Hindawi
Journal of Sensors
Volume 2021, Article ID 3006397, 13 pages
https://doi.org/10.1155/2021/3006397

Research Article

Research on the Advantages of Digital Sensor Equipment in Language Audio-Visual and Oral Teaching

Wu Yufei,[1] **Wang Dandan**,[1] **and Zhu Yanwei**[2]

[1]*School of Foreign Studies, Tangshan Normal University, Tangshan 063000, China*
[2]*School of Mathematics and Computational Sciences, Tangshan Normal University, Tangshan 063000, China*

Correspondence should be addressed to Wu Yufei; 160409128@stu.cuz.edu.cn

Received 8 September 2021; Accepted 18 October 2021; Published 30 November 2021

Academic Editor: Gengxin Sun

Digital sensors use biotechnology and information processing technology to strengthen the processing of relevant visual and auditory information, which is helpful to ensure that the receiver can obtain more accurate information, so as to improve the learning effect and reduce the impact on the environment. This paper designs an experiment to explore the role of digital sensors in language audio-visual teaching, which provides a reference for the application of digital sensors in the future. The impulse response function in sensor technology is introduced. The speech time domain envelope and time-varying mouth area of the sensor device are calculated. The auditory attention transfer detection based on line of sight rotation estimation is carried out through the auditory attention decoding fusion technology and the sensor auditory attention conversion detection method. At the same time, the characteristic of sensor heog signal is analyzed. The results show that the algorithm proposed in this paper has good results.

1. Introduction

With the development of society, the research on image processing, pattern recognition, image and video compression, biometric recognition and information security, three-dimensional visual information processing, and intelligent human-computer interaction in the field of machine vision is gradually increasing. Visual auditory processing is widely used in the field of machine hearing. Machine auditory computing model and speech language information processing system are widely used. Many scholars have carried out artificial neural network and machine learning and carried out computational intelligence research in the field of intelligent information system. Research on neural computing model and physiological and psychological basis of vision and hearing mainly explores the perception mechanism of vision and hearing from the perspective of physiology and psychology, so as to provide basic theories and methods for visual and auditory information processing.

The most important thing in teaching is vision and hearing. Vision and hearing are important perceptual systems that people rely on in verbal communication. But in fact, because of many environmental factors, people in audio-visual and oral teaching cannot fully grasp the information. Digital sensor technology can strengthen information, so digital sensor began to appear in language audio-visual and oral teaching. Digital sensor equipment plays a better role in language audio-visual and oral teaching. Students using digital sensor equipment can learn better, so as to greatly improve the teaching effect.

Strengthening the processing of relevant visual and auditory information helps to ensure that the receiver obtains more accurate information, so as to improve the learning effect and reduce the impact on the environment. This paper creatively puts forward an experiment to explore the role of digital sensors in language audio-visual teaching, so as to provide reference for the application of digital sensors in the future. The impulse response function in sensor technology is introduced. The speech time domain envelope and time-varying nozzle area of the sensor device are calculated. Through auditory attention decoding fusion technology and sensor auditory attention conversion detection method,

auditory attention transfer detection based on line of sight rotation estimation is realized. At the same time, the characteristics of sensor heog signal are analyzed. The results show that the algorithm proposed in this paper has good results.

This paper is mainly divided into five parts. The first part is the background introduction. The second part is literature review, which mainly introduces the relevant research results of digital sensors. The third part is the introduction of related algorithms, mainly introducing the digital sensor and related technologies and algorithms. The fourth part is empirical analysis, which proves the advantages of digital sensor in language audio-visual and oral teaching through various studies. The fifth part is summary and analysis.

2. Related Work

In fact, sensor technology alone refers to information classification and processing technology here. The commonly used classifiers include decision tree (DT), support VEC tor machine (SVM), and neural network. Alhumayani et al. used the correlation coefficient, variance, frequency domain entropy, and mean value of acceleration sensor as behavior characteristics for the first time and combined with DT algorithm to realize the classification of 20 behaviors [1]. Yang realized the recognition of five gait based on FFT coefficient, quartile difference, and SVM algorithm [2]. Feng and Pan proposed using principal component analysis technology to reduce the dimension of mean, variance, and frequency domain features and using decision tree as classifier for classification [3]. In addition, Luo and Xiao use the compressed sensing method to efficiently process low dimensional sampling data for classification [4, 5]. In 2013, researchers from the University of Catalonia in Spain and the University of Genoa in Italy identified six behaviors such as walking, sitting, and standing by using the built-in sensors of mobile phones and disclosed the data set recorded in the experiment to volunteer researchers [6]. Because the traditional methods extract behavior features manually, there will be some empirical deviation in the description of behavior, and a lot of manual intervention is required, so the results are not very ideal. The development of big data technology and deep learning makes it possible to automatically learn the most distinctive behavior characteristics from massive raw data [7–9]. Jiang and Yin proposed a behavior recognition method using deep convolution neural network DCNN [10]. In this method, the data of gyroscope and acceleration sensor are converted into frequency domain signals through sliding window and Fourier transform processing, and the model is trained through supervised learning to obtain a feature extractor to identify human behavior. The experimental results show the superiority of this method on three public data sets (UCI, USC, and SHO) [11–13]. As an interpretable probability graph model, restricted Boltzmann machine is mainly divided into deep belief network (DBN) and deep Boltzmann machine (DBM) [14–17]. Among them, DBN belongs to a generation model, which can extract feature representation from unlabeled high-dimensional sensor data. For example, Fang and Hu and Bhattacharya and Lane used this method to extract irrelevant data in the data, realizing nonlinear dimensionality reduction of high-dimensional data [18, 19]. DBM learns features from sensor data by stacking undirected bipartite graphs. This method mainly uses sparse feature technology to reduce the sensitivity of data and combines cross-correlation feature extraction and sensor fusion methods [20–22]. Deep auto-encoder (DAE) includes two parts: encoder and decoder. First, it uses the encoder to find the correlation characteristics between sensor data, converts the high-dimensional data space to the low-dimensional space, and then uses the error back propagation algorithm to reconstruct the sensor sample data in the decoder [23]. Many research works use this method to reduce the feature dimension and use the method of approximate identity and compressed version to screen the feature vector, using DAE [24, 25]. It can be seen that there are many applications of digital sensors in information processing, but there are still few direct biological information enhancement at present. This paper also explores the enhancement of human information classification and collection by digital sensor technology.

3. Method

3.1. Introduction of Impulse Response Function in Sensor Technology. For the speech time domain envelope $s(t)$ sampled in discrete time $t(t = 1, \cdots, T)$ and the EEG signal $r(t, n)$ sampled in EEG channel $n(n = 1, \cdots, N)$, assuming that the auditory processing of mapping speech features to neural response is a linear time invariant system, the impulse response $w(\tau, n)$ of the forward system can be used to describe the system, as shown in the following formula:

$$r(t, n) = \sum w(\tau, n)s(t - \tau) + \varepsilon(\tau, n), \tag{1}$$

where $\varepsilon(\tau, n)$ represents the system residual. Impulse response $w(\tau, n)$ can be regarded as a set of temporal spatial filters, called TRFs. When solving TRFs, the inverse correlation method and ridge regression strategy can be used to solve the ill posed problem encountered in matrix inversion, see the following formula:

$$w = \left(S^T S + \lambda M\right)^{-1} S^T r. \tag{2}$$

The matrix $S \in R^{T \times \tau_{\text{win}}}$ is composed of the delay sequence of speech time domain envelope $s(t)$, and τ_{win} is the delay length; matrix $r \in R^{T \times N}$ is multichannel EEG data; matrix $M \in R^{T_{\text{win}} \times \tau_{\text{win}}}$ is regularization matrix; λ is a ridge parameter, which can be optimized by cross validation. The optimization index is the correlation coefficient (Pearson correlation coefficient) between the predicted $EEG\hat{r}(t, n)$ and the real $EEGr(t, n)$.

