Journal of Beijing Institute of Technology, 2021, Vol. 30, Suppl.1

Analysis and Classification of Speech Imagery EEG Based on Chinese Initials

Yongsheng Zhao, Ying Liu✉ and Yunlong Gao

(School of Mechanical Engineering, Beijing Institute of Technology, Beijing 100081, China)

Abstract: Brain-computer interfaces (BCI) can provide external information communication for people with normal thinking but impaired motor functions. For patients with language disorders, speech imagery BCIs make it possible to communicate normally. However, there are few studies on Chinese speech imagery at present. Almost all studies employed fixed experimental content, without considering the diversity in subjects. With the purpose of improving the effect of a Chinese speech imagery BCI system, a novel experiment of Chinese initials imagery was designed. The experiment is divided into two parts. A preliminary experiment used to select content for subjects. Formal experiment-specific experimental content was designed for subjects. After preprocessing, feature extraction was carried out by common spatial patterns (CSP) and discrete wavelet transform (DWT), and then a support vector machine (SVM) and extreme learning machine (ELM) were used for classification. Finally, the best performance was obtained by the model using DWT and ELM with a highest accuracy of 73.04%. This study shows that the novel experiment is feasible and can potentially extend the capability of utilizing speech imagery in future BCI applications.

Key words: brain-computer interface (BCI); speech imagery; Chinese initials; diversity of subjects; classification

CLC number: TP312 **Document code:** A **Article ID:** 1004-0579(2021)S1-0044-08

The technology of brain-computer interfaces (BCI) has recently gained increased attention and has great potential in improving quality of life for many people[1]. A BCI represents a communication and control mechanism which can provide new auxiliary functions and means between the human brain and computers. People with normal thinking but impaired motor function, especially paralyzed patients who have lost muscle control, can regain control of sensory and motor functions and carry out their daily life with BCI technology[2].

Electroencephalograms[3], electrocorticography[4], magnetoencephalography[5], functional magnetic resonance imaging[6], and near-infrared spectroscopy[7] are used to read users' intentions and control external devices. In this work, an EEG signal was used to extract the necessary features because it is noninvasive and its application does not require specific medical knowledge.

For patients with language disorders, it is very important to be able to communicate with the outside world normally through a BCI. In the process of language production, semantic, lexical, and phonetic processing are needed first, and the movement of the vocal system is controlled to produce sound[8-9]. Almost one third of the primary motor cortex is allocated to produce speech, which reflects its complexity[10]. Everyone has their own modality of cognition, which leads to a difference in brain activity between individuals, whether in time domain, frequency domain, or spatial pattern[11]. For each individual, phonetic processing activity for specific words

Received 2020-07-23
Supported by the National Natural Science Foundation of China (51575048)
✉ Author for correspondence, associate professor, Ph.D.
E-mail: biggirlliu@bit.edu.cn
DOI: 10.15918/j.jbit1004-0579.20095

does not change over time[12-13] and is not affected by pre-stimulation, cognitive activity, or task frequency[14-15].

Current research on speech imagery mainly focuses on syllables and vocabulary. DaSalla et al.[16-17]created a system apparently capable of recognizing three stages: /a/, /u/, and rest. However, the results of CSP point out that the discriminant channels are Fz, C3, Cz, and C4. This suggests that the results are mainly related to the activation of the motor cortex. Idrees and Farooq[18] proposed an approach in which 11 temporal and spectral features are extracted; then, a linear discriminant analysis classifier is used for those features. The author obtained a classification accuracy from 65% to 82.5%. Salama et al.[19] were able to distinguish between two imagined words, namely "Yes" and "No", using different types of classifiers. Their analysis produced average accuracies between 57% and 59% among seven subjects. Wang et al.[20] distinguished between two Chinese characters that mean "left" and "one", and the rest state respectively. The classification model was composed of a CSP and a SVM, and its highest accuracies were between 73.65% and 95.76%. Yang et al.[21] chose four vowels and four consonants in Mandarin Chinese as well as a no-imagery state as the control. Spectral, temporal, and spatial analyses were used to extract the optimal phoneme imagery features for pairwise classification by SVM. The classification accuracy reached up to 83%, averaged across subjects.

