Effects of acoustic periodicity, intelligibility, and pre-stimulus alpha power on the event-related potentials in response to speech

Kurt Steinmetzger [a) *] **& Stuart Rosen** [a)]

[a)] Speech, Hearing and Phonetic Sciences, University College London, Chandler House, 2 Wakefield Street, London WC1N 1PF, United Kingdom

[*]Author to whom correspondence should be addressed. Electronic mail: kurt.steinmetzger.12@ucl.ac.uk

Abstract

Magneto- and electroencephalographic (M/EEG) signals in response to acoustically degraded speech have been examined by several recent studies. Unambiguously interpreting the results is complicated by the fact that speech signal manipulations affect acoustics and intelligibility alike. In the current EEG study, the acoustic properties of the stimuli were altered and the trials were sorted according to the correctness of the listeners' spoken responses to separate out these two factors. Firstly, more periodicity (i.e. voicing) rendered the event-related potentials (ERPs) more negative during the first second after sentence onset, indicating a greater cortical sensitivity to auditory input with a pitch. Secondly, we observed a larger contingent negative variation (CNV) during sentence presentation when the subjects could subsequently repeat more words correctly. Additionally, slow alpha power (7–10 Hz) before sentences with the least correctly repeated words was increased, which may indicate that subjects have not been focussed on the upcoming task.

Keywords: EEG; speech; intelligibility; periodicity; CNV; pre-stimulus alpha power

1. Introduction

Acoustically degraded noise-vocoded speech has been used extensively to investigate the neural correlates of speech intelligibility in both magneto- and electroencephalographic (M/EEG) studies (e.g. Becker, Pefkou, Michel, & Hervais-Adelman, 2013; Ding, Chatterjee, & Simon, 2014; Obleser & Weisz, 2012; Peelle, Gross, & Davis, 2013) and imaging work (e.g. Davis & Johnsrude, 2003; Evans et al., 2014; Scott, Blank, Rosen, & Wise, 2000). Noise-vocoding has proven a very useful tool because it allows the parametric reduction of the intelligibility of speech signals by reducing the number of channels in the analysis/synthesis process. However, this signal manipulation alters the acoustic properties of the stimuli as well as their intelligibility, and these two factors have so far not been considered independently.

Furthermore, while the reduction in intelligibility can mainly be attributed to the lowered spectral resolution of the vocoded speech signals, other acoustic properties are affected by the signal processing as well. Most notably, due to the use of a broadband noise as sound source, noise-vocoded speech is completely aperiodic (i.e. unvoiced), making it sound like an intense version of a whisper. In natural speech, on the other hand, voiced and unvoiced segments alternate. Importantly, only voiced speech possesses a pitch. Previous studies that have investigated pitch perception reliably found increased neural responses for stimuli that possess a pitch, when compared to a spectrally matched control condition (e.g. Griffiths et al., 2010; Norman-Haignere, Kanwisher, & McDermott, 2013) or a broadband noise (Chait, Poeppel, & Simon, 2006). In particular, studies analysing MEG signals in the time domain (Chait et al., 2006; Gutschalk, Patterson, Scherg, Uppenkamp, & Rupp, 2004) have shown that following a transient pitch onset response peaking after around 150 ms, a sustained neural response can be observed for several hundred milliseconds. Thus, it appears

likely that the neural response elicited by noise-vocoded speech is *per se* attenuated due to the absence of voicing.

In order to address these issues, we have used a vocoding technique that allows the choice between a periodic (voiced) or an aperiodic (unvoiced) source excitation. This technique was used to synthesise speech that is either completely unvoiced (i.e. noise-vocoded, henceforth referred to as the *aperiodic* condition), preserves the natural mix of voiced and voicelessness (henceforth the *mixed* condition; Dudley, 1939), or is completely voiced (henceforth the *periodic* condition). Previous behavioural work (Steinmetzger & Rosen, 2015) has shown that the intelligibility of the aperiodic and mixed conditions is very similar, while the unnatural-sounding fully periodic condition was found to be considerably less intelligible. In order to analyse effects of acoustic periodicity while controlling for differences in intelligibility, the individual trials in the current study were sorted according to the listeners' spoken responses (i.e. the number of correctly repeated key words) obtained after every sentence, and only fully intelligible trials were considered. In summary, the first hypothesis was that speech with more periodicity would lead to more negative event-related potentials (ERPs), reflecting the increased neural sensitivity to auditory input that possess a pitch. This effect was expected to be observed during an early time window following sentence onset, including the auditory evoked potentials (AEPs) and the acoustic change complex (ACC; Pratt, 2011).

Sorting the individual trials according to the behavioural responses was also intended to enable the separate analysis of more or less intelligible trials in the periodic condition. This second analysis additionally included spectrally rotated speech, a completely unintelligible non-speech analogue that has been used in a number of the previously mentioned studies (Becker et al., 2013; Peelle et al., 2013; Scott et al., 2000), as a baseline condition (henceforth the *rotated* condition). In contrast to several recent M/EEG studies that have investigated the

perception of noise-vocoded (Becker et al., 2013; Obleser & Weisz, 2012; Obleser, Wöstmann, Hellbernd, Wilsch, & Maess, 2012) and unprocessed speech (e.g. Kerlin, Shahin, & Miller, 2010; Müller & Weisz, 2012; Wilsch, Henry, Herrmann, Maess, & Obleser, 2015) by analysing neural activity in the frequency domain, the current study focusses on time-domain responses. Few studies to date have investigated ERPs in response to degraded speech (for exceptions see Becker et al., 2013; Obleser & Kotz, 2011; Wöstmann, Schröger, & Obleser, 2015) and it is hence not well understood how they are affected by both the acoustic characteristics and the intelligibility of the speech signals, particularly over the course of whole sentences.

Based on the notion that slow cortical potentials reflect the degree of cortical excitability (Birbaumer, Elbert, Canavan, & Rockstroh, 1990; He & Raichle, 2009), it was hypothesised that ERP amplitudes over the course of the sentences would be larger in response to more intelligible speech. More specifically, slow negative potentials with an anterior scalp distribution have been associated with both working memory load (e.g. Guimond et al., 2011; Lefebvre et al., 2013) and increased attention (e.g. Teder-Sälejärvi, Münte, Sperlich, & Hillyard, 1999; Woods, Alho, & Algazi, 1994) in auditory tasks. A typical slow negative potential is the contingent negative variation (CNV), which emerges in between a warning stimulus and a task-relevant second stimulus, and is larger when subjects expect and prepare to respond to the latter stimulus (McCallum & Walter, 1968; Tecce & Scheff, 1969). Importantly, the second stimulus may also be a response to the first stimulus (Birbaumer et al., 1990; Kononowicz & Penney, 2016), and hence the design of the current experiment, in which subjects are supposed to verbally repeat the stimulus sentence, fits into the CNV framework too.