Similar to the forward system, if the system inversely mapped from EEG to speech time domain envelope is also a linear time invariant system and its impulse response is $g(\tau, n)$, the system can be described by formulas ((1))–((3)):…∗as the following formula

$$\hat{s}(t) = \sum\sum r(t + \tau, n)g(\tau, n). \tag{3}$$

$\hat{s}(t)$ represents the reconstructed speech time domain envelope, which can be regarded as linear regression of speech time domain envelope. Impulse response $g(\tau, n)$ is also a set of spatiotemporal filters, which integrates the neural response of specific delay τ and then sums the integrated signal between channels to obtain the reconstructed speech time domain envelope. The filter is also called a decoder. The solution of the decoder is similar to that of TRFs, as shown in the following formula:

$$g = \left(R^T R + \lambda M\right)^{-1} R^T s. \tag{4}$$

The matrix $R \in R^{T \times N - \tau_{\text{win}}}$ is composed of the delay sequence of multichannel EEG data $r(t, n)$, and τ_{win} is the delay length; matrix $s \in R^{T \times 1}$ is the time domain envelope of speech; matrix $M \in R^{N-\tau_{\text{win}} \times N-\tau_{\text{win}}}$ is the regularization matrix. The ridge parameter λ can be optimized by cross-validation. The optimization index is the correlation coefficient (Pearson correlation coefficient) between the reconstructed speech envelope $\hat{s}(t)$ and the actually noticed speech envelope $s_{att}(t)$, which is called reconstruction accuracy.

3.2. Speech Time Domain Envelope and Time-Varying Mouth Area Calculation of Sensor Equipment. For each clean speech signal, we first filter the speech signal by band through the bandpass filter bank (a total of 8 frequency bands). The center frequency of the bandpass filter bank is evenly distributed on the equivalent rectangular bandwidth (ERB) scale of 150–8000 Hz.

The relationship between ERB scale and frequency F_i (kHz) is shown in the following formula:

$$\text{ERB}_N \text{ number} = 21.4 \log_{10} 4.37 F_i + 1, \tag{5}$$

where F_i represents the passband center frequency of the $i\,(i = 1, \cdots, N)$ th filter. The output $x_i(t)$ of each filter is Hilbert transformed to obtain the analytical signal $H(x_i)$. Considering the nonlinear compression characteristics of the cochlea, we further model and compress the $H(x_i)$ (0.3 power) and conduct 8 Hz low-pass filtering to obtain the speech time domain envelope $e_i(t)$ of the corresponding frequency band, as shown in the following formula:

$$e_i = \text{LP}\left(|H(x_i)|^{0.3}\right), \tag{6}$$

where LP(•) represents low-pass filter operator. Finally, the speech time domain envelope $e\,(t)$ can be obtained by averaging the speech time domain envelope of all frequency bands, as shown in formula (7). For the convenience of writing, we will abbreviate it as speech envelope later.

$$e = \frac{\sum_N^{i=1} e_i}{N}. \tag{7}$$

The manually set features contain people's prior knowledge of the main features of the signal, but they cannot reflect all the features of the signal, such as the slope and small fluctuation of HEOG signal in the impulse process. Therefore, this paper further takes the heog signal waveform as the input and uses the DNN classifier in the sensor technology to automatically learn and extract the features related to the scanning angle in the signal. Considering that the time-domain characteristics of HEOG signal are most related to the scanning angle, and the cyclic neural network based on long short-term memory (LSTM) structure has good time series analysis ability, we use the classifier based on LSTM network. Because the signal waveform in this task is relatively simple, we use the cascade structure of single-layer LSTM (number of neurons: 64), single-layer fully connected network (FCN) (number of neurons: 12), and softmax classifier to map the HEOG waveform (sequence length is 160, corresponding to 5 s) to the scanning angle label.

In the training of LSTM, we use multiclassification cross-entropy as the loss function. Assuming that the one hot code of the sample category label is $y = [y_1, y_2, \cdots, y_c]$, and the prediction result of the LSTM classifier is $\hat{y} = [\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_c]$, where C is the number of categories, the loss function is shown in the following formula:

$$\text{Loss}_{\text{CE}} = -\sum_c^{i=1} y_i \log\left(\hat{y}_i\right) = -\log\left(\hat{y}_c\right). \tag{8}$$

3.3. Fusion Technology of Auditory Attention Decoding and Auditory Attention Conversion Detection Method of Sensor. We can match the output of AAD method based on auditory selective attention neural mechanism (measured by EEG) with the auditory attention object of the listener, but the accuracy of linear decoding algorithm currently used is low and the detection delay of attention conversion is large (5–10 s).

Because the advantages and disadvantages of AAD and aad methods are complementary, the visual behavior-based aasd method (measured by heog and NEMG) has low alarm leakage rate and small detection delay (<2 seconds), but its output can only reflect the conversion of auditory attention and cannot correspond to the object of auditory attention. Therefore, this paper puts forward some "early warning and correction" strategies. Aad issues early warning to AAD. The output of the AAD method will be used to guide the implementation of the AAD method. When line of sight rotation is detected, the AAD algorithm will be executed once to detect the auditory attention object; otherwise, the final detection result will be maintained. This strategy can significantly reduce the amount of computation of the AAD algorithm. Aad directs aad to correct the test results. In order to alleviate the error propagation problem caused by aad errors, when the line of sight rotation is not detected for a continuous period of time, the AAD algorithm is executed once to compare and correct the new and old detection results.

Its formal expression is as follows.

That is, when no attention transition is detected in consecutive M frames, set the correction state $c(k)$ to 1. After calculating the warning and correction state of the k frame, the AAD operation state $p(k)$, that is, the final

detection result of the auditory attention object ID (k), can be calculated, as shown in the following formulas.

$$p(k) = a(k) + c(k), \tag{9}$$

$$\mathrm{ID}(k) = \begin{cases} \mathrm{ID}(k-1), & \text{if } p(k) = 0, \\ \boldsymbol{d}(k), & \text{if } p(k) = 1. \end{cases} \tag{10}$$

Note that the AASD algorithm is the operator $v(\bullet)$, and the AAD algorithm is the operator $d(\bullet)$. Boolean value $v(k)$ is the AASD state of frame k (1 indicates rotation is detected), Boolean value $a(k)$ is the early warning state of frame k (1 indicates early warning), Boolean value $c(k)$ is the correction state of frame k (1 indicates correction should be performed), and constant M is the correction window length; Then, it satisfies the following relationship formulas:

$$a(k) = \boldsymbol{v}(k-1), \tag{11}$$

$$c(k) = \neg\left(\sum_{M}^{i=1} v(k-i) + \sum_{M}^{i=1} a(k-i)\right). \tag{12}$$

In order to solve the problem of gradient saturation, ReLU activation function is selected in this paper. ReLU activation function was introduced into neural network by Nair and Hinton in 2010. It is essentially a piecewise function, which is defined as follows;

$$\mathrm{ReLU}(x) = \max\{0, x\} = \begin{cases} x & x \geq 0, \\ 0 & x < 0. \end{cases} \tag{13}$$

In this paper, L_2 regularization is used as the regularization method of convolutional neural network model. L_2 regularization is a very common model regularization method in traditional machine learning and deep neural network models. Two regularization techniques are also commonly used in depth model to regularize its convolution layer and classification layer. Finally, the overfitting phenomenon is avoided by constraining the model complexity of convolutional neural network. Assuming that the network layer parameter to be regularized is ω, the form of L_2 regularization term is as follows:

$$L_2 = \frac{1}{2}\lambda\|\omega\|_2^2. \tag{14}$$

Among them, λ reflects and controls the complexity of the regular term. The larger the value, the greater the model complexity, and vice versa. In the actual use of L_2 regular term, the regular term is usually added to the objective function set in advance, and the overall objective function is used to complete the error back propagation in the convolutional neural network model, so as to guide the training of convolutional neural network model by changing the regular term. L_2 regularization is commonly called "weight attenuation" in deep learning. In addition, L_2 regularization is also called ridge regression or regularization in machine learning.

