Speech Recognition of Oral English Teaching Based on Deep Belief Network

https://doi.org/10.3991/ijet.v15i10.14041

Jianmei Wang
English Department in Taiyuan University, Taiyuan, China
xhwmail2005@sina.com

Abstract—The oral English teaching faces several common problems: the teaching method is very inefficient, and the learners are poor in oral English. The development of computer-aided language learning offers a possible solution to these problems. Based on techniques of speech recognition, cloud computing and deep learning, this paper applies the deep belief network (DBN) to recognize the speeches in oral English teaching, and establishes a multi-parameter evaluation model for the pronunciation quality of oral English among college students. The model combines the merits of subjective and objective evaluations, and assesses the pronunciation from four aspects: pitch, speech rate, rhythm and intonation. Finally, the proposed model was verified through speech recognition and pronunciation evaluation experiments on 26 non-English majors from a college. The results show that the proposed evaluation model output credible results, which are consistent with those of experts, as evidenced by consistency, neighbourhood consistency and Pearson correlation coefficient. The research provides a feasible way to evaluate the oral English proficiency of learners, laying the basis for improving the teaching and learning efficiency of oral English.

Keywords—Computer-aided language learning, English, speech recognition, deep learning (DL), neural network (NN)

1 Introduction

In the era of economic globalization, trade is booming across borders. English, as an international language, has achieved unprecedented importance globally. English learning is being encouraged in many countries. For various reasons, there are many deficiencies with English teaching as a second language. Many English learners perform well in listening, reading and writing, but have difficulty in speaking. Computer-aided language learning has made it possible for them to overcome the difficulty.

Currently, many try to learn oral English by listening and repeating of audio-visual materials on mobile phones and MP3 players. However, these devices cannot evaluate or instruct the learner's pronunciations. Against this backdrop, many colleges around the world have organized oral communication internships and interactive language programs, and developed speech recognition/scoring techniques, providing learners great

chances to sharpen their oral English [1]. Fruitful results have also been achieved in pronunciation evaluation.

With the emergence of deep neural network (DNN) and deep learning (DL), many DL-based neural networks (NNs) are being applied in speech recognition. Technical giants like Microsoft, Google and Baidu have all designed DNN speech recognition models, which can recognize speeches accurately and rapidly [2].

Based on the above analysis, this paper briefly introduces relevant theories on speech signal pre-processing, feature extraction, and DL networks, and applies the deep belief network (DBN) to recognize the speeches in oral English teaching. Then, a multi-parameter evaluation model was established for the pronunciation quality of oral English among college students, and verified through simulation experiments.

2 Related Theoretical Basis

2.1 Speech signal preprocessing and feature extraction

Speech recognition process: To convert the learner's spoken English into machine-recognizable digital information, firstly the computer's sound card is used to digitize the voice analogue signal, and the voice signal is then pre-processed to extract the characteristic parameters. As a result, a reference model is established for the training of test sentences. After the training, pattern matching and speech recognition are performed, and the final recognition results are obtained through post-processing [3].

Voice signal preprocessing: The process of voice signal preprocessing is shown in Figure 1.

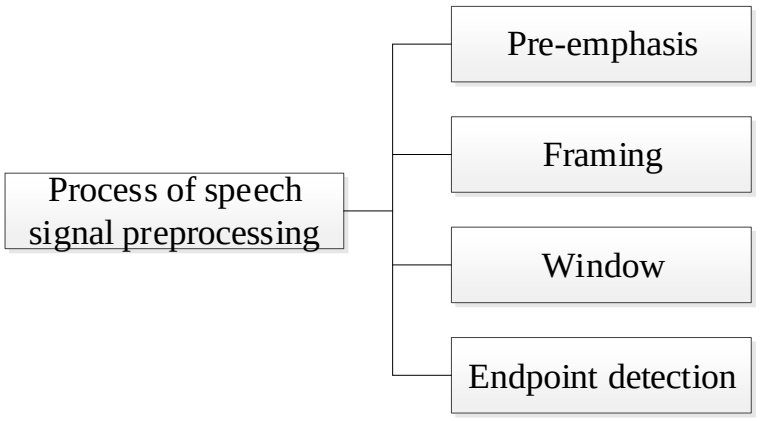

Fig. 1. Process of speech signal preprocessing

Pre-emphasis: To flatten the speech signal, the high-frequency part of the speech signal in this paper was improved by a 6dB/oct high-frequency boost pre-emphasis digital filter. Formula (1) shows the filter response function [4]. The input speech signal

x(n) was used to represent the result y(n) after pre-emphasis processing, as shown in formula (2).

$$H(z) = 1 - \alpha z^{-1}, 0.9 \le \alpha \le 1.0 \tag{1}$$

$$y(n) = x(n) - \alpha x(n-1) \tag{2}$$

where, α is the pre-emphasis coefficient.