In order to further investigate differences between intelligible and unintelligible trials, we additionally analysed the amount of alpha power in the silent baseline interval preceding

the stimulus sentences. Decreased alpha power in the pre-stimulus window has been shown to be a predictor of successful target identification in studies using low-level visual and somatosensory stimuli (e.g. Hanslmayr, et al., 2007; Romei, Gross, & Thut, 2010; Schubert, Haufe, Blankenburg, Villringer, & Curio, 2009; Van Dijk, Schoffelen, Oostenveld, & Jensen, 2008). Strauß, Henry, Scharinger, and Obleser (2015) have recently also reported alpha phase differences before correctly and incorrectly perceived words in a lexical decision task, but no study to date has reported alpha power differences in the baseline window using speech materials presented auditorily. As reviewed by Klimesch (1999, see also Klimesch, Doppelmayr, Russegger, Pachinger, & Schwaiger, 1998), slower alpha frequencies (~7–10 Hz) in particular have been associated with alertness and expectancy, and may thus serve as a measure of the attentional state in the period before sentence onset. We thus additionally hypothesised to observe enhanced slow alpha power, indicating that subjects have not been fully focussed on the upcoming task, before sentences that would turn out to be unintelligible to them.

2. Methods

2.1. Participants

Eighteen normal-hearing right-handed subjects (8 females, mean age = 21.6 years, SD = 2.3 years) took part in the study. All participants were native speakers of British English and had audiometric thresholds of less than 20 dB Hearing Level at octave frequencies from 125–8000 Hz. All subjects gave written consent and the study was approved by the UCL ethics committee.

2.2. Stimuli

The stimulus materials used in this experiment were recordings of the IEEE sentences (Rothauser et al., 1969) spoken by an adult male Southern British English talker with a mean

F0 of 121.5 Hz that were cut at zero-crossings right before sentence onset and normalised to a common root-mean-square (RMS) level. The IEEE sentence corpus consists of 72 lists with 10 sentences each and is characterized by similar phonetic content and difficulty across lists, as well as an overall low semantic predictability (e.g. "*The birch canoe slid on the smooth planks.*"). The individual lists are thus supposed to be equally intelligible. Every sentence contains five key words.

All stimulus materials were processed prior to the experiment using a channel vocoder implemented in MATLAB (Mathworks, Natick, MA). For all three vocoding conditions (aperiodic, mixed, and periodic) the original recordings of the IEEE sentences were first band-pass filtered into eight bands using zero phase-shift sixth-order Butterworth filters. The filter spacing was based on equal basilar membrane distance (Greenwood, 1990) across a frequency range of .1–8 kHz (upper filter cut-offs in Hz: 242, 460, 794, 1307, 2094, 3302, 5155, 8000; filter centre frequencies in Hz: 163, 339, 609, 1023, 1658, 2633, 4130, 6426). The output of each filter was full-wave rectified and low-pass filtered at 30 Hz (zero phase-shift fourth-order Butterworth) to extract the amplitude envelope. The low cut-off value was chosen in order to ensure that no temporal periodicity cues were present in the aperiodic condition.

In order to synthesise aperiodic speech, the envelope of each individual band was multiplied with a broadband noise carrier. In the mixed condition, the envelope of each band was also multiplied with a broadband noise source, but only in time windows where the original speech was unvoiced. Sections that were voiced in the original recordings were synthesised by multiplying the envelopes with a pulse train following the natural F0 contour. The individual pulses had a duration of one sample point, i.e. about 23 μs at a sampling rate of 44.1 kHz. The F0 contours of the original sentences were extracted using ProsodyPro version 4.3 (Xu, 2013) implemented in PRAAT (Boersma & Weenink, 2013), with the F0

extraction sampling rate set to 100 Hz. The resulting F0 contours were corrected manually where necessary and then used to determine the distance between the individual pulses of the pulse train sources. Based on the original intermittent F0 contours, we also produced artificial continuous F0 contours by interpolation through unvoiced sections and periods of silence. These continuous F0 contours were used to produce the pulse train sources for the periodic condition.

Finally, in all three vocoding conditions, the eight sub-band signals were again band-pass filtered using the same filters as in the analysis stage of the process. Before the individual bands were summed together, the output of each band was adjusted to the same RMS level as found in the original recordings.

Spectrally rotated speech was produced using a technique introduced by Blesser (1972) and implemented in MATLAB. Here, the waveforms of the mixed condition described above were first multiplied with an 8 kHz sinusoid, resulting in a spectral rotation around the midpoint frequency of 4 kHz. Note, that this procedure also renders the rotated speech inharmonic, since the frequencies of the component tones will not be multiples of a particular F0 anymore. The rotated waveforms were then filtered (FFT-based FIR filter, order 256) to have the average UK long-term speech spectrum (Byrne et al., 1994) and, finally, scaled to the same RMS level as the original waveforms in the mixed condition.

Figure 1 shows an unprocessed example sentence along with the same sentence processed in the four ways described.

Figure 1 about here

2.3. Procedure

Each participant listened to 80 aperiodic, 80 mixed, 160 periodic, and 80 rotated sentences. There were twice as many trials in the periodic condition because we wanted to ensure a

sufficient number of unintelligible trials. All 4 conditions were presented in blocks of 10 sentences (i.e. 1 complete IEEE sentence list) and the order of the conditions and IEEE lists was randomised. Only the first 40 IEEE lists were used in the main experiment and none of the sentences was presented more than once. Participants were asked to repeat as many words as possible after every sentence. The verbal responses were logged by the experimenter before the next sentence was played and no feedback was given following the responses. The presentation of the stimuli and the logging of the responses was carried out using Presentation version 17.0 (Neurobehavioral Systems, Berkeley, USA). Throughout this study, the term intelligibility will be defined simply as the average number of correctly repeated key words per condition.

Single trials consisted of a silent pre-stimulus interval with random duration (1.5–2.5 s), a stimulus sentence (average duration = 2.04 s, SD = .24 s) followed by a silent interval of .25 s, a short beep sound signalling the participants to start responding, the spoken responses, and the subsequent logging of the responses by the experimenter.

Before being tested, the subjects were familiarised with the materials by listening to 10 aperiodic, mixed, and periodic examples sentences each (IEEE lists 41–43). During the familiarisation phase, every sentence was directly followed by its unprocessed counterpart, and again followed by the processed sentence.

The main part of the experiment took about 70 minutes to complete and subjects were allowed to take breaks whenever they wished to. The experiment took place in a double-walled sound-attenuating and electrically shielded booth, with the computer signal being fed through the wall onto a separate monitor. Participants sat in a comfortable reclining chair during EEG acquisition and told to not move their eyes during sentence presentation. The stimuli were converted with 16-bit resolution and a sampling rate of 22.05 kHz using a Creative Sound Blaster SB X-Fi sound card (Dublin, Ireland) and presented over Sennheiser

HD650 headphones (Wedemark, Germany). The presentation level was about 71 dB SPL over a frequency range of .1–8 kHz as measured on an artificial ear (type 4153, Brüel & Kjær Sound & Vibration Measurement A/S, Nærum, Denmark).

