An Online Classroom Question Answering Evaluation System Based on Voiceprint and Behavior Recognition

Xiangyun Wang,Jieyu Liang,Xiang Xu,Xiang Li,Quanyin Zhu*
Faculty of Xiangyu,Huaiyin Institute of Technology,Huai'an,China
*Corresponding author's email:hyitzqy@qq.com

Abstract—**Faculty of Xiangyu proposed and designed an online class question answering and judgment system based on convolutional neural network behavior recognition, voice pattern recognition and dialect recognition for the wide development of online teaching. This design collects students' voice patterns and behaviors when answering questions through microphone, camera and other equipment, identifies and analyzes students' voice patterns and behaviors through computer algorithm, and judge's students' mastery of knowledge points.**

Keywords- convolutional neural network; action recognition; voiceprint recognition; dialect recognition

I. INTRODUCTION

In recent years, the epidemic situation in China has remained severe. In order to ensure the safety of teachers and students, online teaching activities have been carried out one after another. Online teaching reduces the gathering of staff and alleviates the pressure of epidemic prevention, but the quality of online teaching is often difficult to guarantee. Although teachers can also play a role in urging students, but can not be comprehensive. Moreover, if the students use dialect to answer when the teacher supports teaching, there will often be communication barriers. In order to ensure the teaching quality of online teaching, strengthen teachers' grasp of class, improve students' self-control ability, and optimize the identification accuracy of rural students' answering in dialect, it is necessary to develop a judgment system for online classroom question answering based on voice pattern recognition, behavior recognition and dialect recognition.

II. RELATED WORD

A. DNN-RELIANCE algorithm for voice print recognition

Because there is no need for students to go to the classroom, only need to be at home or other places, so there will be students lazy not in class. In recent years, biometrics based on physiological features such as fingerprints and faces have developed rapidly and are widely used in many fields[1]. Among them, voiceprints have been gradually applied to many fields due to their unique advantages[2]. Using computer voice print recognition can confirm whether it is the student who is answering the question, so as to strengthen the supervision of students and improve their self-management ability and consciousness. In this paper, DNN-RELIANCE algorithm is used for voice print recognition. Dnn-reliance algorithm is divided into deep neural network (DNN) module and RELIANCE module, and the voice print signal should be preprocessed before proceeding[3]:

1) Speech endpoint detection: in the process of voiceprint recognition, no voice or non-voice fragment can not only provide effective identification information, but also increase the amount of computation of the algorithm, so it needs to be eliminated in advance..

2) Pre-weighting: the power of voiceprint information is inversely proportional to the signal frequency, so the pre-weighting technique can avoid the rapid attenuation of high-frequency speech fragments. The relationship between input voiceprint signal $P(x)$ and output voiceprint signal $Q(x)$ is as follows:

$$q(x) = p(x) - \alpha \cdot p(x-1), 0 \le n \le N, 0.9 < \alpha < 1 \quad (1)$$

Where, n is the size of the frame, α is the DNN-Reliance algorithm[4] after pre-weighting coefficient pretreatment, the former is used to extract the depth characteristics of the speaker's mixed feature parameters and the decision of first-order speaker identification. The latter is used to modify and improve the former recognition results. The overall architecture is shown in Figure 1.

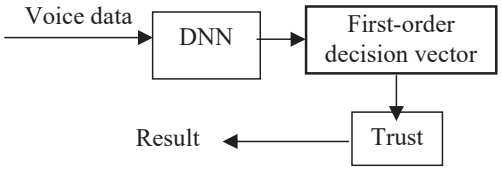

Figure 1. Overall architecture diagram

The RELIANCE module of the overall architecture diagram can detect the approximation between the vector and the label by calculating the distance between the first-order decision vector and the label as shown in Formula (1) and the discretization of the first-order decision vector as shown in Formula (2) to avoid the loss of value of the maximum dimension data.

$$\alpha = \frac{\sqrt{l1^2 + l2^2 + (l_{max} - 1)^2 + \dots \dots + l_n^2}}{n} \quad (2)$$

$$\eta = \frac{l_{max}}{\sum_{i=1}^{n-1} l_i} * 100\% \quad (3)$$

Where n is the number of labels and L_n is the size of the NTH label.

In Formula (2), if α and n meet the threshold conditions of αt and nt respectively, then the trust degree is calculated.

The trust degree is used to judge the distribution of the speaker's voice print recognition results. When the statistical result of the test voice print probability distribution is greater than the trust threshold condition, the recognition success is judged and the recognition result is output; otherwise, the recognition fails. The formula is as follows:

2473-3636/22/$31.00 ©2022 IEEE
DOI 10.1109/DCABES57229.2022.00015

$$x_1 = \frac{\sum_{k=1}^{L_1} l_{1k}}{L} \quad (4)$$

$$x_2 = \frac{\sum_{k=1}^{L_2} l_{2k}}{L} \quad (5)$$

$$x_3 = \frac{\sum_{k=1}^{L_3} l_{3k}}{L} \quad (6)$$

Where, l_{1k}, l_{2K} and l_{3K} represent the first, second and third types of frames respectively, and L is the total number of frames of a voice print data.

