Health, Medicine and Therapeutics
ISSN Print: 3079-5079
ISSN Online: 3079-5087
https://doi.org/10.63313/hmt.9014

Research on Speech Rehabilitation Assessment Models Based on Natural Language Processing

Yifan Zhou[a,*]. Yutong Yang[b] and Lingqin Jiang[c]
Heilongjiang University of Science and Technology, Harbin, Heilongjiang, 150000, China
Email: [*a]13320008806@163.com. [b]YangRainT@outlook.com. [c]18286289304@163.com

How to cite this paper: Zhou, Y. F., Yang, Y. T., & Jiang, L. Q. (2025). Research on Speech Rehabilitation Assessment Models Based on Natural Language Processing. *Health, Medicine and Therapeutics*, 1(2), 58–65. Print ISSN: 3079-5079; Online ISSN: 3079-5087.
https://doi.org/10.63313/hmt.9014
Published: 2025-12-08

Abstract

Artificial intelligence technologies present new developmental opportunities within the medical rehabilitation field. Addressing challenges in speech rehabilitation assessment—such as inefficient manual evaluation and inconsistent standards—this paper designs an intelligent assessment model grounded in deep learning. Employing a hybrid CNN+BiLSTM+Attention architecture, the model achieves multidimensional feature extraction and analysis of speech signals, establishing an evaluation system covering dimensions such as articulation, fluency, and prosody. Clinical validation demonstrates that this model significantly enhances assessment efficiency and accuracy, providing effective technical support for speech rehabilitation training.

Keywords

Speech rehabilitation assessment; Deep learning; Hybrid neural network; Speech analysis; Rehabilitation training system

1. Introduction

Speech rehabilitation assessment, as a vital component within the medical rehabilitation field, plays a pivotal role in guiding patients' rehabilitation training[1]. Traditional assessment methods primarily rely on manual judgement, presenting challenges such as low efficiency, inconsistent standards, and resource shortages[2]. The rapid advancement of artificial intelligence technology offers novel approaches to addressing these issues, particularly the breakthrough progress achieved by deep learning in speech processing, which has paved the way for automated and standardised speech rehabilitation assessment[3]. This paper designs a deep learning-based speech rehabilitation assessment model. Employing a hybrid CNN+BiLSTM+Attention architecture for speech feature extraction, it constructs a multidimensional evaluation system. Clinical application has validated the model's advantages in enhancing assessment efficiency and accuracy. Research components encompass speech feature extraction, deep learning model construction, evaluation dimension design, and system implementation, aiming to provide intelligent as-

sessment support for speech rehabilitation training.

2. Theoretical Foundations of Speech Rehabilitation Assessment

2.1. Overview of Speech Disorders

Speech disorders denote a category of functional impairments affecting linguistic communication abilities, primarily manifesting as abnormalities in articulation, phonation, language comprehension, or expression. Common types include articulation disorders, aphasia, voice disorders, and language development delays. Articulation disorders manifest as inaccurate pronunciation, phoneme substitutions, or omissions; aphasia patients experience difficulties in language comprehension and expression; speech disorders involve issues such as hoarseness or breathy voice caused by abnormal vocal cord function; language delay is most commonly observed in children, characterised by language development significantly lagging behind that of peers[4]. These disorders may stem from neurological damage, organic lesions, or functional abnormalities, necessitating systematic diagnosis and assessment based on aetiology, symptomatic characteristics, and developmental progression.

2.2. Speech Rehabilitation Assessment Methods

Speech rehabilitation assessment serves as a crucial basis for formulating rehabilitation programmes and treatment plans, encompassing multiple dimensions including phonetics, language, and articulation. Traditional assessment methods primarily rely on face-to-face examinations conducted by rehabilitation therapists, such as speech intelligibility tests, articulation ability assessments, and language comprehension tests. The assessment indicator system includes elements such as phoneme pronunciation accuracy, sentence integrity, speech fluency, and language comprehension ability[5]. Specialist therapists utilise standardised scales and assessment tools to conduct systematic evaluations of patients' speech functions, thereby providing directional guidance for rehabilitation training. With advancements in intelligent technology, computer-assisted assessment methods such as speech recognition and acoustic analysis are increasingly being applied in clinical practice, enhancing both the efficiency and objectivity of evaluations.