2.4. EEG recording and analysis

The continuous EEG was recorded using a Biosemi ActiveTwo system (Amsterdam, Netherlands) with 61 Ag-AgCl scalp electrodes mounted on a cap according to the extended international 10-20 system. Four additional external electrodes were used to record the vertical and horizontal eletrooculogram (EOG) by placing them on the outer canthus of each eye as well as above and below the left eye. Two more external electrodes were used to record the reference signal from the left and right mastoids. EEG signals were recorded with a sampling rate of 512 Hz and an analogue anti-aliasing low-pass filter with a cut-off frequency of 200 Hz.

EEG data were processed offline using EEGLAB 12.0.2.5b (Delorme & Makeig, 2004). The continuous waveforms were first down-sampled to 256 Hz, re-referenced to the mean of the two mastoids, and then filtered using zero-phase shift Hamming-windowed sinc FIR filters (EEGLAB firfilt plugin version 1.5.3.; high-pass cut-off 0.01 Hz, transition bandwidth 0.1 Hz; low-pass cut-off 20 Hz, transition bandwidth 0.5 Hz). An independent component analysis (ICA) was used to remove eye artefacts. Epochs ranging from -1000 to 2500 ms were extracted and rejected if amplitudes exceeded ±200 μV, if linear trends exceeded 200 μV in a 1000 ms gliding window, or if the trial was lying outside a ±6 SD range (for single channels) and ±3 SD (for all channels) of the mean voltage probability distribution or the mean distribution of kurtosis values. On average 81% (324/400, SD = 48.3) of the total number of trials passed the rejection procedure.

EEG power spectra were computed by estimating the power spectral density (PSD) using Welch's method. The PSD was calculated with a 256-point Hamming window, an oversampling factor of 40, and a window overlap of 50%, resulting in a frequency resolution of .025 Hz. The EEG power spectra were computed for the single trials and averaged afterwards in order to estimate the total spectral power (i.e. time- but not necessarily phase-locked).

The processed EEG data were sorted according to the spoken behavioural responses. For the analysis of periodicity only trials with all five key words correct were considered, in order to control for the effect of intelligibility. This resulted in an average of 44.2 trials (SD = 8.2) in the aperiodic condition, 44.2 trials (SD = 9.7) in the mixed condition, and 57.9 trials (SD = 17.7) in the periodic condition.

For the analysis of intelligibility, trials in the periodic condition with different numbers of correctly repeated key words and the completely unintelligible rotated condition were separately compared. This resulted in the following average numbers of trials per condition: 8.4 (SD = 4.3) for 0 or 1 key words correct, 12.5 (SD = 5.5) for 2 key words correct, 21.4 (SD = 5.1) for 3 key words correct, 28.9 (SD = 5.9) for 4 key words correct, 57.9 (SD = 17.7) for 5 key words correct, and 67.1 (SD = 10.5) for the rotated condition. In order to obtain the final ERPs, the averaged epochs of each subject were baseline corrected by subtracting the mean amplitude in the -50–0 ms window before averaging across subjects.

Statistical differences between conditions were examined using non-parametric cluster-based permutation tests (Maris & Oostenveld, 2007). Firstly, it was tested whether there was a linear relationship between the amount of periodicity in the stimuli and the ERP amplitude by computing separate two-sided regression *t*-tests for dependent samples with linearly spaced regressors (1–3) at each electrode and for each sample point from 0–1000 ms after sentence onset. The same procedure was applied to test whether there was a linear

relationship between the intelligibility of the sentences and the ERP amplitude, but this time all sample points in the stimulus window (0–2500 ms) were examined and the regressors were set to values ranging from 1–6. Secondly, the individual sample points were merged into clusters if the *t*-values of their regression coefficients were significantly different from 0 at an alpha level of .05, and if the same was true for temporally adjacent sample points and at least 3 neighbouring channels. This procedure provides a weak control for false positive findings due to multiple comparisons by only allowing effects that are coherent in time and space. Next, the *t*-values within a given cluster were summed to obtain the cluster-level statistic. The significance probability of a cluster was then assessed by comparing this cluster-level statistic to the one obtained after randomly re-allocating the individual trials to the conditions. This step was repeated 1000 times and the proportion of these Monte-Carlo iterations in which the cluster-level statistic was exceeded then determined the final cluster *p*-value.

The same statistical technique was applied to test whether there was a linear relationship between pre-stimulus alpha power and sentence intelligibility in the periodic condition, but in this case the EEG power spectrum in the pre-stimulus period (-1000–0 ms) was first averaged over a frequency window from 7–10 Hz in each condition. Here, the regressors were set to values from 1–5, corresponding to the number of correct key words. Consequently, only a single regression coefficient was computed per electrode, and these were subsequently clustered according to their *p*-values and spatial adjacency.

3. Results

3.1. Behavioural data

The averaged spoken responses obtained after each trial (Fig. 2) show that the aperiodic and mixed conditions are equally intelligible (88.8% and 90.0% correct key words on average; $t(17) = -1.60$, $p = .13$), while the rotated condition is completely unintelligible (0%), and

periodic speech is less intelligible (77.4%) than the aperiodic ($t(17) = -8.42$, $p < .001$) and mixed conditions ($t(17) = -11.60$, $p < .001$). Furthermore, we compared the responses to the first and second half of the trials in the periodic condition and found no significant differences (77.8% and 77.0%; $t(17) = .70$, $p = .49$), indicating that there were no learning effects over the course of the 160 trials.

Figure 2 about here

3.2. Periodicity

As shown by the ERP traces recorded at electrode FC2 in Fig. 3A, the three conditions varying regarding the amount of acoustic periodicity (aperiodic, mixed, and periodic speech) all elicited an auditory-evoked P1-N1-P2 complex after sentence onset, followed by an acoustic change complex (ACC, consisting of CP1, CN1, and CP2 components) from about 300–500 ms in response to the onset of the second syllable (Pratt, 2011). Furthermore, all three conditions showed a sustained negativity from about 300–2500 ms past sentence onset.

Crucially, after the initial P1 component, peaking at around 50 ms after sound onset, the ERPs in the three conditions were found to diverge, showing greater negative amplitudes with more periodicity. This parametric effect is observable until about one second after sound onset and thus considerably overlaps with the slow negativity. A cluster-corrected linear regression *t*-test including all three conditions confirmed that there was a significant linear negative relationship during this time window by returning three separate significant clusters in the right fronto-central scalp region: the first one was found during the period of the N1 and P2 components between about 90–230 ms ($p = .034$), the second cluster ranging from about 300–440 ms ($p = .028$) coincided with the ACC, and the third cluster was observed between about 715–840 ms ($p = .03$) after sound onset. The average voltage distributions of

each condition during the three clusters along with *t*-value maps depicting the scalp distributions of statistical differences for each cluster are shown in Fig. 3B.

Figure 3 about here

3.3. Intelligibility

In order to analyse how the ERPs were affected by the intelligibility of the stimulus sentences, trials in the periodic condition were sorted into five categories, according to the spoken responses of the participants (zero or one, two, three, four, and five key words correct). Additionally, spectrally rotated speech was included as a completely unintelligible control condition.

