenating nine 'words' belonging to three different families characterized by an AxC structure (see Materials and methods). Each participant was tested with both streams either in the no-pause condition or in the pause condition. Streams were counterbalanced across participants. The same syllables were used in the Random streams.

In order to track and clearly dissociate brain responses to different units of the continuous speech structure, and to overcome the baseline problem, we propose here to use an alternative approach to ERPs: a "frequency-tagging" analysis. This analysis exploits the property of the brain electromagnetic activity to respond to a visual or auditory stimulus presented periodically at a specific temporal frequency by resonating at the same frequency during the stimulation period (steady-state response, hereafter indicated as SSR) (Picton et al., 2003). This effect is manifested in the electric/magnetic recordings by a sharp peak in the power spectrum of the signal at that specific frequency. Recent studies (Ahissar et al., 2001; Luo and Poeppel, 2007) show that the oscillatory cortical activity related to speech processing reflects the spectro-temporal organization of speech. The syllabic rate is mirrored in the envelope of the cortical responses recorded from the auditory cortices at least at intelligible speech rate (Abrams et al., 2008; Ahissar et al., 2001). These results suggest that our frequency-tagging analysis may be successful in tracking and discriminating brain responses to different units of the speech structure. We hypothesized that the monotonous presentation of regularly concatenated syllables would give rise to a peak of power at the frequency of the syllable occurrence, as described by Abrams et al. (2008), Ahissar et al. (2001) and Luo and Poeppel (2007). If after some exposure, syllables are bound to adjacent syllables and ultimately grouped in tri-syllabic words, we expected to record a peak of power at the frequency of three syllables. Steady-state power responses at the frequencies of occurrence of single syllables, of bi-syllabic and tri-syllabic words were thus evaluated to tag the perception of monosyllabic, bi-syllabic or tri-syllabic structure in the stream, respectively. To have a behavioral control of the task, participants were asked to write the words they had perceived after 3 and 6 min of exposure of each stream, and in order to compare our results to previous behavioral experiments, they were also asked to classify different types of tri-syllabic items as "words" or not in a test phase following the 9 min of exposure to the word streams. Parts of this work have been published previously in abstract form (Buiatti et al., 2007).

Materials and methods

Participants

Thirteen healthy French monolingual university students (6 males, age range: 18–28 years) were tested, but only 9 subjects had sufficient non artefacted data for the analyses of all four streams. All subjects gave written consent to undertake the experiment and received 20 euros in compensation.

Sound stimuli for the learning phase

Two streams for the experimental conditions (hereafter named Word streams) and two for control conditions (hereafter named Random streams) were generated in a text version as the concatenation of a series of consonant–vowel syllables. Word streams were constructed by the concatenation of nine tri-syllabic items (hereafter named 'words'). 'Words' belonged to three 'families', comprising 3 'words' each. A family was characterized by a fixed first and third syllable (AxC structure), while the second syllable was different for each word (Table 1). For instance, the family /FO_BE/ consisted of the 'words' /FOLOBE/, /FODABE/ and /FONUBE/. 'Words' were pseudo-randomly concatenated in two different streams with the restriction that neither the consecutive repetition of the same 'word' nor of two members of the same family was allowed (e.g. FONUBEPULOKI*MINUDA* PUDAKI*MILODA*FONUBE…). Thus, the transitional probability (TP) between non-adjacent syllables (first and third) was equal to 1 within 'words', and was between 0.33 and 0.5 between 'words'. For adjacent syllables, the TP was between 0.33 and 0.5 both within and between 'words'. Random streams were constructed by a pseudorandom concatenation of 36 different syllables constructed by the combination of the same consonants and same vowels that were used in Word streams. Neither repetition of the same syllable within any three syllables, nor repetition of the same tri-syllabic sequence was allowed. Random streams did not contain any word, part-word or rule-word used in Word streams. The TP between adjacent and non-adjacent syllables in the Random streams was lower than 0.1 without systematic changes at the boundaries of the tri-syllabic items.

Word and Random streams did not contain real French words longer than one syllable. Each one of the streams was generated in two versions, with and without pauses. For streams with pauses, a 25 ms silence was added every three syllables. Several previous studies showed that 25 ms pauses are not explicitly perceived, thus both streams with and without pauses are perceived as continuous speech (Echols et al., 1997; Pena, 2002; Phillips, 1999).

Text versions of the streams were divided in three shorter parts containing the same number of tri-syllables in order to obtain around 3 min of coarticulated speech. Each part was transformed to sound with MBROLA, a speech synthesizer based on the concatenation of natural diphones produced here by a French female speaker. All streams (22050 Hz, mono, 16 bits) were monotonous and nonsense. Duration (116 ms) and pitch (200 Hz) were identical for each phoneme, and the intensity range was similar for each syllable. The total duration was 3.13 min for the streams without pauses and 3.24 min for the streams with pauses.

Sound stimuli for the test phase

Three types of 9 tri-syllabic items were synthesized with MBROLA: 1) 'Words' consisted of the 9 tri-syllabic 'words' presented during the learning stream; 2) 'Part-words' comprised 9 tri-syllabic items constructed by concatenation of two syllables of a 'word' with a syllable from an adjacent 'word' in the stream. Part-words were thus also present in the learning streams but spanned word boundaries. Hence, they had higher adjacent TPs but lower non-adjacent TPs than 'words'; 3) 'Rule-words' comprised 9 tri-syllabic items with an AxC structure. Rule-words were constructed by replacing the second syllable of each 'word' by the first or the third syllable of a 'word' from a different family. Thus TPs between non-adjacent syllables was equal to 1 as for words but TPs between adjacent syllables was equal to zero. Contrary to 'words', rule-words never occurred in the learning stream. Items' duration was 696 ms (116 ms*6 phonemes).

Because EEG signals are sensitive to low-level acoustic properties, we controlled for the phonological properties of the different syllables. In Pena et al. (2002), the first syllable of the 'words' was always a plosive, while the second syllable was a liquid, a fricative or a plosive. Newport and Aslin (2004) criticized this phonetic structure suggesting that word recognition was based on a phonetic grouping rather than on statistical computations between non-adjacent syllables. Because of possible auditory confusion between close phonemes and of the small number of phonemes within each phonetic category, we could not restrict the consonants to one category but we controlled that there was no systematic bias between phonemes at the different syllable positions (see Table 1).

Experimental procedure

All participants were tested in a silent room, wearing earphones. The experiment was presented on a Pentium-based PC using the experimental software EXPE6 (Pallier et al., 1997). Each participant was consecutively exposed to each one of four artificial streams representing the two experimental conditions (Word and Random) presented in two variants (with pauses and without pauses). Each condition used a different speech stream. The order of the conditions and the two possible versions of the stream within each condition were counterbalanced across subjects.

Participants were notified that they would listen to samples of an artificial language containing imaginary 'words' that they must discover. In order to have a behavioral report of the ongoing word-learning process and because it is impossible to ask participants to avoid blinking during 9 min, the learning phase was cut in three blocks of roughly 3 min each. After the first and second block, participants were asked to write down the 'words' they had discovered. After 2 min of silence, the paper with the list of written 'words' was removed from the desk, and the stream was started again. No feedback was given. At the end of the 9 min of exposure to Word streams, a recognition test composed by the presentation of 64 tri-syllabic items was proposed to the participants. They were asked to press a yes/no button, as fast as possible, to indicate whether the item was, or not, a 'word' in the artificial language. Button positions were swapped at the 33rd trial. On the contrary, in Random conditions, after the third block of the learning phase participants performed the same task as after the first and second block because of the huge number of possible tri-syllabic "words" contained in the streams.