B. Action recognition

Behavior recognition For online classes with open cameras, computer algorithms can be used to judge students' class status through their class postures to improve class efficiency. In this paper, OpenPose is used to judge the class status of students.

The results of posture feature extraction are shown in Figure 2, with student features corresponding to the vertical positions of each point (eyes , ears , shoulders , nose and neck). Therefore, student feature ($y \in \{0,1\}^8$) are defined as whether each point $y_d \in y$ is missing ($y_d = 1$) or not ($y_d = 0$)

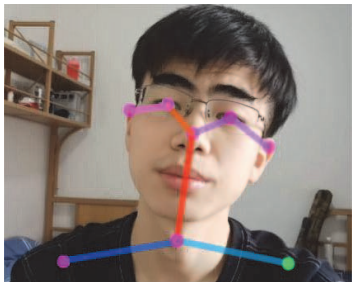

Figure 2. Key points of learner

We employed Logistic regression (LR) as an estimation model[5]. The LR is used to estimate y from x:

$$\hat{y}_d = (1 + \exp(-w_d x - b_d))^{-1} \quad (7)$$

The distance between y and $\hat{y}$ is determined as the binary cross-entropy:

$$D(y, \hat{y}) = -\sum_{d=1}^{8} (y_d \ln \hat{y}_d + (1 - y_d) \ln(1 - \hat{y}_d)) \quad (8)$$

A learning score is calculated by dynamic time warping (DTW) as

$$DTW(Y, \hat{Y}) = C_{I,J} / (I + J) \quad (9)$$

where Y and $\hat{Y}$ are sequences of y and $\hat{y}$ respectively, and i and j are times of the learning sessions ($1 \le I, 1 \le J$).$C_{i,j}$indicates the distance at i and j, which is calculated as

$$C_{i,j} = D(y, \hat{y}) + \min((C_{i-1,j}, C_{i-1,j-1}, C_{i,j-1})) \quad (10)$$

where I and J describe the time in the session The range of i and j is restricted as

$$|(I/J)i + j| < \max(I, J) / \min(I, J) + 5 \quad (11)$$

which means that the system accepts a learner's response delay of 5 s from teaching behavior at *j*.

C. Dialect voiceprint recognition

Dialect recognition The technical difficulties of dialect recognition are that the language used in different places will change, the same meaning will have different names, the same sound will have different tone and weight in different dialects, which brings difficulties to the machine's voice pattern recognition. Therefore, to extract more representative and iconic voiceprint feature parameters and reduce the influence of background speech is a key step to improve dialect voiceprint recognition.

The feature extraction process of dialect voiceprint is as follows:

1) Formant detection based on cepstrum: The formant is one of the important characteristic parameters of voice print recognition. Formant extraction is the acquisition of the spectral envelope of speech, taking the maximum value of the spectral envelope as the formant parameter[6]. In this paper, the formant of dialect is calculated based on cepstrum. Based on homomorphic deconvolution technology, pitch information and channel information can be separated in cepstrum domain, because the formant extracted by the latter is more accurate and effective[7].

If the speech is set as $X(n)$, the channel response is $H(n)$, and the glottic pulse is $E(n)$, the relationship among the three can be expressed as

$$x(n) = e(n) \times h(n). \quad (12)$$

Thus, the cepstrum of the language signal is

$$\hat{x}(n) = \hat{e}(n) + \hat{h}(n). \quad (13)$$

Formula (13) can confirm that pitch information $\hat{e}(n)$ and track information $\hat{h}(n)$ are relatively independent in cepstrum domain.

The characteristic of formant cepstrum obtained from $h(n)$ and $e(n)$ can be divided into the following five steps.

- Set the audio signal frame length as n, preemphasize, windowing and frame segmentation, and finally get the I frame $x_i(n)$ and I representing the sound signal.
- $X_i(k)$ can be obtained from the discrete Fourier transform, as shown in Equation (14).

$$X_i(k) = \sum_{n-0}^{N-1} x_i(n) e^{-\frac{j2\pi kn}{N}}. \quad (14)$$

- Take the logarithm $\hat{X}_i(k)$ of the amplitude of $X_i(k)$, as shown in Equation (15).

$$\hat{X}_i(k) = \log(|X_i(k)|). \quad (15)$$

- Inverse Fourier transform is performed on the results of Equation (14) to obtain the corresponding cepstrum, as shown in Equation (16).

$$\hat{x}_i(n) = \frac{1}{N} \sum_{k=0}^{N-1} \hat{X}_i(k) e^{\frac{j2\pi kn}{N}}. \quad (16)$$

- Set a low-pass window function window(n) on the inverted frequency domain axis, which can generally be set as a rectangular window:

$$window(n) \begin{cases} 1, & if\ n \le n_0 - 1\ and\ n \ge N - n_0 + 1 \\ 0, & if\ n_0 - 1 < n < N - n_0 + 1 \end{cases} \quad (17)$$

- Where,n_0 is the width of the window function, which can be multiplied by cestrum sequence to obtain $H_i(n)$, as shown in Equation (18).

$$h_i(n) = \hat{x}_i(n) \times window(n). \quad (18)$$

- As shown in Equation (19), $H_i(k)$ is obtained from the Fourier transform.

$$H_i(k) = \sum_{n=0}^{N-1} h_i(n) e^{-\frac{j2\pi kn}{N}}. \quad (19)$$

- Through calculation, formant peaks of different frequencies can be obtained.

