9-27-2019

A Novel Audiovisual P300-Speller Paradigm Based on Cross-Modal Spatial and Semantic Congruence

Zhaohua Lu

Changchun University of Science and Technology

Ning Gao

Jingjing Yang

Ou Bai

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 27 September 2019
doi: 10.3389/fnins.2019.01040

A Novel Audiovisual P300-Speller Paradigm Based on Cross-Modal Spatial and Semantic Congruence

Zhaohua Lu[1], Qi Li[1], Ning Gao[1], Jingjing Yang[1] and Ou Bai[2]*

[1] School of Computer Science and Technology, Changchun University of Science and Technology, Changchun, China,
[2] Department of Electrical and Computer Engineering, Florida International University, Miami, FL, United States

Objective: Although many studies have attempted to improve the performance of the visual-based P300-speller system, its performance is still not satisfactory. The current system has limitations for patients with neurodegenerative diseases, in which muscular control of the eyes may be impaired or deteriorate over time. Some studies have shown that the audiovisual stimuli with spatial and semantic congruence elicited larger event-related potential (ERP) amplitudes than do unimodal visual stimuli. Therefore, this study proposed a novel multisensory P300-speller based on audiovisual spatial and semantic congruence.

Methods: We designed a novel audiovisual P300-speller paradigm (AV spelling paradigm) in which the pronunciation and visual presentation of characters were matched in spatial position and semantics. We analyzed the ERP waveforms elicited in the AV spelling paradigm and visual-based spelling paradigm (V spelling paradigm) and compared the classification accuracies between these two paradigms.

Results: ERP analysis revealed significant differences in ERP amplitudes between the two paradigms in the following areas (AV > V): the frontal area at 60–140 ms, frontal–central–parietal area at 360–460 ms, frontal area at 700–800 ms, right temporal area at 380–480 and 700–780 ms, and left temporal area at 500–780 ms. Offline classification results showed that the accuracies were significantly higher in the AV spelling paradigm than in the V spelling paradigm after superposing 1, 2, 5, 6, 9, and 10 times ($P < 0.05$), and there were trends toward improvement in the accuracies at superposing 3, 4, 7, and 8 times ($P = 0.06$). Similar results were found for information transfer rate between V and AV spelling paradigms at 1, 2, 5, 6, and 10 superposition times ($P < 0.05$).

Significance: The proposed audiovisual P300-speller paradigm significantly improved the classification accuracies compared with the visual-based P300-speller paradigm. Our novel paradigm combines spatial and semantic features of two sensory modalities, and the present findings provide valuable insights into the development of multimodal ERP-based BCI paradigms.

Keywords: brain–computer interface, audiovisual, P300-speller, spatial congruence, semantic congruence

OPEN ACCESS

Edited by:
Andrey R. Nikolaev,
KU Leuven, Belgium

Reviewed by:
Hongzhi Qi,
Tianjin University, China
Dong Ming,
Tianjin University, China

****Correspondence:***
Qi Li
liqi@cust.edu.cn

Specialty section:
This article was submitted to Perception Science, a section of the journal Frontiers in Neuroscience

Received: *22 May 2019*
Accepted: *13 September 2019*
Published: *27 September 2019*

Citation:
Lu Z, Li Q, Gao N, Yang J and Bai O (2019) A Novel Audiovisual P300-Speller Paradigm Based on Cross-Modal Spatial and Semantic Congruence. Front. Neurosci. 13:1040. doi: 10.3389/fnins.2019.01040

INTRODUCTION

Brain–computer interfaces (BCIs), which provide a direct method of communication between the brain and external devices (Kubler et al., 2001; Wolpaw et al., 2002), can help severely disabled people, especially patients with amyotrophic lateral sclerosis, to interact with the outside world (Kubler and Neumann, 2005; Nijboer et al., 2008). The P300-speller system is one of the most commonly used BCI applications. P300 refers to an event-related potential (ERP), which occurs around 300 ms after the presentation of a stimulus and is elicited by an oddball event. In the P300-speller system, the user focuses on the desired character (i.e., the target character) and counts the number of times of its intensification; the probability of the target character highlighted each time is small, and this oddball event would elicit a P300 potential. The P300-speller system outputs the target character by detecting the P300 potential; thus, it realizes communication with the outside world by the way of "mental typewriting" (Farwell and Donchin, 1988; Bernat et al., 2001). In the past few decades, studies have investigated many P300-speller systems, including the auditory-based P300-speller (Guo et al., 2010; Halder et al., 2010; Xu et al., 2013), tactile-based p300-speller (Brouwer and Van Erp, 2010; Van der Waal et al., 2012), and visual-based P300-speller (Kaufmann et al., 2011; Jin et al., 2014; Li et al., 2015; Xu et al., 2018). The visual-based P300-speller is the most common P300-speller because its performance is much better than those of the other two types (Belitski et al., 2011; An et al., 2014; Wang et al., 2015; Hammer et al., 2018). However, the visual P300-speller system is still in an exploratory stage because its accuracy and information transfer rates (ITRs) are not satisfactory for practical application. In addition, the visual P300-speller is limited for some patients with neurodegenerative diseases, in which muscular control of the eyes may be impaired or deteriorate with time (Szmidt-Salkowska and Rowinska-Marcinska, 2005; Suminski et al., 2009). Hence, it is necessary to design a multimodal P300-speller based on audiovisual stimuli that is superior to visual-based P300-spellers and can be used more universally.

Some recent studies have investigated the audiovisual P300 BCI systems. Sellers and Donchin (2006) evaluated the effectiveness of a P300-based BCI system involving bimodal audiovisual stimuli and found that its performance was not significantly better than that of the unimodal visual mode system. Belitski et al. (2011) proposed an audiovisual P300-speller that was implemented by combining the row/column number with a spoken number and demonstrated that the effectiveness of the modified audiovisual P300-speller slightly out-performed that of either the visual-based or the auditory-based P300-speller. These findings provide a basis for further exploration of the audiovisual bimodal P300-speller.

The spatial and semantic information of auditory and visual stimuli may affect the integration of the auditory and visual features of these stimuli (Stein and Meredith, 1993). For instance, semantically congruent audiovisual stimuli elicit larger amplitudes between 180 and 210 ms than do audiovisual stimuli without semantic information (Hu et al., 2012) and significantly enhance behavioral performances compared with unimodal visual and auditory stimuli (Laurienti et al., 2004). A functional magnetic resonance imaging (fMRI) study has demonstrated that neural responses are more pronounced for semantically matching audiovisual stimuli than for unimodal visual and auditory stimuli in the multisensory superior temporal sulcus areas (Hein et al., 2007). In addition, ERP amplitudes elicited by multisensory stimuli (i.e., audiovisual stimuli) are larger than the sum of the amplitudes elicited by visual and auditory stimuli at 200–220 ms (central–medial positivity) and 300–450 ms (centro–medial positivity) when the spatial orientation of the auditory and visual stimuli is congruent (Talsma and Woldorff, 2005). Audiovisual stimuli combining spatial and semantic information also influence audiovisual integration. When the auditory sound sources are spatially congruent with the semantically matching visual stimulus, the right middle and superior temporal gyrus areas are more activated, which indicates that spatial congruency appears to enhance the semantic integration of audiovisual stimuli (Plank et al., 2012).

In the present study, we designed a novel audiovisual P300-speller paradigm (AV spelling paradigm) based on spatial and semantic congruence to achieve spelling of multiple characters (i.e., multiple classifications). We compared the performance of the audiovisual P300-speller with that of a unimodal visual P300-speller (V spelling paradigm).

MATERIALS AND METHODS

Subjects

Eighteen subjects (nine males) aged 19–29 years (mean age, 24.8 ± 2.27 years) were recruited in this study. One subject had previously participated in similar BCI studies. These subjects did not have any known neurological disorders and had normal or corrected-to-normal vision and normal hearing. They were undergraduates or master students and were familiar with the alphabet used in this study. They provided written informed consent in accordance with the Declaration of Helsinki, allowed use of the data, and acknowledged their personal rights. All subjects were compensated with 100 RMB after completion of the experiments. The study was approved by the ethics committee of the Changchun University of Science and Technology.