As illustrated in Fig. 4A, which shows the ERPs recorded at electrode FC2, all six conditions elicited a slow negativity from about 300–2500 ms after the beginning of the sentences. This slow negativity, taken to be a CNV, had the smallest amplitude in the rotated condition, followed by slightly larger amplitudes for trials in the periodic condition with zero or one and two correct key words, and substantially larger amplitudes for trials in the periodic condition with three, four and five key words correct. A cluster-corrected regression *t*-test including all six conditions returned one large significant cluster ($p = .004$) from about 470–2250 ms, confirming that there was a linear negative relation between the intelligibility of the sentences and the amplitude of the CNV. The corresponding *t*-map shows that this cluster included a large number of electrodes in the central and right fronto-temporal scalp region (see map at far right in Fig. 4B). The voltage maps showing the ERP amplitudes averaged over the duration of the whole cluster in each condition confirm that the activity was strongest in the fronto-central scalp region and slightly lateralised to the right, particularly for the more intelligible conditions (three or more correct key words, Fig. 4B).

In order to test whether the smaller CNV in the conditions with two or less correct key words were due to the low trial numbers, we computed the Spearman rank correlation coefficients between the number of trials per subject and their CNV amplitudes (averaged over all 61 scalp electrodes and the whole stimulus window). These correlations were in both cases not significant (0/1 Words: r = -.24, p = .34; 2 Words: r = .24, p = .33), indicating that the observed effect was not driven by the subjects with the fewest trials within each condition.

In addition to the finding that CNV amplitudes were larger when the sentences were more intelligible to the subjects, the data in Fig. 4 also show that the six conditions appeared to group into three distinct categories (rotated, maximally two key words, and minimally three key words). In order to follow up this observation, we tested if there were any significant differences within these categories. Firstly, trials with zero or one correct key words were compared to trials with two correct key words using a cluster-based *t*-test. Secondly, trials with three, four, and five correct key words were compared using a cluster-based ANOVA. Both tests revealed no significant differences at any point during the stimulus window (0–2500 ms). Based on this finding, trials in the periodic condition were pooled into a more and less intelligible category (maximally two versus minimally three correct key words, respectively) and separately compared, leaving out the rotated condition to ensure a test that is free of any acoustic confounds. For this *post-hoc* analysis, a cluster-corrected regression *t*-test including all sample points in the significant time window (470–2250 ms) revealed one cluster with a *p*-value of .036 from about 780–1640 ms. The voltage maps averaged over this significant time window show that the activity is lateralised to the right in the more intelligible condition, which is confirmed by the location of the significant cluster of electrodes in the right temporal scalp region (Fig. 4C).

Figure 4 about here

3.4. Pre-stimulus alpha power

In a final analysis, we tested whether the amount of alpha power in the silent pre-stimulus period before sentence onset stands in relation to the intelligibility of the stimulus sentences in the periodic condition. As shown by the line plot in Fig. 5A, depicting the average EEG power spectra in the pre-stimulus window (-1000–0 ms) recorded at electrode FC2, slow alpha power (7–10 Hz) was markedly increased before the least intelligible trials, with maximally one out of five correctly repeated key words. Furthermore, it can be seen that the differences between the five conditions were indeed confined to the slow alpha range. The scalp distributions of the average spectral power in this frequency window show peaks of activity over the occipital scalp region in all five conditions, along with a widespread power increase extending into the anterior scalp region for the least intelligible trials (Fig. 5B). A cluster-based regression *t*-test comparing the averaged pre-stimulus slow alpha power (7–10 Hz/-1000–0 ms) in all five conditions at each electrode revealed a large cluster comprising 18 significant electrodes in the right frontal scalp region ($p = .016$, see *t*-map at far right of Fig. 5B).

Same as for the ERPs, the Spearman rank correlation coefficients between the number of trials per subject and the amount of slow alpha power (averaged over all 61 scalp electrodes, and the whole pre-stimulus window) was not significant for the conditions with the fewest trials (0/1 Words: $r = .07$, $p = .78$; 2 Words: $r = .06$, $p = .83$), showing that the results within these conditions were not biased by the subjects with the lowest numbers of trials.

As the relation between slow alpha power in the pre-stimulus window and intelligibility was not strictly linear, further tests were performed. Firstly, the four conditions with two or more correct key words, who appeared not to differ regarding the amount of slow alpha power, were separately compared using a cluster-based ANOVA. This test did not

reveal any significant differences between the four conditions. However, when all trials with two or more correct key words were pooled into a single condition and compared to the least intelligible trials using a one-tailed cluster-corrected *t*-test, the same significant cluster of electrodes as shown in Fig. 5 was obtained, which confirms that slow alpha power was increased for the least intelligible trials only.

Figure 5 about here

4. Discussion

The purpose of the present study was to tease apart effects of acoustics and intelligibility on the ERPs in response to speech. It was found, firstly, that more acoustic periodicity in the speech signals parametrically rendered the ERP waveforms during the first second after sentence onset more negative. Periodicity thus appears to amplify the evoked cortical response in the early period after sound onset. Secondly, we observed a CNV that was larger when the speech signals were more intelligible to the participants. However, this relationship was not strictly linear, as the amplitude of the negativity differed significantly between trials with less and more than half of the key words correctly repeated, but not within these categories. Additionally, slow alpha power (7–10 Hz) in the silent baseline interval preceding the sentences that turned out to be least intelligible to the participants was found to be markedly increased, while there was no difference between the rest of the trials.

4.1. Periodicity

The finding that more periodicity leads to larger negative ERP amplitudes is in line with pitch perception studies reporting greater neural responses to sound input that possesses a pitch (e.g. Chait et al., 2006; Griffiths et al., 2010; Norman-Haignere et al., 2013). As we have controlled for differences in intelligibility across conditions by only including trials with all five key words correctly repeated, and sentence materials as well as the behavioural task were

the same throughout the experiment, it seems unlikely that any cognitive process can explain this effect. Furthermore, the effect was significant from as early as 90 ms after acoustic onset, a latency which is generally thought to be dominated by responses to the acoustic properties of a stimulus (Picton, Hillyard, Krausz, & Galambos, 1974; Pratt, 2011). However, the effect was not confined to the time window of AEPs and ACC, i.e. until about 500 ms post-onset, but present until almost one second after sound onset, classifying as a sustained pitch response (Gutschalk et al., 2004). The current results thus stress the importance of taking the acoustic properties of the stimuli into account when investigating speech perception, particularly when the duration of the stimuli is relatively short (e.g. single words).

4.2. Intelligibility

As outlined in the introduction, slow cortical potentials may reflect working memory operations, the level of attention spent on a task, and how prepared to respond a subject is. Regarding the task to verbally repeat relatively long auditorily presented sentences, it appears likely that all three factors play a role. Firstly, larger amounts of verbal material have to be retained in working memory when the sentences are more intelligible. Secondly, when the stimulus sentences were less intelligible to them, subjects were presumably paying less attention to a task they realised they could not accomplish. Similarly, the inability to understand the materials is necessarily going along with failing to prepare for the subsequent verbal response. In line with this interpretation, significant differences in CNV amplitude were not observed right after sentence onset, but started to emerge a few hundred milliseconds after, suggesting that the subjects first had to process the initial part of the sentences before these cognitive processes were triggered.