2) Feature extraction: Feature extraction The speech signal is preprocessed and a set of CP-stral characteristic parameters are obtained by using Gammatone recorder according to the auditory characteristics of cochlea. The recognition rate and robustness of MFCC are better than those of traditional MFCC when the background noise of speech signal is different, and this advantage is more obvious when the signal to noise ratio is low. In this paper, first-order difference and second-order difference are selected as dynamic features, and GFCC's characteristic parameters can be obtained by combining GFCC with them.

This paper collected five local words from Yuci of Shanxi Province, Guangdong province, Guizhou Province, Huai 'an city of Jiangsu Province and Yangzhou City of Jiangsu Province, the content is unified for computer related problems. As the research environment of this paper is an online classroom, the students who participated in the collection of voice print are all 18-20 years old students of both genders. The collected voiceprints are shown in Figure 3.

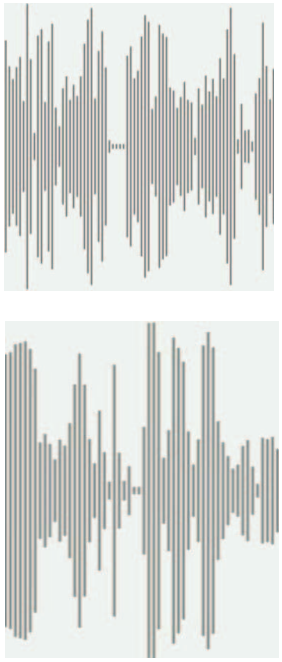

Figure 3. Two local dialects are examples

Different recording processing was marked, their characteristic parameters were extracted and matched with each GMM model, and the similarity of different models was calculated. The highest similarity was selected as the result and compared with the correct result.

The accuracy rate was calculated through multiple experiments. Mathematically, the calculation formula of accuracy is shown in Equation (20).

$$C = \frac{n_x}{n_y} \times 100\% \quad (20)$$

Where, C is the accuracy rate, n_x is the number of correct recognition results, and n_y is the total number to be recognized.

The independent variables are determined, and MFCC, GFCC, GFCC+MFCC and GFCC+ formant are tested as dependent variables. Judging the performance of different methods by the accuracy rate, it can be found that the last method has the highest accuracy rate.

III. SUMMARY

The theme of this project helped me to learn more about the parts and techniques that I had never touched on before, and I also learned a lot of knowledge that I did not have the opportunity to learn in school and books.

Among them, there is the concerted efforts of our team of students, and the careful teaching of the instructor, it is precisely because of the instructor that we can correct the mistakes and successfully complete this thesis. Technology is rolling forward, and various technologies are constantly emerging and updating and iterating, helping human beings achieve greater convenience.

As a college student, I should explode my learning attitude, constantly try to learn continuously, understand the development of technology, pay attention to the emerging new problems, and become a person who can contribute to society.

ACKNOWLEDGMENT

The project name is an online classroom Q&A evaluation system based on voiceprint and behavior recognition. The project number is:202211049080Y.

REFERENCES

[1] Y. H. Zheng. Development and application strategy of voiceprint recognition technology. Technology Wind, no.21, pp.9–10, 2017.

[2] Z. Lian, Y. Li, J. H. Tao, J. Huang, M. Y. Niu. Expression analysis based on face regions in real-world conditions. International Journal of Automation and Computing, vol.17, no.1, pp.96–107, 2020.

[3] V. Hautamaki, T. Kinnunen, I. Karkkainen, et al. Maximum a Posteriori Adaptation of the Centroid Model for Speaker Verification[J]. IEEE SIGNAL PROCESSING LETTERS, 2008, 15: 162-165.

[4] J. Zhang, "The Algorithm of Voiceprint Recognition Model based DNN-RELIANCE," *2020 International Conference on Computer Engineering and Application (ICCEA),* Guangzhou, China, 2020, pp. 250-253..

[5] T. Kawamata and T. Akakura, "Automatic Evaluation of Learning Behaviors for Online Lectures by OpenPose," 2022 IEEE 4th Global Conference on Life Sciences and Technologies (LifeTech), Osaka, Japan, 2022, pp. 384-385..

[6] N. Srivastava, G. Hinton, A. Krizhevsky, A. Sutskever, R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learn-ing Research, vol.15, no.1, pp.1929–1958, 2014. D

[7] N. Jiang, T. Liu. Research on voiceprint recognition of camouflage voice based on deep belief network. International Journal of Automation and Computing, vol.18, no.6, pp.947–962, 2021.