EEG recordings

EEG was recorded from 129 electrodes (EGI, USA) referenced to the vertex. Scalp voltages were amplified, digitized at 125 Hz, low-pass filtered at 40 Hz, and stored on the hard disk of a Power-MacIntosh 7100.

Data analysis

Data analyses were performed with custom-made software based on MATLAB (Natick, MA) and on the Fieldtrip toolbox (http://www.ru.nl/fcdonders/fieldtrip). ERP and normalized power topographies were plotted using the EEGLAB toolbox (Delorme and Makeig (2004), http://www.sccn.ucsd.edu/eeglab/).

Frequency tagging analysis

EEG data from each block of 3 min were segmented in epochs of approximately 12 s overlapping for 5/6 of their length. In order to obtain a high frequency resolution with one bin centered on one-, two-, and three-syllable frequency, epoch lengths corresponded to exactly 18 tri-syllabic words (12.528 s for streams without pauses, 12.960 s for streams with pauses), i.e. an integer number of one-syllable, two-syllables or three-syllables items, resulting in frequency bins of ≈0.078 Hz. Epochs containing amplitudes exceeding ±120 μV were rejected. This relatively high artifact rejection threshold eliminates the epochs with the largest artifacts while keeping epochs containing slow eye movements, since the latter have a broad power spectrum and therefore do not affect narrow-band steady-state responses (Srinivasan and Petrovic, 2006). The resulting signals were mathematically referenced to the average of the 129 channels.

For each electrode, the Fourier transform Fm (f) of each epoch was calculated using a fast Fourier transform (FFT) algorithm (MATLAB, Natick, MA). The power spectrum was calculated from these Fourier coefficients as the average over epochs of the single-epoch power spectrum: $Pm(f)=Fm(f)\times Fm^{*}(f)$. Target frequencies were selected as the inverse of the duration of one syllable ($f=4.31$ Hz), two syllables ($f=2.15$ Hz) and three syllables ($f=1.44$ Hz) for the streams without pauses, and adding to such duration one third of a pause ($f=4.17$ Hz), two thirds of a pause ($f=2.08$ Hz) and a pause ($f=1.39$ Hz) for the streams with pauses. Normalized power (NP) at each target frequency was calculated as the ratio between the power spectrum at the target frequency and the average power spectrum at 14 neighboring frequency bins excluding the ones adjacent to the peak frequency bin. NP of the whole 9 min stream was computed by calculating the NP on the power spectrum of each block and then averaging over blocks. In order to control for the effect of the typical 1/f-like low-frequency background EEG spectrum, NP was compared with the normalized power obtained by subtracting from the power spectrum at the peak frequency the power-law fit of the power spectrum in the neighboring bins, but no significant difference was found. The potential confound of harmonics (components oscillating at frequency values multiple of 1-syllable, 2-syllable and 3-syllable frequencies) was controlled by checking subject by subject in all conditions whether 1-syllable peaks appeared in the same channels where 2-syllable or 3-syllable peaks emerged, and this was never the case. Also, no relevant peak emerged at the first harmonic of the 1-syllable and 3-syllable frequency.

ERP analysis

Continuous EEG data from each block of 3 min were band-pass filtered in the frequency band 0.5–8 Hz and segmented in epochs of 1024 ms starting 100 ms before each tri-syllabic word onset (filtered dataset). To identify epochs containing artifacts, the original continuous data were segmented as above with no prior filtering (non-filtered dataset), and epochs containing amplitudes exceeding ± 70 μV and/or visually evident artifacts were marked as bad. In order to avoid artifacts 'leaking' to neighboring epochs after filtering, both epochs marked as bad in the non-filtered dataset and epochs contiguous to such bad epochs were rejected in the filtered dataset. This procedure was possible because of the large number of epochs (250 in each 3 min block). The band-pass filtered EEG signals were then mathematically referenced to the average of the 129 channels. ERPs were computed by averaging the EEG across all epochs and all blocks in each condition, and (as in De Diego Balaguer et al., 2007) subtracting the baseline over the 100 ms preceding tri-syllabic word onset. A control ERP analysis was also performed by using a standard 0.5–30 Hz band-pass filtering, and following the same procedure described above.

Statistical analysis

The significance of the difference between Random and Word conditions was established both for NP and ERP values by means of a nonparametric randomization test (Nichols and Holmes, 2002) called Cluster Randomization Analysis (CRA) (Maris and Oostenveld, 2007) as implemented in the Fieldtrip toolbox (http://www.ru.nl/fcdonders/fieldtrip). This test effectively controls the type I error rate in a situation involving multiple comparisons (129 channels by 129 time points for ERP spatio-temporal arrays and 129 channels by 1 time point for NP spatial arrays) by clustering neighboring (channel, time)-pairs that exhibit the same effect. The first step of CRA is to identify for each time point channels whose *t* statistics exceeds a critical value when comparing two conditions channel by channel ($p<0.05$, two-sided). The goal of this step is to identify channels with effects exceeding a threshold for the subsequent cluster analysis, i.e. it is not required that the power values to be tested are normally distributed. To correct for multiple comparisons, channel-time points that exceed the critical value and neighboring in the channel array (separated by less than 5 cm) and/or in the time array are grouped as a cluster. Each cluster is assigned a cluster-level statistic whose value equals the sum of each channel-time point *t* statistics. Thus, the cluster-level statistics depends on both the extent of the cluster and the magnitude of each channel-time point *t* statistics that belong to this cluster. The Type-I error rate for the complete spatio-temporal set of channels and time points is controlled by evaluating the cluster-level statistics under the randomization null distribution of the maximum cluster-level statistics. By controlling the Type-I error rate for this single test statistic, the multiple comparison problem is solved, simply because there is only a single test statistic, instead of one test statistic for every channel and time point. The randomization null distribution is obtained by randomizing the order of the data of the two conditions within every participant (in our study, 512 permutations = 2^number of subjects). The *p*-value is estimated as the proportion of the randomization null distribution in which the maximum cluster-level test statistics exceeds the observed maximum cluster-level test statistic.

Results

Behavioral results

Written responses during the learning phase

We expected that participants should report more tri-syllabic items if a correct segmentation occurs. Thus we submitted the number of syllables of the written words reported after 3 and 6 min of exposure to a repeated measure ANOVA with Condition (Word and Random), Pause (pause and nopause), Time (3 and 6 min) and Number of Syllables (3 vs others) as within subject factors. There was a significant interaction Number by Condition ($F(1,8)=12.2$, $p=0.008$) and Number by Pause ($F(1,8)=6.24$, $p=0.037$): Participants reported more tri-syllabic words in Word conditions than in Random conditions and in streams with pauses than in streams without pauses. Post-hoc analyses revealed that participants reported a higher number of tri-syllabic words rather than other word lengths only during exposure to Word stream with pauses ($F(1,8)=5.14$, $p=0.05$; from 0 to 8 words, 2.83 in average; Fig. 1).

