Electrophysiological and behavioral measures of some speech contrasts in varied attention and noise

Morris, David Jackson; Tøndering, John; Lindgren, Magnus

Published in:
Hearing Research

Publication date:
2019

Document version
Peer reviewed version

Electrophysiological and behavioral measures of some speech contrasts in varied attention and noise

Authors: David Jackson Morris[ab], John Tøndering[a] & Magnus Lindgren[c]

a) University of Copenhagen, Department of Nordic Studies and Linguistics, Speech Pathology and Audiology, Emil Holms Kanal 2, 2300 Copenhagen, Denmark.
b) Lund University, Humanities Laboratory, Helgonabacken 12, Lund, 22100, Sweden.
c) Lund University Department of Psychology, Paradisgatan 5, Lund, 22100, Sweden.

Corresponding author: David Morris dmorris@hum.ku.dk +46 734 351 878
John Tøndering johnt@hum.ku.dk
Magnus Lindgren magnus.lindgren@psy.lu.se

Number of tables: 3
Number of figures: 6
Appendix: 1
Supplementary file: 1.m4v

Running head suggestion: Some speech contrasts in varied attention and noise

Keywords: Speech perception; Speech features; Voice onset time; Place of articulation; Vowel length; Danish phonology; Electrophysiology; Global Field Power; EEG microstates

Abbreviations:
EEG Electroencephalograpy
ERP Event Related Potential
FDR False Discovery Rate
GFP Global Field Power
ISI Interstimulus Interval
RMS Root-Mean Square
SNR Signal-Noise Ratio
SINFA Sequential Information Transfer Analysis
VOT Voice-Onset Time

Abstract

This paper investigates the salience of speech contrasts in noise, in relation to how listening attention affects scalp-recorded cortical responses. The contrasts that were examined with consonant-vowel syllables, were place of articulation, vowel length and voice-onset time (VOT) and our analysis focuses on the correspondence between the effect of attention on the electrophysiology and the decrement in behavioral results when noise was added to the stimuli. Normal-hearing subjects (n=20) performed closed-set syllable identification in no noise, 0, 4 and 8 dB signal-noise ratio (SNR). Identification in noise decreased markedly for place of articulation, moderately for vowel length and marginally for VOT. The same syllables were used in two electrophysiology conditions, where subjects attended to the stimuli, and also while their attention was diverted to a visual discrimination task. Differences in global field power between the attention conditions from each contrast showed that that the effect of attention was negligible for place of articulation. They implied offset encoding of vowel length and were early (starting at 117 ms), and of high amplitude (>3 μV) for VOT. There were significant correlations between the difference in syllable identification in no noise and 0 dB SNR and the electrophysiology results between attention conditions for the VOT contrast. Comparison of the two attention conditions with microstate analysis showed a significant difference in the duration of microstate class D. These results show differential integration of attention and syllable processing according to speech contrast and they suggest that there is correspondence between the salience of a contrast in noise and the effect of attention on the evoked electrical response.

1. Introduction

The gain theory of selective attention proposes that the processing of target stimuli is enhanced while simultaneous competing sensory input is gated (Hillyard, et al., 1973). The putative mechanism underlying this is that sensitivity and responsiveness of neuronal populations that are tuned to targets are upregulated, while nontarget relevant populations are downregulated (Johnson and Zatorre, 2006; Treue and Martínez Trujillo, 1999). The upshot of this attentional effect on auditory event-related potentials (ERP) is adjuvant as it bolsters the amplitude and may decrease the latency of response components (for example, Hillyard et al., 1973; Luo and Wei, 1999). Thornton et al. (2007) have shown that for tonal stimuli, latency jitter in the averaged time series accounts for much of this attentional difference. During dichotic listening, the adjuvant effect of attention on electroencephalography (EEG) has been shown to diminish with decreases in spectral richness, suggesting that there is a relationship between stimulus degradation and the neuronal response (Kong, et al., 2015). This study examines the adjuvant effect of attention on electrophysiological measurements elicited by speech contrasts and compares them to behavioral identification of the same stimuli in noise backgrounds.

The three speech contrasts that were examined in both behavioral and electrophysiological testing were place of articulation, vowel length and voice onset time (VOT). Place of articulation describes the point of maximal constriction in the vocal tract where an active meets a passive articulator. The perception of place of articulation is highly sensitive to noise (Miller and Nicely, 1955) and is thought to be intact at birth (Molfese and Molfese, 1979). Contrastive vowel length is not encountered in most dialects of English, but it is a common syllabic feature in Scandinavian languages. This contrast is illustrated by the initial vowel of the Danish word *arme* /ˈɑːmə/ ‘arm’ and *amme* /ˈɑmə/ ‘breastfeed’. Our interest in including the vowel length contrast was that vowels, due to their amplitude, are prominent features in speech and it is a meaningful contrast in Danish. The VOT contrast is longer in /pV/, /tV/ and /kV/ than in /bV/, /dV/ and /gV/ syllables and sensitivity to this contrast has also been observed in infancy (Jusczyk, et al., 1989). In CV syllables, contrastive VOT is principally differentiated by the time from the release of the stop to the onset of vocal fold vibration in the consequent vowel. The consonants used in the present study, /p b k g/, are realized according to aspiration as [pʰ b̥ kʰ g̊] and it should be noted here that in Danish phonology, stops, including /p t k/ and /b d g/, are all unvoiced in syllable initial position.

It is unclear whether the VOT contrast, particularly the voiceless phase prior to the vowel, is rendered in the electrical response of listeners. Intracranially-recorded neural activity arising from different VOT has implicated temporal coding of the contrast as a possible mechanism, whereby short VOTs are associated with responses time-locked to the consonant release, while long VOTs have a 'double-on' response that reflects both the burst and the vowel onset (Steinschneider, et al., 1994; Steinschneider, et al.,1999). Similar responses have been shown in human vertex ERPs where a bifid averaged N1 was observed for long VOTs (Sharma and Dorman, 1999). However, the first negative peak is absent for short VOTs and, unlike categorical perception, peaks do not vary according to the place of consonant production (Sharma, et al., 2000). Electrophysiological measurements of VOT have examined sibilant and plosive onsets (Tremblay, et al., 2003) and also the addition of noise. When noise was added to speech at a SNR of 5 dB, there was a change in N1 latency that corresponded to the VOT difference, and the generator site of N1 was in the right lobe in both syllable initial and post-vowel VOT contexts (Dimitrijevic, et al., 2013). Component metrics associated with N1 elicited with short VOT syllables have also been linked to sentence perception in noise (Billings, et al., 2013), indicating that cortical measures elicited by stimuli that include voicing onset may be related to broader speech perceptual measures.

The purpose of the present study was to investigate whether the chosen speech contrasts are differentially affected by the attention of a listener, and if so, whether the effect of attention on electrophysiological measures of syllable perception and processing is linked to behavioral perception of the contrast in noise. A related purpose was to study the accuracy of syllable identification as represented in the neural-electrical time series and also to examine the net effect of attention on listening. To explore these issues, we probed the automaticity of syllable processing by making EEG recordings in conditions where subjects both attended to the stimuli and also where their attention was diverted to a visual discrimination task. Our analysis is particularly concerned with the relationship between the adjuvant effect of attention on the electrophysiology and behavioral performance differences when the same subjects performed syllable identification in no noise and noise backgrounds. Such a relationship is of potential interest, as it would imply that the neuronal mechanisms involved in the gain theory of attention perform a function that is similar to noise reduction.