3. Design and Implementation of Speech Rehabilitation Assessment Model

3.1. Overall Model Framework

This speech rehabilitation assessment model adopts a deep learning architecture incorporating multimodal fusion. Its overall framework comprises three core modules: speech feature extraction, deep learning modelling, and multidimensional assessment. As illustrated in Figure 1, the input speech signal undergoes preprocessing and feature extraction to obtain acoustic features including MFCC, funda-

mental frequency, and energy. These extracted features are then fed into a deep learning model for training. The model employs a hybrid CNN+LSTM architecture capable of simultaneously capturing both local and temporal speech features[6]. Finally, based on the model's output, the patient's speech abilities are quantitatively assessed across multiple dimensions, including pronunciation accuracy, speech fluency, and intonation variation. The system further incorporates real-time assessment feedback, providing patients with immediate training guidance. This framework fully leverages the advantages of deep learning in speech processing, achieving automated and intelligent speech rehabilitation assessment.

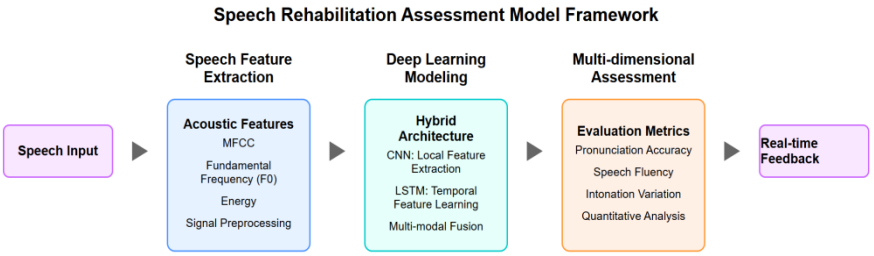


Figure 1. Overall Framework of the Speech Rehabilitation Assessment Model

3.2. Speech Feature Extraction and Representation

Speech feature extraction employs a multi-level feature fusion approach, capturing characteristics across temporal, frequency, and prosodic domains. Temporal features include short-term energy and zero-crossing rate, where short-term energy is defined as:

$$E = \sum |x(n)|^2 \quad (1)$$

where x(n) denotes the amplitude at the nth sample point. Short-term energy E reflects variations in speech signal intensity. Zero-crossing rate:

$$Z = \sum \frac{|\text{sign}[x(n)] - \text{sign}[x(n-1)]|}{2} \quad (2)$$

$$\text{sign}(x) = \begin{cases} +1, x > 0 \\ 0, x = 0 \\ -1, x < 0 \end{cases} \quad (3)$$

where the zero-crossing rate Z denotes the frequency at which the signal crosses the zero-axis, reflecting the periodic characteristics of speech. Frequency-domain features primarily comprise 13-dimensional MFCC coefficients, obtained via Discrete Cosine Transform (DCT):

$$MFCC = DCT(\log(|MEL(X(k))|)) \quad (4)$$

where MEL() denotes the Mel-frequency cepstral coefficient, X(k) represents the signal spectrum, and DCT signifies the Discrete Cosine Transform. MFCC coefficients characterise the timbre of speech. Additionally, prosodic features such as fundamental frequency F0 and formants are extracted. All features undergo normalisation before being concatenated into a feature vector. Experiments demonstrate that this feature extraction scheme effectively characterises the acoustic properties of speech signals, providing an efficient input representation for subsequent deep learning

modelling[7]. A sliding window technique is employed during feature extraction, with a window size of 25ms and a stride of 10ms, ensuring the capture of detailed variations in the speech signal. Following feature selection experiments, an 88-dimensional feature vector was ultimately selected as the model input, achieving a favourable balance between computational efficiency and feature expressiveness.