Framing: Based on the short-term stationary characteristics of the speech signal, framing was performed on the speech signal stream using a half-frame overlap method [5] to analyse the time series of the characteristic parameters in the speech signal.

Window: The purpose of windowing is to strengthen the speech waveform near the sampled speech signal. The specific calculation formula is shown in (3). Hanning window, Hamming window and rectangular window are the three most used window functions [6]. In this paper, the more widely used Hamming window was used to perform windowing on speech signals, and its definition is shown in formula (4).

$$Q_n = \sum_{m=-\infty}^{\infty} T\left[s(n)\right]\omega(n-m) \tag{3}$$

$$\omega(n) = \begin{cases} 0.54 - 0.46\cos\left(\dfrac{2\pi n}{N-1}\right), 0 \le n \le N-1 \\ 0, \end{cases} \tag{4}$$

Endpoint detection: The performance of speech recognition depends directly on the quality of the endpoint detection algorithm. The dual-threshold endpoint detection method [7] is a widely used method at present, which has the advantage of not only accurately detecting the endpoints of valid voice signals, but also improving the effectiveness of system processing [8]. Thus, it's applied in this study for endpoint detection.

Speech feature parameter extraction: The original speech signal contains a lot of interference information, which can only be used for speech processing after removing redundant information. Therefore, it's necessary to extract the speech feature parameters of the original speech signal [9]. Mel frequency cepstrum coefficient (MFCC), linear prediction cepstrum coefficient (LPCC), and fast Fourier transform spectral coefficient (FFT) are common speech frequency characteristics. Among them, MFCC can better improve the performance of speech signal processing with good robustness [10] so that it's used as speech parameters in this study.

2.2 Deep learning and neural networks

Deep learning [11] is a new field in machine learning research and a type of unsupervised learning. Its essence is to improve the accuracy of prediction or classification, and discover the internal connections and characteristics between the data by establishing a multi-hidden layer learning model and simulating the human brain to analyse the training data. The core ideas of deep learning [12] are: (1) unsupervised learning is used

for pre-train of each layer for the network; (2) only one layer is trained each time with unsupervised learning, and its training results is taken as an input to its higher layer; (3) supervised learning is adopted to adjust all layers.

Deep Belief Network (DBN) [13] is an unsupervised greedy layer-wise learning algorithm. It is an efficient stacked deep learning algorithm based on several restricted Boltzmann machines (RBM). The training process is conducted layer by layer from the bottom to the top, and finally the network is fine-tuned by traditional global learning algorithms (such as BP algorithm), so that the model converges to the local optimum, as shown in Figure 2.