2. Material and methods

2.1 Subjects

Twenty-four students and staff (13 female; mean age 25 yrs, SD 7) from the University of Copenhagen participated in this experiment. All subjects reported right-hand dominance and no existing neurological conditions. They all had normal or corrected normal vision and also normal hearing as revealed by audiometric screening (puretone thresholds less than 25 dB HL at 250–4000 Hz, in both ears). All subjects were native Danish speakers and none of them had any prior knowledge of Japanese Kanji orthography, which was used in the visual discrimination task. Informed consent was provided prior to the experiment and subjects received a bottle of wine for their participation. The study was conducted in accordance with the Danish Code of Conduct for Research Integrity and was assigned the protocol number H-4-2014-FSP by the Scientific Ethical Committee for the Central Region, Denmark.

2.2 Stimuli

Stimuli for the behavioral and EEG tests were made from the syllables [pʰɑː] and [kʰɑː] recorded by a 42 yo male. The fundamental voice-pitch frequency of these exemplars was flattened to 105 Hz with the PSOLA algorithm implemented in PRAAT (Broersma and Weenink, 2018). Aspiration and the voiceless phase was removed from the exemplars to yield the syllables [b̥ɑː] and [g̊ɑː]. The resulting VOTs were above the values mentioned in (Sharma, et al., 2000) that yielded consistent behavioral categorization, and were: [pʰɑː] 79 ms; [b̥ɑː] 13 ms; [kʰɑː] 79 ms; and, [g̊ɑː] 18 ms. Because editing of the voiceless phase was performed at zero-crossings after the stop, the VOT of [g̊ɑː] could not be shortened to be less than 18 ms. The vowel of these items was then truncated by removing portions to provide short vowel tokens which were 120 ms and long vowel tokens which were 200 ms, which is consistent with phonological descriptions of contrastive vowel length in Danish. Finally, linear gating was applied to the last 50 ms of all items. The amplitude waveforms of all stimuli are given in figure 1 and can be heard in the online material that is supplemental to this article.

[Insert Fig. 1 approximately here]

The 8 stimuli differed from one another according to the contrastive features of VOT, place of articulation and vowel length. For instance, [pʰa] differed from [b̥a] by VOT; [pʰa] differed from [g̊a] by VOT and place of articulation; and, [pʰa] differed from [g̊aː] by VOT, place of articulation and vowel length. First and second authors, who were both privy to stimulus modifications, listened to the stimuli and deemed them to be representative of the desired feature values in Danish. Table 1 shows the feature attributes of all syllables.

[Insert Table 1 approximately here]

2.3 Addition of Noise

Syllable identification was performed in no noise and 3 noise backgrounds, where unmodulated speech-spectrum shaped random noise from the International Collegium of Rehabilitative Audiology collection (Dreschler et al, 2001) was combined with the syllables. The spectral properties of the selected noise were based on the speech of a male speaker speaking with normal vocal effort. This noise was added to the syllables at 0, 4 and 8 dB SNR measured relative to the long-term root-mean square (RMS) levels. These SNRs were chosen on the basis of results from Studebaker, et al., (1999) that show a sharp decrease in open-set monosyllable identification between 8 dB SNR (58 rau or 65%) and 3 dB SNR (40 rau or 31%). All auditory stimuli were presented in a soundfield that was calibrated so that stimuli were 65 dB (A-weighting) when subjects were seated 1m in front of the loudspeaker.

2.4 Testing

Subjects performed closed-set syllable identification, then EEG testing in the attend followed by the divert conditions, in the course of a test session that took approximately 1 hr 30 mins. Table 2 gives the ordered specifics of testing, all of which was carried out in an electrically shielded and sound-treated room. Prior to behavioral syllable identification subjects completed a training block consisting of 32 items, during which visual correct/incorrect feedback was given. The purpose of the training was to allow participants to gain familiarity with the layout of the response alternatives on the labelled number pad of the computer keyboard on which they responded. After the completion of training, subjects performed syllable identification in no noise, 0, 4 and 8 dB SNR backgrounds. The ordering of the stimuli in no noise and noise backgrounds was randomized in the behavioral block, as was the order of the syllables in both EEG conditions.

[insert table 2 about here]

EEG was recorded in an attend condition where subjects were instructed to identify the syllable presented by responding on the same number pad that was used during behavioral testing. In the attend condition, a black dot was presented on the screen for half a second after the auditory stimulus, and subjects were instructed to respond as soon as

the black dot disappeared from the screen. This was done in an attempt to dissociate syllable-evoked activity from later and larger components that reflect discriminatory and response processes. In the divert condition, no postsimulus black dot was shown on the screen and subjects performed a visual discrimination task simultaneous to the presentation of syllabic stimuli. In this task subjects were instructed to ignore the auditory stimuli and identify a deviant Kanji symbol from a row of three symbols, where two were the same and one was different, thus requiring subjects to closely consider the spatial detail of the symbols on a trial-by-trial basis. Previous experience with this task has shown that it yields reductions in ERP component amplitudes that are greater than those recorded using passive distraction paradigms, like watching a film (Morris, et al., 2016), and response times that index cognitive load when thematic content differs in speech production (Iwarsson, et al., 2016). Up to 1080 trials were presented and, in order to minimize the possibility of rhythmic synchronization between responses to the visual task and the auditory stimuli, the interstimulus interval (ISI) was varied after every 180 presentations from 100, 250 to 500 ms, and this was repeated. Subjects were given a 5 min break between EEG conditions.

2.5 EEG Recording, processing and microstate analysis

Sintered Ag-AgCl electrodes were positioned at 18 scalp locations, which were O1/z/2, P3/z/4, C3/z/4, F3/z/4, Fp1/2, T7/8, and M1/2, according to the extended international 10-20 system. In addition, electrooculogram signals were recorded with four electrodes, one at the outer canthus of each eye and below and above the right eye.

Continuous EEG data was acquired at a sampling rate of 2048 Hz and subsequently downsampled to 256 Hz. These were bandpass filtered (zero-phase) between 0.1-30 Hz and referenced to both mastoid electrodes. The traces were visually inspected to remove artifacts that were coherent across electrodes. Data from 3 male subjects and 1 female subject were excluded due to excessive noise, and data from one subject who completed all 1080 trials in the divert condition was discarded from the point where they stopped performing the visual discrimination. Independent Component Analysis was then performed on the data from the remaining 20 participants using the infomax algorithm implemented in EEGlab (Delorme and Makeig, 2004), after which a mean of 1.9 (SD=1) components characteristic of either eyeblink or cardiac origin were removed. The oculogram channels were then deleted and epochs of -200 to 400 ms, relative to stimulus onset, were extracted for each syllable and baselined to the prestimulus data. Epochs that exceeded +/-50 μV were removed from further analysis.

