International Journal of Swarm Intelligence Research
Volume 15 • Issue 1

Using Deep Learning and Swarm Intelligence to Achieve Personalized English-Speaking Education

Yang Liu, Huanghe Science and Technology University, China*
https://orcid.org/0009-0008-8449-6213

ABSTRACT

This paper presents a pioneering approach to personalized English oral education through the integration of deep learning and swarm intelligence algorithms. Leveraging deep learning techniques, our system offers precise evaluation of various aspects of spoken language, including pronunciation, fluency, and grammatical accuracy. Furthermore, we combine swarm intelligence algorithms to optimize model parameters to achieve optimal performance. We compare the proposed optimization algorithm based on swarm intelligence and its corresponding original algorithm for training comparison to test the effect of the proposed optimizer. Experimental results show that in most cases, the accuracy of the test set using the optimization algorithm based on the swarm intelligence algorithm is better than the corresponding original version, and the training results are more stable. Our experimental results demonstrate the efficacy of the proposed approach in enhancing personalized English oral education, paving the way for transformative advancements in language learning technologies.

KEYWORDS

ACO, Deep Learning, Multi-Feature, Neural Networks, Optimization Algorithm, Personalized English, Speaking Education, Swarm Intelligence

Mastery of the English language, especially in the context of spoken communication, is integral to navigating the complexities of our interconnected world (Bagea, 2023). Conventional language education approaches often encounter challenges in accommodating diverse learning styles, resulting in standardized methodologies that may not effectively cater to the unique needs of individual learners. Recognizing this limitation, the research endeavors to redefine English oral education by seamlessly integrating the capabilities of deep learning models and the adaptive nature of swarm intelligence algorithms.

In the dynamic landscape of language education, the amalgamation of deep learning and swarm intelligence algorithms present a groundbreaking avenue for the realization of personalized English oral instruction (Wu et al., 2024; Zaidi, 2021). This paper embarks on a journey to explore the synergies between deep learning methodologies and swarm intelligence algorithms, aiming to create an intelligent English oral assessment system that adapts and evolves through collective intelligence. To advance language education, the researchers aim to create a cutting-edge system that utilizes deep

DOI: 10.4018/IJSIR.343989

*Corresponding Author

learning for accurate assessments while also incorporating swarm intelligence to enhance and tailor the model for personalized learning experiences.

Deep learning stands as a formidable force in the realm of artificial intelligence, representing a paradigm where machines autonomously unravel intricate patterns from extensive datasets (LeCun, Bengio, & Hinton, 2015; Lopez, 2023). Its prowess lies in the construction of complex neural networks, inspired by the human brain, enabling the model to discern hierarchical features and generate predictions. This groundbreaking technology has led to significant advancements in various fields, including image recognition, speech synthesis, and the comprehension of natural language. Within the context of this research, deep learning serves as the catalyst for an intelligent English oral assessment system, imparting nuanced evaluation capabilities to discern the subtleties of spoken language and instill a tailored learning experience through adaptive feedback mechanisms. Swarm intelligence, drawing inspiration from the harmonious collaboration observed in social insect colonies and flocks of birds, presents a decentralized and self-organizing approach to problem solving. This unconventional paradigm harnesses the collective power of individual entities to make decisions independently, fostering adaptability and efficiency. In algorithmic terms, swarm intelligence mimics the synergy found in nature to optimize complex systems. By integrating this innovative approach into the proposed research framework, the researchers infuse the deep learning architecture with a dynamic layer. Swarm intelligence becomes the orchestrator, steering the continuous evolution and refinement of the English oral assessment system, transforming it into an adaptive and collaborative entity that responds organically to the unique nuances of individual learners.

At the core of this methodology lies the deployment of state-of-the-art deep learning architectures to craft an intelligent English oral assessment system (Alom et al., 2019; Balaban, 2015). Through these advanced neural network models, the researchers aspire to create a system adept at evaluating multiple facets of spoken language, including pronunciation, fluency, and grammatical accuracy. This system promises instantaneous and precise feedback, cultivating a tailored and adaptive learning environment that addresses the distinct linguistic requirements of each learner. To further enhance the adaptability and efficacy of the proposed deep learning-based assessment system, the researchers introduce swarm intelligence algorithms into the framework. Swarm intelligence, derived from the collective behaviors of decentralized entities, is employed to continually refine and optimize the neural network model. This collaborative approach, fueled by the interactions, preferences, and feedback of users, fosters a system that evolves dynamically, catering to the diverse learning styles and preferences of individual users. In this research, swarm intelligence algorithms are utilized to optimize the training process of deep learning models, particularly in parameter tuning and network architecture optimization. Swarm intelligence algorithms, such as Ant Colony Optimization (ACO), mimic the behavior and interaction patterns of biological swarms in nature to solve complex problems. Within the context of deep learning models, these algorithms can efficiently search the parameter space to find configurations that maximize model performance. Moreover, the parallel search mechanism of swarm intelligence algorithms speeds up the optimization process, reducing the computational resources and time required to achieve optimal solutions. This is particularly important for handling large datasets and complex models, making this research more practical and innovative in the field of English speech recognition.

This research not only contributes to the technological landscape of education, but also responds to the growing demand for personalized language learning solutions. As researchers navigate the intricate web of linguistic diversity and varying proficiency levels, the fusion of deep learning and swarm intelligence emerges as a promising paradigm for transforming English oral education into an adaptive, engaging, and personalized journey. Through this exploration, the researchers aim to shed light on the transformative potential of intelligent systems in shaping the future of language education, unlocking new dimensions for customized and efficient learning experiences. The main innovation is as follows. First, the proposed model innovatively combines deep learning techniques to extract a wide array of features from English language inputs. This approach allows for a more nuanced and comprehensive analysis of language components, significantly improving the accuracy of scoring by capturing the intricacies of language use that are often missed by conventional methods. Second, the researchers employ swarm intelligence algorithms as a novel method to optimize the performance of their English scoring model. This approach harnesses the collective behavior of decentralized, self-organized systems to fine-tune the model parameters. By simulating the natural intelligence of swarms, the proposed model dynamically adjusts and evolves, leading to enhanced decision-making capabilities and superior overall performance.

