ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 1 - 2026

LSTM Networks Optimize English Speech Recognition Accuracy

Jingwen Liu[1] *and Yannan Li*[1,*]

[1] School of Humanities and Education, Jinan Preschool Education College, Jinan 250000, Shandong, China

SUMMARY: *In order to solve the problems of insufficient context dependence modeling, high recognition error of long sentences and limited decoding stability in English continuous speech recognition, this paper proposes a recognition accuracy optimization method based on LSTM network. From the perspective of computer implementation, this study constructs an integrated recognition process of "speech preprocessing-feature representation-time series modeling-sequence decoding". After completing pre-emphasis, framing and window-adding, log Mel spectrum extraction and feature normalization, bidirectional LSTM is used to jointly model the context of speech sequence. CTC beam search and language model re-ranking are combined to improve the consistency and readability of the output text. At the same time, joint optimization is carried out around the number of hidden units, the number of network layers, the learning rate, the Dropout rate and the decoding parameters to enhance the adaptability of the model under different speaking rates and sentence lengths. Experimental results show that the word error rate of the optimized LSTM model is reduced to 5.7%, the character error rate is 3.1%, and the sentence recognition accuracy is 84.6% on the English speech test set. The overall performance of the optimized LSTM model is better than DNN, RNN and unoptimized LSTM model. The results show that the LSTM network has strong temporal expression advantages in English speech recognition tasks, and can effectively improve the recognition accuracy and operation stability of the system after combining reasonable parameter adjustment and decoding strategy.*

KEYWORDS: *LSTM network; English speech recognition; Temporal modeling; Recognition accuracy Optimization*

1 Introduction

With the continuous deepening of international communication, the frequency of English is increasing in education, business, scientific research collaboration and cross-border information services. Compared with written text, voice interaction is more in line with real communication scenarios, and easier to be embedded in mobile terminals, online education platforms and intelligent service systems. Because of this, English speech recognition is no longer a single research topic in the field of speech technology, but an important basic link connecting human-computer interaction, natural language processing and intelligent computing applications. However, in practical applications, English speech signals are often affected by various factors such as speech speed, continuous reading and weak reading, accent differences, environmental noise and the change of acquisition equipment conditions. The performance of the same semantic content in the acoustic level is not stable, which makes the

*15628835599@163.com
https://doi.org/10.65102/is2026491 Published: 30 April, 2026

recognition system difficult to maintain high accuracy for a long time. If the model does not adequately describe the temporal changes, the system is easy to deviate from the segments with fuzzy phoneme boundaries, unobvious short pauses or strong context dependence, which will affect the recognition results of the whole sentence.

Computer science perspective The main objective of English speech recognition can be viewed as converting fluctuating speech waveforms into calculable acoustic properties and then developing an association between the acoustic properties and the desired transcriptions. Originally, HMMs and GMMs were used together in recognition systems. However, these models are not very good at representing complex contextual interactions and time relations. The findings made by Mohamad et al. [1] and Dahl et al. [2] demonstrate that the performance of speech recognition jobs has improved significantly due to the use of deep neural networks in acoustic modeling. Therefore, it can be concluded that the application of deep learning algorithms enables the discovery of meaningful correlations in speech data. However, the traditional feedforward neural network is unable to learn temporal correlations in speech. However, ordinary feedforward networks are better at static feature fitting, and it is still difficult to fully retain contextual information when processing continuous speech. Hochreiter and Schmidhuber[3] proposed LSTM network to alleviate the gradient disappearance problem of traditional recurrent network in long sequence training through memory unit and gating mechanism, so that the model can retain effective information in a longer range of time. Schuster and Paliwal[4] further enhanced the context utilization ability from the perspective of bidirectional modeling, which provided a key support for the subsequent research on English continuous speech recognition.

On this basis, recognition methods for unaligned speech sequences have been developed. Graves et al. [5] proposed CTC method to reduce the dependency of manual frame-level labeling, so that more flexible correspondence could be established between input acoustic sequences and output character or word segments. The experiments of Graves et al. [6] and Graves et al. [7] further prove that deep recurrent networks and bidirectional LSTM have strong modeling advantages in speech recognition tasks. Sak et al. [8, 9] engineered the LSTM structure around large-scale acoustic modeling to achieve a better balance between recognition speed and accuracy. With the improvement of public corpora and training strategies, the LibriSpeech corpus published by Panayotov et al. [10] provides a standardized data basis for English speech recognition research, and SpecAugment proposed by Park et al. [11] improves the model's adaptability to acoustic disturbances from the perspective of data enhancement. Together, these results show that the improvement of English speech recognition accuracy does not depend on the local optimization of a single module, but the result of the synergy of feature representation, temporal modeling, training mechanism and parameter adjustment.

This paper has chosen to use this understanding to base its research topic on optimizing the LSTM network to achieve higher accuracy in recognizing spoken English. Its emphasis is placed on speech signal preprocessing, sequential feature representation, time series modeling based on LSTM, and optimal parameters selection to improve the stability and accuracy of English continuous speech recognition within a computer implementation model. The following sections will create a method of LSTM optimization to enhance the accuracy of the recognition by relying on other available techniques in the area of English speech recognition, and will also evaluate the performance of the suggested method through experimental data. By so doing, the research offers a more focused technological foundation that can be used in the design and enhancement of English speech recognition systems.

2 English speech recognition related technology

The basic task of English speech recognition is to transform a continuous input speech signal into a discrete text sequence that can be processed by a computer. Compared with keyboard input or manual transcription, voice input is closer to the real communication environment, and easier to combine with online learning platforms, intelligent terminals and interactive teaching systems. However, speech is not a natural readable data object for computers. It is affected by the gender of the speaker, speaking rate, accent, recording distance and environmental noise in the acquisition stage. Therefore, English speech recognition is not a simple "voice to text", but a multi-step computational process including signal processing, feature extraction, time series modeling, sequence alignment and language decoding. Its overall technology chain is shown in Figure 1.