Although the task used in this study was not typical for eliciting a CNV, the fact that the amplitude of the slow negativity did not increase further when three or more key words per sentence were correctly repeated provides further evidence for this interpretation. CNV amplitudes have often been reported to be limited, or even to have an inverted u-shaped relationship with task demand (Birbaumer et al., 1990; Kononowicz & Penney, 2016). In turn, however, the monotonic but not strictly linear relation of speech intelligibility and CNV amplitude observed in the current study also suggests that the CNV cannot be used as an accurate predictor of speech intelligibility scores.

In a recent study that resembles the current one to some extent, Wöstmann et al. (2015) have reported a slow negativity, which was also taken to be a CNV, in an auditory number comparison task. In their study, subjects had to remember numbers in the presence of a competing talker in the background, and the signal mixture was furthermore acoustically degraded. Crucially, more severe degradations resulted in larger CNV amplitudes, although the intelligibility of the numbers and the task performance decreased somewhat. Wöstmann et al. thus concluded that the CNV amplitude serves as a measure of expected task difficulty and listening effort. Although it remains to be investigated how the CNV in response to speech presented in background noise varies when the intelligibility fluctuates over a wider range, this suggests that slow cortical potentials may reflect different cognitive processes for speech presented in quiet and in noise. Importantly, in the present study subjects could not know whether they would be able to understand a particular sentence in the periodic condition before it was played to them. Hence, the differences in CNV amplitude for the more or less intelligible trials cannot be explained by the expected task difficulty, which was assumed to be constant.

4.3. Pre-stimulus alpha power

The slow alpha power before the least intelligible trials was found to have a broad scalp distribution extending into the anterior scalp region. As summarised by Klimesch (1999), slower alpha frequencies generally have a more anterior scalp distribution than faster ones and the distribution found in the current study also corresponds well with the example scalp map provided in this review paper. As shown by Laufs et al. (2006), there appear to be two distinguishable alpha networks, one that comprises occipital vision areas and a second one in fronto-parietal areas associated with attention. The scalp location of the cluster of significant electrodes found in the current study corresponds well with that of the right-lateralised ventral fronto-parietal attention network, which is deactivated when subjects focus on a task (Corbetta, Patel, & Shulman, 2008; Corbetta & Shulman, 2002). Deactivation of this network has been associated with the prevention of irrelevant task switching (Shulman, Astafiev, McAvoy, d'Avossa, & Corbetta, 2007) and our data suggest that this deactivation may coincide with a decrease in alpha power. The location of this effect is also well in line with the results of Strauß et al. (2015), who have observed the strongest differences in alpha phase before correct and incorrect trials in a lexical decision task in this region.

As described by Mazaheri and Jensen (2008, 2010), slow ERP deflections may be caused by amplitude fluctuations of induced alpha power because the peaks of alpha oscillations appear to be more strongly modulated than the troughs. However, this explanation does not seem to apply to the current results, since the amplitude of the slow negativity varies independently of the pre-stimulus alpha power. That is, the slow alpha power was only increased before the least intelligible trials (zeros or one correct key words), but the CNV had a similar amplitude for these trials and those with two correct key words. Hence, same as for the CNV, the non-linear relationship of pre-stimulus alpha power and intelligibility does not allow the accurate prediction of speech intelligibility rates.

5. Conclusions

The current study investigated cortical EEG responses to auditorily presented sentences with a focus on the differential contributions of acoustics and intelligibility. Firstly, more acoustic periodicity in the stimuli was found to render the ERPs during the first second after speech onset more negative. This demonstrates that acoustic factors should not be disregarded in neuroscientific studies investigating speech perception, even when focussing on cognitive processes. Secondly, we observed a CNV from about half a second after sentence onset, the amplitude of which was larger when the sentences were more intelligible to the participants. Additionally, slow alpha power before the least intelligible sentences was significantly higher than before the rest of the trials. However, as the latter two measures did not vary precisely as a function of the number of correctly repeated key words and did not appear to stand in relation, they both do not appear to serve as accurate predictors of speech intelligibility.

Acknowledgments

This project has been funded with support from the European Commission under Contract FP7-PEOPLE-2011-290000. We thank Natalie Berger and Jyrki Tuomainen for helpful comments.

References

Becker, R., Pefkou, M., Michel, C. M., & Hervais-Adelman, A. G. (2013). Left temporal alpha-band activity reflects single word intelligibility. *Front. Syst. Neurosci.,* 7, Article 121.

Birbaumer, N., Elbert, T., Canavan, A., & Rockstroh, B. (1990). Slow potentials of the cerebral cortex and behavior. *Physiol. Rev.,* 70(1), 1–41.

Blesser, B. (1972). Speech perception under conditions of spectral transformation: I. Phonetic characteristics. *J. Speech Hear. Res.,* 15(1), 5–41.

Boersma, P., & Weenink, D. (2013). Praat: doing phonetics by computer [Computer program] (Version 5.3.49).

Byrne, D., Dillon, H., Tran, K., Arlinger, S., Wilbraham, K., Cox, R., et al. (1994). An international comparison of long-term average speech spectra. *J. Acoust. Soc. Am.,* 96(4), 2108–2120.

Chait, M., Poeppel, D., & Simon, J. Z. (2006). Neural response correlates of detection of monaurally and binaurally created pitches in humans. *Cereb. Cortex*, 16(6), 835–848.
Corbetta, M., Patel, G., & Shulman, G. L. (2008). The reorienting system of the human brain: from environment to theory of mind. *Neuron*, 58(3), 306–324.
Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, 3(3), 201–215.
Davis, M. H., & Johnsrude, I. S. (2003). Hierarchical processing in spoken language comprehension. *J. Neurosci.*, 23(8), 3423–3431.
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(1), 9–21.
Ding, N., Chatterjee, M., & Simon, J. Z. (2014). Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. *Neuroimage*, 88, 41–46.
Dudley, H. (1939). Remaking speech. *J. Acoust. Soc. Am.*, 11(2), 169–177.
Evans, S., Kyong, J., Rosen, S., Golestani, N., Warren, J., McGettigan, C., et al. (2014). The pathways for intelligible speech: multivariate and univariate perspectives. *Cereb. Cortex*, 24(9), 2350–2361.
Greenwood, D. D. (1990). A cochlear frequency-position function for several species – 29 years later. *J. Acoust. Soc. Am.*, 87(6), 2592–2605.
Griffiths, T. D., Kumar, S., Sedley, W., Nourski, K. V., Kawasaki, H., Oya, H., et al. (2010). Direct recordings of pitch responses from human auditory cortex. *Curr. Biol.*, 20(12), 1128–1132.
Guimond, S., Vachon, F., Nolden, S., Lefebvre, C., Grimault, S., & Jolicoeur, P. (2011). Electrophysiological correlates of the maintenance of the representation of pitch objects in acoustic short-term memory. *Psychophysiology*, 48(11), 1500–1509.
Gutschalk, A., Patterson, R. D., Scherg, M., Uppenkamp, S., & Rupp, A. (2004). Temporal dynamics of pitch in human auditory cortex. *Neuroimage*, 22(2), 755–766.
Hanslmayr, S., Aslan, A., Staudigl, T., Klimesch, W., Herrmann, C. S., & Bäuml, K.-H. (2007). Prestimulus oscillations predict visual perception performance between and within subjects. *Neuroimage*, 37(4), 1465–1473.
He, B. J., & Raichle, M. E. (2009). The fMRI signal, slow cortical potential and consciousness. *Trends Cogn. Sci.*, 13(7), 302–309.
Kerlin, J. R., Shahin, A. J., & Miller, L. M. (2010). Attentional gain control of ongoing cortical speech representations in a “cocktail party”. *J. Neurosci.*, 30(2), 620–628.
Klimesch, W. (1999). EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain Res. Rev.*, 29(2), 169–195.
Kononowicz, T. W., & Penney, T. B. (2016). The contingent negative variation (CNV): timing isn’t everything. *Current Opinion in Behavioral Sciences*, 8, 231–237.
Laufs, H., Holt, J. L., Elfont, R., Krams, M., Paul, J. S., Krakow, K., et al. (2006). Where the BOLD signal goes when alpha EEG leaves. *Neuroimage*, 31(4), 1408–1418.
Lefebvre, C., Vachon, F., Grimault, S., Thibault, J., Guimond, S., Peretz, I., et al. (2013). Distinct electrophysiological indices of maintenance in auditory and visual short-term memory. *Neuropsychologia*, 51(13), 2939–2952.
Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG-and MEG-data. *J. Neurosci. Methods*, 164(1), 177-190.
Mazaheri, A., & Jensen, O. (2008). Asymmetric amplitude modulations of brain oscillations generate slow evoked responses. *J. Neurosci.*, 28(31), 7781–7787.
Mazaheri, A., & Jensen, O. (2010). Rhythmic pulsing: linking ongoing brain activity with evoked responses. *Front. Hum. Neurosci.*, 4, Article 177.