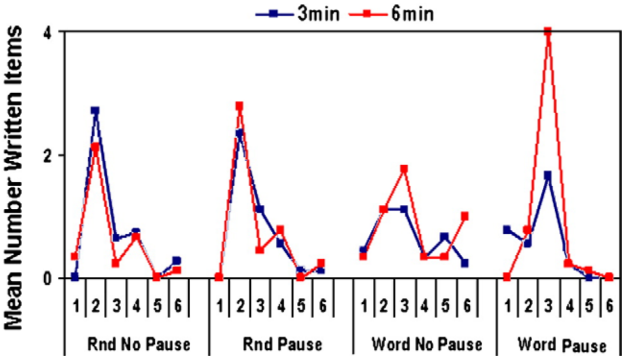

Fig. 1. The mean number of one to six syllables written items reported by free recall after 3 (blue) and 6 (red) min of the learning phase is plotted for the four different conditions: Random without pauses (Rnd No Pause), Random with pauses (Rnd Pause), Word without pauses (Word No Pause) and Word with pauses (Word Pause).

Recognition test

This test was presented only after Word streams. Participants had to classify three types of tri-syllabic items as words or not: 1) Words present in the stream, 2) Rule-words that were structurally similar to words but were not present in the stream, 3) Part-words created by the concatenation of one syllable from a word with two syllables from another word (see Materials and methods for a more detailed description). The percentage of classification of these stimuli as 'word' was submitted to a repeated-measures analysis of variance (ANOVA) with Greenhouse–Geisser correction, with Type of item ('word', 'part-word' and 'rule-word') and Pause (pause and nopause) as within subjects factors. A main effect of Type of item was observed ($F(1.9,13.8)=22.8$; $p<0.00004$) because after both types of stream, 'word' items were significantly more often identified as words than 'rule-words' (Word stream with pauses: 77.8% vs 42.6%, $F(1,7)=13.8$; $p<0.007$; Word stream without pauses: 67.3% vs 43.2%, $F(1,8)=19.8$; $p<0.002$). Words were also significantly more recognized as words than 'part-words' (Word stream with pauses: 77.8% vs 39.4%, $F(1,7)=13.6$; $p<0.008$; Word stream without pauses: 67.3% vs 46.3%, $F(1,8)=9.8$; $p<0.01$). Responses to 'rule-words' and 'part-words' were similar after both types of streams and the interaction Type of item X Pause was not significant.

In summary, subjects were able to discover the artificial words embedded in the structured word streams. Pauses added at word boundaries improve the detection of tri-syllabic items during the learning exposure and improved performance to correctly classified words and non words during the test phase.

Neural correlates of word learning: Frequency-tagging analysis

The power spectrum of the EEG signal was computed for each electrode on narrow-band frequency bins (bin size ≈0.078 Hz) including those corresponding to one syllable ($f\approx4.2$ Hz), two syllables ($f\approx2.1$ Hz) and three syllables ($f\approx1.4$ Hz). As seen in Fig. 2 presenting data from a single subject, peaks are clearly visible at these frequencies. Typically, peaks at a frequency corresponding to a single syllable occur during Random streams, while tri-syllabic peaks mainly emerged in the Word stream with pauses. Each peak was restricted to only one frequency bin, and the predominant peaks were never recorded at non-expected frequency bins. In order to compare the responses of different subjects in different conditions, normalized power (NP) was calculated for each subject and condition at one-syllable, two-syllables and three-syllables tag frequencies (see Materials and methods for details). Typical values of NP associated with the largest peaks are in the (2–4) range.

The topographies of the grand-averaged NP computed over the 9 min of each stream are shown in Fig. 3 (top and middle row) for each frequency of interest. The highest NP values associated to one-syllable frequency (Fig. 3 first and fourth column) occurred in Random streams for both types of streams, with and without pauses. This one-syllable steady-state response clustered around the vertex and above the temporal areas. Intensity and topography of this response was very reproducible across subjects. By contrast, quite surprisingly no response was observed at this frequency in any Word stream, with and without pauses. At the frequency of two-syllables (Fig. 3 second and fifth column), no clear peak of power was visible in any condition; on the contrary, it is worth noting that NP values for the Word stream with pauses are all negative.

Critically, the NP topography at the frequency of three syllables (Fig. 3 third and sixth column) revealed a widespread steady-state response in the Word stream with pauses only, whereas no evident response emerged in the other conditions. Peaks of power in the Word condition with pauses clustered in front of the vertex and at posterior locations. This pattern was more variable across subjects and its peak magnitude somewhat lower than the one-syllable power peaks in Random streams.

In summary, a one-syllable steady-state response was recorded for both Random streams but not for Word streams. A three-syllables steady-state response was present only for the Word stream with pauses whereas no peak at any frequency (not even at one-syllable frequency) was observed for the Word stream without pause.

To compute the statistical significance of the effects described above, we used CRA, a nonparametric randomization test (Maris and Oostenveld, 2007) that effectively controls the type I error rate in a situation involving multiple comparisons (such as 129 channels) by clustering neighboring channel pairs that exhibit the same effect (see Materials and methods). CRA was computed over 9-min NP scalp arrays between Random and Word conditions separately for streams with and without pauses. CRA results show that the aforementioned observations are all statistically significant. During streams without pauses, the one-syllable steady-state response was significantly suppressed around the vertex for Word streams relative to Random streams ($p<0.02$); no significant difference emerged at the two other frequencies of interest (Fig. 3, bottom left row). By contrast, when

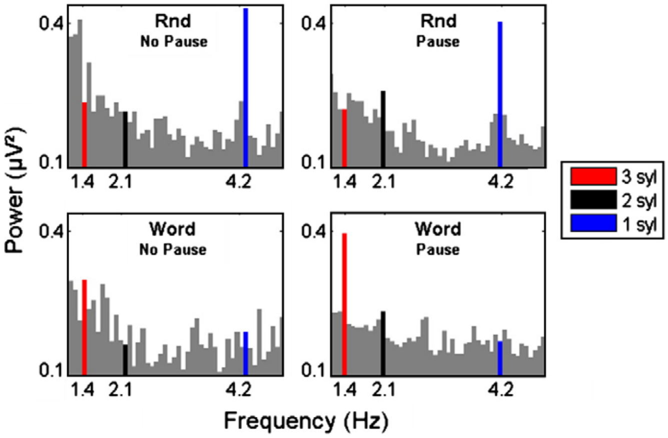

Fig. 2. Power spectrum of the EEG signal in a central midline electrode of one participant (subject 2) calculated from the whole 9 min period of exposure to artificial speech in the four different conditions to illustrate the method and the main results. Power bars at target frequency bins are colored in blue (one-syllable frequency bin ≈4.2 Hz), black (two-syllable frequency bin ≈2.1 Hz) and red (three-syllable frequency bin ≈1.4 Hz). At one-syllable frequency, power peaks clearly emerge in both Random conditions (top row), while they disappear in both Word conditions (bottom row). Conversely, a peak at three-syllable frequency is clearly visible in the Word condition with pauses only (bottom right-hand panel).