The aforementioned works have indicated that speech imagery is a promising approach, and the accuracy can be improved by selecting appropriate features and classifiers. However, almost all studies employed fixed experimental content, without considering the diversity in subjects. It might lead to an inefficient result if the experimental content is difficult to recognize for the subjects. Therefore, a novel experiment was designed to select different content for the subjects. Most Chinese syllables consist of an initial and a final. The initials are in front of finals. Therefore, initials were studied at first. The initials with high discriminability for subjects were found through the preliminary experiment. Then, selected content would be tasked in a formal experiment. In this paper, two different signal extraction methods for BCI systems, CSP and DWT, were compared. Finally, SVM and ELM were responsible for the classification task.

1 Methods

1.1 Subjects

The 8 subjects were all students from universities in Beijing (6 males and 2 females, average age 23.67), whose mother tongue was Chinese. According to the Edinburgh handedness test, they were all right-handed. They were labeled S1-S8 in this experiment. All eight subjects participated in the EEG experiment for the first time and understood the experiment process in detail. They all signed the informed agreement before the experiment began.

Some studies have shown that speech is partially processed by right brain, but most of the best channels cover the left brain. In this paper, EEG signals are recorded according to 10–20 system of the International Federation, as shown in Fig. 1.

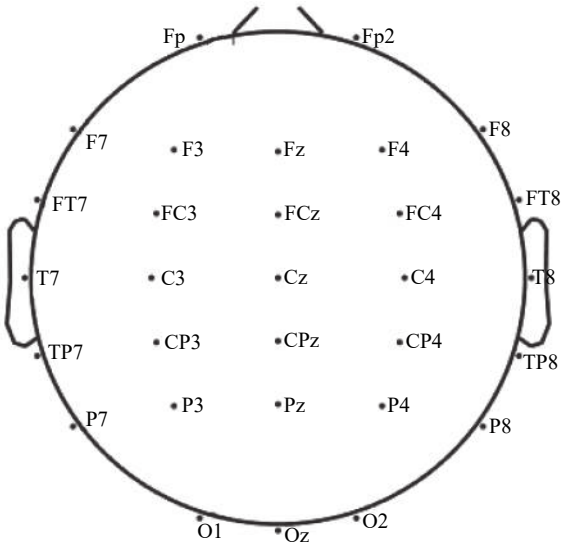

Fig. 1 Electrode positions correspond to positions in the international 10–20 system

1.2 Experimental paradigm

Before the beginning of the experiment, the subjects received Chinese pronunciation training to ensure that their pronunciations were correct. During the experiment, the subjects were asked to imagine the pronunciation of the Chinese initials in the center of the screen. At the beginning of the experiment (t=0 s), a fixed character "+" appeared in the center of the gray screen, indicating that the experiment was about to begin. After 6 seconds, a stimulus symbol appeared randomly in the center of the screen and stayed for two seconds. The subjects imagined the corresponding phoneme until the stimulus symbol disappeared. After that, "+" appeared again on the computer screen for one second for the subjects to rest. The specific experimental process is shown in Fig. 2. A grey screen was chosen as the background to reduce the visual aftereffect.

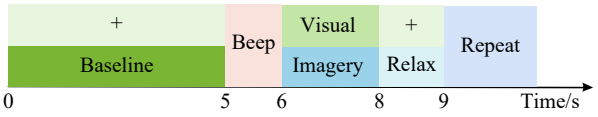

Fig. 2 Experimental paradigm

In the preliminary experiment, 23 Chinese initials were randomly displayed in the center of the screen as the content of pronunciation imagination, and the character "+" was displayed at rest. Initials are used in front of Chinese vowels with which they form a complete Chinese syllable. According to the vocal organ and physiology of vocalization, 21 initials can be divided into groups in advance. In this experiment, two zero initials, y and w, are added to the content of the whole experiment.

1.3 Channel selection

During the generation and collection of EEG signals, there was a lot of interference due to each subject's own activities and external influences. In the pretreatment stage, the EEG signals were first re-referenced by the mean of the left and right earlobes, and the signal to noise ratio (SNR) of the EEG signals was improved by the common averaged reference. The Butterworth filter was used to carry out bandpass filtering in the band of 6–30 Hz. Then, the classification model comprising a CSP and a SVM traversed all the channel combinations and determined the optimal channel combination by comparing the classification accuracy. The selected channels were mainly located in the left hemisphere, where the Broca's area and Wernicke's area are.

