Proceedings of the 3rd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/145/2025.21896

Research on Noise Robustness of Chinese Speech Recognition based on Bidirectional LSTM

Yatai Ji[1,a,*]

[1]*School of Electronics Engineering and Computer Science, Peking University, Beijing, 100871, China*
a. jyt@pku.org.cn
**corresponding author*

Abstract: Automatic speech recognition (ASR) techniques are becoming more and more important in people's daily life as a vital method of human-computer interaction. The ability to maintain a high recognition accuracy in noisy environments is the key for a model to be widely used in ASR. This article researches on the robustness against different kinds of background noise in Mandarin Chinese speech recognition and the baseline model used here is Bidirectional Long Short-Term Memory (BiLSTM). To compare the effects of different kinds of data augmentation method and different model structure, four common used data augmentation methods are applied in the process of training respectively and together, and two model methods, the CNNs and the attention mechanism, are combined with the baseline model, using the method of controlled experiment. After the model is trained, it will be tested within three kinds of background noise (car noise, café noise, white noise) to evaluate the anti-noise ability of different methods applied to the model. Among the four data augmentation methods, the method of random increased/decreased volume performed best in improving recognition accuracy, while the method of time-frequency masking increased Character Error Rate (CER) unexpectedly. As for the model methods, the CNNs performed better than the attention mechanism.

Keywords: deep learning, automatic speech recognition, Mandarin Chinese

1. Introduction

Speech recognition is one of the most important ways of human-computer interaction. Through data processing and machine learning algorithms, computer systems can transform the complex speech into corresponding texts and instructions, which can largely improve the efficiency of human-computer interaction. Thus, speech recognition techniques can improve users' experience and are already used in many fields like the medical care, the financial sector, education, transportation and so on.

The original study about speech recognition goes back to the 1950s. As the development of computing resources and algorithms, speech recognition techniques enable computers to understand not only simple words, but sentences with complex structures and even different accents as well. According to Fan's work, speech recognition techniques mainly include methods based on acoustic, methods based on statistical models and machine learning methods [1]. Acoustic methods started early but are not practicable now because of the high complexity. Methods based on statistical models

DOI: 10.54254/2755-2721/145/2025.21896

like the Hidden Markov Model are in practice now, but their performance is not as good as machine learning. The machine learning methods are the most practicable and have already succeeded in many practices, according to Zhang's description [2]. Meanwhile, deep learning methods perform outstandingly among all the machine learning methods, as deep learning can automatically extract features of speech to learn and deal with more complex speech with background noise and different accents [3]. In other words, deep learning methods are more robust and accurate in coping with complex speeches. For example, as implemented in Liu's and Lian's work, the 1D Convolutional Neural Network (1D-CNN) deep learning model can be used in speech recognition and the accuracy is up to 95.4% without noise, 85.6% with severe noise, which shows the high accuracy and robustness of deep learning methods [4].

This paper focuses on the recognition of Mandarin Chinese speech. The dataset used is THCHS-30 from Tsinghua University, a commonly used Mandarin dataset. The goal of this paper is to implement different models and compare their performances on Mandarin Chinese speech recognition in different kinds of background noise, and try to find ways to improve the performance.

2. Methodology

2.1. Dataset

The data-set used in the experiment is THCHS-30, a Mandarin Chinese speech corpus released by Tsinghua University, designed for general-domain Large Vocabulary Continuous Speech Recognition (LVCSR) tasks [5].

THCHS-30 includes 30 hours of Chinese speech without any accent recorded by 50 participants in a quiet environment, and the audio data is sampled at 16 kHz with 16-bit PCM encoding. The corpus is divided into three main parts, 25 hours for training, 2.5 hours for validation, and 2.5 hours for testing. It also provides three special test sets mixed with car, café, and white noise at 0 dB Signal-to-Noise Ratio (SNR) for additional robustness evaluation.

As one of the earliest open-source Mandarin speech corpora, THCHS-30 has become a vital tool for Chinese speech research. By providing standardized clean and noisy test conditions, it helps researchers develop accurate and noise-resistant models, while supporting replicable studies in speech technology advancements.

2.2. Model Architecture

2.2.1. Baseline Model

Bidirectional Long Short-Term Memory (BiLSTM) model serves as the baseline model in this study.

As traditional Recurrent Neural Network (RNN) model struggles with handling long sequences because of the phenomenon of vanishing gradients or exploding gradients, LSTM is designed as an optimized version of RNNs, to solve the problem of long-term dependency. The key mechanism of LSTM is the cell state, which enables information to be stable throughout the process. Then, three gates called forget gate, input gate and output gate are designed to control and protect the cell state, deciding which information can be discarded from the cell state, added to the cell state or outputted to the hidden state. The gates' computations follow a unified framework below:

$$g_t = \sigma\left(W_g \cdot [h_{t-1}, x_t] + b_g\right) \quad (1)$$

Where W_g is the trainable weight matrix and b_g is the bias term of the specific gate. With the whole structure, LSTM model can deal with extremely long sequences.