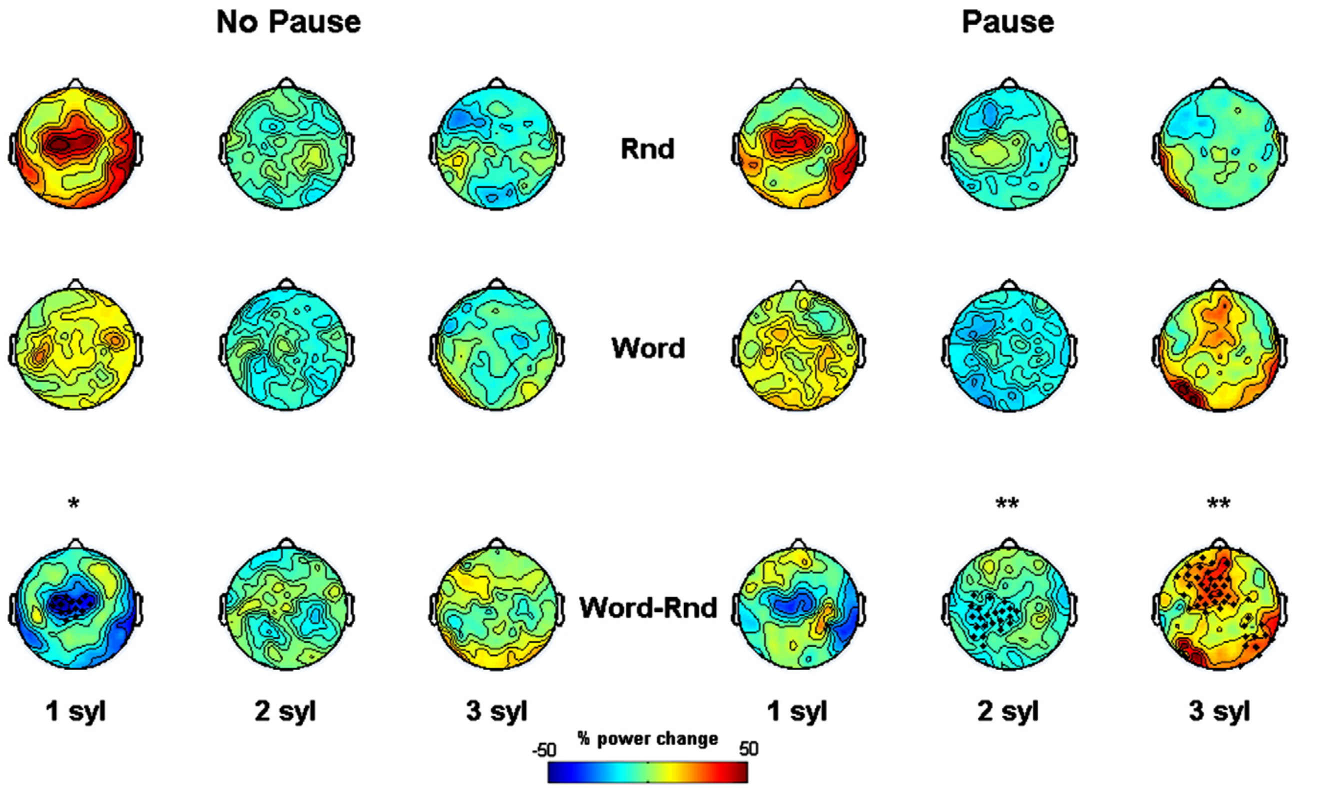

Fig. 3. Topography of grand-averaged NP computed over the whole 9 min period of exposure to artificial speech without pauses (left panel) and with pauses (right panel). Each panel shows the NP topography of the Random (Rnd) stream (first row), of the Word stream (second row) and of the difference (Word–Rnd, third row) for one-syllable ($f\approx4.2$ Hz, first column), two-syllable ($f\approx2.1$ Hz, second column) and three-syllable ($f\approx1.4$ Hz, third column) frequencies. Power peaks at one-syllable frequency are observed in central electrodes in both Random streams, but not in Word streams. Three-syllable power peaks emerge only in the Word stream with pauses. In the third row, channels belonging to significant clusters (CRA, $p<0.05$) are marked with black points. Stars over topographies having significant clusters indicate $p<0.05$ (*) or $p<0.01$ (**).

pauses were present, Word streams elicited a significant suppression of the two-syllables steady-state response in a left parietal cluster ($p<0.01$) and a strong enhancement of the three-syllable steady-state response in a left anterior and a right occipital cluster ($p<0.01$) with respect to Random streams with pauses (Fig. 3, bottom right row). A tendency to suppress the one-syllable steady-state response in the central cluster was also observed ($p<0.07$).

A fast learning

Because a couple of studies (De Diego Balaguer et al., 2007; Pena et al., 2002) suggested that 3 min of exposure were sufficient to discover the word in the stream with pauses, we computed the same analyses as above over the first 3 min of learning phase. Results were very similar to the ones obtained over 9 min of exposure (Fig. 4). A clear three-syllable steady-state response was observed in the Word stream with pauses only. This enhancement emerged in a cluster similar to the one arising after 9 min, and was statistically significant with respect to the Random stream with pauses ($p<0.02$). When pauses were present, suppression of two-syllable steady-state responses was also highly significant (Word stream vs Random stream, $p<0.01$). One-syllable steady-state responses were already visible in both random conditions, and were significantly suppressed in the Word condition without pauses ($p<0.02$), while this suppression was incomplete in the stream with pauses (Word stream vs Random stream, $p>0.14$).

Correlation between EEG power and behavior

In order to directly test the link between tri-syllabic steady-state responses and word learning, correlation between tri-syllabic power peaks and the number of correctly reported words was evaluated. As a global measure of tri-syllabic power peaks, the average of NP corresponding to an increase in power (NP>1) was computed for each subject and for the first (1–3 min) and the second block (4–6 min) of exposure to the Word stream with pauses on the cluster where tri-syllabic NP significantly differed between Word stream and Random stream with pauses (Fig. 3, bottom right-hand topography). Average of NP during both the first (1–3 min) and the second (4–6 min) block significantly correlated ($R=0.56$, $p<0.025$) with the sum of the number of correct words that subjects reported both after the first (1–3 min) and the second (4–6 min) block, respectively (Fig. 5). This correlation tended to be significant when the analysis was restricted to the first 3 min ($R=0.68$, $p<0.095$) and became significant in the second block (filled diamonds in Fig. 5; $R=0.77$, $p<0.016$).

By contrast, the EEG power did not correlate with the performances during the test phase (correlation between tri-syllabic SSR in the relevant cluster and percentage of correct responses to Words in the test phase: $R=0.03$, $p>0.9$).

ERP analysis

Next, we investigated whether the tri-syllabic oscillations arising in Word streams with pauses are characterized by a distinct temporal profile time-locked to tri-syllabic words. EEG data were band-pass filtered in the range 0.5–8 Hz and averaged in synchrony with tri-syllabic word onsets to obtain word-locked ERPs (see Materials and methods). Such ERPs provide a temporal signature of the tri-syllabic oscillations phase-locked to tri-syllabic words, and enable a comparison with previous ERP studies. The low-pass filter value (8 Hz) was chosen because we were mostly interested in the low frequency