McCallum, W., & Walter, W. G. (1968). The effects of attention and distraction on the contingent negative variation in normal and neurotic subjects. *Electroencephalogr. Clin. Neurophysiol.*, 25(4), 319–329.

Müller, N., & Weisz, N. (2012). Lateralized auditory cortical alpha band activity and interregional connectivity pattern reflect anticipation of target sounds. *Cereb. Cortex*, 22(7), 1604–1613.

Norman-Haignere, S., Kanwisher, N., & McDermott, J. H. (2013). Cortical pitch regions in humans respond primarily to resolved harmonics and are located in specific tonotopic regions of anterior auditory cortex. *J. Neurosci.*, 33(50), 19451–19469.

Obleser, J., & Kotz, S. A. (2011). Multiple brain signatures of integration in the comprehension of degraded speech. *Neuroimage*, 55(2), 713–723.

Obleser, J., & Weisz, N. (2012). Suppressed alpha oscillations predict intelligibility of speech and its acoustic Details. *Cereb. Cortex*, 22(11), 2466–2477.

Obleser, J., Wöstmann, M., Hellbernd, N., Wilsch, A., & Maess, B. (2012). Adverse listening conditions and memory load drive a common alpha oscillatory network. *J. Neurosci.*, 32(36), 12376–12383.

Peelle, J. E., Gross, J., & Davis, M. H. (2013). Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cereb. Cortex*, 23(6), 1378–1387.

Picton, T. W., Hillyard, S. A., Krausz, H. I., & Galambos, R. (1974). Human auditory evoked potentials. I: Evaluation of components. *Electroencephalogr. Clin. Neurophysiol.*, 36, 179–190.

Pratt, H. (2011). Sensory ERP Components. In S. J. Luck & E. S. Kappenman (Eds.), *The Oxford Handbook of Event-Related Potential Components* (pp. 89–114). New York: Oxford University Press USA.

Romei, V., Gross, J., & Thut, G. (2010). On the role of prestimulus alpha rhythms over occipito-parietal areas in visual input regulation: correlation or causation? *J. Neurosci.*, 30(25), 8692–8697.

Rothauser, E. H., Chapman, N. D., Guttman, N., Nordby, K. S., Silbiger, H. R., Urbanek, G. E., et al. (1969). IEEE recommended practice for speech quality measurements. *IEEE Trans. on Audio and Electroacoustics*, 17(3), 225–246.

Schubert, R., Haufe, S., Blankenburg, F., Villringer, A., & Curio, G. (2009). Now you'll feel it, now you won't: EEG rhythms predict the effectiveness of perceptual masking. *J. Cognit. Neurosci.*, 21(12), 2407–2419.

Scott, S. K., Blank, C. C., Rosen, S., & Wise, R. J. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain*, 123(12), 2400–2406.

Shulman, G. L., Astafiev, S. V., McAvoy, M. P., d'Avossa, G., & Corbetta, M. (2007). Right TPJ deactivation during visual search: functional significance and support for a filter hypothesis. *Cereb. Cortex*, 17(11), 2625–2633.

Steinmetzger, K., & Rosen, S. (2015). The role of periodicity in perceiving speech in quiet and in background noise. *J. Acoust. Soc. Am.*, 138(6), 3586–3599.

Strauß, A., Henry, M. J., Scharinger, M., & Obleser, J. (2015). Alpha Phase Determines Successful Lexical Decision in Noise. *J. Neurosci.*, 35(7), 3256–3262.

Tecce, J. J., & Scheff, N. M. (1969). Attention reduction and suppressed direct-current potentials in the human brain. *Science*, 164(3877), 331–333.

Teder-Sälejärvi, W. A., Münte, T. F., Sperlich, F.-J., & Hillyard, S. A. (1999). Intra-modal and cross-modal spatial attention to auditory and visual stimuli. An event-related brain potential study. *Cognitive Brain Research*, 8(3), 327–343.

Van Dijk, H., Schoffelen, J.-M., Oostenveld, R., & Jensen, O. (2008). Prestimulus oscillatory activity in the alpha band predicts visual discrimination ability. *J. Neurosci.*, 28(8), 1816–1823.

Wilsch, A., Henry, M. J., Herrmann, B., Maess, B., & Obleser, J. (2015). Alpha Oscillatory Dynamics Index Temporal Expectation Benefits in Working Memory. *Cereb. Cortex*, 25(7), 1938–1946.

Woods, D. L., Alho, K., & Algazi, A. (1994). Stages of auditory feature conjunction: an event-related brain potential study. *J. Exp. Psychol. Hum. Percept. Perform.*, 20(1), 81–94.

Wöstmann, M., Schröger, E., & Obleser, J. (2015). Acoustic detail guides attention allocation in a selective listening task. *J. Cognit. Neurosci.*, 27(5), 988–1000.

Xu, Y. (2013). *ProsodyPro – A tool for large-scale systematic prosody analysis.* Paper presented at the Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013), Aix-en-Provence, France.

Figure 1. Stimuli. Waveforms, wide-band spectrograms, and F0 contours for one example sentence (*Say it slowly but make it ring clear.*). A) The unprocessed version of the sentence. B) The same sentence processed to have an aperiodic source, C) a mixed source, D) a periodic source, or E) a mixed source and spectrally rotated. The four processed conditions (B–E) were all vocoded with eight frequency bands. The unprocessed version of the sentence in panel A) is shown for the purpose of comparison only.