Stimuli and Paradigms

The spatially and semantically congruent audiovisual P300-speller paradigm (AV spelling paradigm) was proposed on the basis of the traditional regional flashing spelling paradigm (Fazel-Rezai et al., 2011) and included two levels: level 1 consisted of several character group-areas and level 2 consisted of single characters from one of the group-areas in level 1, each of which was treated as a sub-area. The stimuli and experimental paradigm designs are shown in **Figure 1**. We divided 36 characters into six group-areas (i.e., level 1), and the six group-areas were arranged in the left and right columns (three in each column). The purpose of this arrangement was to match the pronunciation of the left and right characters with the left and right channels. In order to locate each group-area, we numbered the six group-areas from top to bottom and from left to right, i.e., 1–6 (**Figure 1A**). There

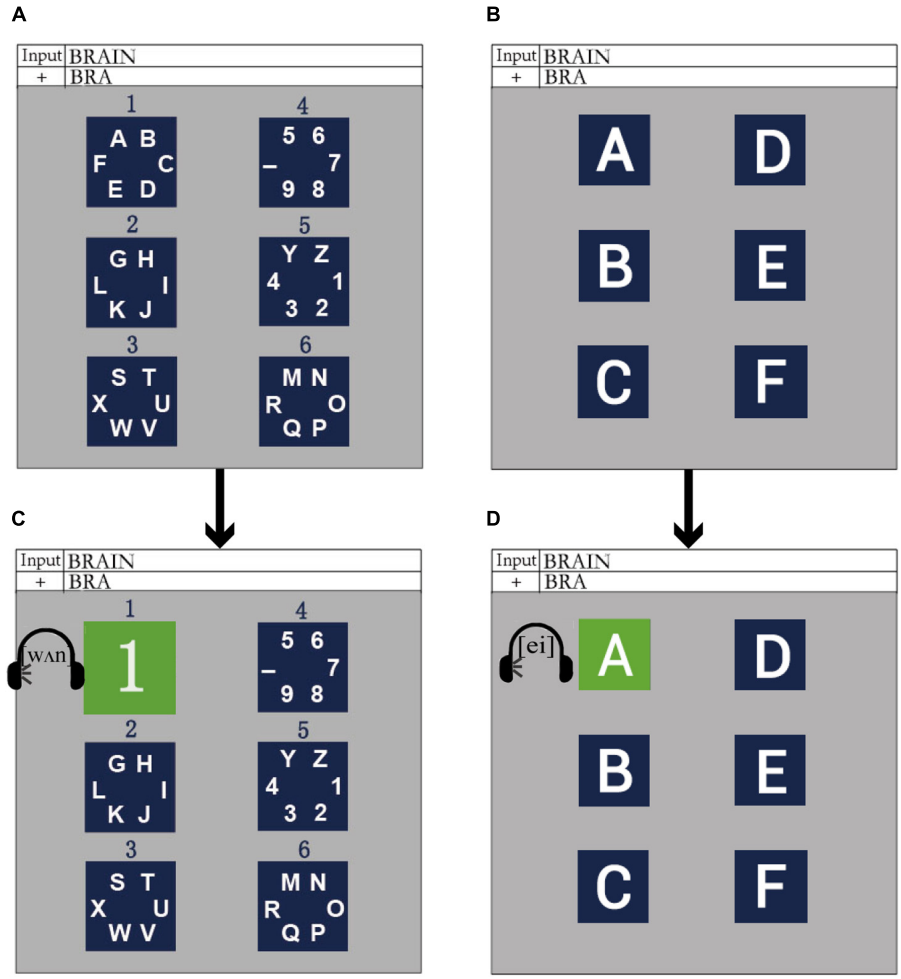

FIGURE 1 | The stimuli and experimental paradigm in the audiovisual P300-speller. **(A)** The layout of level 1 in the AV spelling paradigm; **(B)** an example for the layout of level 2, which corresponded to group-area 1 in level 1 of the AV spelling paradigm; **(C)** an example for a group-area intensified in level 1 of the AV spelling paradigm; **(D)** an example for a sub-area highlighted in level 2 of the AV spelling paradigm.

were six sub-areas in level 2, which corresponded to six characters in a group-area. Similarly, the six sub-areas were arranged in the left and right columns (three in each column). **Figure 1B** shows the layout of level 2, which corresponds to group-area 1 in level 1. To induce obvious ERPs and make the subject feel comfortable, we chose blue as the background color for the group-area and sub-areas (Takano et al., 2009).

The design of the audiovisual P300 spelling paradigm with spatial and semantic congruence was as follows: when a group-area on the left was highlighted (e.g., number 1), it was covered by the corresponding number on a green background (Takano et al., 2009), during which the pronunciation of the corresponding group-area number was played simultaneously in the left earphone with a maximum sound-pressure level of approximately 65 dB (Senkowski et al., 2007; **Figure 1C**); when a group-area on the right was highlighted (e.g., number 4), it was covered with the corresponding number on a green background, and the pronunciation of the corresponding group-area number was played in the right earphone at the same time. Consequently, the spatial and semantic congruence of the group-area was ensured. After a group-area was selected, it transformed to level 2 (i.e., the sub-area), which was the spread of a selected group-area. When a sub-area on the left (or the right) was highlighted (e.g., character "A," **Figure 1D**), the sub-area was covered with the corresponding character on a green background, and the pronunciation of the corresponding character was played in the left (or right) earphone at the same time. Thus, the spatial and semantic congruence of the sub-area was also ensured.

The control paradigm was a unimodal visual P300-speller (V spelling paradigm), in which the presentation of stimuli was the same as in the AV spelling paradigm, except that there was no sound.

Experimental Procedure

The experiment was conducted in a shielded room that was dark and soundproof. After completing the preparation for EEG recording, subjects sat comfortably in front of the monitor,

and their eyes were about 70 cm from the computer monitor. Subjects were familiar with the experimental task prior to commencement, and they were asked to avoid blinking during the stimulus presentation. To specify the target character for a subject's output, the target character with a green background was first presented for 500 ms by audiovisual synchronization (**Figure 2**), and the background was then presented for 500 ms (**Figure 2**). Subsequently, the six group-areas began to flash in a pseudo-random order. The stimulus onset asynchrony (SOA) of the flashing group-area was 250 ms, in which each group-area was highlighted for 180 ms before reverting to the background for 70 ms. The process was referred to as a sub-trial. In a trial, each group-area flashed once (a total of six sub-trials). The trial was repeated 10 times, making up a block (**Figure 2**). After completing a block for a group-area, it reverted to the background for 1 s. The display was then transformed to level 2, and the sub-area flashed in a manner similar to that of the group-area. After the sub-area flashed, it returned to the background for 1 s again before the next target character was presented; the process was then repeated. A sequence consisted of a group-area block and a sub-area block to output a target character (**Figure 2**). The output of a word with five characters was defined as a run. Between each run, subjects were permitted a 5-min break. Each subject took part in five runs (five words) for each spelling paradigm (AV and V spelling paradigms), and a total of 10 runs were presented in a pseudorandom order to avoid learning effects.

Data Acquisition and Processing

Data Acquisition

Electroencephalogram (EEG) signals were recorded with a NeuroScan amplifier (SynAmps 2, NeuroScan Inc., Abbotsford, Australia) from 31 Ag/AgCl scalp electrodes (F7, F3, Fz, F4, F8, FC7, FC3, FCz, FC4, FC8, T7, C3, Cz, C4, T8, TP7, CP3, CPz, CP4, TP8, P7, P3, Pz, P4, P8, PO3, POz, PO4, O1, Oz, and O2; **Figure 3**). The AFz was used as a ground, and the reference electrode was placed on the mastoid of the right ear. Vertical and horizontal eye movements were measured using the VEO and HEO electrodes, respectively. The impedance was maintained below 5 KΩ. All signals were digitized at a rate of 250 Hz. EEG data were digitally filtered with a band-pass filter of 0.01–100 Hz. Presentation of the auditory and visual stimuli was controlled by the E-prime 2.0 software (PST Inc., Savannah, GA, United States).

Data Preprocessing

Original EEG data were first corrected for ocular artifacts using a regression analysis algorithm (Semlitsch et al., 1986) and were digitally filtered using a band-pass filter of 0.01–30 Hz. Data were then divided into epochs from 100 ms before the onset of each stimulus to 800 ms after the onset, and baseline corrections were made against −100–0 ms. Bad stimuli were removed by setting ±80 μV as the threshold for ocular artifacts. ERP data were averaged for each stimuli type (target, non-target stimulus) and used for the ERP waveform analysis. Grand-averaged ERP data were acquired from all subjects for each stimulus type in the two spelling paradigms (AV and V spelling paradigms). The