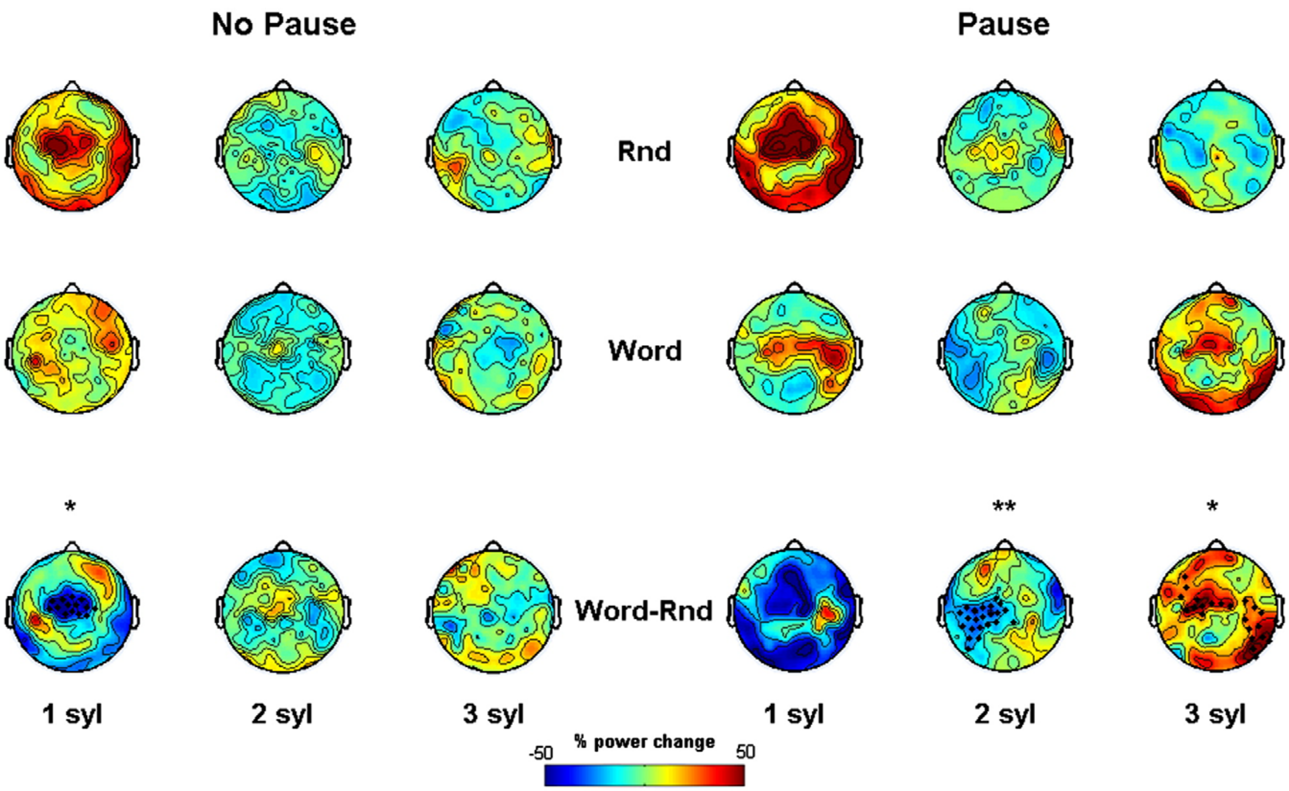

Fig. 4. Topography of grand-averaged NP computed over the first 3 min of exposure to artificial speech without pauses (left panel) and with pauses (right panel). Each panel shows the NP topography of the Random (Rnd) stream (first row), of the Word stream (second row) and of the difference (Word–Rnd, third row) for one-syllable ($f \approx 4.2$ Hz, first column), two-syllable ($f \approx 2.1$ Hz, second column) and three-syllable ($f \approx 1.4$ Hz, third column) frequencies. In the third row, channels belonging to significant clusters (CRA, $p<0.05$) are marked with black points. Stars over topographies having significant clusters indicate $p<0.05$ (*) or $p<0.01$ (**). Results are very similar to those relative to the whole 9 min stream (cf. Fig. 3).

activity that might sustain the three-syllables SSR, and also to avoid interference from the alpha (≈10 Hz) activity, which was relatively high during such a long and repetitive stimulation.

Comparison between Word and Random conditions from streams with pauses revealed a large difference centered at 400 ms after word onset (Fig. 6, right-hand column). Since we had no prior expectation on the time interval of a potential effect, CRA was computed in the whole time window 0–696 ms. CRA revealed that such difference was statistically significant and distributed on an anterior (positive difference, $p<0.03$) and posterior (negative difference, $p<0.05$) cluster, both extending in the range 256–464 ms (Fig. 6, right-hand topography). Importantly, the sum of the t statistics over such clusters reaches its maximum at 400 ms, suggesting a response similar to a N400. No significant difference arose in any other time interval. In particular, there was no significant difference at the N1 or P2 latencies, nor in the whole 0–696 ms time window when comparing Word and Random conditions in streams without pauses (Fig. 4, left-hand column). All these results were confirmed in a control ERP analysis using a standard 0.5–30 Hz filtering (Fig. S1).

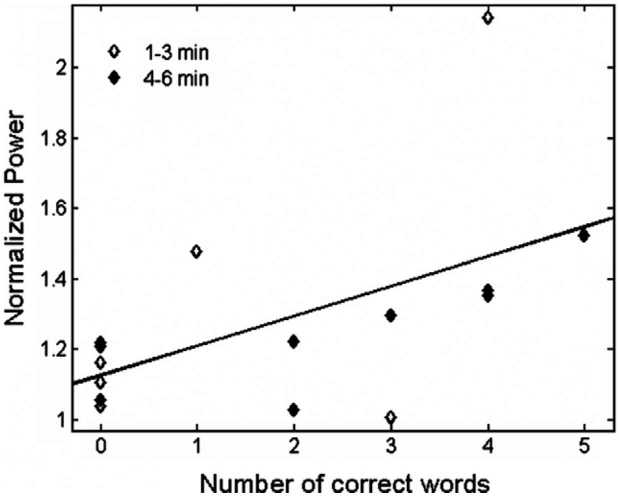

Fig. 5. Subject-by-subject relationship between the average tri-syllabic SSR in the cluster of interest (see Results) during the first (open diamonds) and the second (filled diamonds) block of 3 min of exposure to the Word stream with pauses and the corresponding number of correct written words at the end of each block.

To prove that the ERP peak at 400 ms really identifies the temporal profile of the tri-syllabic oscillations time-locked to word onsets, we calculated the correlation coefficient for the Word stream with pauses between the tri-syllabic NP values averaged over the CRA cluster indicated in Fig. 3 (bottom right-hand topography) and the sum of the absolute values of the ERPs averaged over the two CRA clusters indicated in Fig. 6 (right-hand topography) at 400 ms. Such correlation was indeed significant: $R=0.85$, $p<0.015$.

Discussion

Frequency tagging of pertinent speech units

In this study, SSRs labeled by frequency tags corresponding to one-, two- and three-syllable words were used to investigate on-line the neural mechanisms underlying word learning from continuous speech while manipulating their statistical regularities and acoustic cues. First, we observed that a one-syllable SSR was steadily recorded in Random streams emerging in most subjects on electrodes around