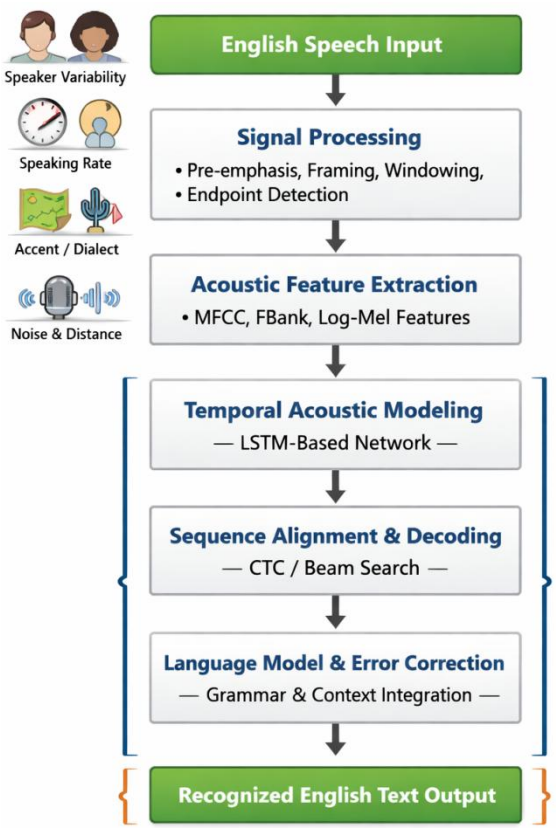

Figure 1: Flowchart of key technologies of English speech recognition

As shown in Figure 1, English speech recognition usually consists of two parts: front-end processing and back-end modeling. The front-end processing is mainly responsible for converting the original speech waveform into a stable feature matrix. The goal is not to preserve all the details of the sound, but to extract as much information as possible related to phoneme discrimination, pronunciation structure and timing variation. Common practice in engineering practice includes processing steps such as pre-emphasis, framing, window-adding, endpoint detection as well as spectrum analysis before further forming feature representations such as MFCC, logarithmic Mel spectrum or FBank. The role of this stage is to transform the high-dimensional, continuous, and fluctuant audio signal into a digital sequence that is easy for the model to learn. If the front-end feature expression is insufficient, it is difficult to obtain stable recognition results even if the subsequent network is complex.

At the level of model construction, English speech recognition technology has experienced a transformation from traditional statistical modeling to deep sequence modeling. Graves proposed the idea of sequence transduction, which emphasized the establishment of end-to-end mapping between input and output without strict alignment, and provided a new technical framework for subsequent continuous speech recognition [6]. Chan et al. proposed the Listen, Attend and Spell model, which integrates the encoding, attention and decoding mechanisms into a unified network to make the conversion from speech to text more direct [12]. Amodei et al. proved the scalability and practical value of Deep models in English continuous Speech recognition through larger scale data training and deeper network structure in Deep Speech 2 [13]. Soltau et al. further proposed the acoustic-to-word LSTM model, attempting to directly map Acoustic features to word-level output, indicating that LSTM is not only suitable for phoneme level modeling, but also can undertake higher-level recognition tasks [14].

However, from the perspective of the development path of related technologies, what really determines the performance of English speech recognition is the ability of the model to grasp the timing dependence of the speech. English pronunciation is clearly contextual, and linking, schismatic, blasting ellipsis, and intonation change the boundary form of local acoustic features. If the model can only focus on short-time intra-frame information, it is easy to produce cumulative errors in similar phonemes, fast speech flow and long sentence recognition. Kim et al. proposed a method of combining CTC and attention mechanism to enhance the global modeling ability while maintaining temporal constraints [15]. On this basis, Hori et al. combined deep encoder and RNN language model to improve the adaptability of recognition network to complex context [16]. Rao et al. studied the structure of RNN-transducers around the task of stream recognition, and showed that the co-design of timing modeling and decoding mechanism is equally important in real-time scenarios [19]. The analysis of external language model fusion by Kannan et al. also shows that the acoustic model cannot solve all errors alone, and the language constraint at the text level is still an important supplement to improve the quality of English recognition [20].

At the same time, English speech recognition technology also shows a development trend of robustness, lightweight and structural integration. Zhang et al. reviewed the research progress of environmentally robust speech recognition and pointed out that noise interference, reverberation conditions and cross-device acquisition differences are still important factors affecting system accuracy [21]. Shangguan et al. optimized the speech recognition process for edge devices, indicating that in resource-constrained scenarios, model structure and computational overhead must be considered simultaneously [23]. Gulati et al. proposed the Conformer model, which combines the convolutional structure with the encoding ability of Transformer, and shows strong performance in speech recognition [24]. These studies have expanded the technical boundaries of speech recognition, but from the perspective of the practical needs of English teaching assistance, mobile speech input and medium-sized recognition system construction, LSTM still has the advantages of clear structure, stable time series modeling ability, and moderate engineering implementation cost. Because of this, it is still of clear research value and application significance to optimize the recognition accuracy around LSTM.

3 English speech recognition accuracy optimization method based on LSTM network

3.1 Speech signal preprocessing and feature representation methods

In the English speech recognition system based on LSTM network, the original speech waveform cannot be directly used as high-quality input. The reason is that the speech signal itself has the characteristics of strong continuity, obvious time-varying and many local disturbances. If the unprocessed waveform sequence is directly fed into the network, it will not only increase the computational burden, but also weaken the ability of the model to recognize key speech patterns. Therefore, before entering the LSTM time series modeling, it is necessary to complete the preprocessing and feature representation of English speech, so that the one-dimensional acoustic signal is transformed into a time series feature matrix with stable statistical significance. Its basic process is shown in Figure 2.

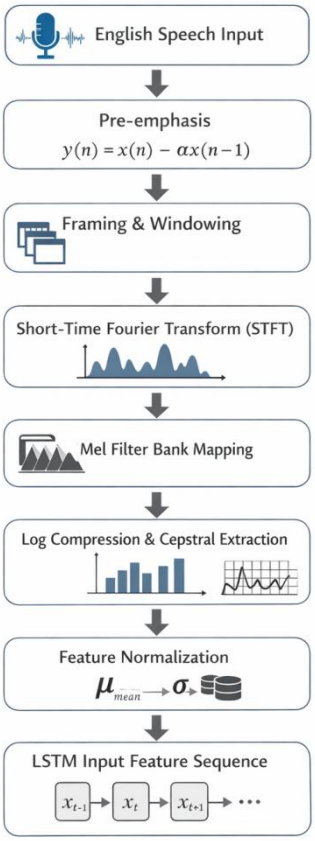

Figure 2: Flow chart of speech signal preprocessing and feature representation

The purpose of the setup of the flow shown in Figure 2 is not just to perform regular signal processing, but more importantly to provide a temporally continuous, spectrally separable, and numerically trainable input representation for subsequent LSTM networks. English speech usually contains more consonant details and pronunciation mutation information in the high frequency part, so the pre-emphasis operation is used to improve the high frequency component in the preprocessing stage, and its expression is as follows:

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

where, x(n) is the original speech signal, y(n) is the output after pre-emphasis, and α is the pre-emphasis coefficient, which is usually taken around 0.95. After this processing, the high frequency part of the speech spectrum will be moderately compensated, which is beneficial to the subsequent feature extraction to retain the clear consonants, fricatives and other details which are key to English recognition.

Since speech belongs to non-stationary signals and is difficult to analyze directly in the global scope, it also needs to be divided into short-time stationary segments. Let the speech signal of frame t be x(n), and the windowed frame signal can be expressed as follows.

$$s_t(n) = x_t(n) \cdot w(n), \quad 0 \le n \le N-1 \tag{2}$$

w(n) is the window function, and Hamming window is used in this paper to reduce the spectral leakage caused by frame boundary truncation. After framing and windowing, the system performs short-time Fourier transform on each frame, and simulates the human ear's nonlinear perception of frequency through Mel filter bank to obtain the frequency band energy representation which is more suitable for speech recognition. The energy at the m filter can be written as follows.

$$E_t(m) = \sum_{k=0}^{K-1} |S_t(k)|^2 H_m(k) \tag{3}$$

where, $S_t(k)$ is the amplitude spectrum of the TTH frame in the frequency domain, and $H_m(k)$ is the response function of the MTH Mel filter. After taking the logarithm of the filter bank energy, the MFCC or logarithmic Mel spectrum features can be further extracted. Considering that LSTM is better at dealing with the dependence between consecutive frames, this paper uses the log Mel spectrum and the first-order and second-order dynamic difference to jointly form the input features, so that the model can not only retain the spectral envelope information, but also capture the change trend of speech over time.