1.4 Clustering analysis

After the channel selection, a 6th-order DWT was used to extract wavelet coefficients of the EEG signals. Then, the 5th and 6th layers of the wavelet coefficients were selected to calculate their mean, variance, skewness, kurtosis, energy, and entropy as the features. Afterwards, the features were input to the model for fuzzy clustering analysis, which divided the experiment content into four groups, and then, according to the vocal organ and physiology of vocalization, selected four representative initials to test again. The chosen initials are shown in Tab. 1.

Tab. 1 Speech imagery content of the subjects

Subject	Experimental content			
S1	m	s	sh	y
S2	m	d	x	zh
S3	f	l	h	c
S4	b	h	z	zh
S5	d	c	zh	ch
S6	n	c	ch	w
S7	p	g	x	s
S8	h	q	sh	y

1.5 Feature extraction: CSP and DWT

The CSP algorithm is widely used to improve the SNR of EEG data, which adopts the method of supervision and creates a public space of the optimal filter. The maximization of variance and minimizing another kind of variance at the same time can be used to calculate the covariance matrix diagonalization of two tasks, in order to get the largest eigenvector that can distinguish degrees. It is suitable for the feature ex-

traction of a binary classification task.

The process of the CSP algorithm is as follows.

① Segment the raw data by category. Two types of sample data $\boldsymbol{E}$ can be classified as $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$.

② Calculate the mean covariance matrices $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ for the two classes. $\boldsymbol{C}_c$ is the sum of two types of data. The eigenvalue decomposition is

$$\boldsymbol{C}_c=\boldsymbol{U}_c\boldsymbol{\Lambda}_c\boldsymbol{U}_c^{\mathrm{T}} \tag{1}$$

③ Calculate the whitening transfer matrix

$$\boldsymbol{P}=\frac{\boldsymbol{1}}{\sqrt{\boldsymbol{\Lambda}_c}}\times\boldsymbol{U}_c^{\mathrm{T}} \tag{2}$$

The Whitening transformation on the mean covariance matrix is

$$\boldsymbol{S}_i=\boldsymbol{P}\boldsymbol{C}_i\boldsymbol{P}^{\mathrm{T}},\ i=1,2 \tag{3}$$

④ $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ share common eigenvectors $\boldsymbol{B}$

$$\boldsymbol{S}_i=\boldsymbol{B}\boldsymbol{\Lambda}_i\boldsymbol{B}^{\mathrm{T}},\ i=1,2 \tag{4}$$

⑤ Each row in the projection matrix $\boldsymbol{W}=\boldsymbol{B}^{\mathrm{T}}\boldsymbol{P}$ is the common spatial pattern of the two classes.

⑥ The pre-processed EEG data $\boldsymbol{E}_{M\times N}$ is projected by the matrix $\boldsymbol{W}_{M\times N}$ to obtain the characteristic matrix

$$\boldsymbol{Z}_{M\times N}=\boldsymbol{W}_{M\times M}\boldsymbol{E}_{M\times N} \tag{5}$$

After normalization, the features are obtained.

DWT decomposes a discrete time signal $\boldsymbol{x}[k]$ into a set of signals (or wavelet coefficients) by scaling and shifting the mother wavelet. The first step in the DWT decomposition lies in picking the appropriate number of wavelet decomposition levels. For the initial scale level1, signal $\boldsymbol{x}[k]$ passes concurrently through both the high-pass filter h and low-pass filter l, followed by the process of downsampling by 2. The output of each level j is represented in the form of two signals: Detail ($\boldsymbol{D}_j$) and Approximation ($\boldsymbol{A}_j$), described as

$$\boldsymbol{D}_j[i]=\sum_k \boldsymbol{x}[k]\ h[2i-k] \tag{6}$$

$$\boldsymbol{A}_j[i]=\sum_k \boldsymbol{x}[k]\ l[2i-k] \tag{7}$$

Progressing to the next level, the approximation $\boldsymbol{A}_j$ is set as $\boldsymbol{x}(k)$ and j is incremented by 1. The aforementioned procedure of generating $\boldsymbol{D}_j$ and $\boldsymbol{A}_j$ repeats as long as j does not surpass the stated number, as shown in Fig. 3.