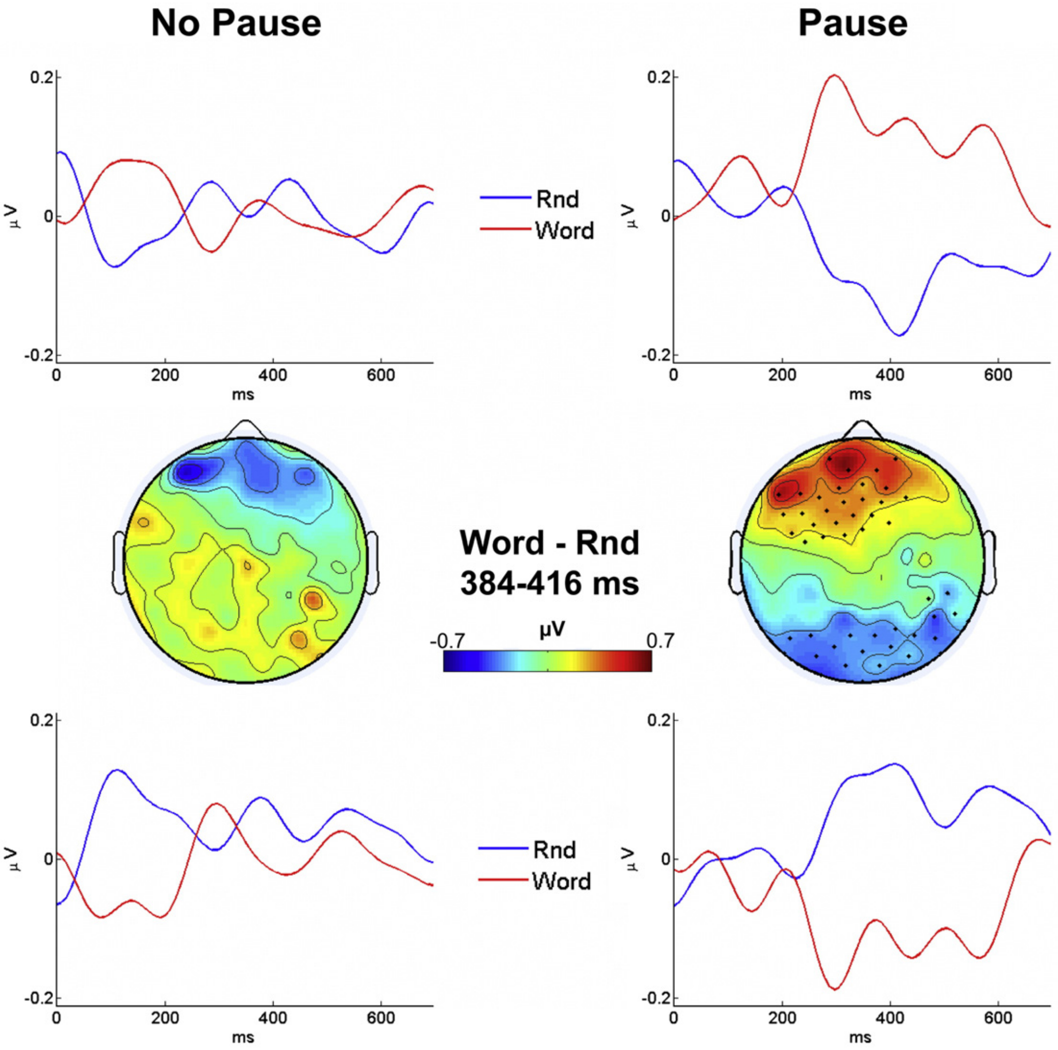

Fig. 6. ERPs of band-pass filtered (0.5–8 Hz) EEG data for Random (blue lines) and Word (red lines) streams without pauses (left column) and with pauses (right column), averaged over the anterior (top row) and posterior (bottom row) cluster emerging from CRA (see Results). The middle row shows the topography of the difference between Word and Random ERPs in streams without pauses (left column) and with pauses (right column) averaged over the time interval 384–416 ms, corresponding to the maximum of the CRA difference statistics for the stream with pauses. CRA anterior and posterior clusters are plotted as black dots in the right-hand topography. The difference between Word and Random streams with pauses is significant in the time interval 256–464 ms in both clusters, while no significant difference arises in streams without pauses.

the vertex and temporal areas (Fig. 3). This observation is in agreement with the results of Ahissar et al. 2001 which showed using MEG that the cortical auditory response oscillated with the speech rate. It is also consistent with Luo and Poeppel (2007), who hypothesized that cortical speech analysis is based on the modulation of inherent cortical rhythms in the theta range (4–8 Hz) (the syllable rate is ≈4.2 Hz in our experiment). Using simultaneous EEG-fMRI recordings, spontaneous oscillations in ranges that match the rhythmic properties of the speech signal have been recorded within the auditory cortices (Giraud et al., 2007). These authors also report spontaneous oscillations in the gamma range in the left auditory cortex and they hypothesize a relation between this frequency and phonemic processing. Here, although consonants and vowels were regularly presented, having the same duration, there was no increase of power at the phoneme frequency (8.6 Hz). As stimulation with clicks at 40 Hz and even 100 Hz easily induce strong SSRs, this absence of power at the phoneme frequency cannot be related to a physiological limit in evoked cortical oscillations but might rather be due to speech perception mechanisms, as syllable, and not phoneme, is the basic perceptive unit for speech perception at least in French native speakers (Bertoncini and Mehler, 1981; Mehler et al., 1981).

By contrast, the presence of tri-syllabic words (i.e. AxC) in the stream induced the suppression of one-syllable steady state response, both when pauses were present and absent. This suppression suggests that frequency tagging is not dominated by low-level processing but can be modulated by higher levels of stimulus integration. This is congruent with the results reported in binocular rivalry stimulation (Tononi et al., 1998) in which the amount of power at the flickering frequency of two different monocular stimuli increased with the conscious perception of one or the other stimulus although the physical properties of the binocular stimulation remained unchanged.

Finally, a tri-syllabic SSR was recorded when subliminal pauses were added in the Word streams. This effect cannot be due to the regular addition of subliminal pauses every three syllables because no tri-syllabic SSR was recorded during Random streams with pauses. On the contrary, we hypothesize that this specific SSR is induced by the on-line segmentation of the embedded words as more tri-syllabic words were reported in that condition, and critically, the amplitude of the tri-syllabic power response was significantly correlated with the number of correctly reported words (Fig. 5). In addition to the tri-syllabic SSR, a suppression of bi-syllabic SSR was observed for that stream, suggesting that word learning induces both an enhancement of power at the frequency of the discovered word, and an inhibition of power at frequencies associated to alternative words with different lengths. The suppression of the one-syllable SSR in word streams as the robust tri-syllabic SSR in AxC with pauses suggests that oscillatory responses are not limited to the most basic unit (syllable) as already reported (Abrams et al., 2008; Ahissar et al., 2001; Luo and Poeppel, 2007), but are sensitive to the different units of the speech structure and are amplified by directed attention to a specific level involving multi-syllabic units.

The topographies of the three-syllable and the one-syllable SSR were different (the former present over more anterior frontal and more posterior occipital electrodes than the latter, see the significant differences between random and word streams in Fig. 4). Although it is uncertain to precisely locate the brain regions involved, this difference suggests at least that different populations of neurons fire in synchrony with the detection of either syllables or words, and that we are not recording with that method a generic response (i.e. conscious detection of segments) but a specific image of the network involved in the syllable or word learning computation. MEG studies (Abrams et al., 2008; Ahissar et al., 2001; Luo and Poeppel, 2007), as the study of Giraud et al. (2007) combining EEG and fMRI, have suggested that the cortical responses modulated by syllabic rate originate from Heschl gyri. McNealy et al. (2006) using fMRI observed that activity increases in the left superior temporal region while adults were listening to artificial language streams relative to random streams. Among the regions sensitive to the structure of the stream, a cluster in the left STG displayed activity correlated with behavioral word recognition suggesting that this region might be involved in word segmentation. Dehaene-Lambertz et al. (2006) observed a temporal gradient in BOLD responses along the superior temporal regions when adults listened to sentences. The authors proposed that this temporal gradient of activation might reflect a nested organization of processing speech units with progressively longer time-window of integration. The different topographies observed here for the syllable and the word SSR might be related to a shifting of activity from Heschl gyrus toward more ventral regions of the superior temporal lobe, reflecting the first stages of this nested structure.