Through the normalization of square difference, the distribution variance of input and output data can be consistent. The specific formula is as follows:

$$\mathrm{Var}(s) = \mathrm{Var}\left(\sum_{n}^{i} \omega_i x_i\right), \tag{15}$$

where s represents the output result of the layer network before nonlinear transformation, ω represents the layer parameters, and x represents the layer input data.

4. Results and Discussion

4.1. Auditory Attention Transition Detection Based on Line of Sight Rotation Estimation under Sensor Operation. This section will analyze the signal characteristics of heog, NEMG, and IMU sensors under the conditions of head fixation and head rotation, respectively, and explore the feasibility of applying these signal characteristics to AASD tasks by establishing the mapping relationship between these signal characteristics and the rotation angle of line of sight. Estimation of line of sight rotation is based on heog under fixed head. First, the experiment of line of sight rotation under the condition of fixed head is carried out in this section. Subjects induced sensor heog by paying attention to visual stimuli at different horizontal angle positions on the display. Through the feature analysis of heog signal, we will design a suitable classification algorithm to estimate the corresponding line of sight rotation angle. Four subjects with normal vision or corrected to normal vision (1 female, age range 22–25 years) participated in the experiment.

The experiment was conducted in the copper mesh shielded sound insulation room (IAC acoustics) of the speech and hearing research center of Peking University. In this experiment, the subject's head was fixed by the head support (wearing the sensor at the same time), and there was a 34.5-inch display (AOC) 0.4 m in front of the head. Visual stimulation is a red dot presented through the display, and the background of the display is black. The possible positions of red dots are located in the horizontal area on the display at the same height as the subject's eyes. There are five positions, which are 0° in front of the subject, 45° and 30° to the left in front of the subject, and 30° and 45° to the right in front of the subject. During the experiment, the red dot will only appear at a certain position at any time, and the subjects are required to always look at the red dot. We control the position change of the red dot to instruct the subjects to make horizontal scanning behavior, so as to induce heog. In one trial, the subjects were instructed to produce all 20 saccade behaviors once by setting the position change order of red dots. At the beginning of each trial, red dots will randomly appear in one position and continue to appear for 5 s and then jump to another position and continue to appear for 5 s. A total of 10 trials were conducted for each subject. The duration of each trial was 1 minute and 45 seconds. The presentation order of red dots in each trial

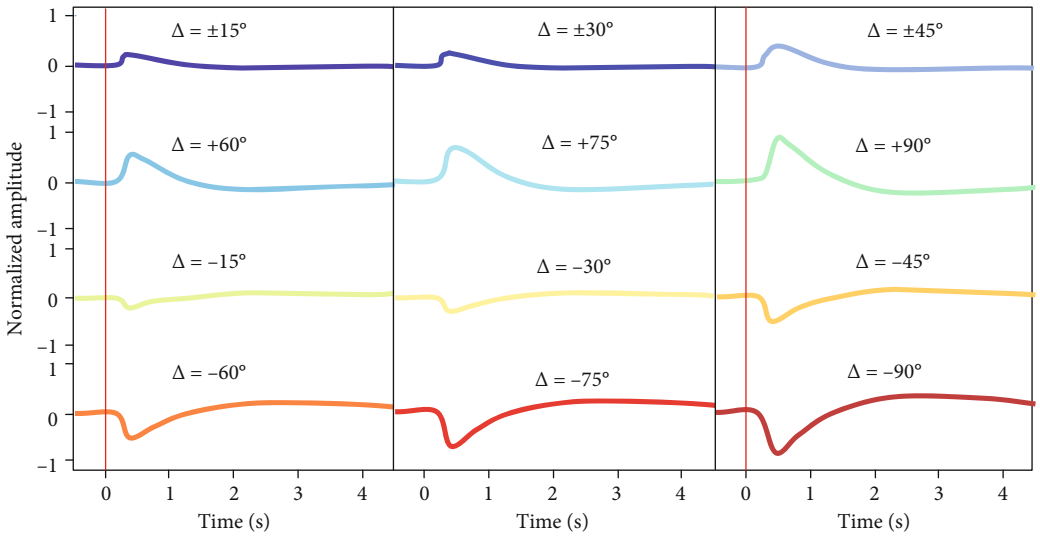

Figure 1: Normalized average sensor HEOG waveform (12 types of glance angles).

was randomly controlled. At the end of each trial, subjects will rest for a period of time to ensure sufficient attention. Before the formal experimental process, each subject should be trained to be familiar with the stimulation materials and experimental process. The whole experiment lasted about 0.5 hours for each subject.

4.2. Sensor Data Acquisition and Preprocessing. The recording and storage of sensor heog data are carried out through the NeuroScan synamps2 system. We used a pair of bipolar electrodes to record heog, which were placed outside the eyes. In addition, the midline of the forehead is used as the ground electrode, and the tip of the nose is used as the reference electrode. In order to ensure the data quality, the impedance between electrodes is kept below 5 k0hm. Heog is amplified (magnification: 20000), online bandpass filtered (frequency range: 0.15–100 Hz) and sampled (quantization accuracy: 16 bits, sampling rate: 250 Hz), and then stored for offline analysis. The generation and recording modes of trigger signal are similar to that in Section 3.2 of this paper. Each time the red dot changes its position, a trigger signal is generated to facilitate subsequent data preprocessing.

We use Matlab for digital preprocessing of heog signal, which is similar to previous work. For each test, we successively extract 20 heog signals according to 20 trigger signals with a window length of 5 s (0.5 s before the trigger signal to 4.5 s after the trigger signal), corresponding to each saccade behavior. Because the basic form of heog signal is slowly changing time-domain fluctuation, and its main components are located in the low-frequency band, the signal is further downsampled to 32 Hz, and 10 Hz low-pass filtering and baseline correction is performed. Finally, in order to ensure that the heog amplitude values of all subjects are comparable, we normalized all data of each subject within subject to ensure that the maximum amplitude is 1, and the minimum is -1 in all heog waveforms of each subject. Through the analysis of preexperiment, when the starting and ending positions of rotation are different but the scanning angle is the same (for example, the rotation amount of line of sight direction is +45° when turning from -45° to 0° and from 0° to 45°), the heog waveform shape is almost the same. Therefore, we classify the same rotation angle into one category, and finally, get the heog data corresponding to 12 types of scanning angles.

4.3. Characteristic Analysis of Sensor Heog Signal. The normalized average heog waveform corresponding to 12 types of scanning angles (Δ = ± 15°, ±30°, ±45°, ±60°, ±75°, and ±90°) is shown in Figure 1 (the error interval represents the standard deviation, and the red vertical line represents the time of trigger signal, which is the same later). It can be seen that in addition to the difference in amplitude, the heog waveform under each scanning angle is very similar. After the indication of saccade is given, the amplitude of heog begins to change significantly after about 0.2 S. At about 0.4 s, these results show that the extreme value of heog amplitude is an important feature of heog signal, which is consistent with previous findings. We further calculated the amplitude extreme value of each heog signal of each subject within the 2S segment at the beginning of the signal, and the average result is shown in Figure 2. It can be seen that the extreme value of heog waveform increases monotonically with the increase of rotation angle, indicating that it can be used as a manually set feature to estimate the change angle of line of sight orientation.