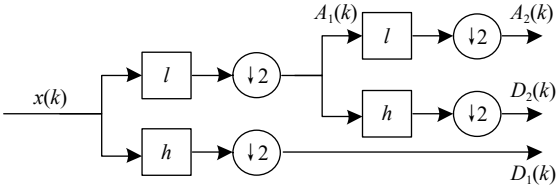

Fig. 3 DWT for scale level 2

After the 6th order of DWT, the mean, variance, skewness, kurtosis, energy, and entropy of the 5th and 6th layers were selected as the features.

1.6 Classification: SVM and ELM

An SVM adopts the principle of structural risk minimization to construct the optimal hyperplane for data classification. It can effectively solve classification problems with small samples and nonlinear and high dimensions, and has the advantages of fast calculation speed and strong generalization ability. For a binary class problem, there is a discriminant function f as the classification rule. The training data $\boldsymbol{x}_{in}$ is judged by $f(\boldsymbol{x}_{in})$. The discriminant function can be written as

$$\begin{cases} \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i-\boldsymbol{b}\le -1, \forall f(\boldsymbol{x}_i)=-1 \\ \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i-\boldsymbol{b}\ge +1, \forall f(\boldsymbol{x}_i)=+1 \end{cases} \tag{8}$$

In a high-dimensional model, the key of SVM classification is the choice of the kernel functions that include liner kernels, polynomial kernels, and Gaussian kernels. In this study, the liner kernel is selected. In order to allow mislabeled samples, error ξ and discipline factor c are introduced as

$$\min \frac{1}{2}\|\boldsymbol{w}\|^2+c\sum_i \xi_i,\ \text{subject to}$$
$$\boldsymbol{y}_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i-b)-1+\xi_i\ge 0,\ \xi_i\ge 0,\ \forall i \tag{9}$$

ELM was proposed as an extension of the single layer feedforward network[22]. In mathematics, ELM with L hidden nodes and a feature mapping $\boldsymbol{h}(\boldsymbol{x})$ is modeled as

$$\sum_{j=1}^{L} \boldsymbol{\beta}_j h(\boldsymbol{x}_i) = \boldsymbol{y}_i, i = 1, 2, \cdots, N \qquad (10)$$

where $\boldsymbol{\beta}_j$ is a weight vector connecting the jth hidden neuron and the output neurons. ELM theory aims to reach not only the smallest training error but also the smallest norm of output weights. The procedure of ELM for EEG is detailed as follows.

① Randomly specify the weight vector of the input layer and the bias of the hidden layer.

② Compute the output matrix between the hidden and output layers.

③ Compute the output weight vector.

④ Compute the predicted label.

2 Results

2.1 Results of experiments

The data of the preliminary experiment and formal experiment were extracted with CSP. Then, the number of CSP components was chosen empirically, varying between 6 and 12, depending on the subjects. Finally, the features were classified in pairs by SVM.

Fig. 4 shows the mean values of accuracy for each subject in the two experiments. The highest classification accuracy of the preliminary experiment was 67.6% and the average was 61.5%, while the highest classification accuracy of the formal experiment was 68.2% and the average was 63.8%. A paired samples T-Test was used to show the classification accuracy differences between the experiments ($p<0.05$). It showed the classification accuracy was increased by using the selected content. The most important thing is that the variance of the formal experiment classification accuracy is smaller, which improves the stability of the classification model. The above results fully illustrate the feasibility and necessity of selecting content rationally.

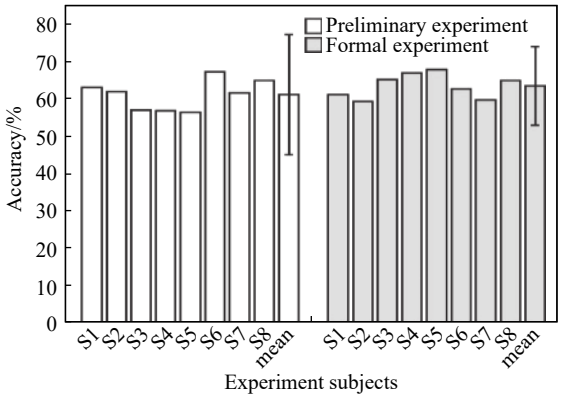

Fig. 4 Averaged accuracies over all pairwise classification

2.2 Time-frequency analysis

After time-frequency analysis, the information of an EEG signal in time and frequency domain could be represented well from the whole time range. The choices of time window and frequency band were vital to the time-frequency representation in the study.