Word segmentation

Our second important result is that brain responses crucially differed when subliminal acoustic pauses were added to the word stream: In that case only, a tri-syllabic SSR was recorded. It was also only in that stream that a N400-like potential was recorded, that was a main contributor to the tri-syllabic SSR. N400-like potentials seem to be triggered by stimuli whose physical forms can be used to access to semantic knowledge about the referent of the stimulus (Halgren, 1990). A N400 increase appears to be a robust index of word access being reported now in numerous experiments having studied words learning in artificial language streams (Cunillera et al., 2006, 2008; De Diego Balaguer et al., 2007; Sanders et al., 2002). As in the experiment of Pena et al. (2002) and De Diego Balaguer et al. (2007), word segmentation appears very early during exposure to artificial streams. EEG and performances were significantly modified relative to other streams since the first block of 3 min of exposure. However, contrary to the report of De Diego Balaguer et al. (2007), no P2 increase was observed in our study, nor N1 increase as reported in Sanders et al. (2002). N1/P2 amplitudes are usually weak in continuous speech, making it difficult to observe an increase especially if the syllables used as word-onset are not plosives whose burst of energy can more easily create a sharp perceptive onset. It is thus possible that the N1/P2 effect might depend on the acoustic characteristics of the chosen syllables in different experiments, either because some consonant categories are perceptively more salient in a continuous stream, or because a choice spanning several phonetic categories with very different acoustic characteristics may spread the response to word onset over a larger interval blurring the sharp N1/P2. Moreover, an increase in N1/P2 amplitude might be significantly amplified if participants are expecting a precise syllable onset. It is the case when words have been previously learned (Sanders et al., 2002) or when the first syllable is acoustically salient facilitating its encoding (Cunillera et al., 2006). In the study of De Diego Balaguer et al. (2007) in which the artificial language corresponds to our Word stream with pauses, the P2 increase appeared after the N400 increase and was correlated to the correct identification of rule-words as possible words in the test phase, suggesting that participants had understood the rule and thus isolated the first and last syllable of the words. The absence in our study of a P2 effect during familiarization, together with the missing discrimination between rule-words and part-words during the test phase suggests that a reliable P2 increase might be present only when subjects become able to analyze the structure of the words and thus may expect precise onsets. The differences between the two studies could depend on differences in the experimental paradigms, such as the number of streams sharing a similar structure to which subjects were exposed to (four in the study of Diego de Balaguer et al., one here) that may favor learning across streams.

By contrast, when pauses were absent, there was neither any peak at any tagged frequency for the Word stream, nor any significant difference in the ERPs between Word and Random streams. Yet, during the test phase following 9 min of exposure, participants significantly distinguished words from part-words in both conditions with and without pauses demonstrating that, like in Pena et al. (2002), they have been able to learn the words exploiting their non-adjacent transition probabilities. Crucially, the significant suppression of the peak at the one-syllable tagged frequency demonstrates that syllables were no more isolated but linked to other syllables contrary to random streams. However, the product of this statistical computation was neither translated into a stable steady-state brain response nor into an explicit word report, as the behavioral measures during exposure to that stream were not different from those in random streams. Transitional probabilities were thus computed but were not sufficient to end up with a reproducible tri-syllabic structure during the familiarization part. The good performance in the test may suggest that segmentation in tri-syllabic items was only explicitly realized during the test when tri-syllabic items were presented in isolation using the transitional probabilities between syllables implicitly learned during exposure.

Acknowledging the radical difference in word recognition performances when a 25 ms gap was introduced at the end of the words, Pena et al. (2002) postulated that different brain mechanisms were involved in both cases. In the light of our results, we can tentatively explain the respective role of statistical and acoustic cues in word learning: When no acoustic cues are given, transitional probabilities between adjacent and non-adjacent syllables are implicitly computed only limited by memory load. The absence of tri-syllabic power peaks in Word streams without pauses and the fact that subjects are unable to report correct tri-syllabic words in this condition, even after 6 min of exposure, suggest that the size of the window on which statistical distribution is computed may vary from time to time, being reset by the subject's explicit search strategy or by superficial contingencies. Although listeners are not able to discriminate streams with pauses from streams without pauses (Pena, 2002), pauses might alter perception in several ways, by breaking coarticulation between syllables involving continuous phonemes, by decreasing the masking effect of the following syllable and by resetting the response to the next syllable. Such cues might act as end- marks that limit the domain of computations to smaller segments, like it is observed in natural speech (Christophe et al., 2004), whose local properties would be easiest to analyze and memorize. They can also add perceptual emphasis around word boundaries facilitating the memorization of the syllables at the words' edges (Endress and Bonatti, 2007). However, pauses by themselves are not sufficient to induce a tri-syllabic segmentation as no tri-syllabic SSR was observed in Random stream with pauses, suggesting that acoustic cues (as 25 ms gaps) are not processed as a first segmentation step before statistical computations are realized, at least in the present case of short pauses.

Pena et al. (2002) and De Diego Balaguer et al. (2007) suggested that pauses not only favor word-learning but also induced rule learning. This interpretation is based on the performances during the test phase when participants have to choose between a rule-word and a part-word as possible words in the artificial language. In our experiment, subjects had to decide for each one of the three possible types of items (words, rule-words and part-words) whether they accepted it as a word or not. In a forced-choice paradigm, it is not possible to distinguish whether the word is recognized as a word or whether it is the part-word which is rejected. Furthermore, when participant had to choose between rule-words and part-words as possible words, the implicit instruction is to choose the closest item to the learned word structure, whereas in our experiment the decision to accept a rule-word as a word depends on whether participants interpret the instructions as a recognition task of heard words or as a classification task of the possible words in this artificial language. Because participants were asked to explicitly learn and write out the words during the learning phase rather than to figure out the structural properties of the stream, they may have been biased toward a recognition strategy. The fact that participants were at chance for rule-words confirms that the instructions were ambiguous and that their performances for rule-words are not conclusive to determine whether they were or not aware of the rule.

Concluding remarks

To conclude, our results underscore the role of prosodic cues (as short pauses) in natural speech to discover words. Infant abilities to perform statistical learning are highly supported (and not limited to language). However, the rich and complex information contained in speech conveys a cognitive resource problem associated to the exponential growth of the number of computations. Limitations on memory or other cognitive mechanisms could contribute to restrict cognitive computations. Sensitivity to prosodic cues could be an essential tool to limit continuous speech processing (Echols and Newport, 1992; Gleitman and Wanner, 1982). Indeed, infants are sensitive to different prosodic units since the first days of life on (Hirsh-Pasek et al., 1987; Jusczyk et al., 1992) and subtle acoustic cues have been shown to be exploitable even by young infants (Christophe et al., 1994). Statistical computations applied within prosodic units (as short speech chunks) can thus be a powerful tool to discover the embedded units of speech. Here we demonstrate that, at least in adults, subliminal prosodic-like cues can rapidly help to discover a set of words by increasing efficiency of segmentation of artificial, monotonous and non-sense continuous speech stream. Moreover, we show that this learning process elicits a SSR at the frequency of word occurrence that is significantly correlated with the successful detection of words in the artificial speech. We therefore propose the frequency-tagging approach as a powerful tool to track the subjective perception of continuous speech and linguistic learning by the underlying neural activity.