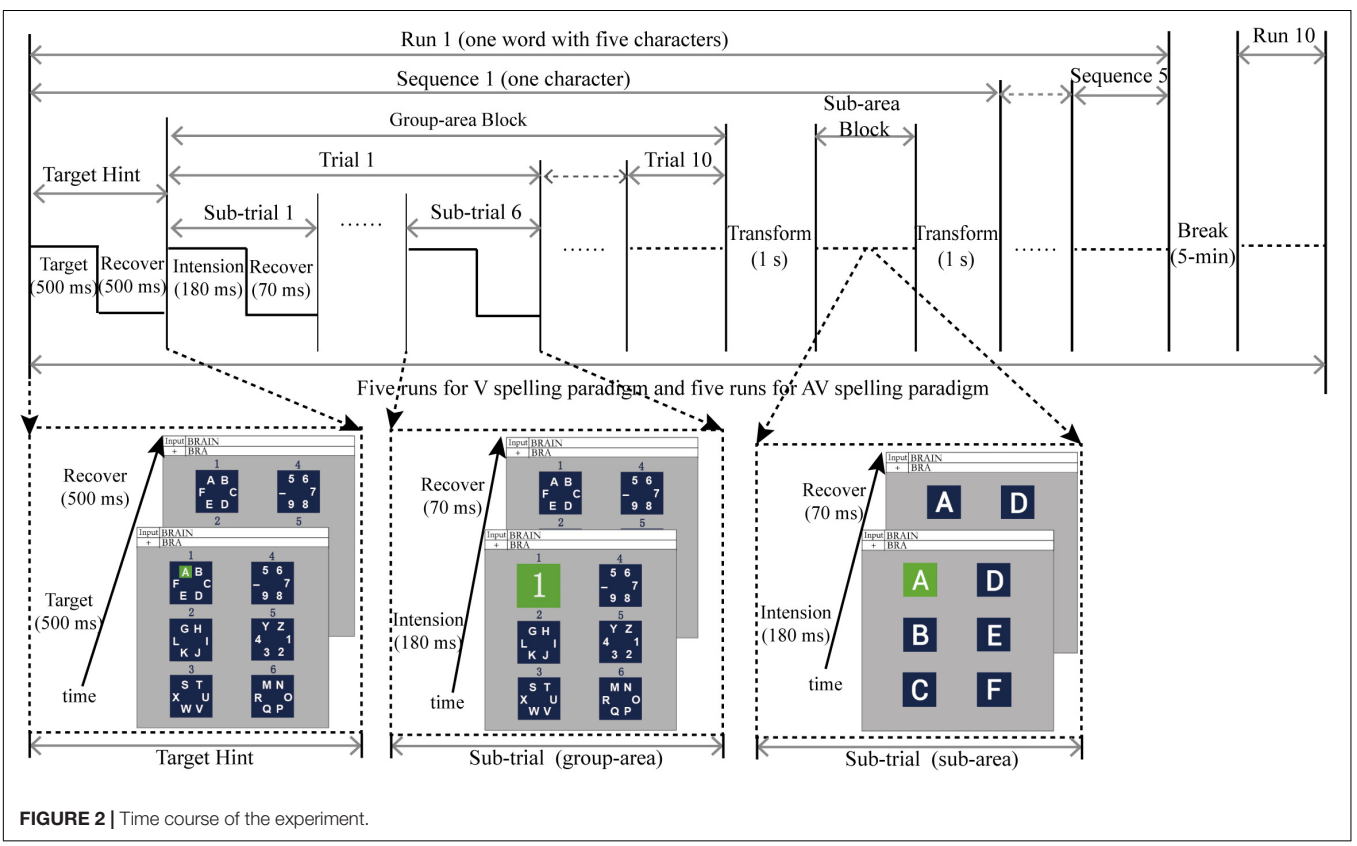

FIGURE 2 | Time course of the experiment.

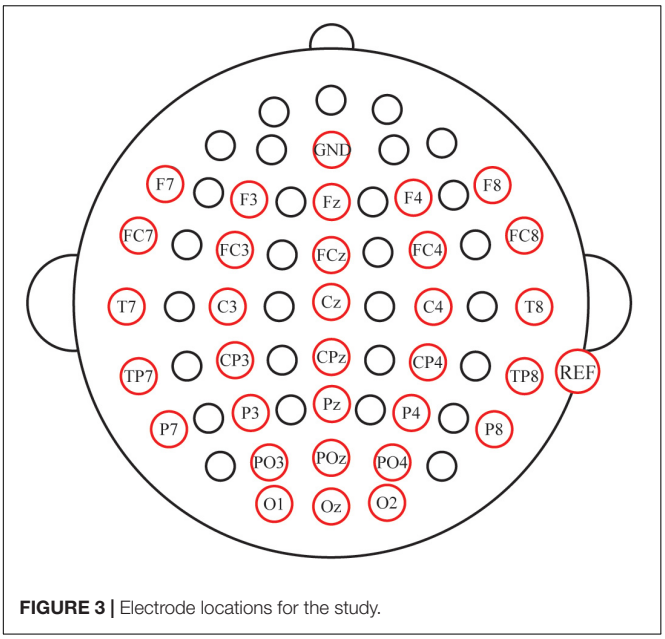

FIGURE 3 | Electrode locations for the study.

pre-processed data, including segmentation, baseline correction, removal of bad stimulus, and filtering, were used for feature extraction and classification.

Feature Extraction and Classification Scheme

For the P300-speller, feature extraction for classification is based on temporal and spatial features of EEG data. For the temporal feature, we selected the time window in which there were obvious ERP amplitudes elicited by target stimuli and those with differences between target and non-target stimuli. Spatial features depended on the electrodes. The r^2-values can provide the mathematic foundation for selecting channels (electrodes) and features of each channel. The r^2 is calculated by formula (1)

$$r^2 = \left(\frac{\sqrt{N_1 N_2}(\text{mean}(x_1) - \text{mean}(x_2))}{(N_1 + N_2)\text{std}(x_1 \bigcup x_2)} \right)^2, \quad (1)$$

where N_1 and N_2 represent the sample sizes of the target and non-target, respectively; x_1 and x_2 are the feature vectors of target and non-target, respectively.

The EEG was then down-sampled from 250 to 50 Hz by selecting every five samples from the epoch. Thus, the size of the feature vector was $C_N \times P_N$ (C_N represents the number of channels, and P_N represents the sample points).

Bayesian linear discriminant analysis (BLDA) was used to classify the EEG data. BLDA is an extension of Fisher's linear discriminant analysis (FLDA) that helps avoid overfitting. The details of the algorithm can be found in a previous report (Hoffmann et al., 2008). We used fivefold cross-validation to calculate the individual accuracy in the offline experiment.

Information Transfer Rate

Information transfer rate is generally used to evaluate the communication performance of a BCI system and is a standard measure that accounts for accuracy, the number of possible selections, and the time required to make each selection (Thompson et al., 2013). ITR (bits min^{-1}) can be calculated as:

$$\text{ITR} = \frac{60\left(P\log_2(P) + (1-P)\log_2\frac{1-P}{N-1} + \log_2 N\right)}{T}, \quad (2)$$

where P denotes the probability of recognizing a character, T is the time taken to recognize a character, and N is the number of classes ($N = 36$).

Data Analysis

Differences in the waveforms between the V and AV spelling paradigms were analyzed using a one-way repeated measures ANOVA with two within-subject factors, i.e., spelling paradigm (V spelling paradigm and AV spelling paradigm) and electrode (the choice of the electrode was based on the difference of ERP waveforms elicited by target stimuli). The latencies of P3a at Fz and P3b at Pz were calculated in the V and AV spelling paradigms, and a pairwise T-test was conducted to analyze the latency difference between them. T-tests were also conducted to compare the accuracies and ITRs at each superposition time in the V and AV spelling paradigms (superposition time represents the time of repeated intensification of the six group-areas/sub-areas, and one superposition time is composed of one trial of group-area and one trial of sub-area). In addition, false discovery rate (FDR) correction was performed for multiple comparisons. Statistical analyses were conducted using SPSS version 19.0 (IBM Corp., Armonk, NY, United States).

RESULTS

ERP Results

Mean ERP waveforms were calculated across all subjects from 31 electrodes in the V and AV spelling paradigms (**Figure 4**). Clear positive deflections in the waveform along with two clear peaks were observed between 200 and 500 ms at F3, Fz, F4, FC3, FCz, FC4, C3, Cz, C4, CP3, CPz, CP4, P3, Pz, P4, PO3, POz, PO4, O1, Oz, and O2, which may be P300 potential. In addition, a clear negative waveform was observed at approximately 200 ms at P7, P3, Pz, P4, P8, PO3, POz, PO4, O1, Oz, and O2, which may be N200 potential.

Scalp topographies were obtained by subtracting the ERP waveforms elicited by the target stimuli in the V spelling paradigm from those elicited in the AV spelling paradigm; time-domain features with statistically significant differences in the waveforms were then analyzed based on these scalp topographies, and the results were corrected by FDR (**Figure 5**). Statistically significant differences between the AV and V spelling paradigms were observed in the waveforms as follows: (1) 60–140 ms at frontal area [$F(1,17) = 10.642$, $P < 0.005$] (**Figure 5A**); (2) 360–460 ms at the frontal–central–parietal areas [$F(1,17) = 11.921$, $P < 0.002$] (**Figure 5B**); (3) 700–780 ms at the right frontal areas [$F(1,17) = 6.031$, $P < 0.05$] (**Figure 5C**); and (4) 340–480 [$F(1,17) = 4.743$, $P < 0.05$] and 720–780 ms [$F(1,17) = 4.021$, $P < 0.05$] at the right temporal areas and 500–780 ms at the left temporal areas [$F(1,17) = 15.16$, $P < 0.001$] (**Figure 5D**).