The researchers arrange the overall structure and content of the entire paper as follows: The "Related Work" section explains the current status of related research. The "Proposed Method" section elaborates on the proposed method, including multiple feature extraction and fusion and model training with swarm intelligence. The "Experiments and Analysis" section systematically validates the performance of the model based on public datasets. The final section, "Conclusions" summarizes the full paper.

RELATED WORK

In the landscape of English oral proficiency assessment, a multitude of methodologies and models have been developed to gauge linguistic competence, reflecting the ever-evolving intersection of technology and language education. This section provides a comprehensive overview of the existing approaches employed in the realm of English oral assessment, ranging from traditional methods to contemporary technological advancements. By examining the current state of the field, the aim is to contextualize this present research within the broader landscape, and highlight the gaps that motivate the exploration into the integration of deep learning and swarm intelligence for a more personalized and effective English oral education experience.

Research on automatic scoring of spoken language was carried out relatively early. As early as 2000, Witt and Young (2000) proposed a classic algorithm for scoring reading questions: the global optimization (GOP) algorithm, and it has been widely used in scoring of spoken pronunciation. However, the limitation of this algorithm is that it is a text-dependent algorithm, so the GOP algorithm is only suitable for those question types where the answers are unique. The most typical application scenario of this algorithm is the pronunciation scoring of reading questions. de Wet, Van der Walt, and Niesler (2009) conducted feature extraction on the data set of reading questions, and extracted three features: speaking speed, pronunciation, and accuracy. The pronunciation feature, leveraging the GOP algorithm, alongside three other features, was utilized to evaluate students' spoken language proficiency, resulting in a record human-machine rating correlation of 0.72. Klaus et al. (Year) have done a lot of work on a variety of speaking question types in the automatic speaking scoring project in the "TOEFL" online test, and proposed that in actual deployment, a 0.17 difference in scoring correlation with human raters is enough. The automatic scoring system can be considered effective (Zechner, Higgins, Xi & Williamson, 2009) and Automated Speech Recognition, (ASR) technology was used in speaking scoring to score two open-ended speaking questions. However, the effect is ideal. The correlation between human and machine scores is only 57.2%, which is still far from actual deployment. Su et al. (Year) proposed a human-computer spoken language-scoring method in 2017. Yoon and Zechner (2017) used a screening system to screen out some speech sounds that were difficult to ensure accurate machine scoring, and then used manual scoring methods. This part of the speech was evaluated, with manual scoring acting as a support to enhance the reliability of the automated scoring system. This approach also introduces an additional concept: employing minimal human assistance to increase the precision of the automatic scoring system. This approach not only reduces the burden caused by manual scoring, but also allows automatic scoring. The system becomes more reliable. Tao, Ghaffarzadegan, Chen, and Zechner (2016) studied a method of using deep learning to improve speech recognition. The human-machine scoring correlation of the spoken language scoring system SpeechRater, based on speech recognition, has been effectively improved. It can be seen that the accuracy of speech recognition is crucial to the spoken language scoring system (Tao, Ghaffarzadegan, Chen & Zechner, 2016).

In studying the relevant literature on automatic oral correction in foreign countries, the researchers found that because foreign language training methods and examination methods are somewhat different from those in China, foreign research on automatic oral correction mainly focuses on reading questions and follow-up questions. The research help for this article was limited, but many classic scoring algorithms are worthy of learning and reference. Based on the study of these classic algorithms, this article will improve some of the algorithms and apply them to the multi-feature intelligent correction model.

Tao et al. (2016) and Cheng, Chen and Metallinou (2015) used hidden Markov models based on deep neural network (DNN) and Gaussian mixture (GMM), respectively, to conduct speech recognition and spoken language scoring experiments. Finally, it was found that the speech recognition system based on DNN had better performance in recognition accuracy and various types of spoken language scoring tasks. Knill et al. (2018) studied the impact of the recognition accuracy of the speech recognition engine on open-ended speaking scores. Among them, the impact of word error rate (WER) on speech feature extraction and text feature extraction was mainly discussed.

Chen, Zechner, and Xi (2009) proposed a new method for extracting pronunciation features for open-ended spoken language. This method forcibly aligned the spoken language recording with the text translated by the speech recognition engine based on the acoustic model trained with standard pronunciation, and calculated the pronunciation likelihood and vowel duration characteristics. Using these two basic features, more pronunciation-related features could be calculated.

The spoken rhythm of non-native speakers may deviate from the rhythm pattern of native speakers. The introduction of rhythm features measures the mastery of English by non-native speakers from another perspective. Chen and Zechner (2011) used the method previously mentioned to extract rhythm features, and conducted comparative experiments between a scoring system without rhythm features and a scoring system containing rhythm features, and found that the introduction of rhythm features could improve the correlation between machine scoring and human scoring. Xie, Evanini, and Zechner (2012) used three different methods to extract similarity features between speech recognition text and reference text. The vector space model (VSM) method can extract lexical similarity features; the latent syntax analysis (LSA) and point mutual information (PMI) methods can extract semantic similarity features. Finally, experiments have proved that the features extracted by these three methods have a high correlation with manual scoring. In order to make the scoring system more portable, Zesch, Wojatzki, and Scholten-Akoun (2015) further divided text features into two categories for research. One type was features that were strongly related to the task, such as: text similarity, semantic similarity, corpus-based features; the other type was features that were weakly related to the task, such as sentence structure, grammatical errors, and article length. Finally, experiments proved that the scoring model trained using the second type of features had better adaptability and portability when facing different scoring tasks.