4.4. Estimation of Line of Sight Rotation Based on Heog Waveform and DNN in Sensor. The classification accuracy of the sensor built-in classifier based on heog amplitude feature is 81.8 ± 2.2%, which shows that the scanning angle can be better estimated by using heog amplitude feature under the condition of fixed head. The confusion matrix of classification results is shown in Figure 3(a). The elements i and j of the confusion matrix represent the proportion of class i samples classified to class j, which is the same later. It can be seen that confusion mainly occurs in the same

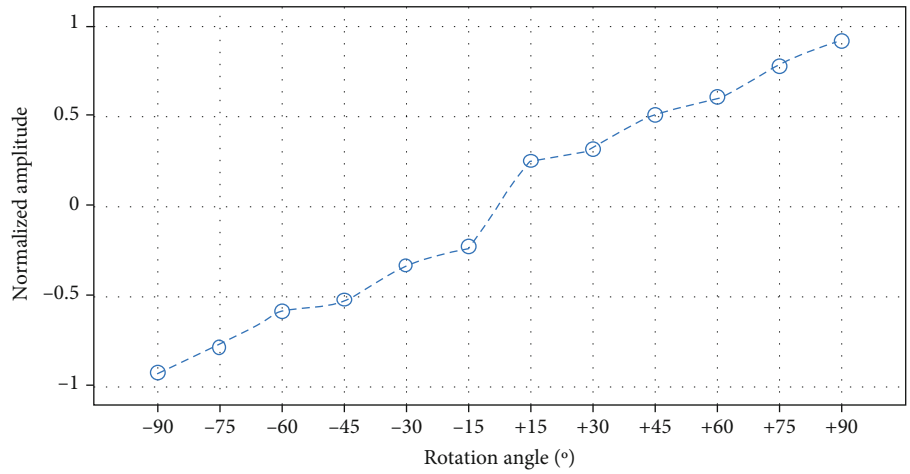

Figure 2: The normalized HEOG amplitude characteristics of the average sensor change with the saccade angle.

rotation direction and between adjacent rotation angles, especially between +60° and +45° and -60° and -45°, which is consistent with the trend in Figure 2. This shows that the spatial resolution of heog using amplitude features is still limited.

The classification accuracy based on heog waveform and LSTM classifier is 90.9 ± 2.0%, which is significantly higher than that of the sensor's own classifier. The confusion matrix of classification results is shown in Figure 3(b), and it can be seen that the degree of confusion is significantly reduced. This shows that compared with manually setting the feature (i.e., the amplitude extreme value of heog), the time-domain feature automatically learned after LSTM network enhancement reflects the information related to the scanning angle more finely and comprehensively and can be used as an enhanced version of the sensor.

4.5. Estimation of Line of Sight Direction Rotation of Sensor Based on HEOG and NEMG under Head Rotation. For example, although the algorithm using heog to estimate the scanning angle has achieved high accuracy under the condition of fixed head, the performance of the algorithm will be significantly reduced under the condition of head rotation. This is because under the condition of head rotation, the change of line of sight orientation no longer depends only on saccade, but also on head rotation, and the strategies of the two kinds of behavior are variable. In this experiment, subjects can perform natural saccade and head rotation when completing the task of line of sight rotation. We will use heog and NEMG to measure saccade and head rotation, respectively. By analyzing the characteristics of several sensor signals, a suitable classification algorithm is designed to estimate the corresponding line of sight rotation angle.

17 subjects with normal vision or corrected to normal vision (4 women, age range 21–28 years) participated in the experiment. All subjects ignored neurological diseases. The experimental process has been approved by the institutional review committee of Peking University, and the informed consent of each subject has been obtained. The experiment was conducted in the copper mesh shielded sound insulation room (IAC acoustics) of the speech and hearing research center of Peking University. The subjects (with sensors) had a 17-inch display (DELL) at 0.85 meters on the left and right in front of them. Visual stimulation is a red dot presented by the two monitors, and the background of the monitors is black. The possible positions of red dots are located in the horizontal area on the display at the same height as the subject's eyes. There are six positions in total, three of which are 45°, 30°, and 15° to the left in front of the subject, and the other three are 15°, 30°, and 45° to the right in front of the subject. During the experiment, red dots will only appear at a certain position of a display at any time. The subjects were asked to always turn their eyes towards the red dot. We instructed the subjects to turn their eyes by controlling the position change of the red dot. Before the change of position, the subjects were asked to minimize head movement and other body movements. When the transition occurred, the subjects were allowed to perform natural saccade and head rotation. In this experiment, in order to ensure that the subjects have obvious saccade and head movement behavior, the position of the red dot will only change between the two displays. We have set up a total of six line of sight orientation changes with different start and end positions: ±30°, ±60°, and ±90° (corresponding to -15° to +15°, +15° to -15°, -30° to +30°, +30° to -30°, -45° to +45°, and + 45° to -45°). In one trial, the red dot is continuously converted between two fixed positions for 40 times. The duration of continuous presentation at a certain position is still 5 s, and the duration of each trial is 3 minutes and 40 seconds. Therefore, two of the six line of sight rotation occurred in one trial, 20 times each. A total of 3 different trials were conducted for each subject to produce all six gaze orientations. At the end of each trial, subjects will rest for a period of time to ensure sufficient attention. Before the formal experimental process, each subject should be trained to be familiar with the stimulation materials and experimental process. The whole experiment lasted about 0.25 hours for each subject.

4.6. Characteristic Analysis of HEOG, NEMG, and IMU Signals in Sensor. In this section, the extreme value of heog waveform is still used as a manually set feature to estimate

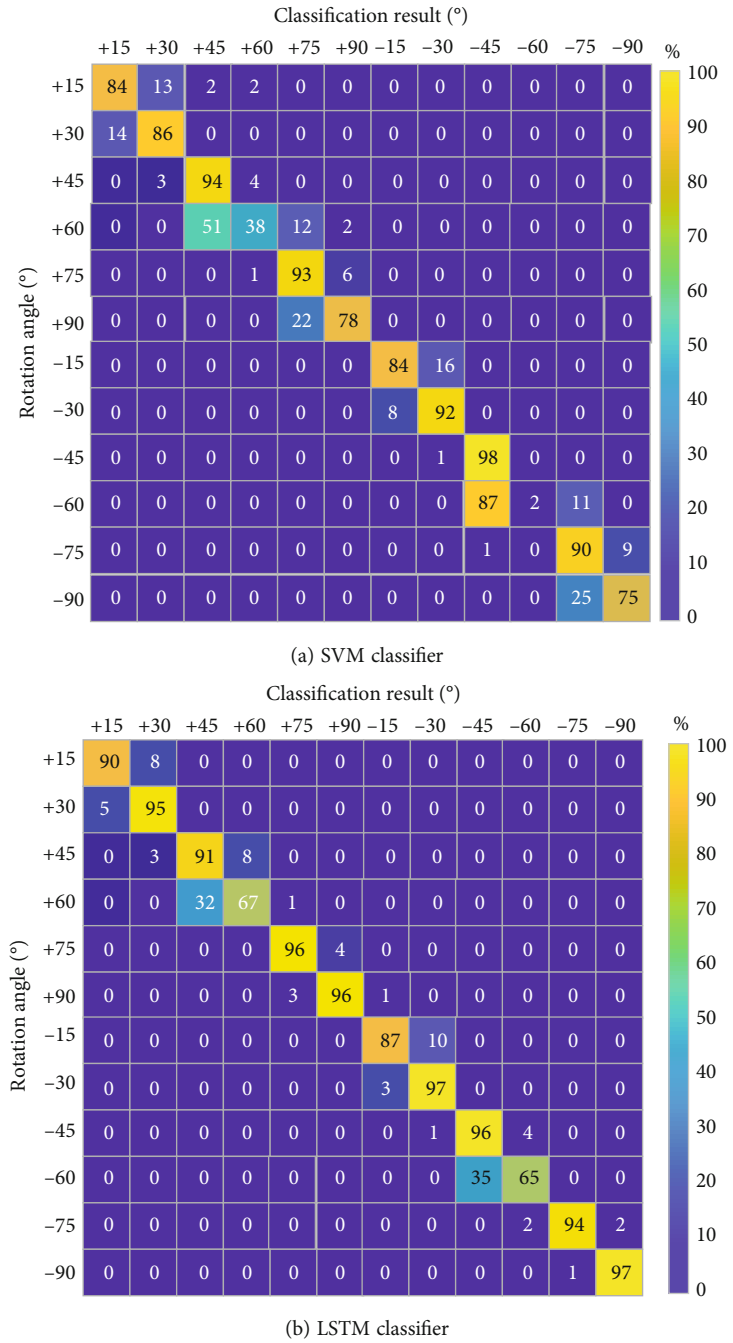

Figure 3: Confusion matrix of the classification results of the sensor's own classifier and the LSTM enhanced sensor classifier under the condition of fixed head (%).