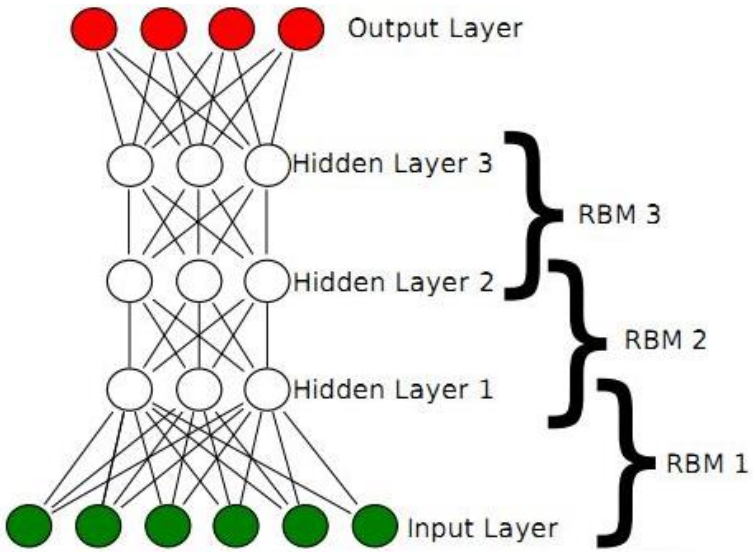

Fig. 2. The main framework of deep learning algorithms

3 Multi-Parameter Pronunciation Quality Evaluation and Simulation Results Analysis

3.1 Multi-parameter pronunciation quality evaluation

Pronunciation quality evaluation process

(1) Subjective evaluation: After listening to the test voice, the language expert discovered the pronunciation errors and the differences between the test voice and standard speech based on the linguistic knowledge, and then evaluated the spoken language level of the test subjects [14]. But limited by the experience and subjective feelings of the language experts, the subjective evaluation method is highly subjective, making it difficult to guarantee the authenticity of the evaluation results.

(2) Objective evaluation: Objective evaluation [15] refers to the use of a computer to perform feature extraction on the spoken English pronunciation of a test subject

through a pronunciation quality evaluation system, make pattern matching with previously extracted standard speech feature parameters to compare the two, and give the evaluation score of the test voice. This objective evaluation method can reduce the evaluation bias and improve the evaluation efficiency.

Evaluation index: Different groups have different requirements and evaluation standards for oral English learning. Taking college students as an example, this paper establishes a multi-parameter pronunciation quality evaluation model for college students' oral English learning. Figure 3 shows the specific evaluation index.

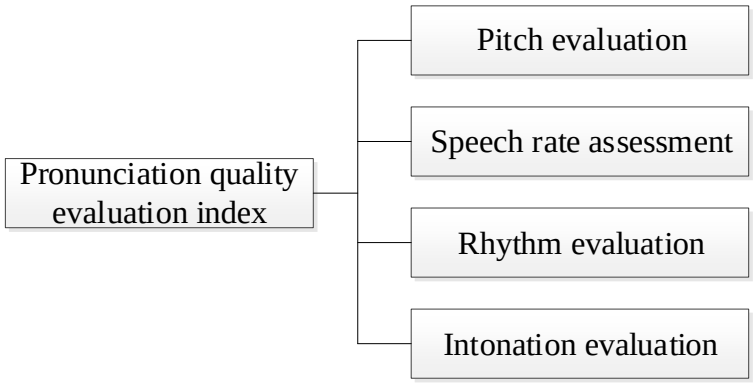

Fig. 3. Multi-parameter pronunciation quality evaluation index for college students

(1) Pitch evaluation: In this study, the MFCC coefficient was used as the evaluation standard of pitch, To be specific, it extracts the MFCC feature parameters of the test speech and the standard speech, and synthesizes the MFCC feature correlation coefficients with the DBN-based speech recognition model to recognize the speech and evaluate the pitch of English learners.

(2) Speech rate evaluation: Speech rate means the measure of how fast a speaker pronounces [16]. This paper uses speech rate based on speech length to evaluate the oral language speed of English learners. The calculation formula is shown in (5). Then, the calculated value φ was compared with the set speech rate threshold, to evaluate the speech rate and obtain the feedback result of the oral English for the learners.

$$\varphi = \frac{Len_{Std}}{Len_{Text}} \tag{5}$$

where, Len_{Text} and Len_{Std} are the length of the test sentences and standard sentences, respectively.

(3) Rhythm evaluation: The rhythm of language can be divided into three types: emphasized stress, incomplete stress, and complete stress [17]. English is a stress-timed language, and the tempo of a sentence is determined by the number of stress syllables. Figure 4 shows the rhythm evaluation mechanism.

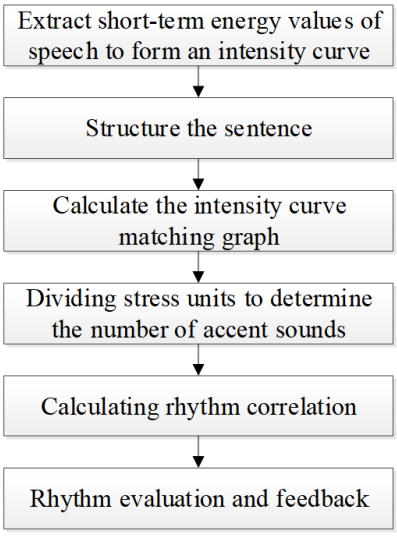

Fig. 4. Specific steps of rhythm evaluation mechanism

a) Extract the short-term energy value to form an intensity curve. The short-term energy is calculated as:

$$E_n = \sum_{m=-\infty}^{\infty} \left[s(n)\omega(n-m) \right]^2 \tag{6}$$

b) Structure the sentence: Despite of different pronunciation characteristics, the pronunciations of individuals follow certain rules. Therefore, before the test, the test sentences were regularized to be close to the standard sentences in order to obtain more objective evaluation results.