Microstate analysis was carried out with the 'Microstates in EEGlab' plugin (Koenig, 2017). Before submitting the EEG to this analysis, the data underwent additional bandpass filtering between 2-20 Hz. Four microstates were then fitted to the entire data from all subjects for each stimulus in each attention condition with atomize-agglomerate hierarchical clustering. This clustering differs from similarity-based methods, like k-means, as it iteratively redistributes members of clusters that substantially detract from the internal correlation of the cluster, to any of the other predominant clusters. Clustering was based on 1000 maps and microstates that were likely to have been truncated were removed. The fitted microstates were then sorted according to the conventional A, B, C and D topographies (see Michel and Koenig, 2017). In distinguishing between microstate class C and D, the mean spread of topographical activation was observed, so that if it was constrained within the ventral hemisphere of the scalp plot, it was classified as type C, and if there was dorsal spread, it was classified as type D. The means of both attention conditions were then calculated and combined to sort the individual data.

2.6 Statistics

All statistical tests were done in R (R Core Team, 2005), and Pearson's product-moment correlations were adjusted for multiple comparisons with the False Discovery Rate (FDR) correction implemented in the Hmisc package (Harell, 2014).

3. Results

This study compared behavioral performance on closed-set syllable identification in different noise backgrounds with EEG recorded while subjects attended to the stimuli and also while their attention was diverted to a visual discrimination task. We report an analysis that focuses on the speech contrasts, how they differ in attention conditions, correlations between EEG and behavioral measures, and the net effect of attention on listening.

3.1 Syllable identification

Mean syllable identification scores were 75.1% in no noise; 62.5% at 8 dB SNR; 52.5% at 4 dB SNR; and, 45.1% at 0 dB SNR. Confusion matrices from all noise backgrounds are given in the appendix. It can be seen that the [b̥ɑ(:)] identification rates are above chance level (12.5%) in the no noise background but are considerably worse than [pʰɑ(:)] identification which is above 90%. The confusion of the syllables [b̥ɑ(:)] with [pʰɑ(:)], may be due to the method that we used in creating short VOT stimuli from naturally-spoken long VOT exemplars. Also, in the noise backgrounds it

can be seen that, the short vowel [b̥ɑ] was more commonly identified as [pʰɑ] than the long vowel [b̥ɑː] with [pʰɑː], which, although it is generally accepted that in Danish there is no mutual complementary relationship between vocalic and consonantal quantity, may reflect that vowels are often shorter after aspirated than unaspirated stops.

Sequential information transfer analysis (SINFA, Wang and Bilger, 1973) was applied to the pooled data and to individual results. SINFA is a data reduction technique that convolves the joint performance (given in the appendix) with the stimuli values (given in table 1) so as to give the isolated transmitted information associated with predefined features. These have been normalized to the total information values from all features and are given in figure 2. Mean transmitted information from the syllable identification between no noise and 0 dB SNR decreased markedly for place of articulation (from 0.99 in to 0.1); moderately for vowel length (from 0.63 to 0.4); and marginally for VOT (from 0.58 to 0.44).

[Insert Fig. 2 approximately here]

3.2 Visual discrimination

Subjects completed a mean of 951 symbol discrimination trials (range 780-1080). Mean reaction time was 777 ms (SD 309) and accuracy was 95% (SD 21). These results indicate that the visual discrimination task was performed adequately so that it occupied the attention of subjects, and we could observe the desired reduction in electrical response to the auditory stimuli.

3.3 Electrophysiology - Vertex response and GFP

Figure 3 shows the group vertex (Cz) and the Global Field Power (GFP) averages for each syllable in both attention conditions. The mean difference in peak vertex amplitudes between attend and divert conditions for all stimuli, which reflects the adjuvant effect of attention, was -3.57 μV (SD=1.02) for N1 and 2.34 μV (SD=1.5) for P2. N1 was calculated as the minimum, and P2 as the maximum in the 100-200 ms and 200-325 ms postimulus windows, respectively. The poststimulus window for calculating P2 extended to 325 ms so as to include peaks from the long VOT and short vowel stimuli [pʰɑ] and [b̥ɑ] in the divert condition, as the group data showed that these occurred at longer poststimulus latencies.

The vertex (Cz) response revealed some systematic stimulus-related variation, e.g., N1 amplitude is larger for syllables with short VOTs, which were [ba(:)] and [ga(:)]. To investigate this, we performed separate ANOVAs on the N1 and P2 amplitude data from all epochs that remained after artefact rejection, with the factors: attention condition (attend and divert); VOT (long and short); place of articulation (bilabial and velar); and, vowel length (long and short). N1 amplitude was significantly effected by attention ($F_{(1,319)}$=128.97, p<0.001) and VOT ($F_{(1,319)}$=20.51, p<0.001) but not place of articulation ($F_{(1,319)}$=0.52, p=0.46) or vowel length ($F_{(1,319)}$=0.2, p=0.65). P2 amplitude was significantly effected by attention ($F_{(1,319)}$=34.7, p<0.001) and vowel length ($F_{(1,319)}$=3.88, p=0.0.5) but not place of articulation ($F_{(1,319)}$=0.65, p=0.79) or VOT ($F_{(1,319)}$=1.05, p=0.3). There were no significant interactions between factors in either of the analyses.

[Insert Fig. 3 approximately here]

3.4 Electrophysiology - contrast-attention differences

To assess the electrophysiological responses to contrastive features of the stimuli we subtracted the GFP time series in the attend and divert conditions for each contrast. This was based on the feature values and was performed according to the following formulas:

$$\text{Length}_{\text{ATTEND}\,|\,\text{DIVERT}} = \frac{\sum([\text{p}^{\text{h}}\text{a:}]+[\text{b̥a:}]+[\text{k}^{\text{h}}\text{a:}]+[\text{g̊a:}])_i}{N} - \frac{\sum([\text{p}^{\text{h}}\text{a}]+[\text{b̥a}]+[\text{k}^{\text{h}}\text{a}]+[\text{g̊a}])_i}{N} \qquad (1)$$

$$\text{Place}_{\text{ATTEND}\,|\,\text{DIVERT}} = \frac{\sum([\text{p}^{\text{h}}\text{a}]+[\text{b̥a}]+[\text{p}^{\text{h}}\text{a:}]+[\text{b̥a:}])_i}{N} - \frac{\sum([\text{k}^{\text{h}}\text{a}]+[\text{g̊a}]+[\text{k}^{\text{h}}\text{a:}]+[\text{g̊a:}])_i}{N} \qquad (2)$$

$$\text{VOT}_{\text{ATTEND}\,|\,\text{DIVERT}} = \frac{\sum([\text{p}^{\text{h}}\text{a}]+[\text{k}^{\text{h}}\text{a}]+[\text{p}^{\text{h}}\text{a:}]+[\text{k}^{\text{h}}\text{a:}])_i}{N} - \frac{\sum([\text{b̥a}]+[\text{g̊a}]+[\text{b̥a:}]+[\text{g̊a:}])_i}{N} \qquad (3)$$

The resulting values from the divert condition were then subtracted from the corresponding attend condition and are given in figure 4. As behavioral syllable identification in no noise showed considerable errors, particularly for the short-VOT stimuli [b̥ɑ(:)] and [g̊ɑ(:)] (see appendix), we excluded all epochs from the attend condition where syllable identification was incorrect or absent. This removed approximately one third of the data and 10420 epochs remained, for which we used formulas 1-3 to calculate contrast-attention differences corrected for accuracy (see figure 4). The VOT contrast shows the earliest and largest amplitude difference which is 3μV for both corrected and all trials and has an initial peak at 117 ms. Differentiation that is >0.5 μV between the correct and all differences, is confined between

180 and 280 ms for VOT. The two peaks included in this time window are at 183 and 273 ms with all attend trials and at 191 and 269 ms when corrected for accuracy. For place of articulation differentiation between correct and all is between 180 and 310 ms; and, for vowel length it begins at 160 ms, peaks for both correct and all responses at 234 ms and continues to the end of the epoch.