On this basis, in order to avoid the interference of different speakers' recording intensity and equipment conditions on the training stability, the feature matrix needs to be normalized by mean and variance. After the above steps, the raw English speech is converted into a chronological sequence of features $X = \{x_1, x_2, \dots, x_T\}$, where each x_T corresponds to a fixed-length acoustic feature vector. The input form obtained in this way is consistent with the calculation mechanism of LSTM that updates the hidden state moment by moment, which can more effectively support the subsequent pronunciation pattern modeling, context dependence capture and recognition accuracy optimization. On the whole, preprocessing and feature representation are not ancillary links, but prerequisites for LSTM to take advantage of temporal modeling, and their quality will directly affect the performance limit of the entire English speech recognition system.

3.2 Optimization Strategy of Time series Modeling and Recognition Based on LSTM network

After the front-end feature extraction is completed, the key problem of English speech recognition turns to temporal modeling. Speech is not a simple concatenation of several independent acoustic frames, and the pronunciation result at a certain moment is often affected by the previous speech flow, the current syllable position and the subsequent linking trend. Although the traditional RNN can process sequence input, when the speech segment is long and the context span is large, the network is prone to gradient attenuation in the back

propagation, the early frame information is difficult to be stably retained, and the model's ability to perceive long-distance dependencies decreases. After Hochreiter and Schmidhuber proposed LSTM, by introducing memory units and gating control into the recurrent structure, the network can selectively retain key information and suppress invalid disturbances, which is more suitable for strong sequential tasks such as English continuous speech recognition [3]. Graves et al. also pointed out in the study of deep recurrent network that LSTM has better stability and recognition performance than ordinary RNN in acoustic modeling [7].

Let the input feature sequence be $X = \{x_1, x_2, \ldots, x_T\}$, the state update process of the LSTM cell at time t can be expressed as follows.

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \tag{4}$$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{5}$$

$$\tilde{c}_t = \tanh(W_c[h_{t-1}, x_t] + b_c) \tag{6}$$

$$c_t = ft \odot ct - 1 + it \odot \tilde{c}_t \tag{7}$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{8}$$

$$h_t = o_t \odot \tanh(c_t) \tag{9}$$

where f_t, i_t and o_t represent the forget gate, input gate and output gate respectively; c_t is the cell state; h_t is the hidden state output; $\sigma(\cdot)$ is the Sigmoid function; $\odot$ represents element-wise multiplication. The significance of the above structure is not just to add a few sets of parameters, but to enable the network to judge at each time step which historical information should be kept and which local fluctuations should be weakened. This mechanism is particularly important for English speech, as linking, loss of blasting, schism, and stress transfer often deform the local spectrum, and without memory control, the model can easily be biased by short-term noise or non-critical changes.

Relying only on unidirectional LSTM still has the problem of insufficient information utilization. Some phoneme judgments in English speech often need to be combined with subsequent pronunciations to be more accurate. Based on this point, this paper adopts bidirectional LSTM structure in the time series modeling stage to introduce forward and reverse contexts into the acoustic encoding process at the same time, and its expression is as follows.

$$\overrightarrow{h_t} = \mathrm{LSTM}_f(x_t, \vec{h}_{t-1}), \quad \overleftarrow{h_t} = \mathrm{LSTM}_b(x_t, \overleftarrow{h}_{t+1}) \tag{10}$$

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}] \tag{11}$$

This bidirectional representation method can make the features of each frame simultaneously obtain the preamble and postamble constraints, and has better adaptability for long sentence recognition, fast speech flow and fuzzy boundary segments. Graves, Jaitly and Mohamed have demonstrated in the study of bidirectional LSTM acoustic modeling that this structure has obvious advantages in large vocabulary speech recognition tasks [8]. The experiments of Sak, Senior and Beaufays around large-scale acoustic models also show that the reasonable design of LSTM layers, hidden unit scale and projection structure can help to achieve a better balance between computational cost and recognition accuracy [9].

In the recognition optimization strategy, this paper does not regard the LSTM output

directly as the final text, but combines the CTC decoding mechanism to complete the sequence mapping without strict alignment. Let the target text be y. The CTC conditional probability can be written as follows.

$$P(y|X) = \sum_{\pi \in \mathcal{B}^{-1}(y)} \prod_{t=1}^{T} p(\pi_t|X) \tag{12}$$

Here, π denotes the alignment path containing whitespace, and $\mathcal{B}^{-1}$ is the set of all paths that can be mapped to the target sequence y. This strategy can avoid the complexity caused by frame-by-frame manual alignment, so that the temporal distribution information of the LSTM output directly serves the character or word segment prediction. Kim, Hori and Watanabe's research on joint CTC framework shows that such methods can effectively improve training stability and output consistency in continuous speech recognition [15]. In order to further reduce the local misjudgment in the decoding stage, this paper also introduces the language model score into the beam search process to reorder the candidate sequences, so that the acoustic probability and the language prior jointly participate in the final decision, which is consistent with the analytical conclusion of Kannan et al.'s external language model fusion [20].

4 Construction and key optimization mechanism of English speech recognition system

4.1 Overall system architecture design for English speech recognition

In order to make LSTM network play a stable role in English speech recognition tasks, the acoustic model alone is not enough. It is also necessary to build a matching system architecture, which organizes speech acquisition, signal processing, model inference, result decoding and interactive feedback into a sustainable computing link. If the system design is too loose, even if a single model achieves high accuracy in the offline test, the recognition results may still fluctuate after entering the real scene due to data transmission delay, unclear module coupling, or resource scheduling imbalance. Based on this consideration, this paper designs the English speech recognition system as a five-level structure of "access layer-processing layer-model layer-service layer-application layer", so that the speech data forms a closed-loop processing path from input to output, as shown in Figure 3.