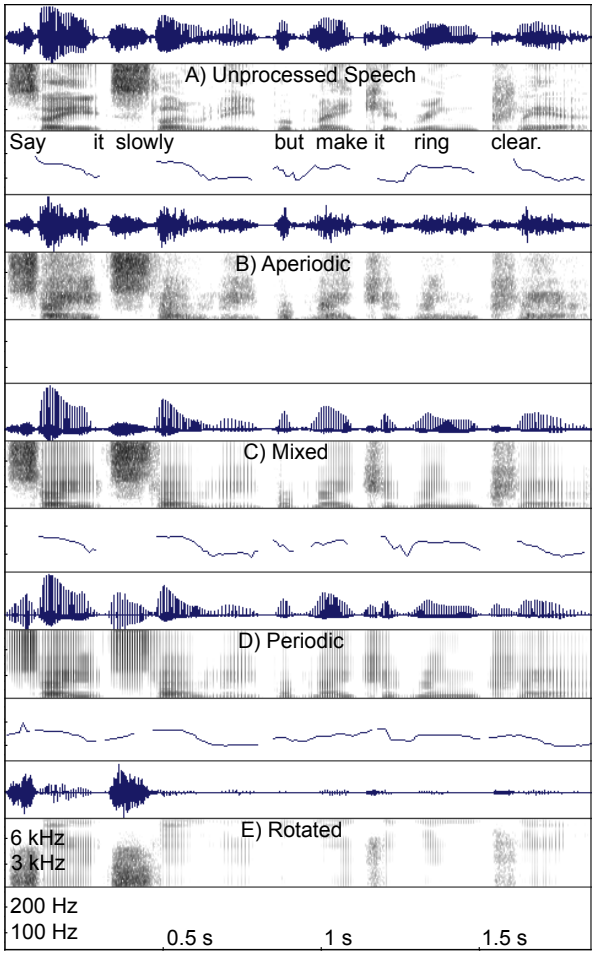
Figure 2. Behavioural data. Boxplots showing the average proportion of correctly repeated key words in each of the four speech conditions. The black horizontal lines in the boxplots represent the median value. *** indicates a *p*-value < .001, *n.s.* stands for not significant.

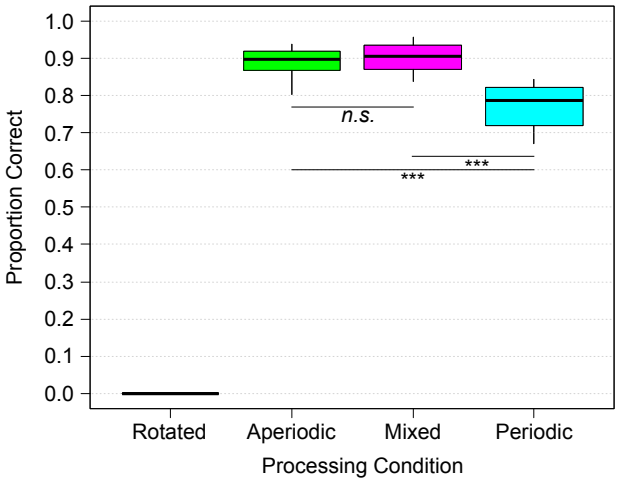
Figure 3. Periodicity. A) Grand average ERPs recorded at electrode FC2 for fully intelligible trials (all 5 key words correctly repeated) in the aperiodic, mixed, and periodic conditions. The three thick black lines below the ERP traces indicate time windows during which there was a significant linear negative relationship between the amount of periodicity in the stimuli and the ERP amplitude ($p < .05$). ERP waveforms were low-pass filtered at 10 Hz for illustration purposes. B) Voltage maps showing the mean activity during the three significant time windows for each condition. In the three *t*-value maps on the far right, black dots indicate electrodes whose *p*-values were < .05 at each sample point during the respective time window.

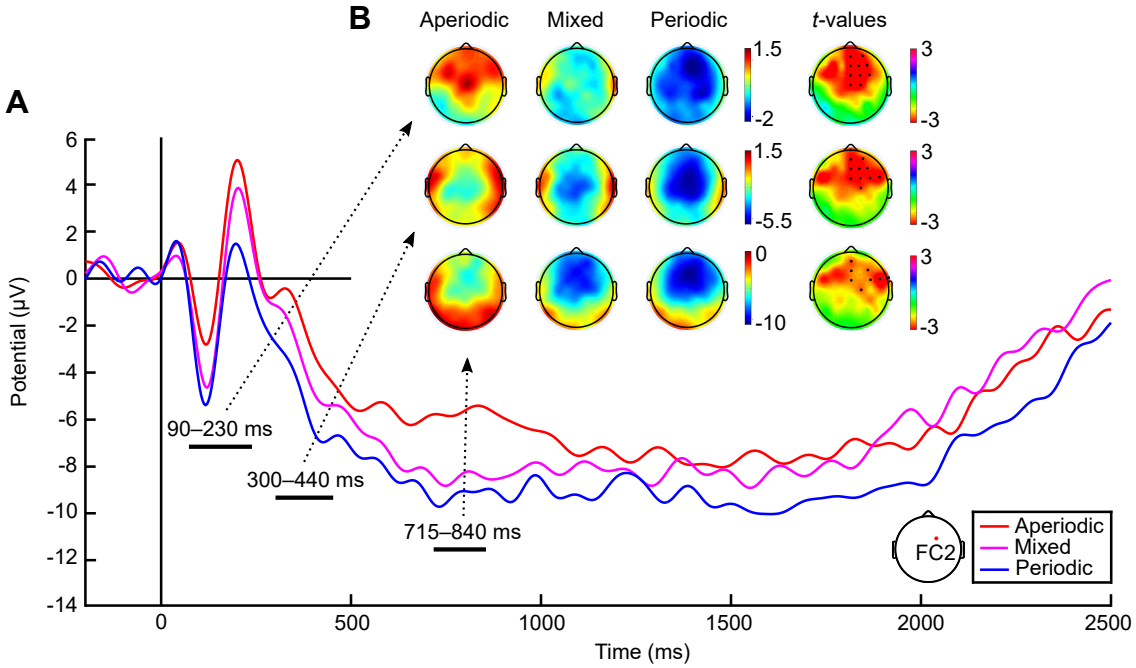
Figure 4. Intelligibility. A) Grand average ERPs recorded at electrode FC2 for the completely unintelligible rotated condition and trials in the periodic condition with 0/1, 2, 3, 4, or 5 correctly repeated key words. The thick black line below the ERP traces indicates the time window during which there was a significant linear negative relationship between the