Zechner, Higgins, Xi, and Williamson (2009 proposed the architecture of the classic open speech scoring system SpeechRater. SpeechRater has a profound influence on the subsequent development of this research field, so that for a long time, various improved SpeechRater scoring systems have been designed with its architecture ideas. SpeechRater is designed for the TOEFL Speaking Online Test, a system that extracts a rich set of features and constructs scoring models using a classified regression tree algorithm and a multiple linear regression algorithm, respectively. Although the correlation of the human-machine score only reached 57.2% in the end, there was still a certain distance from the real practical application, but it provided research ideas for future generations. In recent years, some researchers have tried to use deep neural network techniques to improve the accuracy of scoring models. In the oral scoring system studied by Metalinou and Cheng (2014), a neural network scoring model, based on error backpropagation algorithm and multiple linear regression scoring models, were used for testing, and the results showed that the neural network scoring model performed better in all test data sets. In addition, with the rise of deep learning technology, it is possible to abstract high-level features from the low-level representation of data by using its automatic learning function, which does not require professional feature engineering knowledge to provide richer feature information for the scoring model. Chen, Tao, Ghaffarzadegan, and Qian (2018) studied and designed an "end-to-end" oral scoring system, which utilized a convolutional neural network and a long short-term memory network (a recurrent neural network) to realize automatic feature learning and extraction. However, the final experimental results showed that the human-machine scoring correlation had not made a major breakthrough, and it was even worse than some traditional scoring systems that require manual feature extraction. However, this study provides one with a new research idea, that is, deep learning technology can be used to enable the machine to carry out automatic feature mining of spoken recordings. If the model is sufficiently optimized, the features obtained in this way will better reflect the characteristics of the real data, thus improving the accuracy of the system's scoring. Qian et al. (2019) built a scoring model based on the long short-term memory network. The model carries out feature learning from three dimensions (pronunciation, language use, and spoken content), and scores each dimension separately. In addition, the model is significantly better than the traditional scoring model in the human-machine scoring correlation test results. This also shows that the neural network-scoring model based on deep learning technology can achieve better scoring performance than the scoring system based on feature engineering combined with traditional machine learning algorithms.

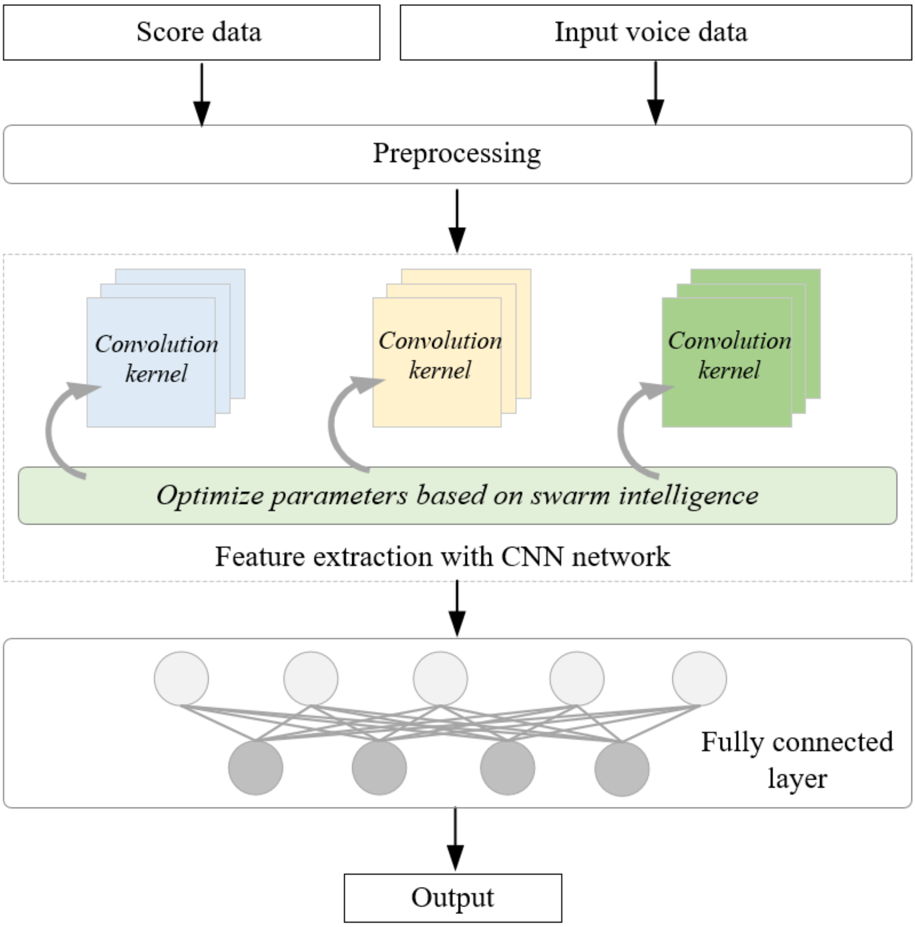
Traditional methods of English oral assessment have long relied on human evaluators, often burdened by subjectivity and resource limitations. Despite their valuable insights, the scalability and objectivity of such approaches remain areas of concern. To address these limitations, automated scoring systems have gained prominence, leveraging a spectrum of techniques ranging from rule-based algorithms to statistical models. However, these conventional automated systems often struggle to capture the intricacies of spoken language, such as nuanced pronunciation variations and contextual understanding.