the change angle of line of sight. When the line of sight rotates, different from the slow fluctuation of heog signal, NEMG signal will show obvious amplitude change in continuous and rapid neural discharge mode. It can be seen that the relationship between NEMG amplitude change trend and head rotation direction is expected, that is, within 1 s after head rotation, NEMG amplitude of contralateral SCM will increase and remain stable, while NEMG amplitude of ipsilateral SCM will decrease and remain stable. Therefore, the amplitude change of NEMG is its most significant signal feature. This characteristic will be described by calculating the short-term energy of NEMG signal. Figure 4 shows the normalized average heog waveform under head rotation. It can be seen that the heog waveform is similar to the heog waveform under the condition of fixed head.

We calculate the root mean square (RMS) of the signal to represent the signal energy and select the frame length of 0.1 s and the frame shift of 0.05 s. On this basis, the short-time energy of each NEMG is normalized, and the average short-time energy in the first 0.5 s is classified as 1. Figure 5 shows the change of normalized NEMG short-time energy with time under various line of sight rotation

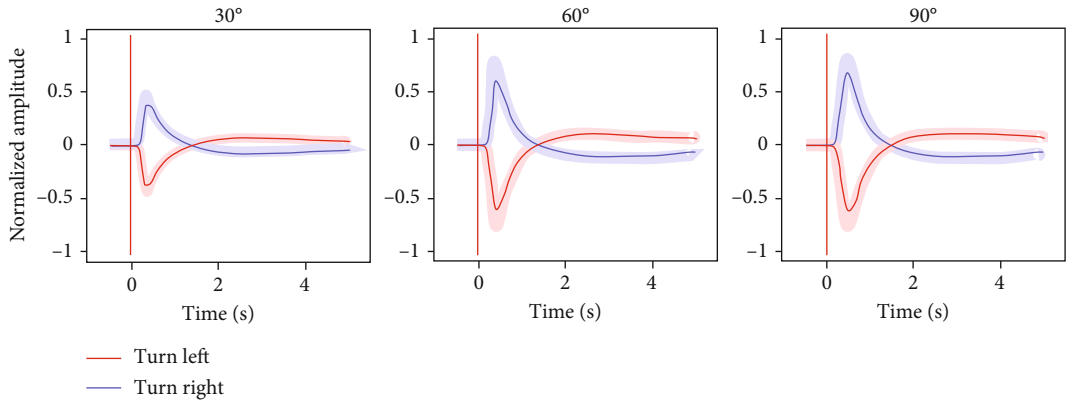

Figure 4: Normalized HEOG waveform of average sensor (6 types of sight rotation).

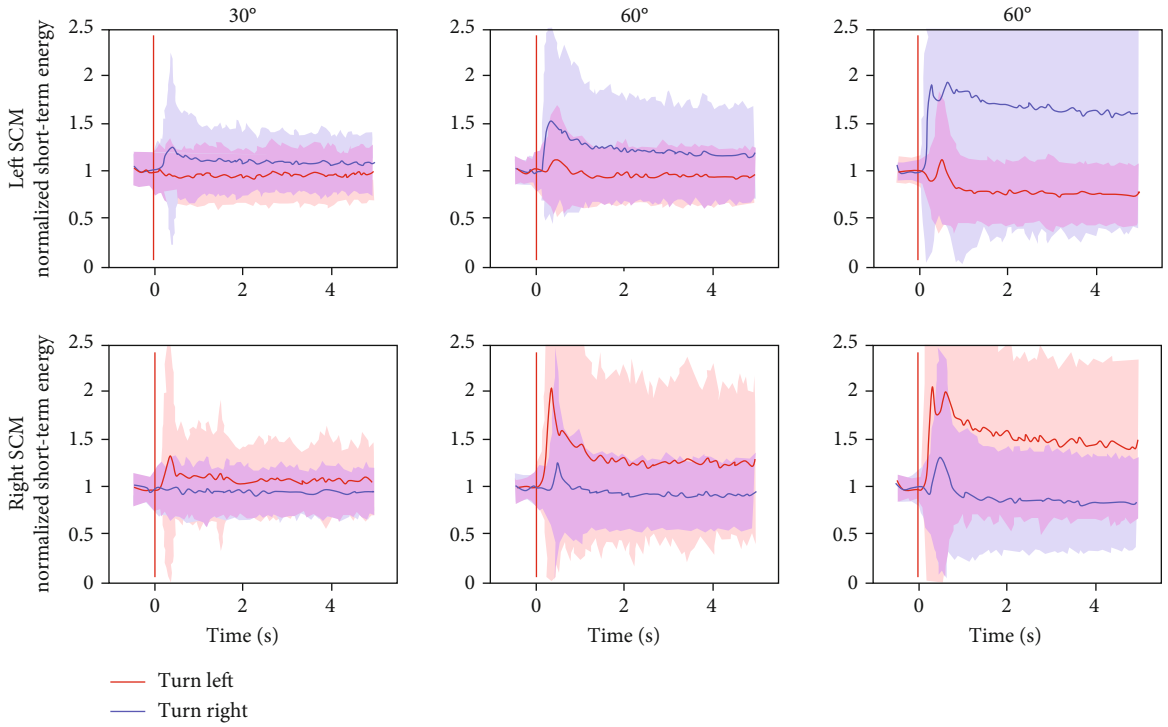

Figure 5: Normalized sensor NEMG short-term energy diagram (6 types of sight orientation changes).

angles. It can be seen that after the indication of line of sight rotation is issued, the NEMG short-time energy of the opposite SCM begins to change significantly after about 0.1 s, reaches the extreme value at about 0.4 s, and then remains at a high level, while the NEMG short-time energy of the ipsilateral SCM has little change. Similar to the trend of HEOG, the increment of NEMG short-time energy value increases with the increase of line of sight rotation angle. The results show that the extreme value of short-term energy can be used as a manually set feature to estimate the change angle of line of sight orientation.

We further analyze the differences between the results of different classifiers. It can be seen that under all input conditions, the accuracy of FCN classifier is not different from that of the sensor's own classifier, which may be because the network structure of FCN is relatively simple, and its automatically extracted signal features are similar to those set manually. Under all input conditions, the accuracy of LSTM classifier is the lowest, which may be because there is certain noise interference in the sensor signal under the condition of head rotation, and LSTM network is sensitive to the time information in the signal waveform, so its ability to extract features will be disturbed by noise. This also explains that the accuracy of univariate heog input (49.3%) is significantly lower than that in head fixation experiment (90.9%). Under the condition of univariate input, the performance of digital optimization classifier is similar to that of SVM and FCN classifier. Under the two bivariate input conditions, the performance of digital optimization classifier is the best, which is 72.6% (HEOG and NEMG) and 93.3%