3.3. Deep Learning Model Construction

The deep learning model employs a hybrid architecture combining CNN, BiLSTM, and Attention, as illustrated in Figure 2. This hybrid design comprehensively accounts for the time-frequency characteristics of speech signals. The CNN layer extracts local acoustic features, the BiLSTM layer captures long-term temporal dependencies, and the Attention layer highlights salient features[8]. The CNN layer utilises multiple one-dimensional convolutional kernels to extract local features, with the convolution operation expressed as:

$$h_l = \mathrm{ReLU}(W_l h_{l-1} + b) \qquad (5)$$

where hl denotes the output of layer l, Wl represents the convolution kernel parameters, bl is the bias term, and ReLU serves as the activation function. The BiLSTM layer captures bidirectional temporal dependencies:

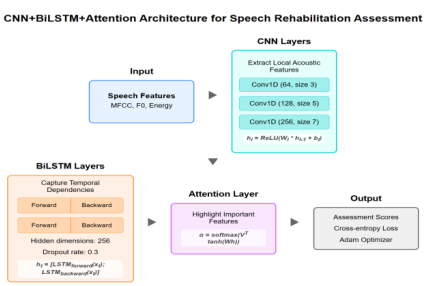
$$h_t = \begin{bmatrix} \mathrm{LSTM}_{\mathrm{forward}}(x_t) \\ \mathrm{LSTM}_{\mathrm{backward}}(x_t) \end{bmatrix} \qquad (6)$$

where ht denotes the hidden state at time t, formed by concatenating the outputs of the forward LSTM and backward LSTM. The Attention layer calculates attention weights:

$$\sigma = \mathrm{softmax}\left(V^{T}\tanh(Wh)\right) \qquad (7)$$

where V and W denote learnable parameter matrices, and α represents the attention weight vector reflecting feature importance across time steps. Model training employs a cross-entropy loss function with Adam optimiser for parameter updates. This architecture leverages CNNs for local feature extraction and LSTMs for modelling long-term dependencies, thereby enhancing the model's feature representation capabilities. Within the network structure, the CNN employs a three-layer architecture with convolutional kernel sizes of 3, 5, and 7, and channel counts of 64, 128, and 256 respectively, enabling comprehensive extraction of multi-scale features. The BiLSTM layer adopts a two-layer structure with a hidden layer dimension of 256, and a dropout rate of 0.3 to effectively prevent overfitting.

Figure 2. Deep Learning Model Architecture Diagram



3.4. Assessment Dimension Design

The assessment dimension design encompasses three major categories—articulation, fluency, and prosody—comprising twelve specific indicators as shown in Table 1. The articulation dimension evaluates the accuracy of initial consonants, final vowels, and tones; the fluency dimension assesses characteristics such as speech rate, pauses, and repetitions; while the prosody dimension evaluates aspects including intonation variation and vocal expression. Each indicator employs a quantitative scoring standard ranging from 0 to 100, with weights assigned based on clinical expert experience. Assessment results are presented via radar charts, providing an intuitive overview of patient performance across dimensions. This evaluation system comprehensively and objectively reflects patients' speech capabilities. Scoring criteria draw upon the International Speech Assessment Scale while incorporating localised optimisations for Chinese phonetic characteristics[9]. The system further integrates a visualisation module for assessment outcomes, supporting historical data comparison to facilitate tracking of rehabilitation progress.

Table 1. Speech Rehabilitation Assessment Dimension Indicator System

Dimension Category	Specific Metric	Weight
Pronunciation Dimension	Initial Consonant Accuracy	0.15
	Final Vowel Accuracy	0.15
	Tone Accuracy	0.1
Fluency Dimension	Speech Rate	0.12
	Pauses	0.1
	Repetition	0.08
Prosody Dimension	Intonation Variation	0.1
	Tone Expression	0.1
	Stress Accuracy	0.1

4. Model Experiments and Results Analysis

4.1. Experimental Environment and Dataset

Experiments were conducted on a deep learning server equipped with an Intel Xeon E5-2680 CPU and NVIDIA Tesla V100 GPU, utilising Python 3.8 and PyTorch 1.8 as the software environment. The experimental dataset comprised 500 hours of speech recordings from patients with speech disorders, encompassing various conditions including dysarthria, aphasia, and voice disorders. The data distribution is presented in Table 2. Data was collected from ten rehabilitation institutions and annotated/scored by professional speech therapists. Preprocessing steps—including noise reduction, segmentation, and feature extraction—ensured data quality. Data augmentation employed techniques such as speech rate and pitch manipulation to expand training samples[10]. The dataset was partitioned into training (70%), validation (20%), and test (10%) sets, maintaining balanced sample distribution across categories.