c) Calculate the intensity curve matching graph: Based on the original dynamic time warping algorithm (DTW) [18], R (reference template) and T (test template) were divided into N and M frames isochronally, and the distance was divided into three path: (1, Xa), (Xa+1, Xb), (Xb+1, N), where Xa and Xb are the closest integers, to be calculated as:

$$X_a = \frac{1}{3}(2M\text{-}N) \tag{7}$$

$$X_b = \frac{2}{3}(2N\text{-}M) \tag{8}$$

Dynamic matching was performed only when 2M-N<3, 2N-M<2.
ymin, ymax are calculated as:

$$y_{\min} = \begin{cases} \frac{1}{2}x, x \in [0, x_b] \\ 2x + (M - 2N), x \in (x_b, N] \end{cases} \tag{9}$$

$$y_{\max} = \begin{cases} 2x, x \in [0, x_a] \\ \frac{1}{2}x + (M - \frac{1}{2}N), x \in (x_a, N] \end{cases} \tag{10}$$

When each frame on the X axis matched with the frame between $[y_{\min}, y_{\max}]$, and the X coordinate moved forward, the corresponding Y axis frames had the same regularity characteristics, and the cumulative distance is:

$$D(x, y) = d(x, y) + \min \begin{Bmatrix} D(x\text{-}1.\ y) \\ D(x-1, y-1) \\ D(x-1, y-2) \end{Bmatrix} \tag{11}$$

d) Divide the stress unit and determine the number of stress: Through experiments and previous research experience, this paper sets the stress threshold and non-stress threshold as shown in formulas (12) and (13). Then, double-threshold comparison method was used to detect stress endpoints. Meanwhile the duration of stress speech was set 100ms, to determine the number of stress in the sentence.

$$T_u = (\max(sig\_in) + \min(sig\_in)) / 2.5 \tag{12}$$

$$T_t = (\max(sig\_in) + \min(sig\_in)) / 10 \tag{13}$$

where, T_u is the stress threshold, and T_t is the non-stress threshold.

e) Calculate rhythm correlation: The Pairwise Variability Index (PVI) can be used to determine the difference in syllable length and measure the correlation of language rhythms. Based on the differences in English pronunciation length, this paper improves the formula for calculating the PVI, as shown in formula (14), which is used as a basis for systematic evaluation.

$$dPVI = 100 * \left(\sum_{k=1}^{m-1} |d1_k - d2_k| + |d1_t - d2_t| \right) / Len \tag{14}$$

where, Len is the standard sentence length, m = min (the number of standard sentence units, the number of test sentence units), and d_k is the duration of the k-th speech unit segment.

f) Rhythm evaluation and feedback: A comprehensive comparison was performed about the intensity curve matching, the number of stresses, and the dPVI parameters

of the standard sentence and the test sentence. Thus, the final evaluation results of oral rhythm were obtained.

(4) Intonation evaluation: The intonation is the preparation and variation in spoken pitch when used in a sentence. For the same sentence, different intonation can result in a difference in meanings. The lexical meaning of a sentence plus the meaning of intonation can be regarded as full meaning. The five basic intonations in English are rising intonation (↗), falling intonation (↙), rising-falling intonation (∧), falling-rising intonation (∨), and flat intonation (→). Pitch is the most basic and important constituent element. In the discourse, the height of the sound is expressed as intonation, and the different rising and falling modes of the intonation is determined by the pitch. This paper uses the autocorrelation function (ACF) in the time domain to extract the pitch in English sentences, then uses a median filter to smooth the pitch, calculates the fit of the fundamental frequency curve through DTW, and finally evaluates the intonation.