The signals of all channels were re-referenced by CAR. Then, channels FC3, CP3, and Cz, near the Broca's area, Wernicke's area, and motor cortex region, respectively, were selected to analyze Event-Related Spectral Perturbation. After wavelet reconstruction, the ERSP could be obtained by superimposing the energy spectrum distributions of single trial. The ERSPs of three channels are presented in Fig. 5, which are from S1 and S3. The averaged results are in frequencies 6–30 Hz. Each row represents the same speech imagery task, and each column represents the same channel. The results of analysis were similar for the other subjects. It shows that the highest activation during the word imagination task was observed in the Alpha band. In addition, major activity that occurred within the first 1s timeframe was also observed. The differences of the ERSP between different imagery tasks makes it possible for the model to achieve higher accuracy.

2.3 Comparison of models

To evaluate the proposed method, we performed a 10-fold-cross validation procedure. The training and testing sets were partitioned randomly. The pairwise classification accuracies ob-

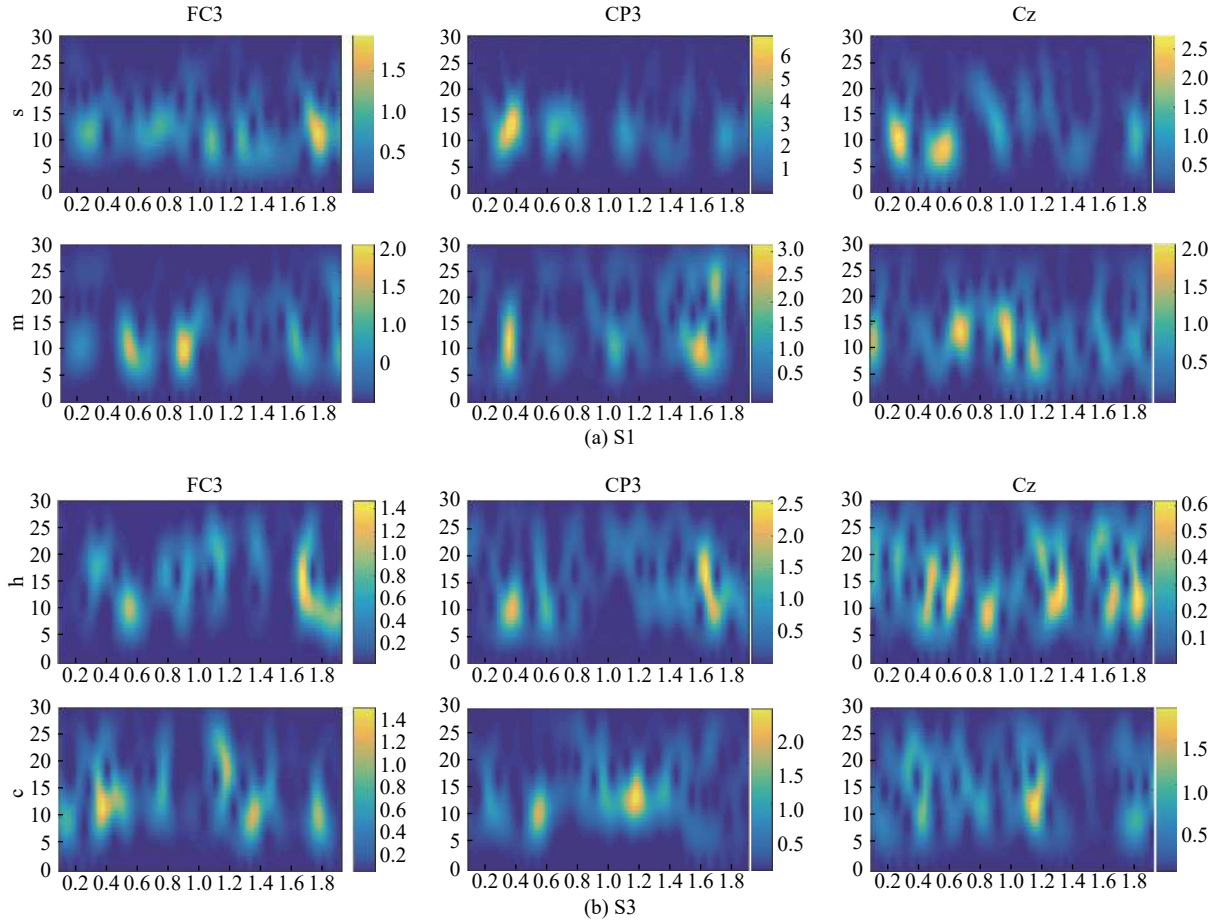

Fig. 5 ERSPs from channels FC3,CP3 and Cz for S1 (a) and S3 (b)

tained by different algorithms respectively are shown in Tab. 2 and Fig. 6. Through the combinations of different feature extraction methods and classification algorithms, the highest accuracy rate was raised to 75.64%, and the average accuracy rate was 73.04%. For the experimental data, DWT, the method of feature extraction, was significantly better than CSP ($p<0.01$), and the performance of classification algorithm ELM was better than SVM ($p<0.01$).

Tab. 2 Classification results of four models %

Subject	CSP+SVM	CSP+ELM	DWT+SVM	DWT+ELM
S1	61.53	67.44	67.78	70.21
S2	59.58	67.17	66.58	74.97
S3	65.56	72.42	70.557	74.61
S4	67.22	75.22	65.24	73.98
S5	68.19	69.83	68.66	72.27
S6	63.06	67.31	71.64	72.07
S7	60.10	68.96	65.67	70.57
S8	65.28	72.97	71.77	75.64
Mean	63.82±10.50	70.17±9.24	68.49±6.81	73.04±4.21

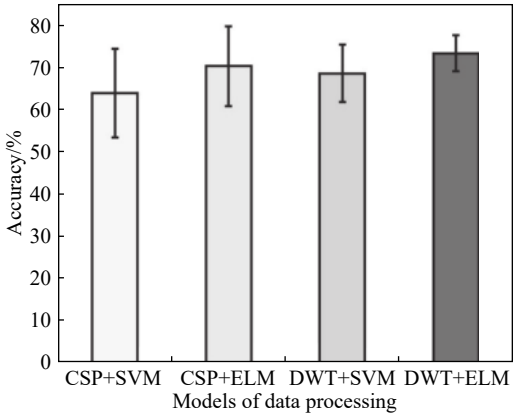

Fig. 6 Averaged accuracies and the variance over all pairwise classification