Furthermore, recent advancements in machine learning, particularly in the domain of deep learning, have ushered in a new era of automated assessment models. These models, fueled by neural network architectures, exhibit enhanced capabilities in recognizing complex patterns and representations in spoken language, providing a more nuanced evaluation. As the researchers delved into the existing literature, they explored the strengths and limitations of these diverse approaches, setting the stage for the innovative fusion of deep learning and swarm intelligence in their pursuit of a truly personalized English oral assessment system.

THE PROPOSED METHOD

This section describes in detail the proposed method for intelligent English scoring model. In this paper, the researchers propose a multi-feature intelligent English scoring model that combines the advantages of feature extraction of deep learning models and uses swarm intelligence algorithms to optimize model performance, ultimately achieving superior performance. The overall structure diagram is shown in Figure 1, which contains a total of two parts. The first part is the multi-feature extraction and fusion part based on deep learning. The second part is based on the swarm intelligence algorithm to optimize the deep learning model to improve the overall performance of the model. The initial part of this approach focuses on the multifaceted process of feature extraction and fusion, employing deep learning techniques. This involves systematically identifying and integrating various linguistic and paralinguistic elements from English language inputs, such as syntax, semantics, pronunciation, and intonation. By leveraging advanced deep learning architectures, such as CNNs, the proposed model is capable of discerning subtle nuances in language use, thereby ensuring a comprehensive and detailed analysis of the input data. The subsequent segment is dedicated to refining and enhancing the deep learning model's efficiency and accuracy through the application of swarm intelligence algorithms. This innovative method

Figure 1. The Overall Structure Diagram of the Proposed Model



draws inspiration from the collective behavior observed in natural systems, such as flocks of birds or colonies of ants, to conduct a decentralized search for optimal model parameters. By simulating these biological processes, the algorithm iteratively adjusts the weights and structures within the deep learning model, aiming to find the most effective combination that maximizes the model's performance. This optimization process not only improves the model's scoring accuracy, but also significantly increases its adaptability and scalability in evaluating English proficiency.

Multiple Feature Extraction and Fusion

For the intelligent scoring model, the selection of scoring features is very important, which determines the reliability and accuracy of the proposed scoring model. The oral questions studied in this paper are personal statement questions based on listening comprehension, which are most similar to the retelling questions. Therefore, the feature extraction of this model draws on the automatic scoring design of the retelling questions of predecessors to a certain extent.

In previous studies, there are mainly the following kinds of retelling questions. The first type is listening-based retelling (Yokouchi, 2015), that is, let the examinee listen to a piece of audio, and then retell according to the content of the audio. This type of question does not require students to reproduce the original text word-for-word, but requires students to fully express the meaning of the original text as far as possible, and does not require students to employ divergent thinking according to the audio theme. The second type is reading-based retelling, which requires candidates to read an article for a certain period of time, then the article is hidden, and students are allowed to describe the content of the passage by memory. Students can either do it all in their own words, or they can recite it. This type of question is basically the same as the listening-based retelling question type, but it is simpler compared to the lack of examination of the candidate's listening. The third is the translation question (Bao, Duan, Zhou, & Zhao, 2014; Jizzakh, 2020), the method of investigation is to give students a Chinese article, let the students translate it into English, and express it. The difference between it and the previous two types of questions is that in the translation process, the model essay is not hidden, but it is similar in technology. Therefore, technically speaking, this type of question can be regarded as a paraphrase.

The research in this paper is based on the systematically collected data of the speaking test, which contains the open-ended speaking questions studied in this paper. The type of question the researchers studied works as follows: Test takers watch a video, which can only be watched once; they then discuss the content of the video based on the topic of the video and a few given questions. The question type combines listening and video, which is closer to reality in form. The main purpose of the video is to help candidates understand the audio text. It can be seen that it is a very flexible question type, and the candidates' answers are ever changing. Candidates not only need to retell the original text to a certain extent, but also need to employ divergent thinking to discuss the original video and state their own views. Thus, it can be seen that although the oral question type studied was very close to the aforementioned three types of retelling questions, it was more open, students' answers were more unpredictable, and teachers could not even provide sample questions. For the automatic scoring of listening retelling questions, the current scoring method needs to rely on the key words of experts, the data of examinee reading questions, and the sample of retelling questions given by experts, which has certain limitations. For the type of questions studied in this paper, the feature extraction method cannot be directly applied. Therefore, this paper will combine the manual correction standard, put forward a new choice of extracting spoken language features, and completely eliminate manual assistance for automatic correction.

According to the selection of scoring features, one needs to carry out feature extraction of speech. The work of feature extraction is mainly divided into four modules, including a pronunciation-scoring module, a fluency scoring module, a grammar and vocabulary scoring module, and a semantic scoring module. The grammar and vocabulary scoring modules will output two eigenvalues, and the other modules will output one eigenvalue. In order to make the feature extraction work proceed smoothly and improve accuracy, the researchers also designed three data processing modules in the scoring model, including a voice noise-reduction module, a voice recognition module, and a text-cleaning module. The scoring flow chart of the entire scoring model is shown in Figure 1.

The researchers chose the convolutional neural network (CNN) as the main skeleton to extract data features. They briefly reviewed convolutional layers, pooling layers, fully connected layers, and batch normalization, which are the main components of CNN models. The convolutional layer is one of the important components of the neural network. Its function is to extract the characteristics of the data through convolution calculation on the data input by the input layer. During each convolution movement, the eigenvalues $h_{j,k}$ at that location will be calculated by summing the channels. The specific calculation formula is shown as Equation 1:

$$h_{j,k} = \sigma\left(\sum_{s=1}^{s} w_j^s * x_k^s + b_{j,k}\right) \tag{1}$$

where $\sigma()$ is the activation function; * represents linear convolution; w_j^s is the weight vector of the j-th filter acting on the local input x_k^s of the s-th input channel; $b_{j,k}$ is the bias.