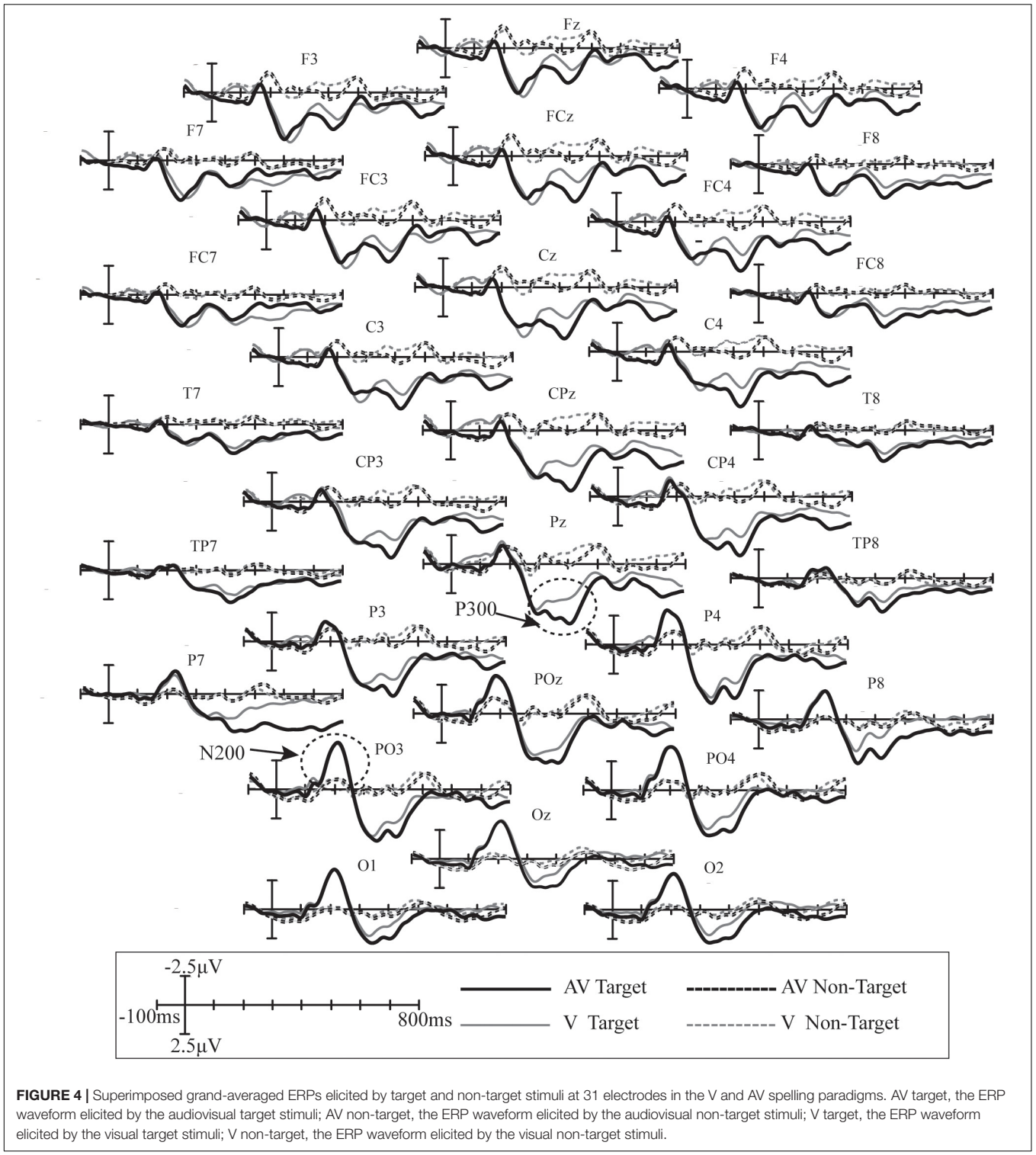

FIGURE 4 | Superimposed grand-averaged ERPs elicited by target and non-target stimuli at 31 electrodes in the V and AV spelling paradigms. AV target, the ERP waveform elicited by the audiovisual target stimuli; AV non-target, the ERP waveform elicited by the audiovisual non-target stimuli; V target, the ERP waveform elicited by the visual target stimuli; V non-target, the ERP waveform elicited by the visual non-target stimuli.

The feature differences between target and non-target stimuli in V and AV spelling paradigms were indicated by the r^2-values (**Figure 6**). As shown in **Figure 6**, the feature differences of ERPs between target and no-target stimuli were mainly found between 200 and 320 ms at F7, F3, Fz, F4, F8, FT7, FC3, FCz, FC4, FT8, C3, Cz, and C4 electrodes and between 300 and 560 ms at CP3, CPz, CP4, P3, Pz, P4, PO3, POz, and PO4 electrodes in V and AV spelling paradigms. In addition, the feature differences of ERPs between target and no-target stimuli at 300-560 ms at CP3, CPz, CP4, P3, Pz, P4, PO3, POz, and PO4 electrodes were larger in the AV spelling paradigm than in the V spelling paradigm. In order to present positive and negative deflections of ERP

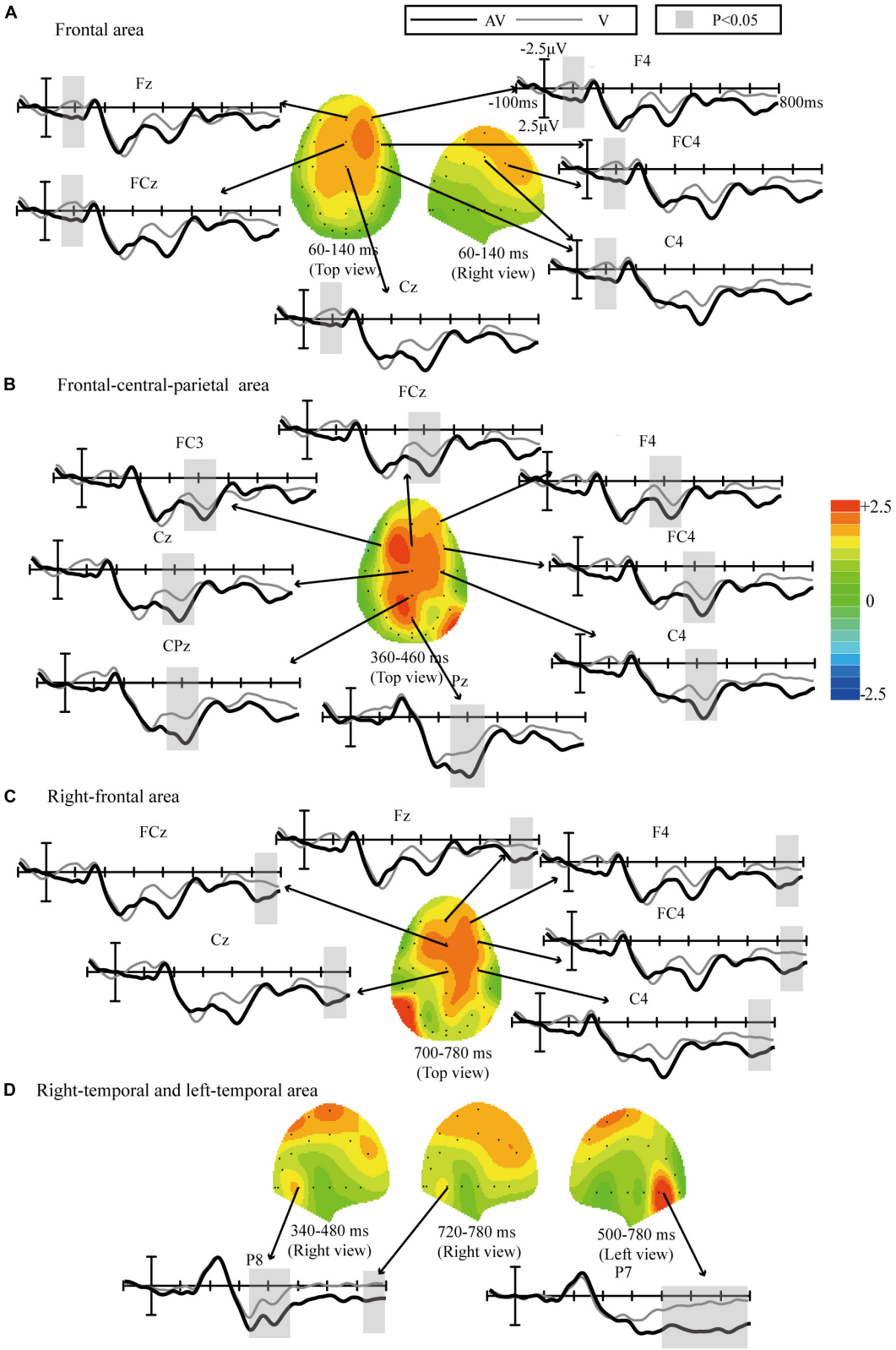

FIGURE 5 | Comparison of waveforms elicited by the target stimuli in the V and AV spelling paradigms and scalp topographies from waveforms with difference formed by subtracting ERPs of the V spelling paradigm from those of the AV spelling paradigm: **(A)** the frontal area at 60–140 ms; **(B)** frontal–central–parietal at 360–460 ms; **(C)** right frontal at 700–780 ms; and **(D)** right temporal at 340–480 and 720–780 ms, and left temporal at 500–780 ms.