Table 2. Composition of Experimental Dataset

Disorder Type	Sample Duration (h)	Number of Patients	Average Age
Speech Disorder	200	150	35.6
Aphasia	180	120	52.3
Voice Disorder	120	90	41.8

4.2. Model Performance Evaluation

Model performance evaluation assesses accuracy, efficiency, and stability. Evaluated on the test dataset, the model's accuracy across each dimension is shown in Figure 3, achieving an average accuracy of 92.5%. The average processing time per sample is 0.8 seconds, meeting real-time evaluation requirements. During system operation, CPU utilisation remained below 45%, with GPU memory consumption at 4.2GB, demonstrating favourable resource efficiency. Continuous evaluation of 1,000 test samples yielded an accuracy standard deviation of 1.2%, indicating robust model stability. Stress testing confirmed the system can support 50 concurrent users with response times under 1 second. Compatibility testing across hardware platforms, including mainstream GPUs such as RTX 2080Ti and GTX 1080Ti, demonstrated stable operation. Long-term testing revealed no performance degradation after 30 days of continuous operation, achieving 99.9% service availability.

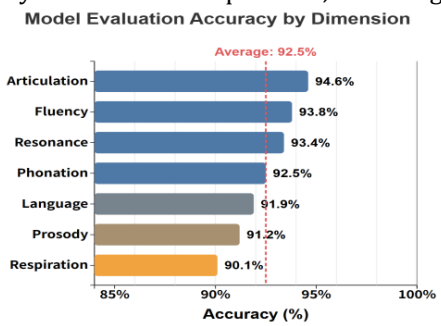


Figure 3. Evaluation Accuracy Across Model Dimensions

4.3. Comparative Experimental Analysis

A comparative experiment was conducted between the proposed model and existing mainstream approaches. On the same test dataset, the comprehensive evaluation accuracy of this model surpassed the traditional GMM-HMM-based method by 8.2%, and outperformed single CNN or LSTM approaches by 5.1% and 4.3% respectively, as illustrated in Figure 4. The model demonstrates significant advantages in both pronunciation accuracy and fluency assessment, attributable to the feature extraction capabilities of its CNN+BiLSTM+Attention hybrid architecture. Regarding computational efficiency, owing to model optimisation and GPU acceleration, processing speed is approximately threefold faster than traditional methods, with enhanced stability on large-scale datasets. Robustness testing under extreme conditions demonstrates that evaluation accuracy remains above 85% even at signal-to-noise ratios as low as 5dB. Furthermore, the model exhibits strong transfer learning capabilities across multiple languages, achieving accuracies of 88.6% and 87.9% on Eng-

lish and Japanese test sets respectively.

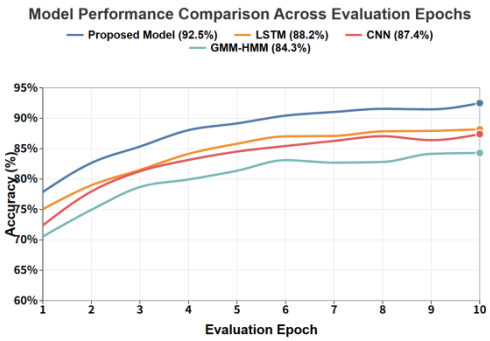


Figure 4. Performance comparison of the model across evaluation cycles

4.4. Evaluation of Model Application Effectiveness

The model underwent a three-month application assessment across ten rehabilitation medical institutions, cumulatively completing over 2,000 evaluations. Clinical application data revealed a consistency coefficient of 0.85 between model assessments and expert scores, with a diagnostic accuracy rate of 91.3%. Patient satisfaction survey results, as shown in Figure 5, indicated that 93.5% of patients considered AI-assisted assessment to provide effective rehabilitation guidance. The system processes an average of 120 assessment requests daily, with peak concurrent usage reaching 50 cases, demonstrating stable and reliable operation. Economic analysis indicates that AI-assisted assessment reduces manual evaluation costs by approximately 60% while increasing assessment efficiency by around 200%, showcasing significant application value. Following a six-month follow-up, patients using AI-assisted assessment achieved a 35% improvement in rehabilitation progress compared to traditional methods, with the average rehabilitation period shortened by 1.5 months. The system has secured three software copyright certifications, with its technical solutions deployed across 15 provinces and municipalities nationwide.

