Proceedings of the 5th International Conference on Education Innovation and Philosophical Inquiries
DOI: 10.54254/2753-7048/59/20241721

Application of Deep Learning Technology in Speech Recognition and Language Teaching

Luciano Li[1,a,*]

[1]*University of Macau, FST, Macau, 999078, China*
a. Lucianoli2004@outlook.com
**corresponding author*

Abstract: In the process of learning a second language, modern Chinese learners are faced with many challenges due to the limitations of the learning environment and teaching conditions. The teachers are usually non-native speakers, and their language level is uneven, resulting in unsatisfactory teaching results. At the same time, the size of the group learning a second language continues to expand, and the requirement for teaching level is getting higher and higher. The topic of this paper is the application of deep learning to speech recognition and language teaching. The results and effects of deep learning in speech recognition and language teaching are summarized by combining the currently available information. It can be found that the current deep learning technology is more and more advanced, the accuracy rate of speech recognition technology has been greatly improved compared with the past, and the teaching effect has also made no small achievements. This technology can not only help learners better master English pronunciation, but also provide more efficient and effective learning experience.

Keywords: Deep learning technology, speech recognition, English teaching, convolutional neural networks

1. Introduction

Due to the increasing level of globalization and internationalization in today's society, the demand for second language learning is growing rapidly. And with the material life needs of the members of today's society being fully satisfied, the staff in various fields of production also pay more and more attention to the importance of multilingual teaching. In recent years, the application of deep learning technology in language teaching has made remarkable progress. Multi-layer residual convolutional neural networks (CNN) based on deep convolutional neural networks and sentence level visual speech recognition lip-reading structures based on multiple convolutional neural networks have been widely used in the field of spoken English pronunciation recognition. The research theme of this paper is to explore how to use these advanced technical means to improve the achievement and effect of language learning. It explores how to optimize the deep learning model to improve accuracy and real-time English pronunciation recognition, proposes pronunciation correction schemes, and achieves effective real-time feedback in practical teaching applications. This study can not only help learners to better master English pronunciation, but also provide an important theoretical and practical basis for the development of language teaching technology in the future. These research results are

© 2024 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-7048/59/20241721

expected to significantly improve the overall level of language teaching in the future, and bring more efficient and effective learning experience to learners.

2. Introduction of speech recognition methods

2.1. Method

Due to the interference of glottic excitation and oral-nasal radiation, the voice signals emitted by humans are attenuated by 6 db/OCT when the high frequency end of the average power spectrum reaches 800 Hz. Therefore, a 6 db /OCT digital filter can be used to enhance the high-frequency part of the speech signal, so that the speech signal spectrum appears flat to support subsequent analysis. At the same time, the spectrum filtering response function from low frequency signal to high frequency signal can be calculated using the same SNR, as shown in Formula 1:

$$H(z) = 1 - \alpha z^{-1} \quad 0.9 \leq \alpha \leq 1 \tag{1}$$

where α is the pre-emphasis coefficient. In this way, the result $y(n)$ after pre-emphasis processing can be expressed by the input speech signal x(n) as follows:

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

Then framing and adding windows are used to eliminate interference factors such as high frequency and high harmonic distortion, so that the voice signal is smoother and more uniform. Then, the dual-threshold comparison method is used to detect the endpoint of the speech signal, and the start and end points of the speech signal are effectively detected through the two features of short-term energy and short-term average zero crossing rate [1].

2.2. Improved Method

With the deepening of the convolutional layer, the traditional CNN is more and more prone to the loss of feature information, which leads to the slow down of the training convergence element and the difficulty in improving the recognition rate. Therefore, a Multilevel Residual Convolutional Neural Network (CNN) is derived, which contains multiple convolutional pooling layers and multi-layer residual structures, and can transmit original information across multiple convolutional layers to compensate for missing features.The residual structure is used to connect the original information of the first n convolutional layers with the current layer to improve the model. By using this algorithm, deep convolutional neural networks can model the speech signal, which improves the defects of traditional speech recognition algorithms, and greatly improves recognition accuracy and convergence speed [2].