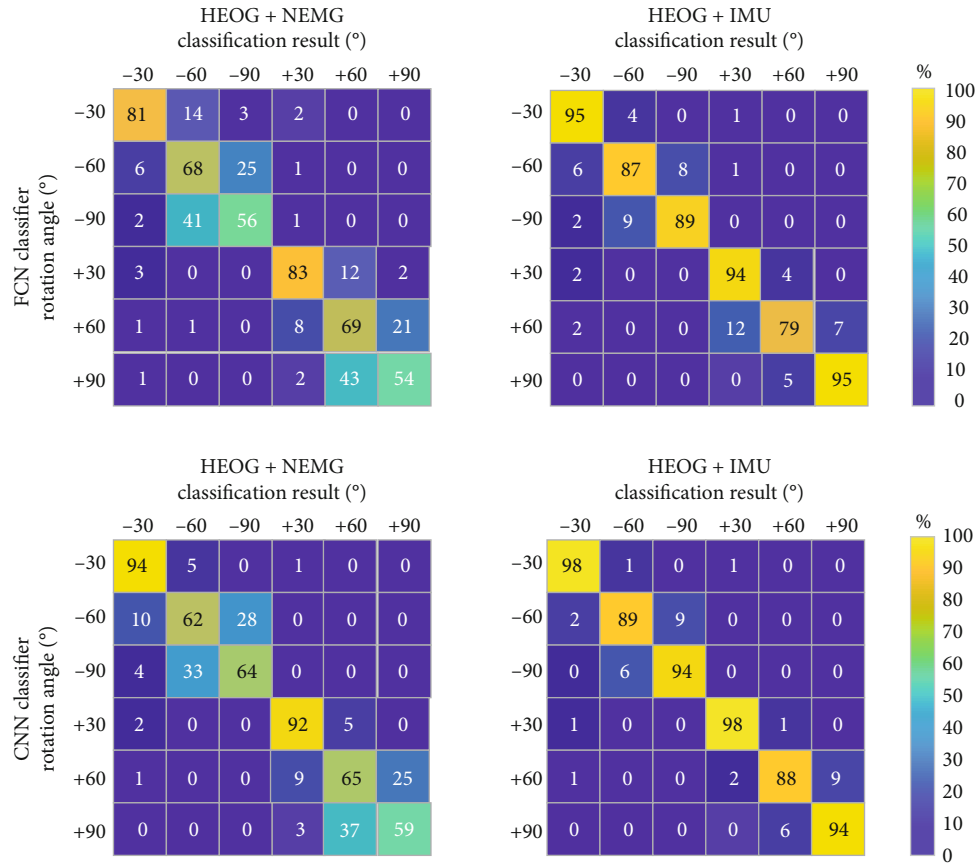

Figure 6: Confusion matrix of classification results of sensor with classifier and digital optimized classifier under head rotation condition (%).

(HEOG and IMU), respectively, indicating that digital optimization has better feature extraction ability than FCN and LSTM. This may be because the convolution kernel can observe the signal waveform in a certain time window at one time and has stronger ability to integrate the features between channels, so it reduces the possibility of noise interference to the digital optimization classifier. The confusion matrix of digital optimization classification results under two bivariate input conditions is shown in Figure 6, and the confusion mode is similar to the classifier provided by the sensor. The results also show that the algorithm can correctly judge the rotation direction under the condition of bivariate input. Compared with IMU, NEMG has poor resolution when measuring head rotation behavior.

In general, these results show that under the condition of audio-visual matching, the Los rotation estimation based on heog and NEMG can more accurately reflect the information related to auditory attention conversion, such as rotation time and rotation angle. Considering that the AASD task does not require accurate line of sight rotation angle information, we can reduce the line of sight rotation angle estimation to line of sight rotation detection, that is, the rotation label output by the classifier (class 7: 0°, ±30°, ±60°, and ±90°) is changed to rotation label (class 2: rotation, no rotation), so that while meeting the requirements of AASD task, the advantage of low detection delay (2 s) is also retained. Based on this change, this section will continue to explore the feasibility of AASD task based on line of sight rotation detection.

Based on the results in Figure 7, after remapping the output label of the classifier into rotation (±30°, ±60°, and ±90°) and nonrotation (corresponding to 0°), the experimental results of AASD task can be obtained, and the confusion matrix is shown in Figure 8. It can be seen that when two variables heog and NEMG are input, FCN is easier to misjudge the rotation condition as no rotation than digital optimization classifier (FCN: 6.3%, digital optimization: 3.1%) (see Figure 9). This missing alarm means that AASD algorithm fails to guide the calculation of AAD, which will affect the detection of auditory attention objects. On the contrary, false alarm has less impact, because AAD algorithm can correct it. In order to further quantify the performance of each algorithm in the AASD task under various input conditions, we calculate three indicators: $F1$ value, missed alarm rate and false alarm rate according to the confusion matrix. $F1$ value is the harmonic average of accuracy rate (the proportion of samples divided into positive examples) and recall rate (the proportion of samples divided into positive examples). The model can be comprehensively evaluated. The missed alarm rate is the proportion of the missed positive cases in all the positive cases. False alarm rate is the proportion of samples judged as positive cases, which are actually negative cases. It can be seen that the performance of the digitally optimized classifier is better, which has higher

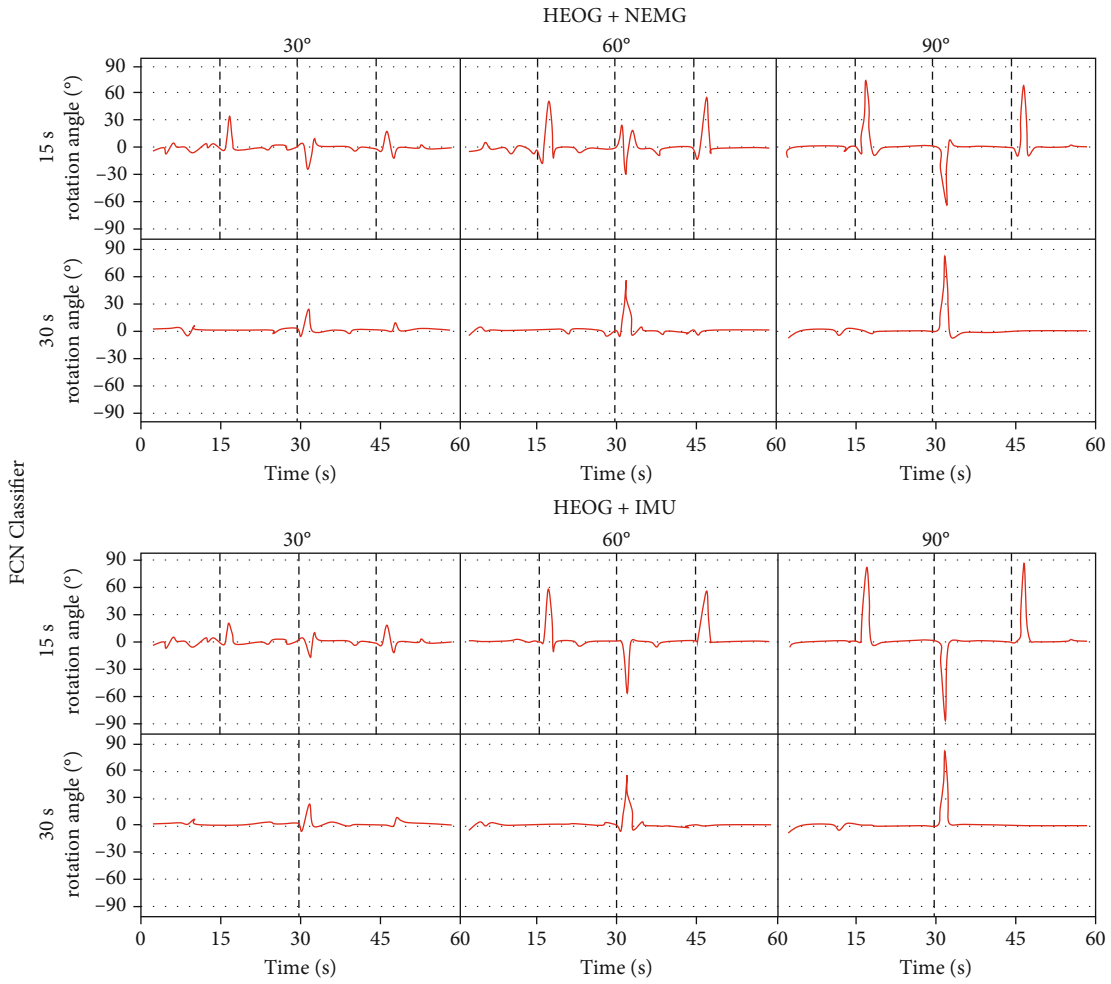

Figure 7: Result of continuous line-of-sight rotation angle estimation results during sensor trials.