[Insert Fig. 4 approximately here]

3.5 Correlations – syllable identification and electrophysiology

Pearson's product-moment correlations with FDR were used to examine the relationship between electrophysiology and the behavioral data. The electrophysiology data were the RMS of the individual contrast-attention GFPs for each stimulus in the 100-250 ms poststimulus window. This window was chosen as it encompassed considerable variation in the GFP between attention conditions. We also used the difference between the vertex peak amplitudes for N1 and P2 in the attend and divert conditions for each contrast, calculated according to the difference formulas 1-3. The behavioral data was the difference in individual syllable identification between the no noise and the 0 dB SNR backgrounds. This behavioral measure was used because the interquartile ranges of individual transmitted information values from the SINFA analysis were the lowest for this SNR, and it was anticipated that parity between signal and noise would be similar to the diversion of attention from the auditory stimuli. These results are given in table 3 and show correlations between the VOT contrast-attention differences and behavioral data (plotted in figure 5), but no significant correlations for place of articulation or vowel length.

[Insert Table 3 approximately here]

[Insert Fig. 5 approximately here]

3.6 Microstate analysis of listening conditions

EEG microstates represent bundles of temporally overlapping but spatially synchronized rectified topographies (for an overview, see Michel and Koenig, 2017). We used the fitted microstate data to examine the net difference between attend and divert conditions according to occurrence, duration and transition probabilities. Pairwise comparisons

revealed a significant difference only in the duration of microstate class D between the attend and divert conditions ($t_{(300)}$=-3.15, p<0.001, see figure 6), and this difference remained after the exclusion of one subject who had class D durations of less than 40 ms ($t_{(293)}$=-4.72, p<0.001). There were no significant differences in the microstate transition data, i.e., the number of transitions between microstate classes, neither in the original form nor when the transitions were adjusted for the frequency of microstate occurrence.

[Insert Fig. 6 approximately here]

4. Discussion

Electrophysiological results from this study show that the speech contrasts tested were differentially affected by whether subjects attended to the stimuli or whether they diverted their attention to a competing visual discrimination task. The largest difference in surface-recorded cortical responses between attention conditions was elicited with the VOT contrast. Information transmitted values showed that VOT was the contrast that was most impervious to the addition of noise, and differences between derived measures from N1, P2 and GFP correlated with the decrement in syllable identification between no noise and 0 dB SNR. There was little change for place of articulation between the attention conditions in the electrophysiological results, and behavioral results showed what Miller and Nicely (1955) first described many decades ago; that transmitted information according to place of articulation decreases considerably in noise. Electrophysiological and behavioral results from contrastive vowel length were found to be between the other two contrasts, as transmitted information declined in increasing noise backgrounds, and the difference in the GFPs between the two attention conditions showed that this contrast is probably encoded as an offset response.

The effect of VOT on the vertex (Cz) N1 and the multisensor GFP shows that the electrophysiological indices of VOT perception vary according to attention condition. The VOT contrast-attention GFP difference varied by more than 3 μV between 117 and 180 ms poststimulus, which is comparable to the time window in which other ERP investigations of VOT continua report distinct stimulus-driven changes (Sharma and Dorman, 1999). This could lead to speculation as to what acoustic features of the stimulus drive the preferential and early allocation of attentional resources to VOT, but not the other contrasts that we investigated? One possibility is that the temporal distribution of elements within VOT features prime attended phonological processing to monitor the time period between the release and the onset of glottal vibration. In this explanation the consonant burst is an initial attentional marker that signals the beginning of a period

that will subsequently reveal the VOT category of the syllable onset when voicing occurs. Such a take on VOT processing is congruent with accounts of speech perception that place importance on early processing that encodes serial elements with high temporal fidelity (Kotz and Schwartze, 2010). However, this explanation does not account for negative VOTs, i.e., voiced stops, where voicing leads with the burst, or subtle VOT characteristics encountered in other languages (for a discussion of these, see Horev, et al., 2007).

The contrast-attention differences (figure 4) show that there is poststimulus differentiation >0.5 μV between all trials and those after the attend condition was corrected for accuracy, which for VOT was 180-280 ms; for place of articulation was 180-310 ms; and, for vowel length, begins at 160 ms and continues to the end of the epoch. As this differentiation is due only to the accuracy of the attended stimuli it is likely that it reflects some form of postperceptual decision processing, and it is therefore of interest to consider these temporal windows in relation to when the contrastive features of the stimuli become perceptually available, which was at 18 ms after stimulus onset for place of articulation; 18-79 ms for VOT; and 120-260 ms for vowel length. In comparing these poststimulus time windows to the differentiation observed after correction for accuracy, place of articulation information is available after the initial stop. For this contrast, differentiation between accurate and all responses starts at the same time as that of the VOT but it continues for 30 ms longer, which may indicate that the repair processing that resolves categorization of this contrast is more demanding. The difference between the two peaks at 191 and 269 ms, in the VOT contrast-attention difference for the correct attended trials, corresponds to the duration of the unvoiced phase of the long VOT stimuli, indicating that there may be a linear temporal relationship between the stimuli and the electrical timeseries recorded during syllable perception. It is also relevant to note that this temporal congruence between stimuli and EEG response, which plausibly occurs due to the previously documented ‘double-on’ response (Steinschneider, et al., 1994), is most apparent after correction for accuracy in the attend condition, a condition that was not investigated in previous studies, including Sharma and Dorman (1999) and Sharma et al. (2000).

We report negligible GFP contrast-attention differences for place of articulation and no correlation between these differences and behavioral performance. Explanations for this may be that the stop consonant that signals place of articulation is processed as a constituent part of VOT. It also may be that the second formant transitions that cue bilabial and velar places of articulation are processed by preattentative perceptual machinery that are obligatory and resistant to the effect of attention. A related explanation is that neuronal coding of the contrast is outside the

measurement limits of ERPs, including the repetitive nature of stimulus presentation and the poor sensitivity to high spectro-temporal detail.