3.2 Simulation experiment and result analysis

Data source: In this paper, 26 college students of non-English majors were selected as test subjects through posters, online solicitation and other means, including 15 males and 11 females. The 10 common spoken English sentences were recorded by CoolEdit recording software.

Evaluation of speech recognition: Using the pronunciations of 55 people in the data set of the UCI machine learning database as the training set and the pronunciations of 33 people as the test set, a comparative analysis method was adopted to compare the speech recognition rates of the model proposed in this study and other model.

The time warping is a problem that exists in most neural networks. In this paper, the segmentation and averaging methods were used to pre-process the speech signal. First, the feature parameters were divided into N segments according to formula (15), and then M(i) into M segments. Next, the mean vector of each sub-segment was obtained through calculation, and the output value of the dimensionality-reduced feature parameter was a K × M × N matrix $\overline{S(K,T)}$ composed of the averages of the respective segments.

$$\begin{aligned} &M(i) = S(K, J), \\ &J = \left[\frac{T}{N}(i-1)+1\right], \cdots \left[\frac{T}{N} i\right] \end{aligned} \tag{15}$$

where, M (i) is the speech feature parameter of the i-th segment after segmentation.

Figure 5 shows the comparison results of recognition rates under different models. It can be seen from the figure that compared with Continuous Density Hidden Markov Model (CDHMM), Traditional Discrete Hidden Markov Model (DHMM), TDA-MWST, BP-AdaBoost algorithm, and the KASWT algorithm, the DBN model proposed in this paper has the highest recognition rate, indicating the rationality and validity of the model. Thus, it can be used in the evaluation of speech pronunciation quality.

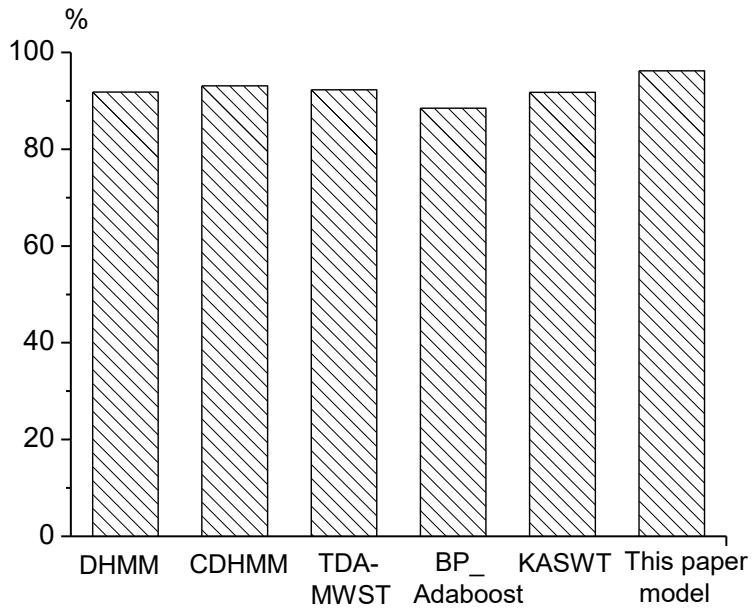

Fig. 5. Comparison of recognition rates under different models

Pronunciation evaluation experiment: In order to validate the English pronunciation quality evaluation model and its method, this paper uses the three indexes of Pearson correlation coefficient, consistency and neighbourhood consistency to test the consistency of manual evaluation and machine evaluation

According to English pronunciation characteristics of college students, this paper combines with previous research results, and selects four indicators of pitch, rhythm, intonation, and speech speed to evaluate college students' oral English pronunciation. They're scored on a four-level system: A, B, C, and D (i.e., 4 points, 3 points, 2 points, and 1 point respectively). Also, two outstanding college English teachers were selected to evaluate the recorded spoken English of the test subjects. The two-sided test results showed that the evaluation results of the two teachers were basically consistent, indicating that the data of the manual evaluation results were valid.

As described above, a comparative analysis was performed on the results of machine evaluation and manual evaluation. Figure 6 shows the experimental results of the evaluation index (the number of samples), and Figure 7 shows the experimental results of the evaluation index (the statistical index). It can be seen that among the 260 English sentences, the consistency rate and neighbourhood consistency rate between the machine and manual evaluation were all above 80% in terms of the four evaluation indicators, and the Pearson correlation coefficients were 0.85, 0.496, 0.557, and 0.631, which indicates that the proposed evaluation method is credible.