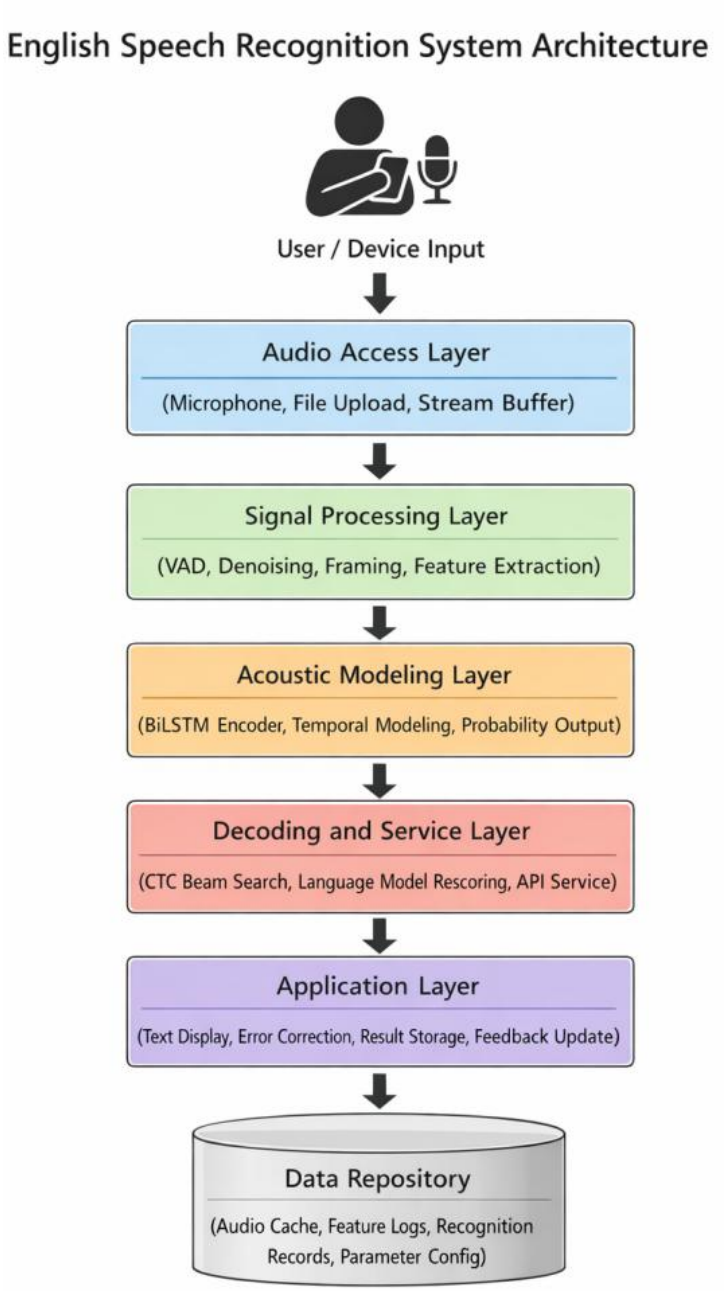

Figure 3: Overall system architecture for English speech recognition

As shown in Figure 3, the access layer is responsible for receiving the microphone real-time stream, audio files and the corpus uploaded by the teaching platform, and maintaining the input continuity through the buffer queue. The focus of this layer is not complex calculation, but to ensure the consistency of different source speech in sampling rate, bit depth and channel format, avoiding the upstream data differences directly transmitted to the model side. The processing layer undertakes tasks such as endpoint detection, noise reduction, framing, windowing and feature extraction, and its output results are fixed-length temporal feature tensors. Since English speech recognition is sensitive to the frame-level continuity, the processing layer adopts a pipelining scheduling mechanism, so that the speech block does not have to wait for the complete end of the whole audio before entering the next stage, so as to give consideration to both recognition real-time and computational stability.

The model layer is the core computing area of the system, which mainly completes the

time-series acoustic modeling based on LSTM. Considering the obvious context dependence in English speech flow, we deploy a bidirectional LSTM encoder inside the system and encapsulate it as an independent inference service, so that the front-end feature sequence can be directly mapped into a character-level or subword-level posterior probability distribution. In order to reduce the impact of model calls on the overall service performance, the layer uses GPU reasoning and batch scheduling in parallel. In the offline test scenario, multiple voices are batch processed. In the real-time input scenario, the sliding time window is continuously fed into the model to ensure that the output result can be gradually updated with the advancement of speech.

The service layer is responsible for converting the model output into readable text. Since the acoustic model produces a time-by-time probability distribution instead of the final sentence, the system introduces CTC beam search decoding and language model re-ranking mechanism here. The objective function can be written as:

$$\hat{Y} = \arg\max_{Y}[\log P_{ac}(Y|X) + \lambda \log P_{lm}(Y) + \beta \cdot |Y|] \tag{13}$$

where X represents the input speech feature sequence, $P_{ac}(Y|X)$ is the candidate text probability given by the acoustic model, $P_{lm}(Y)$ is the language model score, λ is the language model weight, β is the length compensation coefficient, and $|Y|$ represents the output sequence length. This formula shows that the system output is not completely dependent on the local discrimination results of LSTM itself, but completes the optimal search under the combined effect of acoustic information and linguistic constraints, which is of obvious significance for reducing homophone confusion, short word missegmentation and long sentence truncation.

The application layer is oriented to end users and undertakes functions such as recognition text display, error marking, logging and feedback writeback. In order to facilitate the use in teaching AIDS or oral English training scenarios, the system also stores the recognition results, confidence scores and timestamps in a unified database to form a traceable recognition record. If the user modifies the recognized text manually, the corrected results will be sent back to the data warehouse for subsequent parameter update and domain vocabulary expansion. In this way, the system is no longer a static program with one-time output, but a dynamic recognition platform that can be gradually optimized based on usage feedback.

From the perspective of overall operation, system response efficiency is also an important factor affecting availability. In order to describe the contribution of each module to the total delay, the total processing time of an identification request is expressed as follows.

$$T_{sys} = T_{in} + T_{pre} + T_{model} + T_{dec} + T_{out} \tag{14}$$

where T_{in} is the audio access time, T_{pre} is the preprocessing time, Tmodel is the LSTM inference time, T_{dec} is the decoding time, and T_{out} is the result output time. This expression helps to locate the performance bottleneck in the system implementation phase: if the proportion of T_{model} is too high, the network size or inference strategy should be optimized. If T_{dec} increases significantly, the bundle width should be adjusted or the language model should be streamlined. Through this hierarchical design and modular analysis, the system can ensure the recognition accuracy while maintaining good engineering deployability, which lays a foundation for the subsequent key parameter optimization and experimental verification.

4.2 Key parameter optimization methods affecting recognition accuracy

In the English speech recognition system, the model structure has decided the general

direction of the performance limit, but what really affects the recognition accuracy is often whether the parameter configuration is reasonable. With the same LSTM framework, the recognition results may be significantly different under different conditions of hidden layer scale, network depth, learning rate, regularization strength and decoding parameters. If the parameter setting is conservative, the model can not capture the long-term dependence in English continuous speech flow. If the parameters are set too aggressively, it is easy to produce shock or overfitting in the later training period, which will decrease the accuracy of the test set. Therefore, in this study, parameter optimization is regarded as an independent link in system construction, and joint adjustment is carried out around the three dimensions of "training stability, generalization ability and decoding consistency".