One of the aims of this study was to examine whether the effect of noise was comparable to the effect of listening attention. In so doing, we combined the stimuli with speech-spectrum shaped noise that we believed would adequately interfere with the stimuli, and thus render a noisy neural response. This study may have benefitted from using another noise type, for instance one that mimicked neural noise, yet, this would also have been transformed through the efferent auditory system. In terms of procedure, it may have been beneficial to counterbalance the order of the behavioral and both EEG conditions across test subjects. This is due to previous work that has shown that repeated exposure to syllabic stimuli in the course of serial testing, particularly when paired with a subsequent identification task, results in enhanced neural-electrical activity within a range of poststimulus latencies around those in which P2 is observed (Tremblay, et al., 2010). However, the chosen order of the present study was deemed procedurally preferable for subjects as they could use their familiarity with the response alternatives in consecutive behavioral identification and attend conditions. We also took a structuralist approach to deriving the syllabic stimuli where minimal modifications were made to naturally spoken exemplars. This was done in order to yield isolated contrasts that would facilitate orthogonal treatment of the electrophysiological results. The present study shows that caution must be exerted when editing speech sounds for use in electrophysiological and perceptual studies. Specifically, the method that we employed of creating short VOT stimuli by removing the voiceless phase from long VOT exemplars may have meant that burst characteristics and, probably more importantly, formant transitions associated with the long alternative, biased syllable identification. Splicing the bursts from long VOT stimuli with short VOT bursts, may circumvent this issue, but in the present study one could infer that it could introduce further bias whereby short VOTs would be identified as long VOTs. This methodological issue is not easily resolved without compromising the orthogonal nature of the contrast.

Microstate analysis was used to examine the net affect of attention on listening, and this showed that the duration of class D differed between the attention conditions. Microstate class D has a fronto-central topographic maxima and is thought to be functionally linked to the dorsal attention network (Michel and Koenig 2017). Simultaneous resting-state MRI and EEG has shown linkages between microstate class D and right-lateralized frontal and parietal cortex (Britz, Van De Ville and Michel 2010), which are brain regions that are involved in controlling the spatial direction of attention (Corbetta and Shulman 2002). Durations of class D have been observed to increase during the fifth decade

(Koenig et al. 2002) and during resting (Milz et al. 2016), and decrease in acute schizophrenia (Lehmann et al. 2005). Along with our results these observations indicate that the durational properties of microstate class D may reflect behavioral activation levels that are modulated by task modality. What remains unclear from our results is whether the class D durational differences that we observed are due to the modality difference involved in the attention conditions (auditory vs. visual) or the task demands (closed-set identification vs. discrimination), or a combination of these. Microstate analysis of EEG data recorded during more conditions, including with competing dichotic signals, may disentangle these effects.

There is a clinical imperative to improve our understanding of brain-electrical responses to contrasts that are important to speech perception, as speech and speech-like stimuli is being used in hearing assessment and may be useful in assessing the efficacy of rehabilitative and amplification strategies (Carter, Dillon, Seymour, Seeto, and Van Dun 2013; Martin, Tremblay and Korczak 2008). Furthermore, data on the salience of contrasts in noise may complement attempts to replicate the distribution of speech sounds in test material, like phonemic balancing (Lehiste and Peterson 1959), as performance with material that is rich in VOT contrasts is likely to be higher than with material where there are few VOT contrasts. The net effect of the two attention conditions as revealed by microstate analysis, may also have implications for situations where attention to an auditory or a visual task must be modulated by the listener, for instance, the vigilance of task or teacher monitoring by students in a classroom.

Acknowledgements

The authors would like to thank Nicolai Pharao for his help in recording the stimuli.

Declaration of Interests

There are no conflicts of interest and the authors have not received any specific grant that could have influenced the outcome of this work.

References

Billings,C.J.,McMillan,G.P.,Penman,T.M.,Gille,S.M., 2013. Predicting perception in noise using cortical auditory evoked potentials. J Assoc Res Otolaryngol 14, 891-903.

Britz,J.,Van De Ville,D., Michel,C.M., 2010. BOLD correlates of EEG topography reveal rapid resting-state network dynamics. NeuroImage 52, 1162–1170.

Boersma,P.,Weenink,D., 2018. Praat: doing phonetics by computer [Computer program]. Version 6.0.40 retrieved from http://www.praat.org/

Carter,L.,Dillon,H.,Seymour,J.,Seeto,M.,Van Dun,B., 2013. Cortical auditory-evoked Ppotentials (CAEPs) in adults in response to filtered speech stimuli. J Am Acad Audiol 24, 807-22.

Corbetta,M.,Shulman, G. L., 2002. Control of goal-directed and stimulus-driven attention in the brain. Nat Rev Neurosci 3, 201–215.

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.

Dimitrijevic,A.,Pratt,H.,Starr,A., 2013. Auditory cortical activity in normal hearing subjects to consonant vowels presented in quiet and in noise. Clin Neurophysiol 124, 1204-1215.

Dreschler,W.A.,Verschuure,H.,Ludvigsen,C.,Westermann,S. 2001. ICRA Noises: Artificial noise signals with speech-like spectral and temporal properties for hearing aid assessment. Audiol 40, 148-157.

Harell,F. E., 2014. Hmisc: Harrell Miscellaneous, R package version 3.14-0.

Hillyard,S.A.,Hink,R.F.,Schwent,V.L.,Picton,T.W., 1973. Electrical signs of selective attention in the human brain. Science 182, 177–180.

Horev,N.,Most,T.,Pratt, H., 2007., Categorical perception of speech (vot) and analogous non-speech (fot) signals: behavioral and electrophysiological correlates. Ear Hear 28, 111–128.

Iwarsson,J.,Morris,D.J.,Balling,L.W., 2016. Cognitive load in voice therapy carry-over exercises. J Speech Lang Hear Res 60, 1-12.

Johnson,J.A.,Zatorre,R.J., 2006. Neural substrates for dividing and focusing attention between simultaneous auditory and visual events. NeuroImage 31, 1673–1681.

Jusczyk,P.W.,Rosner,B.S.,Reed,M.A.,Kennedy,L.J., 1989. Could temporal order differences underlie 2-month-olds' discrimination of English voicing contrasts? J Acoust Soc Am 85, 1741–1749.

Koenig,T.,Prichep,L.,Lehmann,D.,Sosa,P.V.,Braeker,E.,Kleinlogel,H.,Isenhart,R.,John,E.R., 2002. Millisecond by millisecond, year by year: normative EEG microstates and developmental stages. NeuroImage 16, 41–48.

Koenig,T., 2017. Microstates in EEGLAB. MATLAB Plugin.

Kong,Y-Y,Somarowthu,A.,Ding, N., 2015. Effects of spectral degradation on attentional modulation of cortical auditory responses to continuous speech. J Assoc Res Otolaryngol 16, 783-796.

Kotz,S.A.,Schwartze,M., 2010. Cortical speech processing unplugged: a timely subcortico-cortical framework. Trends Cogn Sci 14, 392-399.

Lehiste,I.,Peterson,G. E., 1959. Linguistic considerations in the study of speech intelligibility. J Acoust Soc Am 31, 280–286.