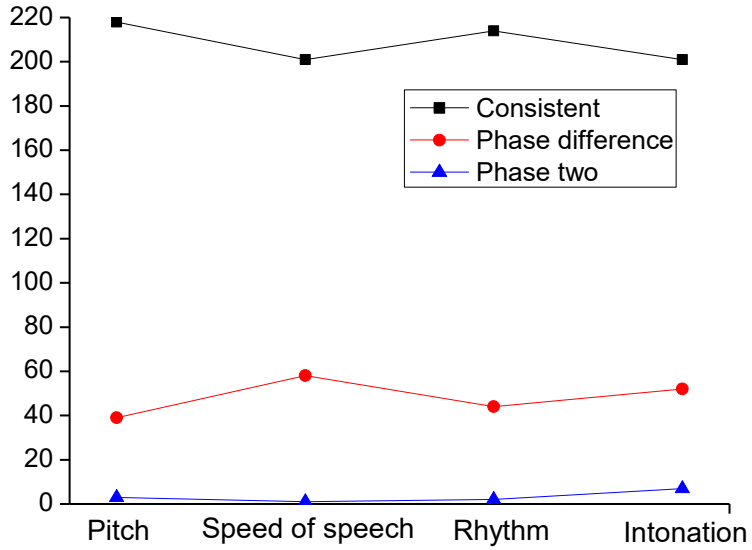

Fig. 6. Evaluation index experimental results-number of samples

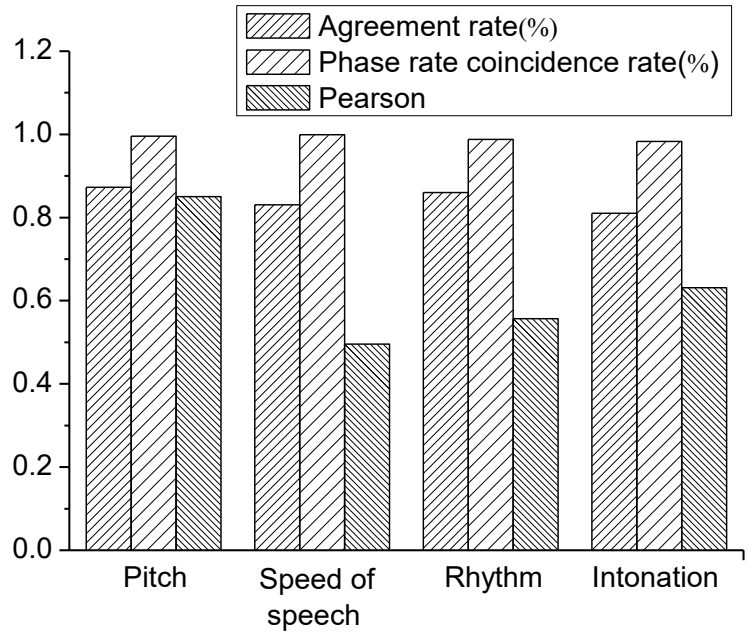

Fig. 7. Evaluation index experiment results-statistical index

4 Conclusion

Speech recognition and pronunciation evaluation technology are the core of computer-aided language learning. This paper studies the speech recognition of oral English teaching based on deep belief network. The specific conclusions are as follows:

- The DBN was applied to recognize the speeches in oral English teaching, and a multi-parameter evaluation model for the pronunciation quality of oral English was established among college students.
- The simulation results of speech recognition experiments show that the speech recognition rate of the model based on the DBN was 96.21%, which is better than other models.
- The simulation results of the pronunciation evaluation experiments show that the consistency rate and the neighbourhood consistency rate of the machine and manual evaluation were all above 80% in terms of the four evaluation indicators, and the Pearson correlation coefficients were 0.85, 0.496, 0.557, and 0.631, which explains the credibility of evaluation method.