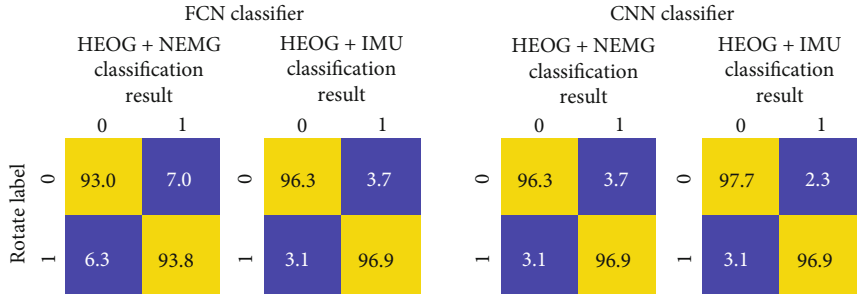

Figure 8: Confusion matrix of classification results of FCN and digital optimization classifiers (%).

$F1$ value, lower false alarm rate, and lower false alarm rate than the FCN classifier. This is similar to the comparison results of the two classifiers in the previous section, which proves that in this task, the digital optimization network has stronger feature extraction ability and is more robust to noise interference. In addition, although the results when using bivariate heog and NEMG are still worse than bivariate heog and IMU on the whole, the gap is not large, especially the missing alarm rate (both 3.1%) under the condition of using digital optimized classifier is almost no difference. The results show that the proposed AASD algorithm based on heog and NEMG is feasible.

One advantage of the fusion strategy is that it avoids the EMG artifact interference caused by saccade and head rotation. Figure 10 shows the signal waveforms of EEG, EOG, and NEMG recorded in a trial with an attention conversion interval of 15 s. In order to show the interference of EMG artifact on EEG, we selected some of the most representative EEG channels for display. They are either closer to the EMG artifact source in spatial position (for example, FPZ is close

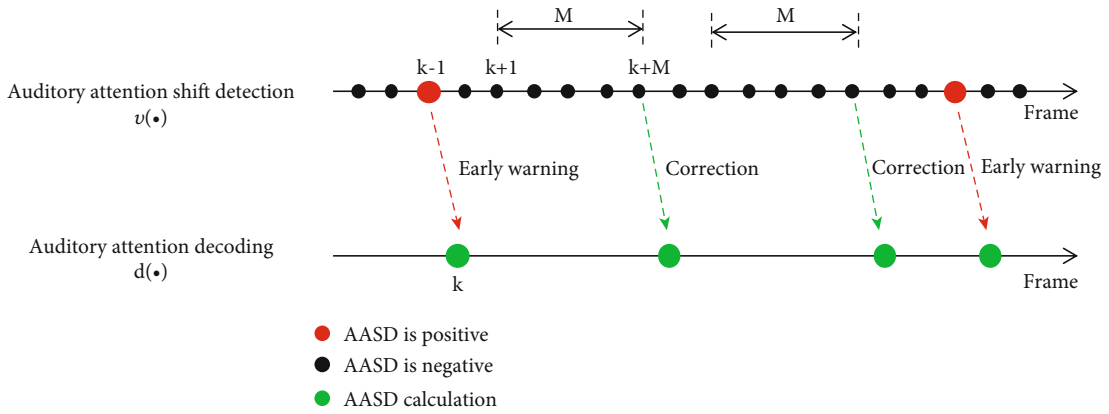

Figure 9: Schematic diagram of the fusion strategy of auditory attention decoding and conversion detection methods.

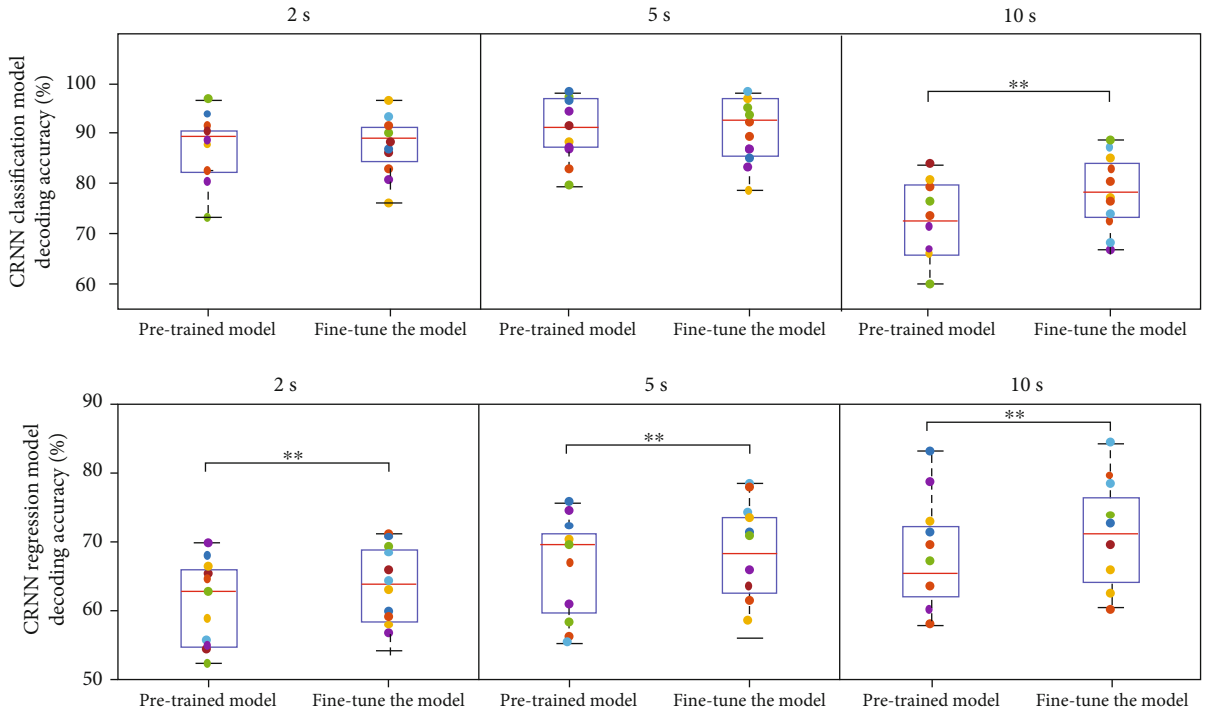

Figure 10: The decoding accuracy of CRNN classification and regression model before and after fine-tuning.

to the saccade and blink artifact source—eye muscle, T7, and T8 are close to the head rotation artifact source—SCM), or they contribute greatly to AAD (for example, FC5 and FC6 are located in the temporal lobe, and Oz is located in occipital lobe). In the channel FPZ close to the eye muscle, obvious interference from heog and VEOG can be seen; obvious interference from NEMG can be seen in channels T7, T8, and OZ close to SCM; FC5 and FC6 channels located in the temporal lobe may be disturbed by both heog and NEMG; the channel CZ located at the top of the skull is hardly disturbed by these artifacts due to its long distance. In general, almost all EEG electrodes are subject to artifact interference at the time of saccade and head rotation, so EMG artifact interference mainly comes from heog and NEMG induced by saccade and head rotation. In addition, the delay introduced by their volume conduction through the head is very small. This shows that the moment when EEG is disturbed by EMG artifact is the moment when AASD output is positive. Under the fusion strategy proposed in this paper, EEG segments interfered by EMG artifact will not participate in the calculation of AAD, so EMG artifact interference has no effect on the detection results of auditory attention objects.