Acknowledgments

We thank S. Dehaene, J. Mehler and M. Piazza for the fruitful discussions. This study was supported by the McDonnell foundation, ECOS C04B02, ANR NIBB, and a VolksWagen grant to MB.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2008.09.015.

References

Abrams, D.A., Nicol, T., Zecker, S., Kraus, N., 2008. Right-hemisphere auditory cortex is dominant for coding syllable patterns in speech. J. Neurosci. 28 (15), 3958–3965.

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

Bertoncini, J., Mehler, J., 1981. Syllables as units in infant speech perception. Infant Behav. Dev. 4, 247–260.

Bonatti, L.L., Pena, M., Nespor, M., Mehler, J., 2005. Linguistic constraints on statistical computations: the role of consonants and vowels in continuous speech processing. Psychol. Sci. 16 (6), 451–459.

Buiatti M., Pena M., Mehler J., Dehaene-Lambertz G., 2007. Neural correlates of continuous speech computation investigated by means of frequency-tagged neuroelectric responses. In: Annual Meeting of the Cognitive Neuroscience Society, Book of Abstracts, New York, U. S. A., p. 253.

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.

Christophe, A., Peperkamp, S., Pallier, C., Block, E., Mehler, J., 2004. Phonological phrase boundaries constrain lexical access: I. Adult data. J. Mem. Lang. 51, 523–547.

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 (1), 168–178.

Cunillera, T., Gomila, A., Rodríguez-Fornells, A., 2008. Beneficial effects of word final stress in segmenting a new language: evidence from ERPs. BMC Neurosci. 18, 9–23.

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 (11), e1175.

Dehaene-Lambertz, G., Dehaene, S., Anton, J.L., Campagne, A., Ciuciu, P., Dehaene, G.P., Denghien, I., Jobert, A., Lebihan, D., Sigman, M., Pallier, C., Poline, J.B., 2006. Functional segregation of cortical language areas by sentence repetition. Hum. Brain Mapp. 27 (5), 360–371.

Delorme, A., Makeig, S., 2004. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J. Neurosci. Methods 134, 9–21.

Echols, C.H., 1993. A perceptually-based model of children's earliest productions. Cognition 46, 245–296.

Echols, C.H., Newport, E.L., 1992. The role of stress and position in determining first words. Lang. Acquis. 2, 189–220.

Echols,, C.H., Crowhurst,, M.J., Childers,, J.B., 1997. The perception of rhythmic units in speech by infants and adults. J. Mem. Lang. 36, 202–225.

Endress, A.D., Bonatti, L.L., 2007. Rapid learning of syllable classes from a perceptually continuous speech stream. Cognition 105 (2), 247–299.

Giraud, A.L., Kleinschmidt, A., Poeppel, D., Lund, T.E., Frackowiak, R.S., Laufs, H., 2007. Endogenous cortical rhythms determine cerebral specialization for speech perception and production. Neuron 56 (6), 1127–1134.

Gleitman, L., Wanner, E., 1982. The state of the state of the art. In: Wanner, E., Gleitman, L. (Eds.), Language acquisition: The state of the art. Cambridge University Press, Cambridge, U. K., pp. 3–48.

Halgren, E., 1990. Insights from evoked potentials into the neuropsychological mechanisms of reading. In: Scheibel, A., Weschsler, A. (Eds.), Neurobiology of Cognition, Guilford, New York, pp. 103–150.

Hayes, J.R., Clark, H.H., 1970. Experiments on the segmentation of an artificial speech analogue. In: Hayes, J.R. (Ed.), Cognition and the development of language. Wiley, New York, pp. 221–234.

Hirsh-Pasek, K., Kemler Nelson, D.G., Jusczyk, P.W., Cassidy, K.W., Druss, B., Kennedy, L., 1987. Clauses are perceptual units for young infants. Cognition 26, 269–286.

Jusczyk, P.W., Hirsh-Pasek, K., Nelson, D.G., Kennedy, L.J., Woodward, A., Piwoz, J., 1992. Perception of acoustic correlates of major phrasal units by young infants. Cogn. Psychol. 24 (2), 252–293.

Luo, H., Poeppel, D., 2007. Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. Neuron 54 (6), 1001–1010.

Maris, E., Oostenveld, R., 2007. Nonparametric statistical testing of EEG- and MEG-data. J. Neurosci. Methods 164 (1), 177–190.

McNealy, K., Mazziotta, J.C., Dapretto, M., 2006. Cracking the language code: neural mechanisms underlying speech parsing. J. Neurosci. 26 (29), 7629–7639.

Mehler, J., Dommergues, J.Y., Frauenfelder, U., Segui, J., 1981. The syllable's role in speech segmentation. J. Verbal Learn. Verbal Behav. 20, 298–305.

Newport, E.L., Aslin, R.N., 2004. Learning at a distance I. Statistical learning of nonadjacent dependencies. Cogn. Psychol. 48 (2), 127–162.

Nichols, T., Holmes, A., 2002. Nonparametric permutation tests for functional neuroimaging: A primer with examples. Hum. Brain Mapp. 15, 1–25.

Pallier, C., Dupoux, E., Jeannin, X., 1997. EXPE: an expandable programming language for on-line psychological experiments. Behav. Res. Meth. Ins. C. 29, 322–327.

Pena M., 2002. Rôle du calcul statistique dans l'acquisition du langage. PhD thesis, EHESS, France.

Pena, M., Bonatti, L.L., Nespor, M., Mehler, J., 2002. Signal-driven computations in speech processing. Science 298 (5593), 604–607.

Phillips, D.P., 1999. Auditory gap detection, perceptual channels, and temporal resolution in speech perception. J. Am. Acad. Audiol. 10 (6), 343–354.

Picton, T.W., Sasha, J.M., Dimitrijevic, A., Purcell, D., 2003. Human auditory SSRs. Int. J. Audiol. 42, 177–219.

Pilon,, R., 1981. Segmentation of speech in a foreign language. J. Psycholinguist. Res. 10, 113–122.

Pisoni, D.B., Luce, P.A., 1986. Speech perception: research, theory and the principal issues. In: Schwab, E.C., Nusbaum, H.C. (Eds.), Pattern Recognition by Humans and Machines: Speech Perception, Volume 1. Academic Press, New York, pp. 1–50.

Sanders, L.D., Neville, H.J., 2003a. An ERP study of continuous speech processing. I. Segmentation, semantics, and syntax in native speakers. Brain Res. Cogn. Brain Res. 15 (3), 228–240.

Sanders, L.D., Neville, H.J., 2003b. An ERP study of continuous speech processing. II. Segmentation, semantics, and syntax in non-native speakers. Brain Res. Cogn. Brain Res. 15 (3), 214–227.

Sanders, L.D., Newport, E.L., Neville, H.J., 2002. Segmenting nonsense: an event-related potential index of perceived onsets in continuous speech. Nat. Neurosci. 5 (7), 700–703.

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.

Shukla, M., Nespor, M., Mehler, J., 2007. An interaction between prosody and statistics in the segmentation of fluent speech. Cogn. Psychol. 54 (1), 1–32.

Srinivasan, R., Petrovic, S., 2006. MEG phase follows conscious perception during binocular rivalry induced by visual stream segregation. Cereb. Cortex 16, 597–608.

Tononi, G., Srinivasan, R., Russell, D.P., Edelman, G.M., 1998. Investigating neural correlates of conscious perception by frequency-tagged neuromagnetic responses. Proc. Natl. Acad. Sci. U. S. A. 95 (6), 3198–3203.