5 References

[1] Zhang, R., Kikui, G. (2006). Integration of speech recognition and machine translation: speech recognition word lattice translation. Speech Communication, 48(3-4): 321-334. https://doi.org/10.1016/j.specom.2005.06.007

[2] Sukhbaatar, O., Usagawa, T., Choimaa, L. (2019). An artificial neural network based early prediction of failure-prone students in blended learning course, International Journal of Emerging Technologies in Learning, 14(19): 77-92. https://doi.org/10.3991/ijet.v14i19.10366

[3] Goodarzi, M. M., Almasganj, F. (2016). Model-based clustered sparse imputation for noise robust speech recognition. Speech Communication, 76: 218-229. https://doi.org/10.1016/j.specom.2015.06.009

[4] Schlüter, R., Macherey, W., Müller, B., Ney, H. (2001). Comparison of discriminative training criteria and optimization methods for speech recognition. Speech Communication, 34(3): 287-310. https://doi.org/10.1016/s0167-6393(00)00035-2

[5] Santos, S. C. B. D., Alcaim, A. (2000). Role of syllables and function-words in continuous portuguese speech recognition. Electronics Letters, 36(12): 1083. https://doi.org/10.1049/el:20000749

[6] Cai, M., Liu, J. (2015). Maxout neurons for deep convolutional and lstm neural networks in speech recognition. Speech Communication, 77: 53-64. https://doi.org/10.1016/j.specom.2015.12.003

[7] Dibazar, A. A., Namarvar, H. H., Berger, T. W. (2004). Continuous speech recognition using dynamic synapse neural network. The Journal of the Acoustical Society of America, 115(5): 2612-2612. https://doi.org/10.1121/1.4784787

[8] Deng, L., Hassanein, K., Elmasry, M. (1994). Analysis of the correlation structure for a neural predictive model with application to speech recognition. Neural Networks, 7(2): 331-339. https://doi.org/10.1016/0893-6080(94)90027-2

[9] Kurogi, S. (1991). Speech recognition by an artificial neural network using findings on the afferent auditory system. Biological Cybernetics, 64(3): 243-249. https://doi.org/10.1007/bf00201985
[10] Ye, H., Wang, S. R., Robert, F. (1990). A pcmn neural network for isolated word recognition. Speech Communication, 9(2): 141-153. https://doi.org/10.1016/0167-6393(90)90067-j
[11] Patil, P., B. (1998). Multilayered network for lpc based speech recognition. IEEE Transactions on Consumer Electronics, 44(2): 0-438. https://doi.org/10.1109/30.681960
[12] Chirra, V.R.R., Uyyala, S.R., Kolli, V.K.K. (2019). Deep CNN: A machine learning approach for driver drowsiness detection based on eye state. Revue d'Intelligence Artificielle, 33(6): 461-466. https://doi.org/10.18280/ria.330609
[13] Gonzalez-Dominguez, J., Lopez-Moreno, I., Moreno, P. J., Gonzalez-Rodriguez, J. (2015). Frame-by-frame language identification in short utterances using deep neural networks. Neural Networks, 64: 49-58. https://doi.org/10.1016/j.neunet.2014.08.006
[14] Plotz, T., Guan, Y. (2018). Deep learning for human activity recognition in mobile computing. Computer, 51(5): 50-59. https://doi.org/10.1109/mc.2018.2381112
[15] Park, C., Choi, K. H., Lee, C., Lim, S. (2016). Korean coreference resolution with guided mention pair model using the deep learning. Etri Journal, 38(6): 1207-1217. https://doi.org/10.4218/etrij.16.0115.0896
[16] Erfani, S. M., Rajasegarar, S., Karunasekera, S., Leckie, C. (2016). High-dimensional and large-scale anomaly detection using a linear one-class svm with deep learning. Pattern Recognition, 58: 121-134. https://doi.org/10.1016/j.patcog.2016.03.028
[17] Oh, S. W., Kim, S. J. (2016). Approaching the computational color constancy as a classification problem through deep learning. Pattern Recognition, 61: 405-416. https://doi.org/10.1016/j.patcog.2016.08.013
[18] Li, S., Tang, M., Guo, Q., Lei, J., Zhang, J. (2017). Deep neural network with attention model for scene text recognition. IET Computer Vision, 11(7): 605-612. https://doi.org/10.1049/iet-cvi.2016.0404
[19] Sukhbaatar, O., Usagawa, T., Choimaa, L. (2019). An artificial neural network based early prediction of failure-prone students in blended learning course, International Journal of Emerging Technologies in Learning, 14(19): 77-92. https://doi.org/10.3991/ijet.v14i19.10366
[20] Goodarzi, M. M., Almasganj, F. (2016). Model-based clustered sparse imputation for noise robust speech recognition. Speech Communication, 76: 218-229. https://doi.org/10.1016/j.specom.2015.06.009
[21] Schlüter, R., Macherey, W., Müller, B., Ney, H. (2001). Comparison of discriminative training criteria and optimization methods for speech recognition. Speech Communication, 34(3): 287-310. https://doi.org/10.1016/s0167-6393(00)00035-2
[22] Santos, S. C. B. D., Alcaim, A. (2000). Role of syllables and function-words in continuous portuguese speech recognition. Electronics Letters, 36(12): 1083. https://doi.org/10.1049/el:20000749
[23] Cai, M., Liu, J. (2015). Maxout neurons for deep convolutional and lstm neural networks in speech recognition. Speech Communication, 77: 53-64. https://doi.org/10.1016/j.specom.2015.12.003
[24] Dibazar, A. A., Namarvar, H. H., Berger, T. W. (2004). Continuous speech recognition using dynamic synapse neural network. The Journal of the Acoustical Society of America, 115(5): 2612-2612. https://doi.org/10.1121/1.4784787