3 Conclusions

In this study, a novel method was proposed to classify Chinese speech imagery. A preliminary experiment was conducted first and the specific content was selected from the Chinese initials according to the performance of the sub-

jects. Different content was designed for the subjects. After preprocessing, CSP was used for feature extraction and SVM for classification. The results showed that the average classification accuracy of the formal experiment was 63.8%, slightly higher than the preliminary experiment, and the variance of the formal experiment was smaller. This demonstrates the feasibility of experimental content selection.

CSP and DWT were used to extract the features as the input of classification algorithms SVM and ELM. By analyzing the accuracy of the models, it can be concluded that DWT was better than CSP ($p<0.01$) and ELM was better than SVM ($p<0.01$) for the experimental data. Compared with other approaches, the combined model of DWT and ELM had the highest accuracy rate of 75.64%.

This is the first use of different content selected from all of the Chinese initials. This study highlights the importance of different content for subjects. Once a cognitive pattern is formed, it is hard to change over time. According to the ERSPs of subjects, it was found that different tasks would motivate corresponding brain activities, which made it possible to classify the tasks correctly.

The experiment selecting the task content was a simple trial to improve classification accuracy. The process may still have some problems that need to be solved. In future work, we would like to improve the method of content selection and classification. It can be predicted that with the improvement of data processing, the classification results of the model may be further improved.

In summary, a unique classification model of Chinese initials imagery was designed, which selected specific experimental content for different experimental subjects to achieve a better classification effect. The model using DWT and ELM would have better performance. Future work would focus on data processing and classification.