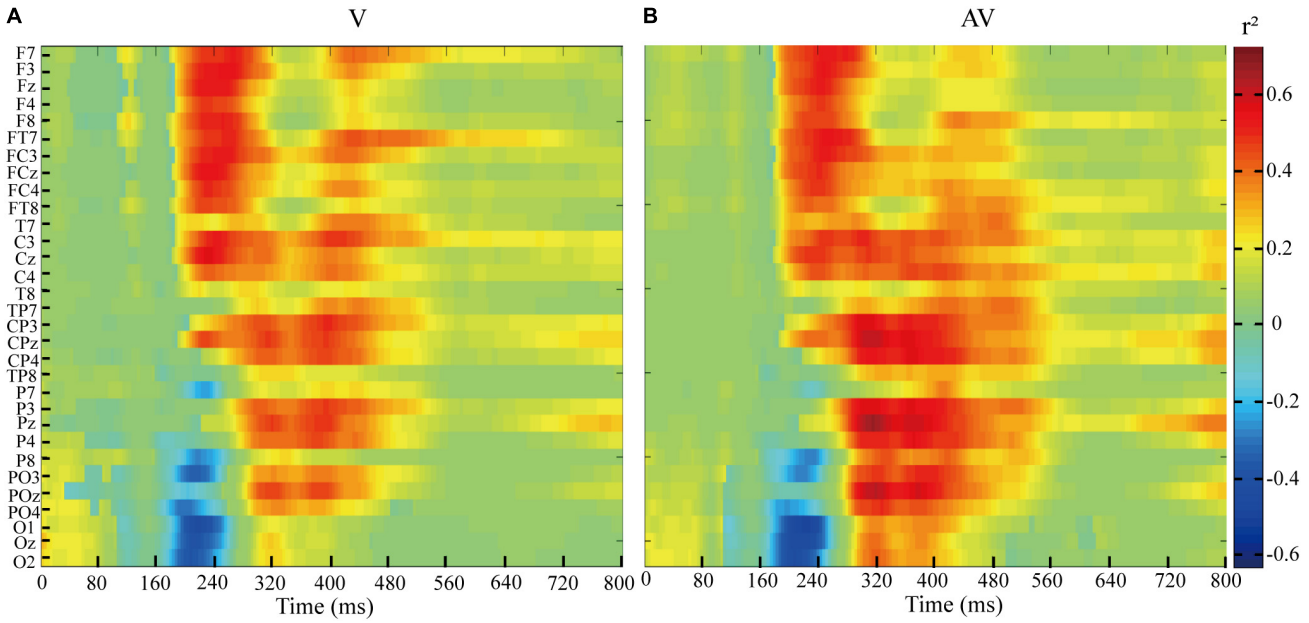

FIGURE 6 | r^2-values of ERPs amplitudes elicited by target and non-target stimuli at 0–800 ms based on EEG data of all subjects in the AV and V spelling paradigms. **(A)** r^2-values of ERPs for the V spelling paradigm. **(B)** r^2-values of ERPs for the AV spelling paradigm.

amplitude and reflect richer information by the graph, we set the r^2-value corresponding to the negative ERP amplitude value as a negative value.

Latency

We computed the latencies of P3a at Fz and P3b at Pz in the V and AV spelling paradigms. The average latencies of P3b and P3a were shorter in the AV spelling paradigm than in the V spelling paradigm. There was no significant difference in the latencies of P3b at Pz between the two paradigms [(AV, V): $t = -1.949$, $P = 0.067$] (**Figure 7**). The latency of P3a at Fz was significantly shorter in the AV spelling paradigm than in the V spelling paradigm [(AV, V): $t = -4.001$, $P = 0.001$] (**Figure 7**).

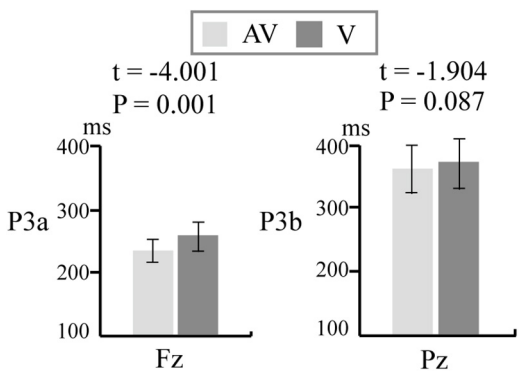

FIGURE 7 | Comparison of latency (means ± SEM) of P3a at Fz and P3b at Pz between the V and AV spelling paradigms. V, visual spelling paradigm; AV, audiovisual spelling paradigm.

Offline Accuracy

According to the results of the r^2 values and ERP analysis, we selected the feature vector for classification as 40 × 22 (40 represents the sample points between 0 and 800 ms in which there was difference for ERP amplitudes between target and non-target stimuli, and the amplitude of ERPs and latencies of P3a also differed between V and AV spelling paradigms; 22 represents channels F7, F3, Fz, F4, F8, FT7, FC3, FCz, FC4, FT8, C3, Cz, C4, CP3, CPz, CP4, P3, Pz, P4, PO3, POz, and PO4). The individual and average accuracies of the AV and V spelling paradigms for the 18 subjects with different superposition times are shown in **Figure 8**. The average accuracies were higher in the AV spelling paradigm than in the V spelling paradigm at each superposition. The best results, the accuracy of 100% at two superpositions, were found for subject 3 and subject 14 in the AV spelling paradigm. In this paradigm, the average superposition time was 3.83 for 12 subjects when accuracies reached 100%. In the V spelling paradigm, the average superposition time was 3.63 for eight subjects when accuracies reached 100%.

Accuracies at each superposition between the V and AV paradigms were compared (**Table 1**). There were significant differences between the V and AV spelling paradigms when superposing from 1 to 10 times ($P < 0.05$), except for superposing 3, 4, 7, and 8 times. However, the accuracies of the AV spelling paradigm had an increasing trend compared with those of the V spelling paradigm at three, four, seven, and eight superpositions ($P = 0.06$). FDR correction was performed for the results.

We compared the ITR at each superposition time for all subjects between V and AV spelling paradigms. **Figure 9** shows the average ITR at each superposition time. Average ITR of AV was larger than that of V at all superposition times. A paired

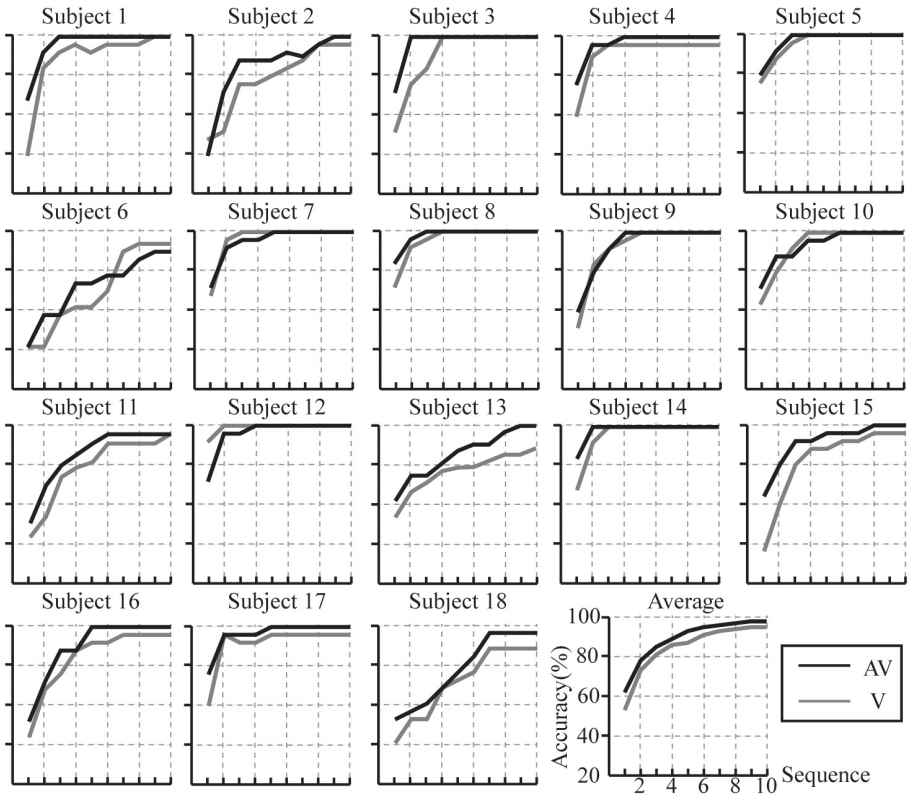

FIGURE 8 | Individual and average accuracies of the P300-speller system with V and AV spelling paradigms for the 18 subjects.