[25] Deng, L., Hassanein, K., Elmasry, M. (1994). Analysis of the correlation structure for a neural predictive model with application to speech recognition. Neural Networks, 7(2): 331-339. https://doi.org/10.1016/0893-6080(94)90027-2
[26] Kurogi, S. (1991). Speech recognition by an artificial neural network using findings on the afferent auditory system. Biological Cybernetics, 64(3): 243-249. https://doi.org/10.1007/bf00201985
[27] Ye, H., Wang, S. R., Robert, F. (1990). A pcmn neural network for isolated word recognition. Speech Communication, 9(2): 141-153. https://doi.org/10.1016/0167-6393(90)90067-j
[28] Patil, P., B. (1998). Multilayered network for lpc based speech recognition. IEEE Transactions on Consumer Electronics, 44(2): 0-438. https://doi.org/10.1109/30.681960
[29] Chirra, V.R.R., Uyyala, S.R., Kolli, V.K.K. (2019). Deep CNN: A machine learning approach for driver drowsiness detection based on eye state. Revue d'Intelligence Artificielle, 33(6): 461-466. https://doi.org/10.18280/ria.330609
[30] Gonzalez-Dominguez, J., Lopez-Moreno, I., Moreno, P. J., Gonzalez-Rodriguez, J. (2015). Frame-by-frame language identification in short utterances using deep neural networks. Neural Networks, 64: 49-58. https://doi.org/10.1016/j.neunet.2014.08.006
[31] Plotz, T., Guan, Y. (2018). Deep learning for human activity recognition in mobile computing. Computer, 51(5): 50-59. https://doi.org/10.1109/mc.2018.2381112
[32] Park, C., Choi, K. H., Lee, C., Lim, S. (2016). Korean coreference resolution with guided mention pair model using the deep learning. Etri Journal, 38(6): 1207-1217. https://doi.org/10.4218/etrij.16.0115.0896
[33] Erfani, S. M., Rajasegarar, S., Karunasekera, S., Leckie, C. (2016). High-dimensional and large-scale anomaly detection using a linear one-class svm with deep learning. Pattern Recognition, 58: 121-134. https://doi.org/10.1016/j.patcog.2016.03.028
[34] Oh, S. W., Kim, S. J. (2016). Approaching the computational color constancy as a classification problem through deep learning. Pattern Recognition, 61: 405-416. https://doi.org/10.1016/j.patcog.2016.08.013
[35] Li, S., Tang, M., Guo, Q., Lei, J., Zhang, J. (2017). Deep neural network with attention model for scene text recognition. IET Computer Vision, 11(7): 605-612. https://doi.org/10.1049/iet-cvi.2016.0404

6 Author

Jianmei Wang, has master of arts, foreign linguistics and applied linguistics. At present, she is working as an associate professor in taiyuan university. She has been teaching English courses for more than 6 years, including college English courses and CAT. Jianmei has published many papers and works on foreign linguistics and applied linguistics in English teaching and translation, and has presided over many state-level scientific research projects and municipal projects.

Article submitted 2020-02-06. Resubmitted 2020-03-06. Final acceptance 2020-03-07. Final version published as submitted by the authors.