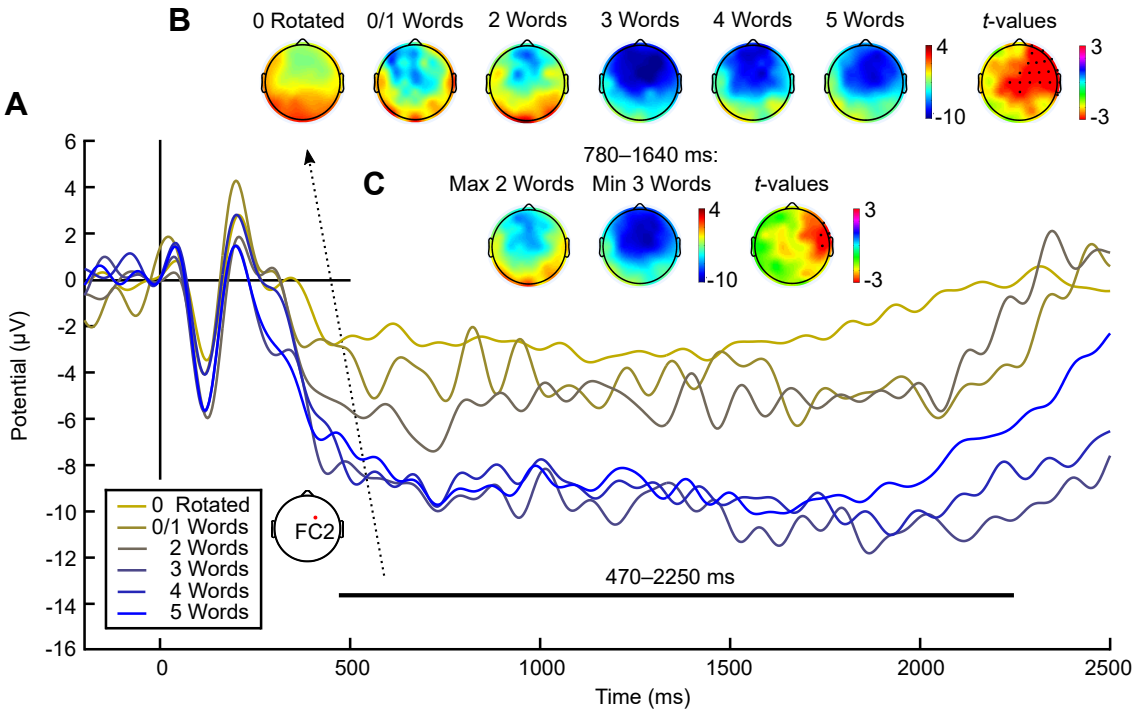
intelligibility of the stimuli and the ERP amplitude ($p < .01$). ERP waveforms were low-pass filtered at 10 Hz for illustration purposes. B) Voltage maps showing the mean activity during the significant time window for each condition. In the *t*-value map on the far right, black dots indicate electrodes whose *p*-values were < .01 at each sample point during the respective time window. C) Voltage distributions and *t*-map showing the mean activity during the time window in which the ERP amplitudes of the pooled less (maximally 2 key words) and more (minimally 3 key words) intelligible trials in the periodic condition differed significantly ($p < .05$).

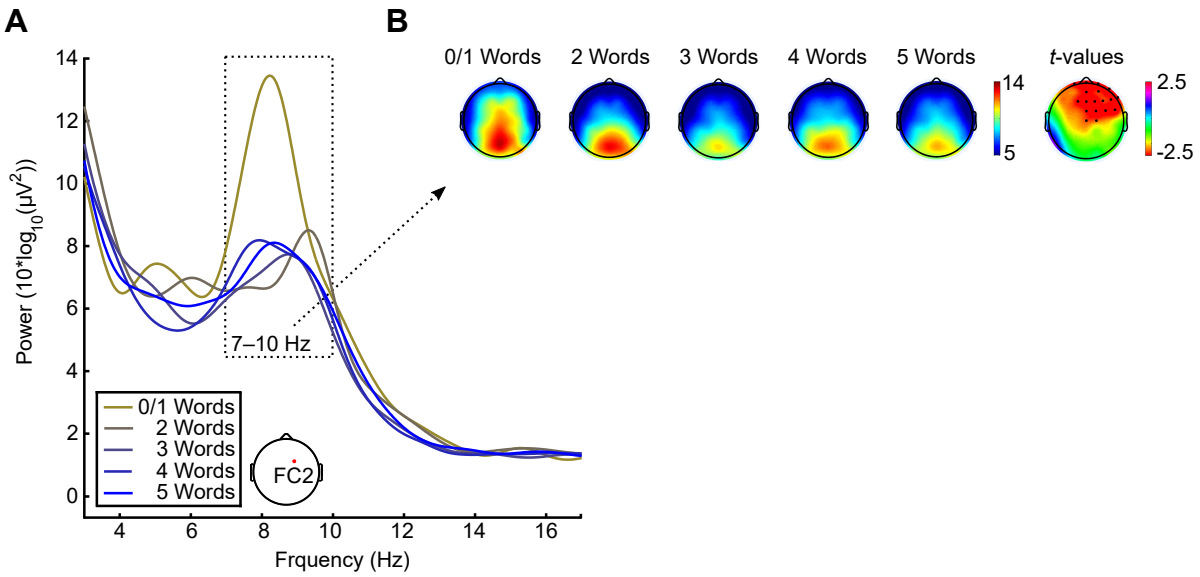
Figure 5. Pre-stimulus alpha power. A) Line plot showing the averaged EEG power spectra in the silent pre-stimulus window (-1000–0 ms), recorded at electrode FC2, for trials in the periodic condition with 0/1, 2, 3, 4, or 5 correctly repeated key words. B) Scalp maps of the mean alpha power in the 7–10 Hz frequency window for each of the five conditions. In the *t*-map on the far right, black dots indicate electrodes with *p*-values < .05.


A) Unprocessed Speech
Say it slowly but make it ring clear.
B) Aperiodic
C) Mixed
D) Periodic
E) Rotated
6 kHz
3 kHz
200 Hz
100 Hz
0.5 s
1 s
1.5 s


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Proportion Correct
n.s.
***
***
Rotated
Aperiodic
Mixed
Periodic
Processing Condition


A
B
Aperiodic
Mixed
Periodic
t-values
1.5
-2
1.5
-5.5
0
-10
3
-3
3
-3
3
-3
Potential (μV)
6
4
2
0
-2
-4
-6
-8
-10
-12
-14
90–230 ms
300–440 ms
715–840 ms
0
500
1000
1500
2000
2500
Time (ms)
FC2
Aperiodic
Mixed
Periodic


A
B
0 Rotated
0/1 Words
2 Words
3 Words
4 Words
5 Words
t-values
4
-10
3
-3
C
780–1640 ms:
Max 2 Words
Min 3 Words
t-values
4
-10
3
-3
Potential (μV)
6
4
2
0
-2
-4
-6
-8
-10
-12
-14
-16
0 Rotated
0/1 Words
2 Words
3 Words
4 Words
5 Words
FC2
470–2250 ms
0
500
1000
1500
2000
2500
Time (ms)


A
B
0/1 Words
2 Words
3 Words
4 Words
5 Words
t-values
14
5
2.5
-2.5
Power (10*log10(μV2))
7–10 Hz
FC2
Frquency (Hz)