Lehmann,D.,Faber,P.L.,Galderisi,S.,Herrmann,W.M.,Kinoshita,T.,Koukkou,M.,Mucci,A.,Pascual-Marqui,R.D.,Saito, N.,Wackermann,J.,Winterer,G.,Koenig, T., 2005. EEG microstate duration and syntax in acute, medication-naïve, first-episode schizophrenia: a multi-center study. Psychiatry Res 138, 141–156.

Luo,Y.J.,Wei,J.H., 1999. Cross-modal selective attention to visual and auditory stimuli modulates endogenous ERP components. Brain Res 842, 30–38.

Molfese,D.L.,Molfese,V.J., 1979. Hemisphere and stimulus differences as reflected in the cortical responses of newborn infants to speech stimuli. Dev Psychol 15, 505-511.

Morris,D.J.,Steinmetzger,K.,Tøndering, J., 2016. Auditory event-related responses to diphthongs in different attention conditions. Neurosci Lett 626, 158-163.

Martin,B.A.,Tremblay,K.L.,Korczak,P., 2008. Speech evoked potentials: From the laboratory to the clinic. Ear Hear 29, 285-313.

Michel,C.M.,Koenig,T., 2018. EEG microstates as a tool for studying the temporal dynamics of whole-brain neuronal networks: A review. NeuroImage 180, 577-593.

Miller,G.A.,Nicely,P.E., 1955. An analysis of perceptual confusions among some english consonants. J Acoust Soc Am 27, 338–352.

Milz,P.,Faber,P.L.,Lehmann,D.,Koenig,T.,Kochi,K.,Pascual-Marqui,R. D., 2016. The functional significance of EEG microstates—Associations with modalities of thinking. NeuroImage 125, 643–656.

R Core team., 2005. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Sharma,A.,Dorman,M.F., 1999. Cortical auditory evoked potential correlates of categorical perception of voice-onset time. J Acoust Soc Am 106, 1078-1083.

Sharma,A.,Marsh,C.M.,Dorman,M.F., 2000. Relationship between N1 evoked potential morphology and the perception of voicing. J Acoust Soc Am 108, 3030-3035.

Steinschneider,M.,Schroeder,C.E.,Arezzo,J.C.,Vaughan,H.G., 1994. Speech-evoked activity in primary auditory cortex: effects of voice onset time. Electroencephalogr Clin Neurophys 92, 30-43.

Steinschneider,M.,Volkov,I.O.,Noh,M.D,Garell,P.C.,Howard,M.A., 1999. Temporal encoding of the voice onset time phonetic parameter by field potentials recorded directly from human auditory cortex. J Neurophysiol 82, 2346-2357.

Studebaker,G.A.,Sherbecoe,R.L.,McDaniel,D.M.,Gwaltney,C.A., 1999. Monosyllabic word recognition at higher-than-normal speech and noise levels. J Acoust Soc Am 105, 2431-2444.

Thornton,A.R.D.,Harmer,M.,Lavoie,B.A., 2007. Selective attention increases the temporal precision of the auditory N100 event-related potential. Hear Res 230, 73–79.

Tremblay,K.L.,Friesen,L.,Martin,B.A.,Wright,R., 2003. Test-retest reliability of cortical evoked potentials using naturally produced speech sounds. Ear Hear 24, 225–232.

Tremblay,K.L.,Inoue,K.McClannahan,K.,Ross,B., 2010. Repeated stimulus exposure alters the way sound is encoded in the human brain. PLOS one 5, 1-11.

Treue,S.,Martínez Trujillo,J.C., 1999. Feature-based attention influences motion processing gain in macaque visual cortex. Nature 399, 575–579.

Wang,M.D.,Bilger,R.C., 1973. Consonant confusions in noise: A study of perceptual features. J Acoust Soc Am 54, 1248-1266.

Figure 1. Stimulus waveforms of syllables used in both behavioral and EEG blocks. A is [b̥ɑ]; B is [b̥ɑː]; C is [pʰɑ]; D is [pʰɑː]; E is [g̊ɑ]; F is [g̊ɑː]; G is [kʰɑ]; and, H is [kʰɑː].

Figure 2. Proportion of transmitted information from the SINFA analysis in the no noise and noise backgrounds. Boxplots show the median (line), 25 and 75 percentiles (box), range (whiskers) and circles (outliers).

Figure 3. Responses recorded at electrode Cz (left) and GFPs (right) in the attend (upper panels) and divert (lower panels) conditions.

Figure 4. GFP contrast-attention differences from all trials and after correction of the attend condition for accuracy, for VOT (left panel), place of articulation (middle panel) and vowel length (right panel).

Figure 5. Correlations between the behavioral difference in the no noise and 0 dB SNR scores and the VOT contrast-attention differences for N1 (left), P2 (middle) and the RMS of the GFP within the 100-250 ms poststimulus window (right).

Figure 6. Mean microstate topographies from the attend (upper panels) and divert (lower panels) conditions. Mean duration of each microstate class in the attention conditions (middle panels). Each non-overlapping point is the mean data from one subjects response to one syllable, group mean (thick line), 25 and 75 percentile (thin lines).

	[b̥ɑ]	[pʰɑ]	[g̊ɑ]	[kʰɑ]	[b̥ɑː]	[pʰɑː]	[g̊ɑː]	[kʰɑː]
place	labial	labial	velar	velar	labial	labial	velar	velar
length	short	short	short	short	long	long	long	long
VOT	short	long	short	long	short	long	short	long