From the perspective of the implementation process, parameter optimization is not a one-time experience setting, but an iterative search process based on the feedback of the validation set. For the LSTM acoustic model, the number of hidden units directly affects the feature representation capacity, the number of layers determines the depth of temporal abstraction, the learning rate controls the step of parameter update, and Dropout is related to the adaptability of the network to noise disturbance and sample differences. At the same time, the beam search width and the language model weight will change the ranking results of candidate sequences at the decoder. In other words, the recognition accuracy is not only controlled by the training phase; the parameters of the inference phase also have a substantial impact on the final output. To this end, this paper adopts the hierarchical optimization idea. The front-end first fixes the feature extraction scheme, and searches the core training parameters inside the acoustic model. After the model converged, the decoder side parameters were corrected twice to avoid the distortion of the result judgment caused by the simultaneous fluctuation of multiple parameters. In order to quantify the pros and cons of the parameters, the comprehensive optimization objective of the verification phase is defined as follows.

$$\theta^{*} = \arg\min_{\theta} J(\theta) \tag{15}$$

$$J(\theta) = \mathrm{WER}_{\mathrm{val}}(\theta) + \lambda \cdot \mathrm{RTF}(\theta) + \mu \cdot \Delta_{\mathrm{gen}}(\theta) \tag{16}$$

where θ represents the combination of parameters to be optimized, $\mathrm{WER}_{\mathrm{val}}(\theta)$ is the word error rate of the validation set, RTF is the real-time factor used to measure the inference efficiency, Δ_{gen} represents the generalization gap between the training set and the validation set, and λ and μ are the weight coefficients. The implication of this objective function is that the parameter selection should not only pursue the local minimum error rate, but also take into account the system running speed and the model generalization performance. For teaching platforms or real-time input systems, if only relying on large-scale networks for a small accuracy gain, the actual deployment value is not high. Combined with the results of multiple rounds of experiments, this paper screens out the parameters that have a significant impact on the recognition accuracy, and gives the corresponding optimization range, as shown in Table 1.

Table 1: Key parameters and optimization Settings that affect the accuracy of English speech recognition

Parameter	Mechanism of Action	Preset Search Range	Optimized Setting
Number of LSTM layers	Determines the abstraction depth of temporal features	1–4 layers	3 layers
Number of hidden units	Affects the capacity of contextual modeling	128, 256, 384, 512	256
Batch size	Affects gradient stability and GPU memory usage	16, 32, 64	32
Initial learning rate	Determines the convergence speed in the early stage of training	0.0005–0.005	0.001
Dropout rate	Suppresses overfitting and enhances generalization	0.1–0.5	0.3
Gradient clipping threshold	Prevents gradient explosion	1, 3, 5	5
Beam width	Controls the search range of decoding candidates	5, 10, 15, 20	10
Language model weight	Balances acoustic probability and textual prior	0.1–0.8	0.4

As shown in Table 1, the number of network layers and the number of hidden units have the most direct impact on the recognition accuracy. When the number of layers is too small, the model can only learn shallow acoustic temporal relations, and has insufficient ability to discriminate linking and fast speech flow. Too many layers will increase the difficulty of training and increase the risk of overfitting on small-scale English corpus. Experiments show that the three-layer LSTM performs more balanced between accuracy and stability. In terms of the number of hidden units, 256 units can well support the context modeling of English speech. Expanding to 512 units can slightly reduce the training error, but the benefit of the validation set is limited, and the inference time is significantly increased.

The learning rate and regularization parameter mainly determine the training quality. When the initial learning rate is too high, the loss function decreases rapidly, but it is prone to oscillation in the middle and late stages. Too low, in turn, will make the convergence process too slow to fully release the model power. In this paper, a piecewise decay strategy is adopted to make the learning rate dynamically updated with training rounds:

$$\eta_t = \eta_0 \cdot \rho^{\lfloor t/s \rfloor} \tag{17}$$

Here, η_0 is the initial learning rate, ρ is the decay coefficient, and s is the decay step. This strategy can maintain a strong search ability in the early stage of training, and gradually refine the parameter adjustment in the later stage. At the same time, setting moderate Dropout and gradient clipping thresholds helps to reduce the instability of deep LSTM in long sequence training.

Parameter optimization at the decoder can also not be neglected. If the Beam width is too small, the candidate path will shrink prematurely and increase the risk of local optimum. Too large will drive up the computational overhead and may introduce more low-quality candidates. Larger language model weights are not always better, and if the text constraint is too strong, the system may tend to output common word sequences, which will weaken the response to real pronunciation details. After comprehensive comparison, this paper chooses

medium beam width and moderate language model weight configuration to keep the acoustic evidence and language prior relatively balanced.

5 Experimental Analysis

5.1 Experimental Environment

In order to verify the actual effect of LSTM network in English speech recognition accuracy optimization, this paper completes model training, parameter adjustment and test analysis on a more stable deep learning experimental platform. The configuration of the experimental environment is not only related to the training efficiency, but also affects the numerical stability of the model in the convergence process. Especially when dealing with long-term speech features and batch sample input, the adequacy of computing resources will directly affect the inference speed of hidden state update, gradient backpropagation and decoding stage. Therefore, this study takes into account the computing power, storage, software compatibility and operability of reproduction experiments when deploying the system to ensure that the speech preprocessing, BiLSTM acoustic modeling, CTC decoding and language model reordering proposed in the previous section can run continuously under a unified framework. The specific configuration of the experimental platform is shown in Table 2.

Table 2: Configuration of the experimental environment

Configuration Item	Specific Content
Operating System	Ubuntu 22.04 LTS 64-bit
Processor	Intel Core i7-12700
Memory	32 GB DDR4
Graphics Processing Unit	NVIDIA RTX 3080 10 GB
Programming Language	Python 3.10
Deep Learning Framework	PyTorch 2.1
Audio Processing Libraries	Librosa, torchaudio
Data Processing Libraries	NumPy, Pandas
Decoding and Evaluation Tools	CTC Beam Search, JiWER
Storage Environment	1 TB SSD

As shown in Table 2, the experimental system uses Python to build the overall process, PyTorch is used for the construction, training and parameter update of LSTM network, Librosa and torchaudio are responsible for speech resampling, pre-emphasis, framing and windowing, and Mel spectrum feature extraction. NumPy and Pandas are mainly used for sample organization, feature caching, and result statistics. Considering that English speech recognition is a typical sequence modeling task, and a large number of matrix operations and gradient propagation need to be performed repeatedly in the training phase, we use GPU to accelerate model training to shorten the single iteration time and improve the efficiency of parameter search. At the same time, in order to ensure the repeatability of the experimental results, the random seed is uniformly set to 42, the batch size, the initial value of learning rate and the verification frequency are fixed in the training process, and the loss value, word error rate and model weight change of each round of experiment are recorded. Such an experimental environment can not only support the offline training of medium scale English speech data, but also facilitate the subsequent comparative analysis of key parameters, recognition effects and system response performance.

5.2 Experimental data

In order to guarantee that the experiment's outcomes can truly demonstrate the impact of optimization by an LSTM neural network on the recognition accuracy of English speech, the study incorporates considerations relating to the speech intelligibility, variability in speakers, and the trainability based on the quantity of samples into the data selection process. In order to provide the experimental data, the publicly available English speech database named LibriSpeech is chosen as the dataset. Developed by Panayotov et al., LibriSpeech has been widely used in tasks involving English automatic speech recognition [11]. Compared with small-scale closed datasets, the variety of speaking rates and styles among the speakers in LibriSpeech enables this dataset to become a better choice for evaluating the ability of LSTM models to fit in diverse conditions of speech modeling. In accordance with the experimental needs of this study, a part of the dataset that includes relatively clear pronunciation, well-written annotations, and high-quality audios is selected for training and testing purposes. The speech samples are then transformed to be monophonic audio at a sampling rate of 16 kHz with 16-bit quantization resolution.