TABLE 1 | *T*-test results of accuracies at each superposition time between the V and AV paradigms.

(V, AV)	**Superposition times**									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
t	−3.205	−2.593	−2.25	−2.001	−3.111	−3.002	−2.097	−2.06	−2.557	−2.557
p	**0.03**	**0.04**	0.06	0.06	**0.03**	**0.03**	0.06	0.06	**0.04**	**0.04**

The bolded values (P < 0.05) represent there were significant difference for accuracy between V and AV spelling paradigms.

t-test showed significant differences in ITR between V and AV at superposing 1, 2, 5, 6, and 10 times ($P < 0.05$), as shown in **Table 2**. The result was corrected by FDR.

DISCUSSION

ERP Analyses

In this study, we proposed a novel audiovisual P300-speller system based on temporal, spatial, and semantic congruence. We assessed grand-averaged ERP waveforms elicited by target stimuli in the AV and V spelling paradigms and analyzed the differences in the waveforms of the ERPs elicited by target Trials. In addition, we compared the performance of the P300-speller system between the AV and V spelling paradigms.

In both spelling paradigms, there was a positive waveform with two peaks elicited by target stimuli between 200 and 500 ms at the frontal, central, and parietal areas (**Figure 4**). The component with a peak between 200 and 300 ms may be P3a potential. The P3a component occurs after a novel event with more frontal distribution, and its latency is usually between 220 and 400 ms (Polich, 2007). The component with a peak between 300 and 500 ms may be the P3b potential because P3b with a more parietal distribution and longer latency is usually between 280 and 600 ms (Polich, 2007). In addition, a negative waveform was observed at approximately 200 ms in the parietal and occipital areas, and it might be the N200 potential (**Figure 4**). The N200, around 200 ms in the temporal–occipital area, is related to conscious attention to the stimuli (Folstein and Van Petten, 2008).

We analyzed the difference in ERP amplitudes elicited by target stimuli between the AV and V spelling paradigms. First, the amplitudes of target stimuli were significantly larger in the AV spelling paradigm than in the V spelling paradigm between 60 and 140 ms in the frontal area (**Figure 5A**). Talsma and Woldorff (2005) observed two enhanced positive

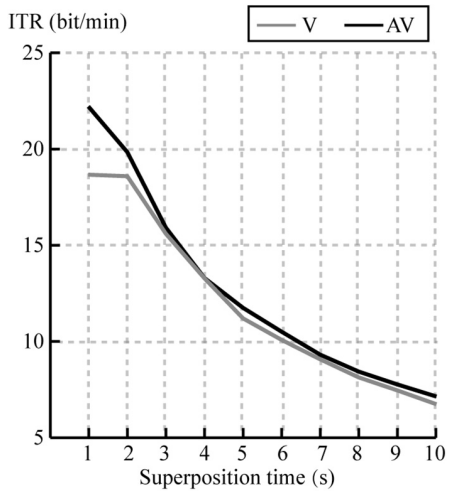

FIGURE 9 | Mean ITR of all subjects at each superposition time for V and AV spelling paradigms.

waveforms for AV (audiovisual stimuli) compared with A + V (the sum of ERPs elicited by unimodal visual and auditory stimuli), one at approximately 100 ms at the frontal area and the other at approximately 160 ms at the central–medial area when the spatial location of the visual and auditory stimuli was matched, indicating the enhanced audiovisual integration. Similarly, Senkowski et al. (2007) observed a positive ERP from audiovisual integration at approximately 60 ms at frontal area, and Teder-Salejarvi et al. (2002) found an audiovisual integration effect with a positive waveform between 130 and 170 ms in the frontal–central area, suggesting that the top-down spatial attention of the audiovisual stimuli may have enhanced the subjects' responses to the target (Talsma and Woldorff, 2005). Thus, the increased amplitude in our study at the earlier stage (60–140 ms) in the AV spelling paradigm compared with the V paradigm may be because the feature information of audiovisual stimuli increased the attention of the subjects to the target, which enhanced the audiovisual integration effect.

The second ERP waveform with a significant difference between AV and V spelling paradigms was P3b (**Figure 5B**). P3b, a sub-component of P300, reflects the cognitive demands during task processing (Polich, 2007), and the P3b amplitude will increase when the cognitive demands increase (Horat et al., 2016; Li et al., 2019). In our audiovisual spelling paradigm, the stimulus included information about the following aspects: first, the stimulus was from two sensory channels; second, the congruence of the spatial location and semantic information resulted in more cognitive demands than did the unimodal visual stimulus when subjects recognized the target stimulus. Therefore, the increased P3b amplitude in the AV spelling paradigm compared with the V spelling paradigm in our study may be due to the increased cognitive demands for the audiovisual stimulus. Our findings are consistent with those in the study by Talsma and Woldorff (2005), in which the audiovisual stimulus elicited a larger P3 amplitude (a positive waveform between 350 and 450 ms) than did the visual stimulus when the spatial position of the visual stimulus and the direction of sound source were closely matched.

The third ERP component with a significant difference was between 700 and 780 ms in the frontal and central areas, and it might be a late positive component (LPC) related to the semantically congruent stimuli. Xie et al. (2018) also observed an enhanced LPC amplitude induced by audiovisual stimuli compared with a unimodal stimulus in the semantically congruent condition at electrodes Fz, Cz, and Pz between 700 and 900 ms. The LPC component has been often observed in studies about semantics and in tasks such as memorizing congruous or incongruous auditory sentences (McCallum et al., 1984), memorizing word lists (Martin and Chao, 2001), and making decisions on congruency (Kounios and Holcomb, 1992). Our findings (increased amplitudes between 700 and 780 ms) are consistent with those in previous studies about semantic processing.

Moreover, we also found waveforms with significant differences in the right temporal at 340–480 and 720–780 ms and in the left temporal at 500–780 ms. The enhanced activation areas may be related to the bimodal audiovisual stimulation. Similarly, Raij et al. (2000) investigated the human brain's audiovisual integration mechanisms for letters and found that auditory and visual brain activations were integrated in the right temporal around 300 ms and in the left and right superior temporal sulci at a later stage (after 500 ms) in a phoneme and grapheme matching task, and Ghazanfar et al. (2005) also observed an increased neural response at the primary auditory cortex caused by semantically matching stimulation. Notably, we found that the difference in ERP amplitudes in the left temporal area was greater than that in the right temporal area at 680–780 ms. Consistently, Campbell (2008) demonstrated that the processing of audiovisual natural speech was activated in the superior temporal regions, in which activation of the left temporal area is usually higher than that of the right temporal area because speech-reading tends to generate left-lateralized activation. Therefore, our findings are in line with those in previous studies.

TABLE 2 | Comparison of ITRs at each superposition time between the V and AV paradigms by *T*-tests.

	Superposition times									
(V, AV)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
t	−2.863	−2.493	−1.962	−1.773	−3.008	−2.897	−1.597	−1.576	−2.089	−2.523
p	**0.04**	**0.049**	0.06	0.08	**0.04**	**0.04**	0.1	0.1	0.08	**0.049**

The bolded values (P < 0.05) represent there were significant difference for ITR between V and AV spelling paradigms.

Although the amplitudes of P3a in the AV spelling paradigm were smaller than those in the V spelling paradigm in the frontal and central areas, there was no significant difference between these two areas. The latency of P3a at Fz in the AV spelling paradigm was significantly shorter than that in the V spelling paradigm. P3a is associated with attention processing, and the latency is related to the speed of allocating attentional resources, in which shorter latencies are related to superior cognitive performance (Polich, 2007). A study about attention effects on the integration of auditory and visual syllables found that the latency of P3a during an audiovisual attention task was shorter than that during a visual attention task, indicating that letter–speech sound integration helps subjects to respond quickly to stimuli (Mittag et al., 2013). Therefore, the shorter P3a latency in the AV spelling paradigm than in the V spelling paradigm may have resulted from faster responses to audiovisual stimuli than to unimodal visual stimuli.