2.2.1. Time delay neural networks and long short-term memory networks

At the same time, in order to improve the ability of the model to capture the context of the speech sequence, this section introduces a time delay neural network (TDNN), which can better capture the context-dependent features by introducing a time delay mechanism in the calculation of the model state. Next, it introduces long short-term memory networks (LSTMS) to solve the time-dependent problem that exists in traditional recurrent neural networks (RNNS). LSTM can effectively capture and retain the long time dependency features in speech sequences through its input gate, forget gate and output gate. Finally, combines TDNN and LSTM to c onstruct a hybrid deep neural network structure for oral English pronunciation recognition tasks. In this hybrid structure, TDNN first extracts the features of the input speech sequence and captures the context information in a short time. The LSTM then further processes these features to capture dependencies over a longer time horizon.

DOI: 10.54254/2753-7048/59/20241721

Through this combination, the model can effectively extract short-time contextual features and fully capture long-time dependent information, thus performing well in spoken English pronunciation recognition tasks, significantly improving the accuracy and robustness of the model [3].

2.2.2. Lipreading Architecture Based on Multiple Convolutional Neural Networks

This method refers to the use of visual information (lips, teeth, tongue) movements by VSR to transcript speech. The speech recognition method first extracts detailed lip movement features from video through a 3D convolutional neural network (3D CNN), a 3D dense-connected convolutional neural network (DenseNet) and a multi-layer feature fusion 3D CNN. These networks can capture information in both spatial and temporal dimensions and enhance feature extraction through multi-level feature fusion. Next, a bidirectional gated cycle unit (Bidirectional GRU) was used for serial modeling to capture contextual information about lip movement. The entire network is trained by the end-to-end connection timing fraction (CTC) loss function, without the need for accurate frame-level labeling, and finally realizes the lip-reading task by converting the lip movement in the video into the corresponding speech text. This speech recognition method can not only recognize speech in a noisy environment, but even help some patients with speech disorders [4].

2.3. Select the most appropriate deep learning model

The input speech file is fed into a speech recognition system, transformed using different deep learning techniques, and the word error rate for each technique is calculated. Then, according to the size of the word error rate, the best performing deep learning model is selected as the final recognition method. For details, see Figure 1. This approach ensures that the most appropriate deep learning model is selected and improves the accuracy of speech-to-text [5].

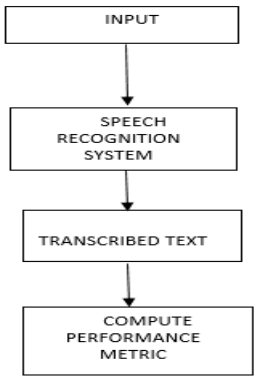

Figure 1: Proposed Framework [5]

3. The application of deep learning in language teaching

3.1. The present situation and advantages of language teaching

Due to the limitations of the learning environment and teaching conditions, learners have many difficulties learning a second language. Most of the teachers are also second language learners. Therefore, especially in the teaching of oral English, both the level of teaching and the accuracy of teaching are not higher than those of native speakers. Therefore, in this case, deep learning plays a very important role in language education. Deep learning can help learners find the difference between their pronunciation and the standard pronunciation through repeated listening and comparison, correct pronunciation mistakes, and improve the efficiency of language learning. It can

DOI: 10.54254/2753-7048/59/20241721

also help teachers understand the emotional state of learners, so as to carry out personalized teaching [6].

3.2. language education improvement with deep learning

3.2.1. Speech evaluation

There are three main aspects of speech assessment: pronunciation standard assessment, speech speed assessment and rhythm assessment. First, the pronunciation standard assessment detects pronunciation errors by extracting speech signal features using Maier frequency cepstrum coefficient (MFCC) technology and combining them with a speech recognition model (SR model) built by a deep belief network (DBN). Articulation clarity and fluency were assessed by comparing the MFCC features of standard sentences and input sentences, and by calculating the correlation coefficient between the two. Second, speed assessment measures a speaker's pronunciation speed by counting the number of syllables pronounced per unit of time, including pause time. The assessment method takes into account the impact of emotional states on speech speed, such as faster speech when angry and happy, and slower speech when sad. Finally, rhythm assessment analyzes the stress and syllable rhythm of language, emphasizing the importance of stressed syllables in sentence organization and semantic expression. It has been found that the number and distribution of stressed syllables directly affect speech speed and syllable articulation. The specific characteristics include: the bigger the number of stressed syllables, the slower the sentence rhythm, the clearer the syllables; unaccented syllables are compact and blurred between accented syllables. These assessment methods demonstrate how modern phonological technologies can be used to improve the accuracy, fluency and overall phonetic quality of English pronunciation, and are of great significance for language learning and teaching [7].