In the process of preprocessing data, first, the time length of the original speech is analyzed and the silences are removed. Then, duplicate sampling is conducted and the data set is divided based on the number of speakers and text annotations. In order to reduce similarity among the data used for training and test in speaking style, training data, validation data, and test data are separately divided according to the number of speakers in the current experiment.The overall information of the experimental data is presented in Table 3.

Table 3: Composition of the experimental data set

Dataset	Number of Speech Samples	Number of Speakers	Total Duration / h	Main Purpose
Training Set	12000	210	26.8	Model parameter learning
Validation Set	1500	32	3.4	Parameter tuning and early stopping
Test Set	1500	31	3.5	Speech recognition performance evaluation

The training set as illustrated in table 3 finishes the primary stage of the LSTM based acoustic model learning, whereas the validation set serves to monitor the loss development tendency and determine the optimal parameters. The test set is used to generate the last recognition output. The corpus includes typical phrases and sentences, short sentences and continuous sentences such as Good morning, full sentences and questions. The examples are highly effective in demonstrating rhythm variations and connecting the articulations, which is typical of the English speech recognition system. In this kind of data framework, the model acquires various acoustic features during the training stage and will be able to identify the speech samples of unfamiliar speakers in the testing stage.

5.3 Experimental Scheme

To evaluate the way the improvement effect of the LSTM model can be applied to improve the English speech recognition more scientifically, the experimental design of the current study takes the form of the model comparison and proof of optimization. Each of the models will be put through an identical experimental procedure. In other words, the training speech will be preprocessed in terms of pre-emphasis, framing, windowing, and log Mel spectrogram representation and the output will be employed to train the relevant recognition model. The

validation set will be employed to tune hyperparameters and monitor convergence and testing set will be utilized to test the end results. The highest number of epochs is 60. Early stopping will be applied when the word error rate on the validation set remains the same over eight consecutive epochs and the best model weights would be saved.

In the experiment of comparing different models, four recognition schemes are presented. Scheme 1 involves the use of a traditional DNN acoustic model for establishing the baseline performance level of non-sequential models on the task of English speech recognition. Scheme 2 uses a standard RNN model for evaluating the recognition abilities of standard recurrent models in the processing of longer speech sequences. Scheme 3 presents an unmodified unidirectional LSTM model, which contains two hidden layers of the LSTM network with 256 hidden units each. Scheme 4 introduces the improved LSTM recognition model proposed in this paper. It consists of three layers of the bidirectional LSTM model with the incorporation of Dropout, gradient clipping, connectionist temporal classification (CTC), and language model re-ranking. The dimensions of the input features are kept identical for all the models. The batch size is fixed at 32, Adam is used as the optimizer, and the initial learning rate is 0.001 with dynamic decay depending on the validation error. This approach is adopted to minimize differences resulting from the training process.

In addition to the overall comparison, this paper also sets up ablation experiments to identify the specific contribution of each optimization link to the final accuracy. Specifically, on the basis of the complete model, the bidirectional structure, language model reordering, SpecAugment data augmentation and gradient clipping are removed respectively, and the fluctuation of word error rate, character error rate and inference delay are observed after the change of each setting. At the same time, in order to investigate the adaptability of the model under different speech lengths, this paper also counts the recognition results of the test set according to three levels: short sentences, medium sentences and long sentences. On the one hand, this experimental arrangement can verify the overall advantages of the optimized LSTM model over other methods, and on the other hand, it can clearly explain whether the improvement of English speech recognition accuracy comes from the enhancement of time series modeling, the improvement of training strategy, or the optimization of decoder constraints, so as to provide a reliable basis for subsequent evaluation indicators and analysis of experimental results.

5.4 Evaluation Metrics

To explore how LSTM network affects the recognition accuracy of English speeches with greater accuracy, we need to define a set of criteria to measure the outcomes in two ways: recognition accuracy and accessibility of the system throughout the experiment. The process of speech recognition in English is more complex than classification since its output is a sequence of text. So, the evaluation must consider not only the prediction errors by the word or frame but also the mean discrepancy between the generated text and the reference text. With this feature considered, the research employs such measures as WER, CER, Sentence Recognition Accuracy, and Real-Time Factor as the primary evaluation criteria.

Word Error Rate is one of the most commonly used measures in speech recognition. It indicates the deviation between identified text and the text of the standard transcription. Its formula is expressed as:

$$\mathrm{WER} = \frac{\mathrm{S} + \mathrm{D} + \mathrm{I}}{\mathrm{N}} \times 100\% \tag{18}$$

where, S denotes the number of replacement error words, D denotes the number of deletion

error words, I denotes the number of insertion error words, and N denotes the total number of words in the reference text. The lower this index is, the closer the recognition result of the model at the word level is to the real text. For English continuous speech, WER can more directly reflect the overall discrimination ability of the model under the conditions of linking, abstention and schwa, so it is used as the main evaluation criterion in this paper.

Considering that character-level confusion still exists in English speech recognition, especially in the recognition of short words, proper nouns and words with similar pronunciation, this paper further introduces the Character Error Rate (CER) as a supplementary index, and its expression is:

$$\mathrm{CER} = \frac{\mathrm{S_c} + \mathrm{D_c} + \mathrm{I_c}}{\mathrm{N_c}} \times 100\% \tag{19}$$

where $\mathrm{S_c}, \mathrm{D_c}, \mathrm{I_c}$ represent the number of substitutions, deletions, and insertions at the character level, respectively, and $\mathrm{N_c}$ is the total number of characters in the reference text. Compared with WER, CER is more sensitive to local spelling bias and can more finely reveal the stability of the model in subword or character-level output.

The error rate alone is not enough to describe the system's acceptance level at the whole Sentence level, so we also calculate the Sentence Accuracy (SA), which describes the proportion of sentences in the test set that are completely correctly recognized. It is calculated as follows.

$$\mathrm{SA} = \frac{\mathrm{N_{correct}}}{\mathrm{N_{all}}} \times 100\% \tag{20}$$

where, $\mathrm{N_{correct}}$ is the number of sentences whose recognition results are completely consistent with the standard text, and $\mathrm{N_{all}}$ is the total number of sentences in the test set. Although this metric is strict, it can effectively reflect the direct usability of the model in real application scenarios. For oral English practice, speech input and teaching assistance systems, whether the whole sentence is complete and correct is often of more practical significance than the number of local errors.