The experimental results of model fine-tuning of CRNN regression and CRNN classification model under different decoding window lengths are shown in Figure 10, which shows the average decoding accuracy (box diagram) and individual decoding accuracy (scatter diagram) of the data of 12 subjects. It can be seen that under all decoding window length conditions, the average decoding accuracy of the fine-

tuning model is higher than that of the pre training model, but the improvement is significant only under the condition of 10 s ($t = -3.411$, $P = 0.006$). For the other two window length conditions, there may be two reasons why the gain caused by fine tuning is not significant. First, the accuracy of the pre training model itself is high and has ceiling effect; second, the classification model mainly relies on the spatial selective attention feature in EEG, which may have high consistency among subjects, so the pretraining model is also suitable for new subject data. On the contrary, the gain of fine tuning on CRNN regression model is significant under all window length conditions (2 s, $t = -3.688$, $P = 0.004$; 5 s, $t = -3.612$, $P = 0.004$; 10 s, $t = -3.597$, $P = 0.004$). This may be because the regression model relies on more complex spatiotemporal features in EEG related to the processing of audiovisual stimuli, which are less consistent among subjects. For example, different subjects may have different dependence on visual and auditory stimuli, which makes their EEG feature space different. It can be seen that compared with the subjects without sensors, the subjects with digital sensors have obvious advantages in language audiovisual teaching.

5. Conclusion

Digital sensor replacement equipment has obvious advantages in people's language audio-visual and oral teaching. Once in a noisy and complex environment, the effect of language audio-visual and oral teaching will decline. The help of digital sensors can help people improve their ability to accept information and reduce environmental obstacles. To some extent, sensor technology will also strengthen people's audio-visual response. In addition, daily face-to-face verbal communication is a scene of audio-visual matching. However, people's senses are often disturbed, which affects the learning in teaching. Under this condition, digital sensors can be selected to strengthen the processing of the obtained information, so as to get better learning and teaching results.

Data Availability

The experimental data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The author declared that they have no conflicts of interest regarding this work.

Acknowledgments

This study was funded by Research and Practice Project of English Teaching Reform in Higher Education of Hebei Province in 2020, Research on the Construction of Practical Curriculum System for Business English Major under the Background of Integration between Industry and Education (project no.: 2020YYJG059) and Project Supported by Scientific Research Fund of Tangshan Normal University, Research on the Construction Path of English Cloud Platform of Jidong Culture under the Background of New Infrastructure Construction (project no.: 2021B12).

References

[1] M. Alhumayani, M. Monir, and R. Ismail, "Machine and deep learning approaches for human activity recognition," *International Journal of Intelligent Computing and Information Sciences*, vol. 21, no. 3, pp. 44–52, 2021.

[2] Y. Xue, *Human Motion Pattern Recognition Based on a Single Acceleration Sensor*, South China University of Technology, Guangzhou, China, 2011.

[3] L. Feng and J. Pan, "Human motion recognition based on three-axis acceleration sensor," *Computer Research and Development*, vol. 53, no. 3, pp. 621–631, 2016.

[4] Y. Luo, S. M. Coppola, P. C. Dixon, S. Li, J. T. Dennerlein, and B. Hu, "A database of human gait performance on irregular and uneven surfaces collected by wearable sensors," *Scientific Data*, vol. 7, no. 1, pp. 1–9, 2020.

[5] L. Xiao, R. F. Li, and J. Luo, "Recognition of human activity based on compressed sensing in body sensor networks," *Journal of Electronics & Information Technology*, vol. 35, no. 1, pp. 119–125, 2013.

[6] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, *A Public Domain Dataset for Human Activity Recognition Using Smartphones*, ESANN, Bruges, Belgium, 2013.

[7] A. Mcafee and E. Brynjolfsson, "Big data: the management revolution," *Harvard Business Review*, vol. 90, no. 10, p. 60, 2012.

[8] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[9] J. Schmidhuber, "Deep learning in neural networks: an overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.

[10] W. Jiang and Z. Yin, "Human activity recognition using wearable sensors by deep convolutional neural networks," in *Proceedings of the 23rd ACM International Conference on Multimedia*, pp. 1307–1310, Brisbane, Australia, 2015.

[11] M. Zhang and A. A. Sawchuk, *USC-HAD: A Daily Activity Dataset for Ubiquitous Activity Recognition Using Wearable Sensors*, ACM Conference on Ubiquitous Computing, 2012.

[12] M. Shoaib, S. Bosch, O. Incel, H. Scholten, and P. Havinga, "Fusion of smartphone motion sensors for physical activity recognition," *Sensors*, vol. 14, no. 6, pp. 10146–10176, 2014.

[13] D. Alberici, P. Contucci, and E. Mingione, *Deep Boltzmann Machines: Rigorous Results at Arbitrary Depth, Annales Henri Poincaré*, Springer International Publishing, 2021.

[14] L. A. Passos and J. P. Papa, "A metaheuristic-driven approach to fine-tune deep Boltzmann machines," *Applied Soft Computing*, vol. 97, article 105717, 2020.

[15] D. Alberici, A. Barra, P. Contucci, and E. Mingione, "Annealing and replica-symmetry in deep Boltzmann machines," *Journal of Statistical Physics*, vol. 180, no. 1-6, pp. 665–677, 2020.

[16] R. Salakhutdinov and H. Larochelle, "Efficient Learning of Deep Boltzmann Machines," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 693–700, Sardinia, Italy, 2010.

[17] M. A. Alsheikh, D. Niyato, S. Lin, H. P. Tan, and Z. Han, "Mobile big data analytics using deep learning and apache spark," *IEEE Network*, vol. 30, no. 3, pp. 22–29, 2016.

[18] H. Fang and C. Hu, "Recognizing human activity in smart home using deep learning algorithm," in *Proceedings of the*

33rd Chinese Control Conference, pp. 4716–4720, Nanjing, China, 2014.

[19] S. Bhattacharya and N. D. Lane, "From smart to deep: robust activity recognition on smartwatches using deep learning," in *2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*, pp. 1–6, Sydney, Australia, 2016.

[20] V. Radu, N. D. Lane, S. Bhattacharya, C. Mascolo, M. K. Marina, and F. Kawsar, "Towards multimodal deep learning for activity recognition on mobile devices," in *ACM International Joint Conference on Pervasive and, Ubiquitous Computing: Adjunct*, pp. 185–188, Heidelberg, Germany, 2016.

[21] Y. Zhao and L. He, "Deep learning in the EEG diagnosis of Alzheimer's disease," in *Computer Vision - ACCV 2014 Workshops*, pp. 340–353, Springer, 2014.

[22] C. Y. Liou, W. C. Cheng, J. W. Liou, and D. R. Liou, "Autoencoder for words," *Neurocomputing*, vol. 139, no. 139, pp. 84–96, 2014.

[23] M. M. Al Rahhal, Y. Bazi, H. AlHichri, N. Alajlan, F. Melgani, and R. R. Yager, "Deep learning approach for active classification of electrocardiogram signals," *Information Sciences*, vol. 345, pp. 340–354, 2016.

[24] B. Jokanovic, M. Amin, and F. Ahmad, "Radar fall motion detection using deep learning," in *2016 IEEE Radar Conference (RadarConf)*, pp. 1–6, Philadelphia, PA, USA, 2016.

[25] M. Munoz-Organero and R. Ruiz-Blazquez, "Time-elastic generative model for acceleration time series in human activity recognition," *Sensors*, vol. 17, no. 2, p. 319, 2017.