Classification Accuracy and ITR of AV and V Spelling Paradigms

We compared and analyzed the classification accuracies and ITRs between the AV and V spelling paradigms. As expected, the average offline accuracies of the AV spelling paradigm were higher than those of the V spelling paradigm at each superposition. Significantly better accuracies were observed at 1, 2, 5, 6, 9, and 10 superposition times in the AV spelling paradigm than in the V spelling paradigm (**Table 1**). As shown in **Table 1**, there was an improvement trend at three, four, seven, and eight superpositions, although this improvement was not significant ($P > 0.05$). Studies have shown that more ERPs and ERPs with larger amplitudes contribute to improved classification accuracy of the P300-speller (Kaufmann et al., 2011; Li et al., 2015). In our study, the spatially and semantically congruent audiovisual stimuli elicited larger ERP amplitudes at 60–140, 360–460 (P3b), and 700–800 ms than the unimodal visual stimuli, indicating significant improvement in the accuracies of the P300-speller paradigm. In addition, the latency of the P300 can be used as a measure of information processing speed for cognitive functions and contributes to detection of target stimuli (Yagi et al., 1999). Some psychophysical studies have shown that behavioral responses to audiovisual stimuli are typically faster and more accurate than those to unimodal stimuli alone (McDonald et al., 2000; Teder-Salejarvi et al., 2002). In our ERP analyses, the latency of p3a for the audiovisual stimulus was shorter than that for the visual stimulus, indicating that the processing speed of audiovisual information is faster than that of visual information and the faster processing speed may help improve the classification accuracy.

Information transfer rates are an important index to measure the performance of the BCI system, which depends on both classification accuracy and character output speed, and the speed of character output depends on the length of SOA and the times of stimuli repeating. The increased SOA would result in larger P300 amplitude to improve the classification accuracy (Lu et al., 2013), but would lead to a longer time for character selection. Therefore, classification accuracy and the speed of character selection must be weighted for obtaining higher ITR in the design of the P300-speller paradigm. In our study, the set of SOA (250 ms) was to ensure the stable classification accuracy and the pronunciation integrity for each character, which may bring losses more or less to ITR. We compared ITRs at each superposition time between V and AV spelling paradigm (**Figure 9** and **Table 2**), and the results showed that ITRs at superposing 1, 2, 5, 6, and 10 times were significantly larger for the AV spelling paradigm than for the V spelling paradigm. Future studies are needed to determine how to reduce SOA to improve ITR while ensuring the stability of accuracy and the integrity of character pronunciation to further optimize the performance of the audiovisual p300-speller.

The proposed novel audiovisual P300-speller paradigm significantly improved the performance of the P300-speller compared with the visual-based paradigm. The spatially and semantically congruent audiovisual P300-speller was based on a traditional region flashing paradigm. This paradigm requires time to transition between the group-area and sub-area, which reduces the speed of character spelling. Therefore, further investigations are needed to adjust the transformation time between group-area and sub-area to further improve the ITR of the audiovisual P300-speller.

Subject Reports

After collecting EEG data from each subject, he/she was asked about his/her comfort with the AV and V spelling paradigms. Fourteen subjects stated that the audiovisual stimuli from the left or right positions helped them focus on the target stimuli, and the auditory stimuli had a certain hint effect when the superposition times increased. Notably, when the spelling paradigm ran into double flashing problems, it was difficult for them to judge whether the stimulus had flashed once or twice in the unimodal visual condition. However, in the audiovisual condition, the subjects could judge the times of stimulus intensification according to the auditory stimulus of the target. The remaining four participants initially felt that the unimodal visual stimulus would make them more focused on the target. However, as the spelling characters increased, they were more comfortable focusing on the target with audiovisual stimuli because the auditory and visual channels hinted each other. This may be one of the reasons why the accuracies were still significantly improved for the AV spelling paradigm compared with the V spelling paradigm when the superposition times increased. Although the subjects' reports only reflected subjective feelings regarding the spelling paradigm, they have significant implications for the development and improvement of the audiovisual P300-speller paradigm, especially for the application of BCI for patients with neurodegenerative diseases, in which muscular control of the eyes may be impaired or deteriorate over time.

CONCLUSION

In this study, we proposed a novel P300-speller paradigm based on audiovisual spatial and semantic congruency to investigate

whether spatial and semantic matching of audiovisual stimuli can improve the classification accuracy of visual-based P300-spellers. We found that the novel audiovisual P300-speller had significantly improved performances compared with the visual-based P300-speller. Our findings enhance the versatility of the P300-speller system because it is not only suitable for patients with limited hearing but also for those with impaired or deteriorating vision over time. In addition, subjects' reports on the comfort of the paradigm are of great value in helping further develop the audiovisual-based P300-speller system.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the manuscript/supplementary files.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Changchun University of Science and Technology. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

The manuscript was written with contributions from all authors. All authors have approved the final version of the manuscript. QL and ZL designed the experiments. ZL and JY performed the experiments. ZL and JY analyzed the experimental results. NG and OB checked and verified the experimental results. QL and ZL wrote the manuscript.

FUNDING

This work was financially supported by the National Natural Science Foundation of China (Grant Numbers 61773076 and 61806025), the Jilin Scientific and Technological Development Program (Grant Numbers 20190302072GX and 20180519012JH), and the Scientific Research Project of Jilin Provincial Department of Education during the 13th 5-Year Plan Period (Grant Number JJKH20190597KJ).

ACKNOWLEDGMENTS

The authors thank all of the individuals who participated in this study.

REFERENCES

An, X. W., Hohne, J., Ming, D., and Blankertz, B. (2014). Exploring combinations of auditory and visual stimuli for gaze-independent brain–computer interfaces. *PLoS One* 9:e111070. doi: 10.1371/journal.pone.0111070

Belitski, A., Farquhar, J., and Desain, P. (2011). P300 audio-visual speller. *J. Neural Eng.* 8:025022. doi: 10.1088/1741-2560/8/2/025022

Bernat, E., Shevrin, H., and Snodgrass, M. (2001). Subliminal visual oddball stimuli evoke a P300 component. *Clin. Neurophysiol.* 112, 159–171. doi: 10.1016/s1388-2457(00)00445-4

Brouwer, A. M., and Van Erp, J. B. (2010). A tactile P300 brain–computer interface. *Front. Neurosci.* 4:19. doi: 10.3389/fnins.2010.00019

Campbell, R. (2008). The processing of audio-visual speech: empirical and neural bases. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 363, 1001–1010. doi: 10.1098/rstb.2007.2155

Farwell, L. A., and Donchin, E. (1988). Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalogr. Clin. Neurophysiol.* 70, 510–523. doi: 10.1016/0013-4694(88)90149-6

Fazel-Rezai, R., Gavett, S., Ahmad, W., Rabbi, A., and Schneider, E. (2011). A comparison among several P300 brain–computer interface speller paradigms. *Clin. EEG Neurosci.* 42, 209–213. doi: 10.1177/155005941104200404

Folstein, J. R., and Van Petten, C. (2008). Influence of cognitive control and mismatch on the N2 component of the ERP: a review. *Psychophysiology* 45, 152–170. doi: 10.1111/j.1469-8986.2007.00602.x

Ghazanfar, A. A., Maier, J. X., Hoffman, K. L., and Logothetis, N. K. (2005). Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *J. Neurosci.* 25, 5004–5012. doi: 10.1523/JNEUROSCI.0799-05.2005

Guo, J., Gao, S., and Hong, B. (2010). An auditory brain–computer interface using active mental response. *IEEE Trans. Neural. Syst. Rehabil. Eng.* 18, 230–235. doi: 10.1109/TNSRE.2010.2047604

Halder, S., Rea, M., Andreoni, R., Nijboer, F., Hammer, E. M., Kleih, S. C., et al. (2010). An auditory oddball brain–computer interface for binary choices. *Clin. Neurophysiol.* 121, 516–523. doi: 10.1016/j.clinph.2009.11.087

Hammer, E. M., Halder, S., Kleih, S. C., and Kubler, A. (2018). Psychological predictors of visual and auditory P300 brain–computer interface performance. *Front. Neurosci.* 12:307. doi: 10.3389/fnins.2018.00307

Hein, G., Doehrmann, O., Muller, N. G., Kaiser, J., Muckli, L., and Naumer, M. J. (2007). Object familiarity and semantic congruency modulate responses in cortical audiovisual integration areas. *J. Neurosci.* 27, 7881–7887. doi: 10.1523/JNEUROSCI.1740-07.2007

Hoffmann, U., Vesin, J. M., Ebrahimi, T., and Diserens, K. (2008). An efficient P300-based brain–computer interface for disabled subjects. *J. Neurosci. Methods* 167, 115–125. doi: 10.1016/j.jneumeth.2007.03.005

Horat, S. K., Herrmann, F. R., Favre, G., Terzis, J., Debatisse, D., Merlo, M. C. G., et al. (2016). Assessment of mental workload: a new electrophysiological method based on intra-block averaging of ERP amplitudes. *Neuropsychologia* 82, 11–17. doi: 10.1016/j.neuropsychologia.2015.12.013