In addition to recognition accuracy, this paper also focuses on the running efficiency of the model in the computer system, so the Real Time Factor (RTF) is introduced to evaluate the inference speed:

$$\mathrm{RTF} = \frac{\mathrm{T_{decode}}}{\mathrm{T_{audio}}} \tag{21}$$

where, $\mathrm{T_{decode}}$ represents the time required for the system to complete the recognition, and $\mathrm{T_{audio}}$ represents the input audio duration. When RTF<1, it indicates that the system has real-time processing capability. This index helps to determine whether the LSTM optimization scheme is suitable for deployment in the actual English speech recognition platform. In summary, WER and CER mainly measure the output error of text, SA reflects the quality of sentence-level recognition, and RTF reflects the efficiency of system operation. Through the joint analysis of these indicators, the actual effect of LSTM network in English speech recognition accuracy optimization can be more comprehensively judged.

5.5 Experimental Results

After the training, validation, and tuning stages for the models, the performance of each of the

four schemes on the test data set was evaluated based on the same criteria. As shown by these results, the LSTM scheme significantly outperforms the non-sequence model in modeling English continuous speech. Better performance can be expected if contextual bidirectional modeling, SpecAugment, CTC beam search, and language model re-ranking methods are added to the LSTM architecture. In tests, the proposed model obtains a word error rate of 5.7%, a character error rate of 3.1%, an average sentence accuracy rate of 84.6%, and an RTF of 0.61, while the conventional RNN provides 8.1% word error rate, 10.4% for the standard RNN, and 12.8% for the DNN baseline. This shows that the improvement of English speech recognition performance does not only depend on whether the network is "deeper", but more importantly, whether the model can stably retain long-term dependence information and effectively suppress local misjudgment in the decoding stage.

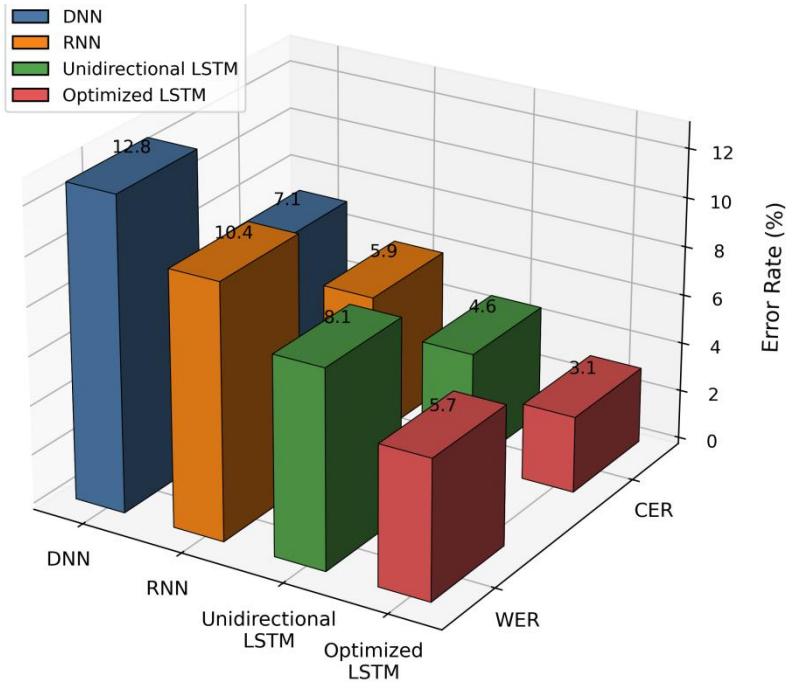

Figure 4: Error comparison of different recognition models on the test set

From Figure 4, we can see that the error rate continues to decrease with the enhancement of the temporal modeling ability of the model. Although DNN can complete the nonlinear mapping from acoustic features to label space, it is still biased towards static discrimination in nature, and does not make sufficient use of the context in continuous speech flow, so it is more prone to substitution errors in sentences with obvious linking, phonation and speech rate fluctuations. RNN has begun to have temporal memory ability, and the recognition results have been improved. However, when the input sequence is long, the information retention of early frames is unstable, resulting in obvious fluctuations in long sentence recognition. The unidirectional LSTM mitigates this problem with a gated unit, and the word error rate decreases by 2.3 percentage points compared to RNN. On the basis of the unidirectional LSTM, the model in this paper continues to use the bidirectional structure, so that each frame can combine the preceding text and reference the following text, and then overlay the language model reordering, which further compresses the insertion error and deletion error in the decoding stage, so the overall performance is the best.

It is still not sufficient to observe only from the population average, and it is necessary to investigate the stability of the model under different sentence lengths. The test results show that the recognition difficulty of each model increases with the increase of sentence length,

but the increase is not consistent. To reflect this more clearly, the word error rate is counted separately for short sentences, medium length sentences and long sentences, and the results are shown in Figure 5.

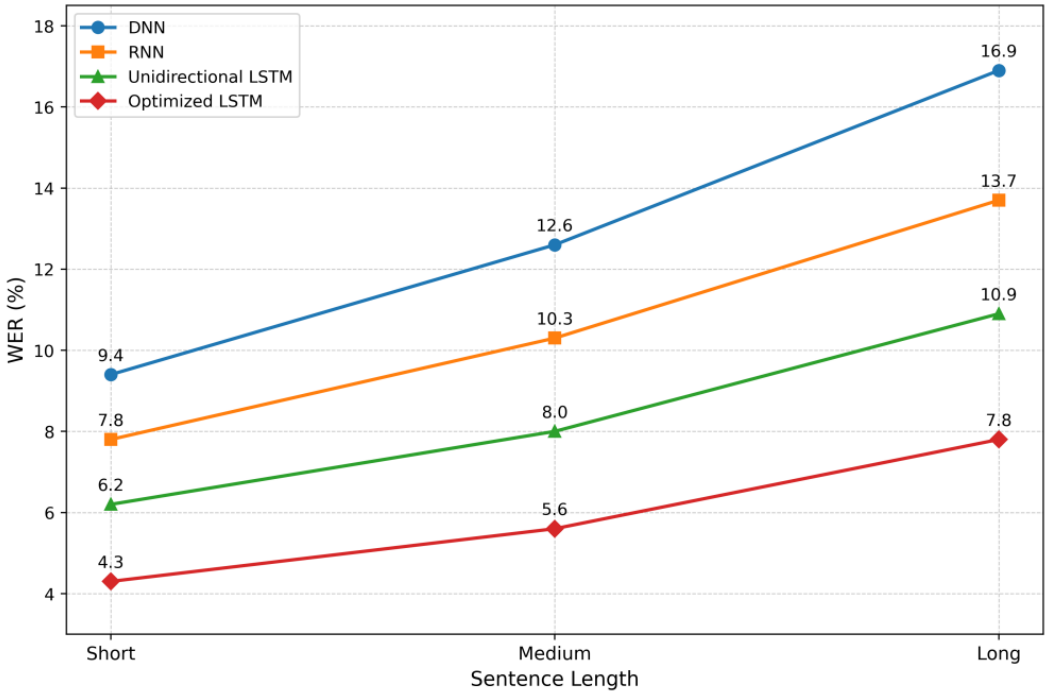

Figure 5: Word error rate variation of each model for different sentence lengths

From Figure 5, it can be found that the gap is more obvious in the long sentence scenario than in the short sentence scenario. The word error rate of DNN from short sentences to long sentences increases by 7.5 percentage points, indicating that it is almost unable to model long-distance dependencies. Although RNN is improved, there are still more substitutions and omissions due to the attenuation of the previous information in the long sentence condition. One-way LSTM has been able to deal with sentences of general length more stably, but the model still has deviation in boundary location when the pause is not obvious, there are many clauses or the linking phenomenon is strong. The word error rate of the proposed model on long sentences is controlled at 7.8%, which indicates that bidirectional temporal coding has stronger adaptability to long English speech flow. In other words, the optimization strategy proposed in this paper is effective for short sentence recognition, and the improvement is more obvious for long sentence recognition, which is consistent with the structural advantages of LSTM in long-term dependency modeling.