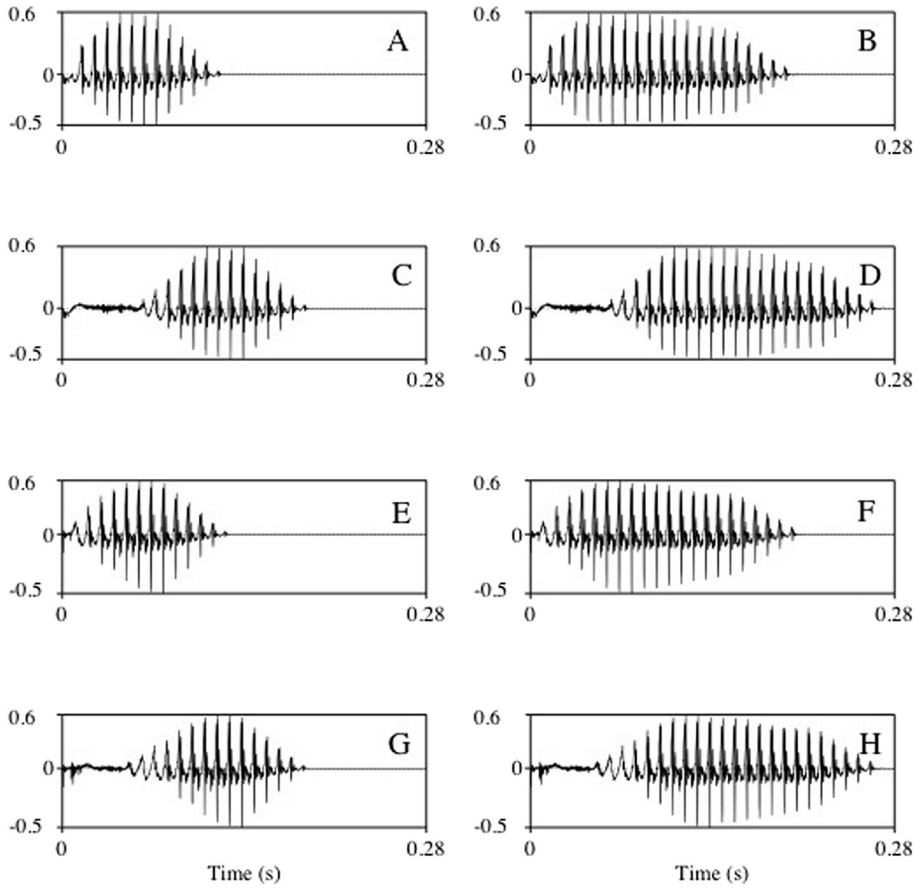
Table 1. Stimuli with feature values

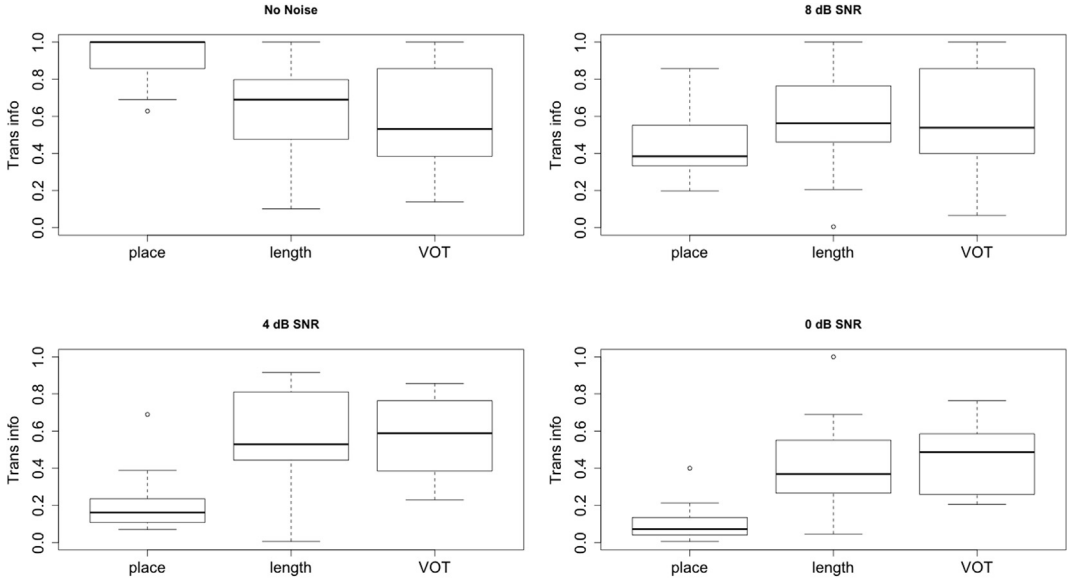
Mode		SNR	No. trials/syllable	ISI (ms)	Task
Behavioral	training	no noise and 6 dB	2	Roved 225-300	Syllable identification
	testing	no noise	10		
		8 dB	10		
		4 dB	10		
		0 dB	10		
EEG	Attend	no noise	100	Roved 900-1100 + fixation dot (500)	Syllable identification
	Divert training	training	32*	100-500*	Visual task
	Divert testing	no noise	100	Roved 900-1100	≤1080 trials-concurrent*

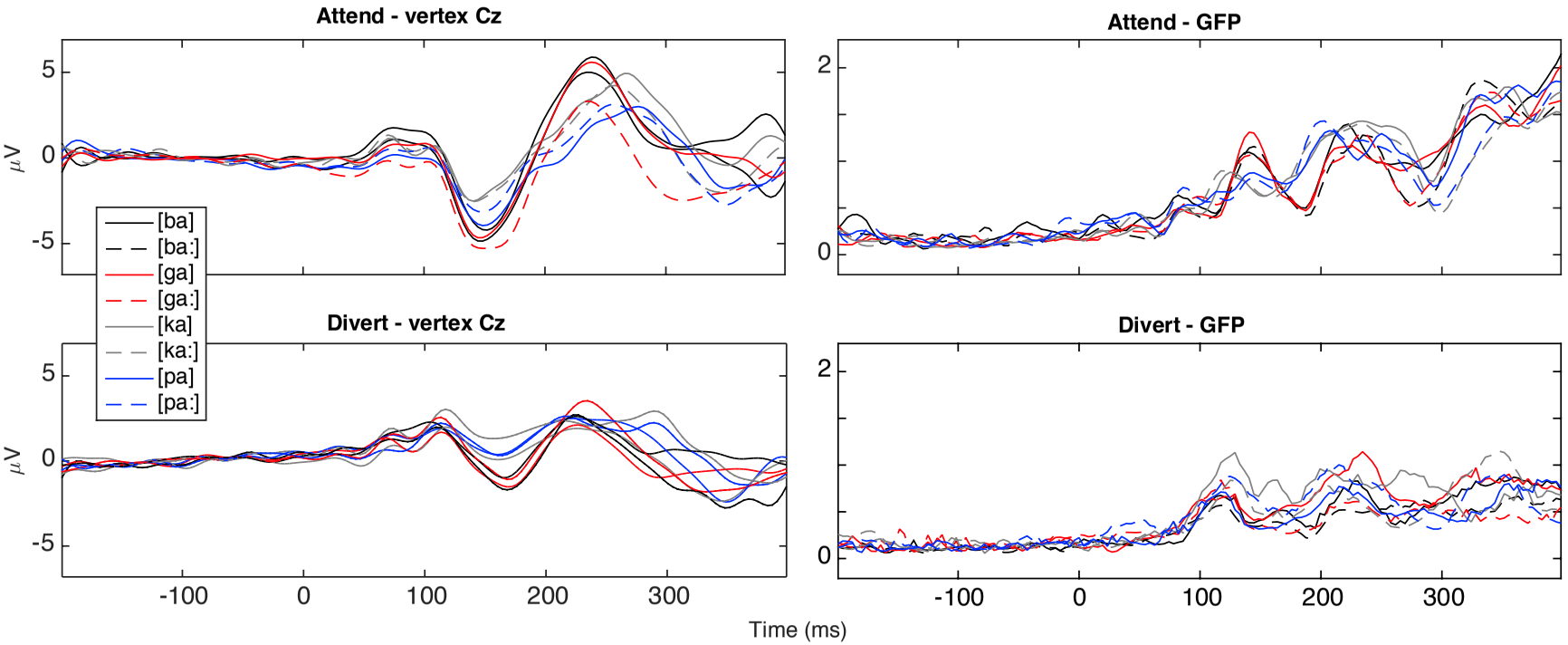
Table 2. Details of the behavioral and EEG testing. *specifics from only the visual discrimination task