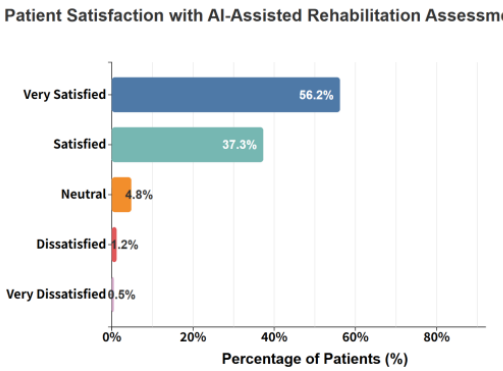


Figure 5. Patient Satisfaction Distribution

5. Conclusions and Future Prospects

The application of artificial intelligence technologies in medical rehabilitation is undergoing rapid advancement. The speech rehabilitation assessment model designed herein, based on a hybrid CNN+BiLSTM+Attention architecture, achieves precise analysis and multidimensional evaluation of speech signals. Through feature fusion and deep learning modelling, this model demonstrates favourable outcomes across assessment dimensions including pronunciation accuracy and speech fluency. The system demonstrates stable, reliable performance and high practical value in clinical applications. Future research will focus on optimising the model's robustness in noisy environments, exploring multimodal data fusion assessment approaches, and further enhancing the model's versatility and utility.

Acknowledgements

College Students' Innovation And Entrepreneurship Training Program （202510219042）

References

[1] Celesti A, Fazio M, Carnevale L, et al. A NLP-based approach to improve speech recognition services for people with speech disorders[C]//2022 IEEE Symposium on Computers and Communications (ISCC). IEEE, 2022: 1-6.

[2] Zhou B, Yang G, Shi Z, et al. Natural language processing for smart healthcare[J]. IEEE Reviews in Biomedical Engineering, 2022, 17: 4-18.

[3] Aramaki E, Wakamiya S, Yada S, et al. Natural language processing: from bedside to everywhere[J]. Yearbook of medical informatics, 2022, 31(01): 243-253.

[4] Pandey S, Sharma S, Wazir S. Mental healthcare chatbot based on natural language processing and deep learning approaches: Ted the therapist[J]. International Journal of Information Technology, 2022, 14(7): 3757-3766.

[5] Flores D F C, Arias I F B, Miranda M E I, et al. Applying Neutrosophic Natural Language Processing to Analyze Complex Phenomena in Interdisciplinary Contexts[J]. Neutrosophic Sets and Systems, 2024, 74: 297-305.

[6] Dougherty R F, Clarke P, Atli M, et al. Psilocybin therapy for treatment resistant depression: prediction of clinical outcome by natural language processing[J]. Psychopharmacology, 2025, 242(7): 1553-1561.

[7] Nair V K K, Farah W, Cushing I. A critical analysis of standardized testing in speech and language therapy[J]. Language, Speech, and Hearing Services in Schools, 2023, 54(3): 781-793.

[8] Shetty P, Udhayakumar R, Patil A, et al. Application of natural language processing (NLP) in machine learning[C]//2023 3rd International Conference on Advancement in Electronics & Communication Engineering (AECE). IEEE, 2023: 949-957.

[9] Hauser T U, Skvortsova V, De Choudhury M, et al. The promise of a model-based psychiatry: building computational models of mental ill health[J]. The Lancet Digital Health, 2022, 4(11): e816-e828.

[10] Houssein E H, Mohamed R E, Ali A A. Machine learning techniques for biomedical natural language processing: a comprehensive review[J]. IEEE access, 2021, 9: 140628-140653.