Further checking the ablation results, after removing the bidirectional structure, the word error rate of the model rebounded to 6.8%. After removing the language model reordering, the word error rate is 6.5%. Without SpecAugment, the word error rate increases to 6.3%. After removing gradient clipping, we show slight oscillations later in training, resulting in a final word error rate of 6.7%. These results show that the performance improvement of the proposed model is not caused by a single module independently, but the result of multiple optimization mechanisms. Among them, bidirectional LSTM contributes the most to the integration of temporal information, language model reordering mainly improves the text-level output consistency, and data augmentation and gradient control enhance the training stability and generalization ability.

In summary, the experimental results clearly verify the effectiveness of the proposed

method. Compared with DNN, RNN and unoptimized LSTM, the optimized LSTM model performs better in recognition error, whole sentence accuracy and long sentence robustness, while the real-time factor remains within the deployable range. This shows that the method is not only suitable for offline test environment, but also has a realistic basis for further embedding into English speech input, teaching assistance and oral training systems.

6 Conclusions

Focusing on the problem of time series modeling and accuracy optimization in English speech recognition, this paper constructs a recognition method with LSTM network as the core, and systematically optimizes the speech preprocessing, feature representation, system architecture design, parameter adjustment and result decoding. Studies show that English speech recognition does not rely solely on acoustic feature input to obtain high-quality results, and its recognition performance is closely related to the quality of front-end signal processing, the ability of temporal dependence modeling, and the constraint mechanism in the decoding stage. To start with, considering the modeling concern, LSTM network is better than the DNN and the conventional RNN in extracting contextual information on English continuous speech using gated memory units, hence maintaining the pronunciation information of interest strongly without being affected by local noise, reading continuously, and various rates of speaking, and enhancing the temporal discrimination to achieve recognition. In terms of the system architecture, this paper proposes a hierarchical computational scheme of English speech recognition comprising voice access, signal processing, BiLSTM acoustic modeling, CTC decoding and language-model reranking. The framework defines the scope of every functional module, providing a practical technological foundation of their future optimization and parameter implementation. Experimental findings show that the optimized LSTM network has improved performance in both recognition and outperforms the DNN, RNN and non-optimized LSTM models in particular in the aspect of words error rate of 5.7% and character error rate of 3.1% and 84.6 percent in sentence recognition accuracy. Also, the optimized network can be used to obtain stable recognition performance under the conditions of recognition of long sentences which confirms the effectiveness of the developed approach to enhance the accuracy of English speech recognition.

References

[1] Mohamed A. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups[J]. IEEE Signal Processing Magazine, 2012.

[2] Dahl G E, Yu D, Deng L, et al. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 20(1): 30-42.

[3] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[4] Schuster M, Paliwal K K. Bidirectional recurrent neural networks[J]. IEEE Transactions on Signal Processing, 1997, 45(11): 2673-2681.

[5] Graves A, Fernández S, Gomez F, et al. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks[C]//Proceedings of the 23rd

International Conference on Machine Learning. 2006: 369-376.

[6] Graves A, Mohamed A, Hinton G. Speech recognition with deep recurrent neural networks[C]//2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013: 6645-6649.

[7] Graves A, Jaitly N, Mohamed A. Hybrid speech recognition with deep bidirectional LSTM[C]//2013 IEEE Workshop on Automatic Speech Recognition and Understanding. IEEE, 2013: 273-278.

[8] Sak H, Senior A W, Beaufays F. Long short-term memory recurrent neural network architectures for large scale acoustic modeling[C]//Interspeech. 2014: 338-342.

[9] Sak H, Senior A, Rao K, et al. Fast and accurate recurrent neural network acoustic models for speech recognition[J]. arXiv preprint arXiv:1507.06947, 2015.

[10] Panayotov V, Chen G, Povey D, et al. LibriSpeech: an ASR corpus based on public domain audio books[C]//2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2015: 5206-5210.

[11] Park D S, Chan W, Zhang Y, et al. SpecAugment: A simple data augmentation method for automatic speech recognition[J]. arXiv preprint arXiv:1904.08779, 2019.

[12] Graves A. Sequence transduction with recurrent neural networks[J]. arXiv preprint arXiv:1211.3711, 2012.

[13] Chan W, Jaitly N, Le Q, et al. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition[C]//2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016: 4960-4964.

[14] Amodei D, Ananthanarayanan S, Anubhai R, et al. Deep Speech 2: End-to-end speech recognition in English and Mandarin[C]//International Conference on Machine Learning. PMLR, 2016: 173-182.

[15] Soltau H, Liao H, Sak H. Neural speech recognizer: Acoustic-to-word LSTM model for large vocabulary speech recognition[J]. arXiv preprint arXiv:1610.09975, 2016.

[16] Kim S, Hori T, Watanabe S. Joint CTC-attention based end-to-end speech recognition using multi-task learning[C]//2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017: 4835-4839.

[17] Hori T, Watanabe S, Zhang Y, et al. Advances in joint CTC-attention based end-to-end speech recognition with a deep CNN encoder and RNN-LM[J]. arXiv preprint arXiv:1706.02737, 2017.

[18] Rao K, Sak H, Prabhavalkar R. Exploring architectures, data and units for streaming end-to-end speech recognition with RNN-transducer[C]//2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2017: 193-199.

[19] Kannan A, Wu Y, Nguyen P, et al. An analysis of incorporating an external language model into a sequence-to-sequence model[C]//2018 IEEE International Conference on

Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018: 1-5828.

[20] Zhang Z, Geiger J, Pohjalainen J, et al. Deep learning for environmentally robust speech recognition: An overview of recent developments[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2018, 9(5): 1-28.

[21] Shangguan Y, Li J, Liang Q, et al. Optimizing speech recognition for the edge[J]. arXiv preprint arXiv:1909.12408, 2019.

[22] Gulati A, Qin J, Chiu C C, et al. Conformer: Convolution-augmented transformer for speech recognition[J]. arXiv preprint arXiv:2005.08100, 2020.

[23] Xiong W, Droppo J, Huang X, et al. Achieving human parity in conversational speech recognition[J]. arXiv preprint arXiv:1610.05256, 2016.

[24] Weiss R J, Chorowski J, Jaitly N, et al. Sequence-to-sequence models can directly translate foreign speech[J]. arXiv preprint arXiv:1703.08581, 2017.