Hu, Z., Zhang, R., Zhang, Q., Liu, Q., and Li, H. (2012). Neural correlates of audiovisual integration of semantic category information. *Brain Lang.* 121, 70–75. doi: 10.1016/j.bandl.2012.01.002

Jin, J., Daly, I., Zhang, Y., Wang, X. Y., and Cichocki, A. (2014). An optimized ERP brain–computer interface based on facial expression changes. *J. Neural Eng.* 11:036004. doi: 10.1088/1741-2560/11/3/036004

Kaufmann, T., Schulz, S. M., Grunzinger, C., and Kubler, A. (2011). Flashing characters with famous faces improves ERP-based brain–computer interface performance. *J. Neural Eng.* 8:056016. doi: 10.1088/1741-2560/8/5/056016

Kounios, J., and Holcomb, P. J. (1992). Structure and process in semantic memory: evidence from event-related brain potentials and reaction times. *J. Exp. Psychol. Gen.* 121, 459–479. doi: 10.1037//0096-3445.121.4.459

Kubler, A., Kotchoubey, B., Kaiser, J., Wolpaw, J. R., and Birbaumer, N. (2001). Brain-computer communication: unlocking the locked in. *Psychol. Bull.* 127, 358–375. doi: 10.1037//0033-2909.127.3.358

Kubler, A., and Neumann, N. (2005). Brain–computer interfaces–the key for the conscious brain locked into a paralyzed body. *Prog. Brain Res.* 150, 513–525. doi: 10.1016/S0079-6123(05)50035-9

Laurienti, P. J., Kraft, R. A., Maldjian, J. A., Burdette, J. H., and Wallace, M. T. (2004). Semantic congruence is a critical factor in multisensory behavioral performance. *Exp. Brain Res.* 158, 405–414. doi: 10.1007/s00221-004-1913-1912

Li, Q., Liu, S., Li, J., and Bai, O. (2015). Use of a green familiar faces paradigm improves P300-speller brain–computer interface performance. *PLoS One* 10:e0130325. doi: 10.1371/journal.pone.0130325

Li, Q., Lu, Z. H., Gao, N., and Yang, J. J. (2019). Optimizing the performance of the visual P300-speller through active mental tasks based on color distinction and modulation of task difficulty. *Front. Hum. Neurosci.* 13:130. doi: 10.3389/fnhum.2019.00130

Lu, J., Speier, W., Hu, X., and Pouratian, N. (2013). The effects of stimulus timing features on P300 speller performance. *Clin. Neurophysiol.* 124, 306–314. doi: 10.1016/j.clinph.2012.08.002

Martin, A., and Chao, L. L. (2001). Semantic memory and the brain: structure and processes. *Curr. Opin. Neurobiol.* 11, 194–201. doi: 10.1016/s0959-4388(00)00196-3

McCallum, W. C., Farmer, S. F., and Pocock, P. V. (1984). The effects of physical and semantic incongruities on auditory event-related potentials. *Electroencephalogr. Clin. Neurophysiol.* 59, 477–488. doi: 10.1016/0168-5597(84)90006-6

McDonald, J. J., Teder-Salejarvi, W. A., and Hillyard, S. A. (2000). Involuntary orienting to sound improves visual perception. *Nature* 407, 906–908. doi: 10.1038/35038085

Mittag, M., Alho, K., Takegata, R., Makkonen, T., and Kujala, T. (2013). Audiovisual attention boosts letter-speech sound integration. *Psychophysiology* 50, 1034–1044. doi: 10.1111/psyp.12085

Nijboer, F., Sellers, E. W., Mellinger, J., Jordan, M. A., Matuz, T., Furdea, A., et al. (2008). A P300-based brain–computer interface for people with amyotrophic lateral sclerosis. *Clin. Neurophysiol.* 119, 1909–1916. doi: 10.1016/j.clinph.2008.03.034

Plank, T., Rosengarth, K., Song, W., Ellermeier, W., and Greenlee, M. W. (2012). Neural correlates of audio-visual object recognition: effects of implicit spatial congruency. *Hum. Brain Mapp.* 33, 797–811. doi: 10.1002/hbm.21254

Polich, J. (2007). Updating P300: an integrative theory of P3a and P3b. *Clin. Neurophysiol.* 118, 2128–2148. doi: 10.1016/j.clinph.2007.04.019

Raij, T., Uutela, K., and Hari, R. (2000). Audiovisual integration of letters in the human brain. *Neuron* 28, 617–625. doi: 10.1016/s0896-6273(00)00138-0

Sellers, E. W., and Donchin, E. (2006). A P300-based brain–computer interface: initial tests by ALS patients. *Clin. Neurophysiol.* 117, 538–548. doi: 10.1016/j.clinph.2005.06.027

Semlitsch, H. V., Anderer, P., Schuster, P., and Presslich, O. (1986). A solution for reliable and valid reduction of ocular artifacts, applied to the P300 ERP. *Psychophysiology* 23, 695–703. doi: 10.1111/j.1469-8986.1986.tb00696.x

Senkowski, D., Saint-Amour, D., Kelly, S. P., and Foxe, J. J. (2007). Multisensory processing of naturalistic objects in motion: a high-density electrical mapping and source estimation study. *Neuroimage* 36, 877–888. doi: 10.1016/j.neuroimage.2007.01.053

Stein, B. E., and Meredith, M. A. (1993). The merging of the senses. *J. Cogn. Neurosci.* 5, 373–374. doi: 10.1162/jocn.1993.5.3.373

Suminski, A. J., Tkach, D. C., and Hatsopoulos, N. G. (2009). Exploiting multiple sensory modalities in brain-machine interfaces. *Neural Netw.* 22, 1224–1234. doi: 10.1016/j.neunet.2009.05.006

Szmidt-Salkowska, E., and Rowinska-Marcinska, K. (2005). Blink reflex in motor neuron disease. *Electromyogr. Clin. Neurophysiol.* 45, 313–317.

Takano, K., Komatsu, T., Hata, N., Nakajima, Y., and Kansaku, K. (2009). Visual stimuli for the P300 brain–computer interface: a comparison of white/gray and green/blue flicker matrices. *Clin. Neurophysiol.* 120, 1562–1566. doi: 10.1016/j.clinph.2009.06.002

Talsma, D., and Woldorff, M. G. (2005). Selective attention and multisensory integration: multiple phases of effects on the evoked brain activity. *J. Cogn. Neurosci.* 17, 1098–1114. doi: 10.1162/0898929054475172

Teder-Salejarvi, W. A., McDonald, J. J., Di Russo, F., and Hillyard, S. A. (2002). An analysis of audio-visual crossmodal integration by means of event-related potential (ERP) recordings. *Brain Res. Cogn. Brain Res.* 14, 106–114. doi: 10.1016/s0926-6410(02)00065-4

Thompson, D. E., Blain-Moraes, S., and Huggins, J. E. (2013). Performance assessment in brain–computer interface-based augmentative and alternative communication. *Biomed. Eng. Online* 12:43. doi: 10.1186/1475-925X-12-43

Van der Waal, M., Severens, M., Geuze, J., and Desain, P. (2012). Introducing the tactile speller: an ERP-based brain–computer interface for communication. *J. Neural. Eng.* 9:045002. doi: 10.1088/1741-2560/9/4/045002

Wang, F., He, Y. B., Pan, J. H., Xie, Q. Y., Yu, R. H., Zhang, R., et al. (2015). A novel audiovisual brain–computer interface and its application in awareness detection. *Sci. Rep.* 5:9962. doi: 10.1038/srep12592

Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G., and Vaughan, T. M. (2002). Brain–computer interfaces for communication and control. *Clin. Neurophysiol.* 113, 767–791.

Xie, Q. Y., Pan, J. H., Chen, Y., He, Y. B., Ni, X. X., Zhang, J. C., et al. (2018). A gaze-independent audiovisual brain–computer Interface for detecting awareness of patients with disorders of consciousness. *BMC Neurol.* 18:144. doi: 10.1186/s12883-018-1144-y

Xu, H., Zhang, D., Ouyang, M., and Hong, B. (2013). Employing an active mental task to enhance the performance of auditory attention-based brain–computer interfaces. *Clin. Neurophysiol.* 124, 83–90. doi: 10.1016/j.clinph.2012.06.004

Xu, M. P., Xiao, X. L., Wang, Y. J., Qi, H. Z., Jung, T. P., and Ming, D. (2018). A Brain–computer interface based on miniature-event-related potentials induced by very small lateral visual stimuli. *IEEE Trans. Biomed. Eng.* 65, 1166–1175. doi: 10.1109/Tbme.2018.2799661

Yagi, Y., Coburn, K. L., Estes, K. M., and Arruda, J. E. (1999). Effects of aerobic exercise and gender on visual and auditory P300, reaction time, and accuracy. *Eur. J. Appl. Physiol. Occup. Physiol.* 80, 402–408. doi: 10.1007/s004210050611

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.