3.2.2. Multi-dimensional feature extraction and evaluation

The recordings of the test subjects were comprehensively analyzed in five aspects: pronunciation, fluency, vocabulary, grammar and semantics, and a deep neural network model was used to model the feature values to obtain the final score. The Latent Dirichlet Allocation (LDA) topic model is introduced to analyze semantics instead of the word frequency method, and the scoring process is automated. Text cleaning after speech recognition and speech noise reduction technology based on deep learning are added to the scoring model, which improves the accuracy of speech recognition and overall scoring. Through innovative application and improvement of key technologies, a new open oral scoring model is proposed and implemented, which combines a multi-feature fusion algorithm to realize automatic detection and evaluation of pronunciation errors in oral English tests [8].

3.3. Experimental cases

Fei Wen and Yanfeng Yue's experiment of the speech recognition model through deep learning: 400 sentences from 40 college students were comprehensively evaluated. There were 370 samples of the same level of machine evaluation and manual evaluation, 30 samples of the first level differences, and 0 samples of the second and third level difference. The overall evaluation consistency between machine and human was 92.5%, the adjacent evaluation consistency was 100%, and the Pearson correlation coefficient was 0.86. The results show that there is a strong correlation between the machine evaluation and the manual evaluation, which verifies the feasibility of the English pronunciation evaluation model [2]. The Lipreading Architecture of Sanghun Jeon, Ahmed Elsharkawy and MunSang Kim is in the experiment. For non-overlapping speakers, the proposed model achieves a word error rate of 2.853% (CER) and a word error rate of 5.059% (WER). For

DOI: 10.54254/2753-7048/59/20241721

overlapping speakers, the proposed model achieves 1.004% CER and 1.011% WER, both of which are better than the existing state-of-the-art models [4].

3.4. Application cases

Zinan Su has built a teaching model combining speech recognition technology with English teaching, which is divided into four stages: (1) Teaching resource preparation stage; (2) Student independent learning stage; (3) Classroom teaching stage; (4) after-school consolidation stage, the specific content is shown in Table 1 [8]. The results show that the effectiveness of these stages are 0.8734, 0.8537, 0.7502 and 0.7706, respectively, and the impact on students' English proficiency is 0.7526, 0.8115, 0.7039 and 0.8126, respectively [9]. The model not only optimizes the student's learning experience, but also provides students with more opportunities to practice their language skills and cross-cultural communication.

Table 1: Rural English teaching reform mode [9]

Teaching mode stage	Teaching resources preparation stage	Students learn in the autonomous stage	Classroom stage	After-school consolidation stage
Specific way	Use the network technology to prepare the teaching resources and book the teaching content	Students use speech recognition technology to evaluate their spoken English, so as to improve their participation	Combined with network technology, English scenario simulation to improve students' learning effect	Keep the classroom courseware and teaching content developed so that students can review at any time
Validity	0.8734	0.8537	0.7502	0.7706
The influence degree on the students	0.7526	0.8115	0.7039	0.8126

4. Challenges faced by speech recognition systems

4.1. Noise interference problem

Although it is currently possible to use lip movement to ignore noise interference for speech recognition, this technology is not yet mature. At present, there are still some problems in solving the noise interference problem in speech recognition systems. If there is noise around the speaker, or if the tone, mood, and intonation of the speaker make the pronunciation inaccurate or unclear, the speech system cannot effectively recognize the speech information [10].

4.2. Endpoint detection level is not high

In the process of signal recognition, endpoint detection technology is very important. In addition to noise interference, speech signal recognition errors can occur even in a quiet state, mainly due to endpoint detectors. To some extent, in order to improve speech recognition technology, it is most important to optimize endpoint detection technology, and in order to meet this requirement, more stable speech parameters need to be explored.