References:

[1] Tong J, Zhu D. Multi-phase cycle coding for SSVEP based brain-computer interfaces[EB/OL]. (2015-05-14). https://doi.org/10.1186/1475-925X-14-5

[2] Blankertz B, Schäfer C, Dornhege G, et al. Single trial detection of EEG error potentials: A tool for increasing BCI transmission rates[C]// International Conference on Artificial Neural Networks. New York: Springer-Verlag, 2002.

[3] Vidaurre C, Schlögl A, Cabeza R, et al. Adaptive on-line classification for EEG-based brain computer interfaces with AAR parameters and band power estimates [J]. Biomedical Engineering, 2005, 50(11): 350-354.

[4] Pistohl T, Schulze A, Aertsen A, et al. Decoding natural grasp types from human ECoG [J]. NeuroImage, 2012, 59(1): 248-260.

[5] Stepp C, Oyunerdene N, Matsuoka Y. Kinesthetic motor imagery modulates intermuscular coherence [J]. IEEE Trans Neural Systems & Rehabilitation, 2011, 19(6): 638-643.

[6] Berman B, Horovitz S, Venkataraman, et al. Self-modulation of primary motor cortex activity with motor and motor imagery tasks using realtime fMRI-based neurofeedback [J]. NeuroImage, 2012, 59(2): 917-925.

[7] Fazli S, Mehnert J, Steinbrink J, et al. Enhanced performance by a hybrid NIRS–EEG brain computer interface [J]. NeuroImage, 2012, 59(1): 519-529.

[8] Indefrey P, Levelt W J M. The spatial and temporal signatures of word production components [J]. Cognition, 2004, 92(1-2): 101-144.

[9] Leuthardt E C, Pei X M, Breshears J, et al. Temporal evolution of gamma activity in human cortex during an overt and covert word repetition task [J]. Fronters in Human Neuroscience, 2012, 6: 1-12.

[10] Schott G D. Penfield's homunculus: A note on cerebral cartography [J]. Neurol Neurosurg Psychiatry, 1993, 56(4): 329-333.

[11] Indefrey P. The spatial and temporal signatures of word production components: a critical update [J]. Front Psychol, 2011, 2: 1-16.

[12] Flinker A, Chang E F, Kirsch H E, et al. Single-trial speech suppression of auditory cortex activity in humans [J]. Neuroscience, 2010, 30(49): 16643-16650.

[13] Creutzfeldt O, Ojemann G, Lettich E. Neuronal

activity in the human lateral temporal lobe [J]. Experiment Brain Research, 1989, 77: 451-475.

[14] Martin R C, Lesch M F, Bartha M C. Independence of input and output phonology in word processing and short-term memory [J]. Journal of Memory & Language, 1999, 41(1): 3-29.

[15] Leinenger M. Phonological coding during reading [J]. Psychology, 2014, 140(6): 1534-1555.

[16] DaSalla C S, Kambara H, Sato M, et al. Single trial classification of vowel speech imagery using common spatial patterns [J]. Neural Network, 2009, 22(9): 1334-1339.

[17] DaSalla C S, Kambara H, Koike Y, et al. Spatial filtering and single-trial classification of EEG during vowel speech imagery [J]. Convention Rehabilitation Engineering and Assistive Technology, 2009, 13(2): 5-27.

[18] Idrees B M, Farooq O. Vowel classification using wavelet decomposition during speech imagery[C]// The 3rd International Conference on Signal Processing and Integrated Networks, Noida, India, 2016.

[19] Salama M, Elsherif L, Lashin H, et al. Recognition of unspoken words using electrode electroencephalograhic signals[C]//The 6th International Conference on Advanced Cognitive Technologies and Applications, Venice, Italy, 2014.

[20] Wang L, Zhang X, Zhong X, et al. Analysis and classification of speech imagery EEG for BCI [J]. Biomed Signal Process Control, 2013, 8(1): 901-908.

[21] Yang Xiaofang, Jiang Minghu. Study on brain-computer interface based on Chinese phoneme imagery [J]. Journal of Chinese Information Processing, 2014, 28(05): 13-23.

[22] Huang G B, Zhu Q, Siew C. Extreme learning machine [J]. Theory and Applications Neuro Computing, 2006, 70(1): 489-501.