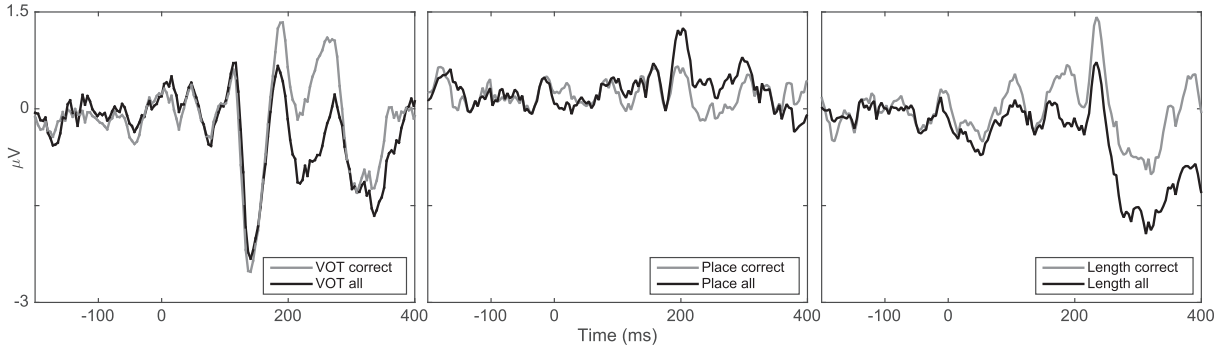
	N1		P2		GFP	
	r	p	r	p	r	p
VOT	0.5	0.03*	0.53	0.02*	0.58	0.01*
Place	0.36	0.18	0.04	0.86	0.17	0.7
Length	-0.19	0.64	-0.18	0.65	-0.01	0.96

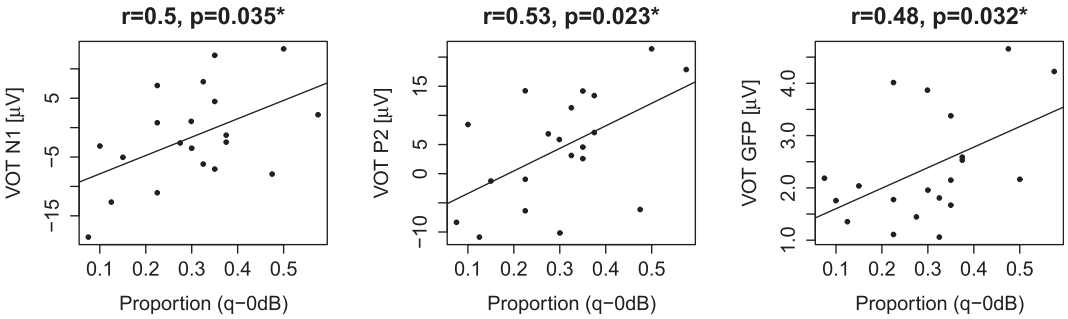
Table 3. Pearson's product-moment correlations, corrected with FDR, between peak amplitude data for N1, P2 and GFP contrast differences, and the proportion correct difference between the no noise and the 0 dB SNR backgrounds.

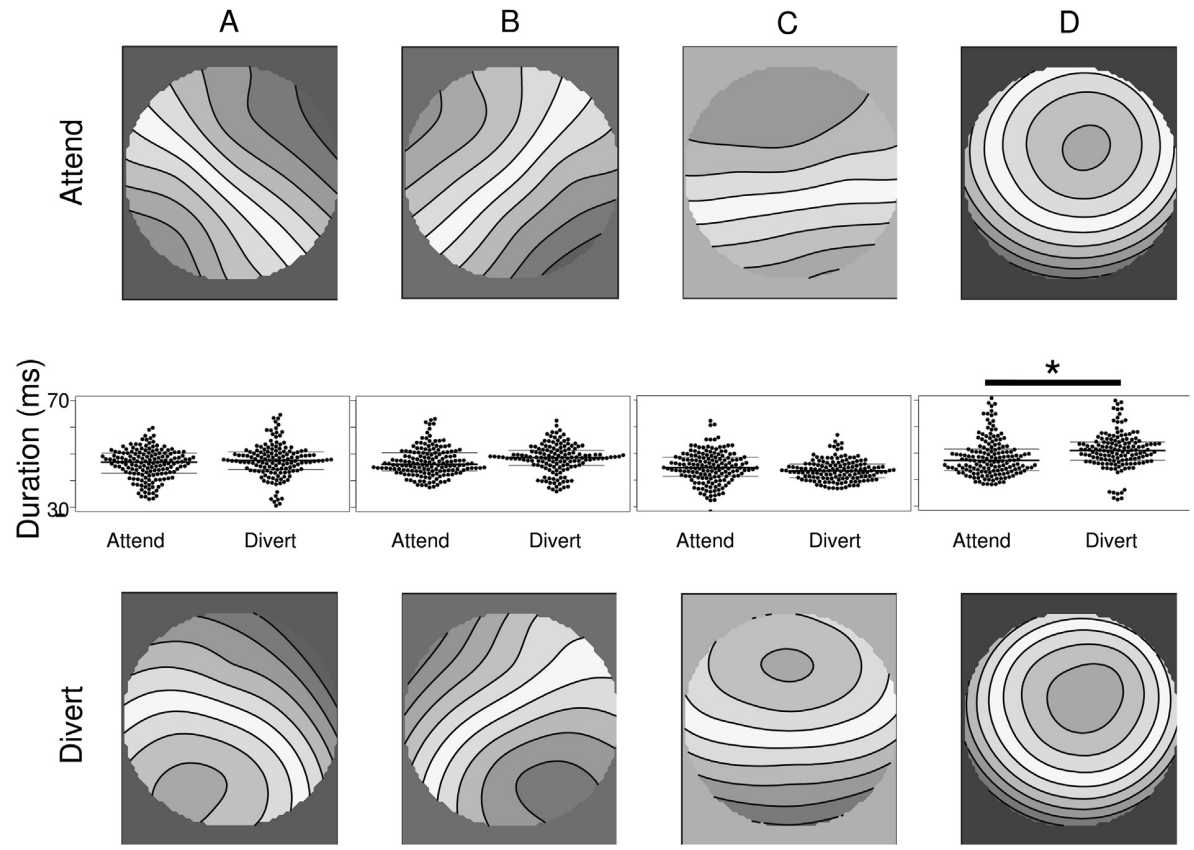

A
B
C
D
E
F
G
H
0.6
0
-0.5
0
0.28
Time (s)


No Noise
8 dB SNR
4 dB SNR
0 dB SNR
Trans info
0.0
0.2
0.4
0.6
0.8
1.0
place
length
VOT


Attend - vertex Cz
Attend - GFP
Divert - vertex Cz
Divert - GFP
[ba]
[ba:]
[ga]
[ga:]
[ka]
[ka:]
[pa]
[pa:]
μV
5
0
-5
2
0
-100
0
100
200
300
Time (ms)


1.5
0
-3
μV
VOT correct
VOT all
Place correct
Place all
Length correct
Length all
-100
0
200
400
Time (ms)


r=0.5, p=0.035*
r=0.53, p=0.023*
r=0.48, p=0.032*
VOT N1 [μV]
VOT P2 [μV]
VOT GFP [μV]
Proportion (q−0dB)


A
B
C
D
Attend
Divert
Duration (ms)
70
30
*
Attend
Divert