The function of the pooling layer is to select and filter the feature information extracted by the convolutional layer to prevent excessive data from affecting the running speed of the network and, at the same time, remove noise interference. After the features pass through the pooling layer, their size will be reduced. The function of the fully connected layer is to integrate the features obtained in the previous steps and transform the input feature vector z by applying affine transformation and nonlinear mapping, as shown in Equation 2:

$$h_i = \sigma\left(W_i \times z + b_i\right) \tag{2}$$

where Wi and bi are the i-th weight vector and bias, respectively. In the last layer, softmax(·) is used to obtain the probability of class s for the input feature vector x_t, as shown in Equation 3:

$$p(s \mid x_\iota) = softmax\left(x_\iota\right) = \frac{\exp\left(w_s y^L\right)}{\sum_{n=1}^{N} \exp\left(w_s y^L\right)} \tag{3}$$

where L is the number of hidden layers; N is the number of output units. Different from layer normalization, which is a horizontal normalization, batch normalization is a vertical normalization. The goal of this normalization is to make these neural units have different means and variances for each neural unit in the same layer in the batch direction to prevent gradient explosion. Batch normalization approximates whitening by normalizing the intermediate representation using the statistics of the current mini-batch to eliminate the undesirable effects of internal covariate shifts.

Model Training With Swarm Intelligence

Once the model design phase was finalized, the next step involved training the neural network to acquire the most suitable parameters. While the training approach for convolutional neural networks bears resemblance to that of traditional neural networks, its unique architecture results in distinct error backpropagation and forward activation across various layers. In this paper, the researchers used swarm intelligence algorithms to design and optimize network parameters to achieve optimal training of the model.

The stochastic gradient descent algorithm is a commonly used optimization algorithm to update the parameters in the network model based on the gradient of each training sample during neural network training to gradually reduce the loss function and minimize it. This algorithm is widely used because of its advantages such as fast convergence speed and simple operation. However, this algorithm also has some shortcomings. For example, an excessively big learning rate may cause the algorithm to cross the optimal point and fail to converge, while an excessively small learning rate may easily fall into a saddle point and local optimum. In order to solve these problems, researchers have successively proposed some new optimization algorithms; for example: Momentum (Chan, Jegadeesh & Lakonishok, 1996), Adam (Kingma & Ba, 2014), Adaptive Gradient Clipping (AdaGrad) (Ward, Wu & Bottou, 2020), and Nesterov gradient descent algorithms (Qu & Li, 2019). However, there are still some unresolved problems in the above improvement methods, such as Momentum, Nesterov, AMSGrad, and Adam have insufficient stability. Due to the advantages of swarm intelligence algorithms, such as global optimization and strong robustness, relevant scholars have begun to combine swarm intelligence algorithms with gradient descent to optimize the training of DCNN. However, as the dimension increases, the search space of swarm intelligence geometrically doubles, causing it to exhibit low search efficiency and a large amount of calculation. In view of the problems that classic optimizers are prone to localization or lack of stability, and existing optimizers based on swarm intelligence have poor performance, the researchers used gradient information as additional guidance, combined with the idea of continuous domain ant colony optimization, and proposed a method based on the Swarm intelligence optimization algorithm NACO-SGD. In this optimizer, each sample batch is associated with an ant, and ant collaboration is exploited to increase the diversity of optimization parameters and, thereby, jump out of the local optimum.

In the ACO algorithm, each ant represents a candidate solution. The search process of ants is similar to the ant colony algorithm, and each ant will move according to the pheromone concentration near its location. Different from traditional discrete domain ACO, the continuous domain ACO algorithm uses continuous parameter values in the search space, which can search the solution space more flexibly. In neural network training, weights and biases are continuous real values, which is a continuous optimization problem. Compared with ACO, continuous domain ACO is more suitable for neural network training. When training a feedforward neural network, the ACO algorithm can be used to optimize the weights and biases of the neural network. The weight and bias term of each neuron can be viewed as a vector, and the ACO algorithm can optimize the values of these vectors to minimize the loss function of the neural network.

The NACO-SGD algorithm combines the continuous domain ant colony optimization algorithm ACO and the gradient descent algorithm SGD. It uses the ACOR algorithm for optimization at a certain frequency between different Mini-Batches. The object sampled by the ants is the gradient information under the historical Mini-Batch. Specific steps are as follows. First, the data set is divided into N mini-batches of equal size, with K samples in each mini-batch. In the SGD optimization part, a Mini-Batch is randomly selected, denoted as S_t, the gradient of the loss function of each sample is calculated, and the average value of the gradient is calculated. Then, the calculated average gradient is used to update the parameters of the model. The calculation formula is shown in Equation 4:

$$\omega_{t+1} = \omega_t - \eta \frac{1}{K} \sum_{\left(x,y\right)\in S_t} \frac{\partial \mathcal{L}\left(y, f\left(x;\omega\right)\right)}{\partial \omega} \tag{4}$$

where $\mathcal{L}(y, f\left(x;\omega\right)$ is the loss function, ω_t represents the weight parameter of the t step gradient descent, η represents the learning rate.