4.3. Difficulty in accurately identifying emotional information

If learners only rely on speech to identify people's emotions when speaking, there are great limitations, but they also need to recognize facial expressions, vocal organ data and movement, in order to more accurately identify voice emotions. This requires collecting large amounts of data and incorporating this data into speech recognition systems to improve the accuracy of speech recognition [10].

5. Conclusion

The speech recognition method of deep learning in artificial intelligence in the future will refer to the algorithms of deep neural networks and be closer to the processes and patterns of the human brain to obtain, analyze and process information, thus creating a more powerful engine that significantly improves perception and cognitive ability. Let students have a more real conversation experience when using artificial intelligence to learn, rather than just a cold machine conversation [10].

In general, this paper first describes the basic ideas and processes of deep learning, and briefly explains the pre-processing process of speech signals to ensure the normal use of speech signals and the robustness of English speech signals. Secondly, some improved methods are proposed to improve the accuracy of speech recognition, including more advanced models and algorithm optimization. Then, the paper summarizes the speech evaluation indicators, covers the evaluation criteria of sound quality (such as intonation, speech speed, intonation, etc.), and explains the specific process of speech evaluation in detail. Then, the paper discusses the challenges faced by the current technology in practical applications, such as noise processing, low levels of endpoint detection. At the same time, in order to solve these problems, the paper also puts forward potential solutions and research directions. Finally, this paper looks forward to the future development trend of speech recognition and evaluation technology, emphasizing that with the continuous advancement of deep learning and artificial intelligence technology, speech technology will play an important role in a wider range of applications, which is expected to achieve higher accuracy, and wider adaptability in the future.

References

[1] *Geng, L. (2021). Evaluation Model of College English Multimedia Teaching Effect Based on Deep Convolutional Neural Networks. Mobile Information Systems, pp.1–8. https://doi.org/10.1155/2021/1874584.*

[2] *Wen, F., & Yue, Y. (2021). A Study of Deep Learning Combined with Phonetic Models in Foreign Language Teaching. Wireless Personal Communications, 119(1), 275–288. https://doi.org/10.1007/s11277-021-08207-7.*

[3] *Li, H., & Liu, X. (2022). A Deep Learning-Based Assisted Teaching System for Oral English. Security and Communication Networks, pp.1–10. https://doi.org/10.1155/2022/1882662.*

[4] *Jeon, S., Elsharkawy, A., & Kim, M. S. (2021). Lipreading Architecture Based on Multiple Convolutional Neural Networks for Sentence-Level Visual Speech Recognition. Sensors (Basel, Switzerland), 22(1), 72. https://doi.org/10.3390/s22010072.*

[5] *C R, R. (2020). Speech Recognition using Deep Learning Techniques. International Journal for Research in Applied Science and Engineering Technology, 8(6), 2199–2201. https://doi.org/10.22214/ijraset.2020.6358.*

[6] *Yang, Y., & Yue, Y. (2020). English speech sound improvement system based on deep learning from signal processing to semantic recognition. International Journal of Speech Technology, 23(3), 505–515. https://doi.org/10.1007/s10772-020-09733-8.*

[7] *Zhao, X., & Jin, X. (2022). Standardized Evaluation Method of Pronunciation Teaching Based on Deep Learning. Security and Communication Networks, 1–11. https://doi.org/10.1155/2022/8961836.*

[8] *Wang, Y. (2021). Detecting Pronunciation Errors in Spoken English Tests Based on Multi-feature Fusion Algorithm. Complexity (New York, N.Y.), 1–11. https://doi.org/10.1155/2021/6623885*

[9] *Su, Z. (2024). Research on the Reform of the Teaching Mode of Rural English Education Assistance Based on the Technical Support of Network Technology. Applied Mathematics and Nonlinear Sciences, 9(1). https://doi.org/10.2478/amns.2023.2.01373*

[10] *Leini, Z., & Xiaolei, S. (2021). Study on Speech Recognition Method of Artificial Intelligence Deep Learning. Journal of Physics. Conference Series, 1754(1), 12183. https://doi.org/10.1088/1742-6596/1754/1/012183.*