After each Mini-Batch training, the loss function value of the Batch and the gradient of each parameter in the network are stored into the archive table T until the archive table stores the corresponding loss value and gradient information of k Mini-batches. The gradient of the algorithm is updated after each Mini-Batch processing. The traditional continuous domain ACO algorithm performs roulette selection on the historical solutions in the archive table, generates the corresponding Gaussian distribution for the selected solution, and then samples the Gaussian distribution to obtain a new solution. Different from the traditional ACO algorithm, the proposed algorithm uses the mean and variance of the historical gradients of the past k batches as the mean and variance of the Gaussian distribution to generate a Gaussian distribution. Specifically, the formula for generating a Gaussian distribution is shown in Equation 5:

$$f\left(x\right) = \frac{1}{\sigma_{S_{t,t+k}} \sqrt{2\pi}} e^{-\frac{\left(x - \nabla \mathcal{L}_{S_{t,t+k}}\right)^2}{2\zeta \sigma_{S_{t,t+k}}^2}} \tag{5}$$

where $\nabla\mathcal{L}_{S_{t,t+k}}$ represents the mean of the historical gradient of the k batches, $\sigma_{S_{t,t+k}}$ represents the standard deviation of the historical gradient, ξ is shown in the standard deviation coefficient, and k represents the number of historical batches stored in the archive table T. The gradient obtained by ACO method is generated by sampling the Gaussian distribution. Finally, the newly generated gradient is used for gradient descent. After this round of Mini-Batch training, the archive table is cleared and the next round of k Mini-Batch training is reentered until the end of the training. In this step, the mean and variance are calculated iteratively to reduce the storage space in the actual program.

EXPERIMENTS AND ANALYSIS

In this section, the researchers evaluate their proposed multi-feature intelligent English scoring model based on the deep learning model and swarm intelligence discussed in this paper. The experimental section of this study serves as a pivotal component, elucidating the methodology employed to validate the efficacy and performance of the proposed deep learning-based English oral assessment system augmented with swarm intelligence. This section encapsulates a comprehensive overview of the experimental framework, encompassing aspects such as data preprocessing, evaluation metrics, experimental environment, model validation, and comparative analysis.

Dataset Processing

To ensure the robustness and reliability of the findings, meticulous attention was devoted to the preprocessing phase of the dataset. This entailed data cleaning, normalization, and feature extraction to optimize the input data for model training and evaluation. By standardizing the dataset, the researchers mitigated potential biases and inconsistencies, thereby enhancing the generalizability of their results. The data used in this article came from the oral examination of a university's situational English course. From the examination data from 2014 to 2020, the researchers extracted the recordings of 800 candidates' answers to the same open-ended speaking question (each recording was about 60 seconds) and the teacher's manual scoring data. The method designed in this article requires separate scoring of spoken pronunciation and spoken content, so the researchers also asked teachers to separately score candidates' speaking from these two aspects. In addition, before training and testing the model, the researchers also need to convert the recording format. The recording files collected from the oral exam are all in MP3 format, and the audio attributes are 16 bit and 16 kHz sampling rate. FFmpeg is an open source tool that specializes in processing video and audio streams. The researchers used this tool to convert recordings in MP3 format to PCM format. Finally, the researchers divided the entire data set into two groups for training and testing, respectively. The training set contained 500 pieces of data, and the test set contained 300 pieces of data.

Evaluation Metrics

In tandem with data preprocessing, the selection of appropriate evaluation metrics was paramount in quantifying the performance of the proposed model accurately. The researchers employed a suite of metrics tailored to assess various aspects of English oral proficiency, including accuracy, fluency, pronunciation accuracy, and grammatical correctness. These metrics provided a holistic understanding of the system's effectiveness in capturing the nuances of spoken language and facilitating personalized learning experiences. In order to comprehensively evaluate and compare the scoring performance of the two models, some evaluation indicators were used to quantitatively analyze the experimental results. First of all, one of the most commonly used performance test indicators in the field of automatic scoring is the Pearson correlation coefficient, which is mainly used to reflect the linear correlation between two sequences. Its mathematical expression is shown in Equation 6:

Table 1. The Comparison Results of Different Methods

Method	ρ	d	Accuracy
BiLSTM	0.43	0.68	86.4%
Seq2Seq	0.54	0.66	89.1%
SDEN	0.63	0.59	91.5%
Ours	0.71	0.53	93.5%

$$\rho_{X,Y} = \frac{cov\left(X,Y\right)}{\sigma_X \sigma_Y} \tag{6}$$

where $cov\left(X,Y\right)$ is the rhombic difference of X and Y, and both σ_X and σ_Y are standard differences. Here, X and Y represent system ratings and teacher ratings, respectively. The value range of P is [-1, 1]. If the result is a positive number, it means there is a positive correlation between the two sequences. If it is a negative number, it means there is a negative correlation between the two sequences. The closer the absolute value of this coefficient value is to 1, the higher the correlation between the two sequences. The second evaluation index is the average score difference between human and machine scoring, which is mainly used to describe the deviation between machine scoring and manual scoring. At the same time, the researchers used the third evaluation metric accuracy. They use the letter d to represent second metric, and its calculation formula is shown in Equation 7:

$$d = E\left|\mathcal{S}_{Machine} - \mathcal{S}_{Human}\right| \tag{7}$$

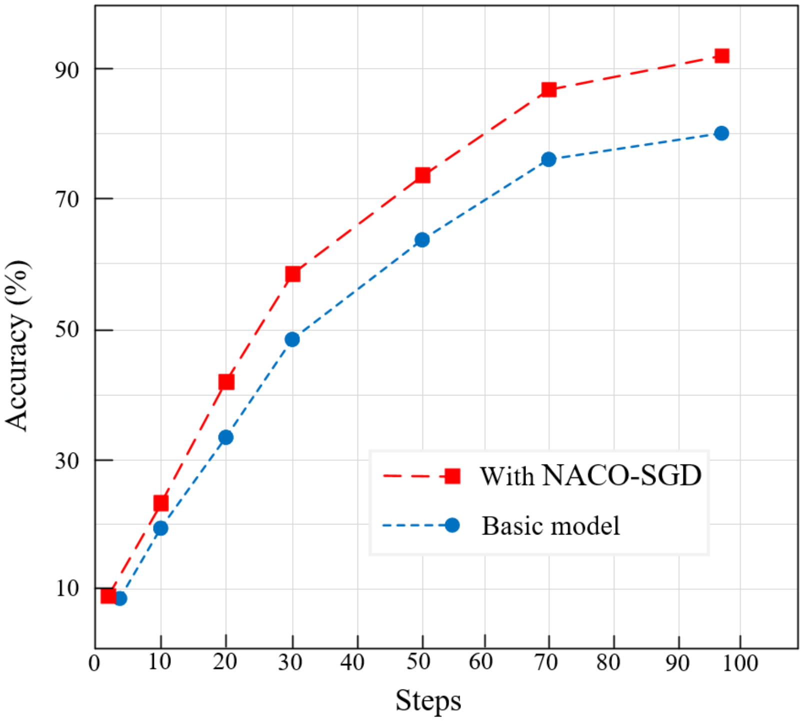
Performance Comparison

Following model training and validation, rigorous performance assessment was conducted to ascertain the efficacy of the proposed system. Through comprehensive validation and cross-validation techniques, the researchers meticulously evaluated the system's performance across processed datasets. Additionally, comparative analyses were conducted against baseline models and existing state-of-the-art approaches to benchmark the performance gains facilitated by the integration of deep learning and swarm intelligence. To verify the effectiveness of the model proposed in this article in the English spoken comprehension-scoring task, the researchers selected multiple spoken language comprehension models for comparison, including the two-layer BiLSTM model, the sequence-to-sequence model (Seq2Seq) (Zhou, Sun, Liu & Lau, 2015), and the hierarchical coding LSTM model of historical information (SDEN) (Bapna, Tur, Hakkani-Tur & Heck, 2017).

Table 1 shows the comparison results of different methods. One can see two conclusions from the table: (1) The proposed personalized English speaking model based on deep learning and swarm intelligence optimization achieved the best results; (2) In addition to the proposed model, the SDEN model achieved optimal results, because it incorporated current-level conversation-level coding and added attention weights between historical information and the current conversation.

In addition, the researchers designed comparative experiments to verify the optimization algorithm NACO-SGD based on swarm intelligence proposed in this article. Figure 2 shows the comparison results in the two cases. The first case is the basic optimization model, that is, the basic optimization method is directly used for training without using the optimization strategy proposed in this article. The second case is to use the swarm intelligence-based optimization method NACO-SGD proposed

Figure 2. The Comparison Results in Two Cases of With or Without NACO-SGD Method



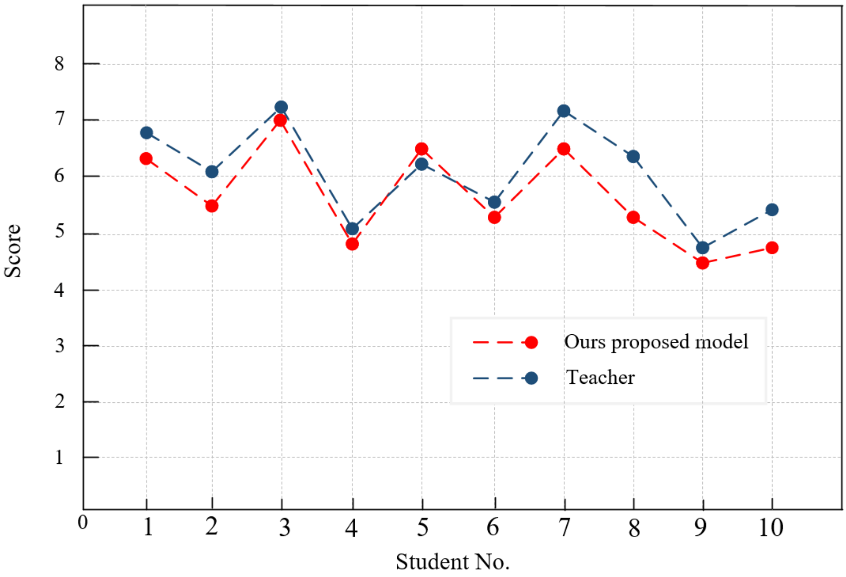
in this article to optimize model parameters. By comparing the model accuracy during the training process, one can see that with the support of the NACO-SGD method, the performance of the model has been greatly improved, which shows that the proposed method can effectively optimize the model parameters.

Furthermore, in order to verify the significance and role of the model proposed in this article for English speaking education, the researchers compared the teacher evaluation results and the model scoring results. This experiment determined whether the scoring results of the proposed model were consistent with the teacher's scoring results, so that the effectiveness of the proposed model could be seen more clearly. The researchers selected 10 students as test samples. First, the researchers used the proposed model for scoring. At the same time, they arranged for three teachers to conduct grading. Finally, the researchers averaged the ratings of the three teachers to get the teacher's final rating. Figure 3 shows the comparison of the scoring gap between the model and the teacher. It can be seen from the results in the figure that the scores predicted by the model are basically consistent with the teacher's ratings, which shows that the model proposed in this article has certain application value. This also demonstrates the effectiveness of using deep learning and swarm intelligence algorithms for English speaking education.

CONCLUSION

In conclusion, this research marks a significant stride towards personalized English oral education by harnessing the synergies between deep learning and swarm intelligence. Through the development and implementation of an intelligent English oral assessment model, the researchers have demonstrated the feasibility and efficacy of leveraging advanced technologies to cater to individual learner needs. Their experiment provides a methodology for assessing the effectiveness of their proposed approach

Figure 3. The Comparison of the Scoring Gap Between the Model and the Teacher



in enhancing personalized English oral education. Through meticulous data preprocessing, rigorous evaluation metrics, robust experimental environments, and comprehensive performance validation, the researchers aim to elucidate the transformative potential of their innovative model in revolutionizing English oral assessment and fostering tailored learning experiences. Their experimental results highlight the system's ability to provide nuanced evaluation and adaptive feedback, thereby enhancing the overall learning experience. Looking ahead, the researchers envision further advancements in the integration of intelligent systems into language education, facilitating tailored and efficient learning experiences for learners worldwide. As researchers continue to explore the frontiers of technology-enhanced education, the fusion of deep learning and swarm intelligence holds immense promise in revolutionizing language learning paradigms and empowering learners to achieve fluency and proficiency in English oral communication.

CONFLICT OF INTEREST

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

FUNDING

No funding was received for this work.

PROCESS DATES

Received: 2/20/2024, Revision: 3/12/2024, Accepted: 3/12/2024

CORRESPONDING AUTHOR

Correspondence should be addressed to Yang Liu; 201503012@hhstu.edu.cn

REFERENCES

Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., Hasan, M., Van Essen, B. C., Awwal, A. S., & Asari, V. K. (2019). A state-of-the-art survey on deep learning theory and architectures. *Electronics (Basel)*, *8*(3), 292. doi:10.3390/electronics8030292

Bagea, I. (2023). Cultural influences in language learning in a global context. *Indo-MathEdu Intellectuals Journal*, *4*(2), 630–645. doi:10.54373/imeij.v4i2.248

Balaban, S. (2015). Deep learning and face recognition: The state of the art. *Biometric and Surveillance Technology for Human and Activity Identification XII*, *9457*, 68–75.

Bao, J., Duan, N., Zhou, M., & Zhao, T. (2014, June). Knowledge-based question answering as machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (*Volume 1*: Long Papers)* (pp. 967-976).

Chan, L. K. C., Jegadeesh, N., & Lakonishok, J. (1996). Momentum strategies. *The Journal of Finance*, *51*(5), 1681–1713. doi:10.1111/j.1540-6261.1996.tb05222.x

Chen, L., Tao, J., Ghaffarzadegan, S., & Qian, Y. (2018, April) End-to-end neural network based automated speech scoring. In *2018 IEEE International Conference on Acoustics, Speech and signal Processing (ICASSP)* (pp. 6234-6238). IEEE. doi:10.1109/ICASSP.2018.8462562

Chen, L., & Zechner, K. (2011). Applying rhythm features to automatically assess non-native speech. In *Twelfth Annual Conference of the International Speech Communication Association.* doi:10.21437/Interspeech.2011-506

Chen, L., Zechner, K., & Xi, X. (2009, June). Improved pronunciation features for construct-driven assessment of non-native spontaneous speech. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 442-449). doi:10.3115/1620754.1620819

Cheng, J., Chen, X., & Metallinou, A. (2015). Deep neural network acoustic models for spoken assessment applications [J]. *Speech Communication*, *73*, 14–27. doi:10.1016/j.specom.2015.07.006

de Wet, F., Van der Walt, C., & Niesler, T. R. (2009). Automatic assessment of oral language proficiency and listening comprehension. *Speech Communication*, *51*(10), 864–874. doi:10.1016/j.specom.2009.03.002

Jizzakh, A. Z. (2020). Marshak and Shakespeare: The question of new content in translation. *Mental Enlightenment Scientific-Methodological Journal,* 149-158.

Knill, K., Gales, M., Kyriakopoulos, K., Malinin, A., Ragni, A., Wang, Y., & Caines, A. (2018, September). Impact of ASR performance on free speaking language assessment. In *Interspeech 2018 International Speech Communication Association* (pp. 1641–1645). ISCA. doi:10.21437/Interspeech.2018-1312

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. doi:10.1038/nature14539 PMID:26017442

López, C. (2023). Artificial intelligence and advanced materials. *Advanced Materials*, *35*(23), 2208683. doi:10.1002/adma.202208683 PMID:36560859

Metallinou, A., & Cheng, J. (2014). Using deep neural networks to improve proficiency assessment for children English language learners. In *Fifteenth Annual Conference of the International Speech Communication Association*. doi:10.21437/Interspeech.2014-358

Qian, Y., Lange, P., Evanini, K., Pugh, R., Ubale, R., Mulholland, M., & Wang, X. (2019, May). Neural approaches to automated speech scoring of monologue and dialogue responses. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 8112-8116). doi:10.1109/ICASSP.2019.8683717

Qu, G., & Li, N. (2019). Accelerated distributed Nesterov gradient descent. *IEEE Transactions on Automatic Control*, *65*(6), 2566–2581. doi:10.1109/TAC.2019.2937496

Tao, J., Ghaffarzadegan, S., Chen, L., & Zechner, K. (2016, March). Exploring deep learning architectures for automatically grading non-native spontaneous speech. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (pp. 6140-6144). IEEE. doi:10.1109/ICASSP.2016.7472857

Ward, R., Wu, X., & Bottou, L. (2020). Adagrad stepsizes: Sharp convergence over nonconvex landscapes. *Journal of Machine Learning Research*, *21*(1), 1–30.

Witt, S. M., & Young, S. J. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication*, *30*(2-3), 95–108. doi:10.1016/S0167-6393(99)00044-8

Xie, S., Evanini, K., & Zechner, K. (2012, June). Exploring content features for automated speech scoring. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 103-111).

Yokouchi, Y. (2015). Do input text length and presentation mode affect speaking performance in retelling tasks?[J]. *JLTA Journal, 18*, 115-133.

Yoon, S. Y., & Zechner, K. (2017). Combining human and automated scores for the improved assessment of non-native speech. *Speech Communication*, *93*, 43–52. doi:10.1016/j.specom.2017.08.001

Zaidi, A. (2021). The role of machine learning in personalised instructional sequencing for language learning. (Doctoral dissertation).

Zechner, K., Higgins, D., Xi, X., & Williamson, D. M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication*, *51*(10), 883–895. doi:10.1016/j.specom.2009.04.009

Zesch, T., Wojatzki, M., & Scholten-Akoun, D. (2015, June). Task-independent features for automated essay grading. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 224-232). doi:10.3115/v1/W15-0626