RNN and LSTM models can only generate outputs based on the previous information, but in some problems, the present state is not only relevant to the previous state, but also relevant to the future state to some extent. BiLSTM model is thus developed to solve such problems. By combining

DOI: 10.54254/2755-2721/145/2025.21896

separate forward and backward LSTM layers, BiLSTM model can capture bidirectional temporal dependencies and construct both past-to-present and present-to-future dependency patterns. After extracting features and encoding bidirectional contextual constraints, BiLSTM model concatenates the hidden states of two directions and then feeds them into a trainable linear projection layer to generate the final results. Due to the bidirectional structure, BiLSTM model performs well solving context-dependent problems, especially Chinese with a complex tone system.

2.2.2. CNN + BiLSTM

Convolutional Neural Networks (CNNs) excel at local feature abstraction due to their convolutional mechanism, so CNNs are widely adopted in hybrid models to improve the performance. According to Ossama Abdel-Hamid et al. (2014), CNN can reduce the error rate by 6%-10% compared with Deep Neural Network (DNN) model in the TIMIT (Texas Instruments/Massachusetts Institute of Technology) phone recognition [7].

As for BiLSTM model for speech recognition task, CNN model can also help. The CNN specializes in extracting consistent spectral features across speakers, and the subsequent BiLSTM model captures temporal dependencies to deal with context-dependent problems. As a result, BiLSTM model combined with CNN can better suppress speaker-specific variations and enhance robustness across genders and accents, thus reducing the error rate.

In this study, two 1D CNN layers are applied along the temporal dimension of the training data to extract local features and improve the accuracy of recognition.

2.2.3. Attention Mechanism + BiLSTM

Similar to humans' behavior looking at a picture or listening to a speech, attention mechanism makes the model focus on the key information but not all the information, to quickly obtain the most effective information and conserve computing resources. Formally, given encoder hidden states $\{h_i\}_{i=1}^{T}$ and a query vector q, the attention weight α_i for the i-th timestep is computed via additive scoring [8]:

$$e_i = v^T tanh\left(W_q q + W_h h_i + b\right) \tag{2}$$

$$\alpha_i = \frac{exp(e_i)}{\sum_{j=i}^{T} exp(e_j)} \tag{3}$$

Where v, W_q, and b are learnable parameters. The model consistently evaluates the relevance between every timestep of the input and current decoding target, and normalizes weights through the SoftMax function to make the model focus over 80% of its attention on key frames that account for 5% or less.

In Chinese speech recognition tasks, the model combined with attention mechanism automatically enhance the attention weights near the turning points of tone contours (such as the rising segment of the rising-then-falling tone), and adjust the sharpness of the weight distribution through a learnable scaling factor to suppress the interference from non-critical steady segments.

3. Experimental Design

3.1. Data Augmentation

The raw data is first processed to extract Mel-Frequency Cepstral Coefficients (MFCCs). Then, to research on anti-noise robustness of Chinese speech recognition model, this study utilizes different kinds of data augmentation methods.

DOI: 10.54254/2755-2721/145/2025.21896

Random speed perturbation performs time-axis stretching or compression (±10%) on an audio waveform while adjusting the sampling rate to be stable. This kind of data augmentation can simulate the speed fluctuation of speech in the real world, forcing the model to learn rate invariance.

Random instantaneous silence inserts a 50-200 ms silence in a random position and cover the original audio. This augmentation method simulates interruptions caused by sudden noises in real situations like coughing or door-closing noises. It can train the model to construct missing segments depending on contextual information, enhancing its tolerance to impulse noise.

Random volume scaling method implements random linear scaling to the amplitude of the audio within a range of ±6 dB. Increased or decreased volume can simulate the difference of recording device's distance to the speaker, or the environmental noise affecting the original audio. It can improve the model's robustness against stationary noise (such as white noise), and make the model focus on relative spectral features rather than absolute energy.

The last data augmentation method is time-frequency masking, which randomly masks a contiguous sequence of T frames where T ranges from 0 to 15 or F frequency bands where F ranges from 0 to 5 on the Mel spectrogram. Simulating broadband noise and attenuation in specific frequency bands, this method helps strengthen the model's generalization ability to local time-frequency noise by employing structured dropout.

After implementing all the methods mentioned above respectively, they will be applied to the audio data together for the final training. According to the respective characteristics and the expected effects of the augmentation methods, random volume scaling may perform best among the methods, and the results will be presented in the next part.

3.2. Comparative Settings and Experiment Pipeline

To compare the performances of different models on the noise test, this study conducts two controlled trials using BiLSTM model combined with CNN and BiLSTM model combined with attention mechanism. Theoretically, CNN layers can help extract spectral features of the audio to eliminate speaker-specific variations, therefore enhancing the model, while attention mechanism enables the model to focus more on key information rather than all information. The hyperparameters are the same in all three trials and the results will be shown in part 4 as well.

Firstly, the original BiLSTM model without data augmentation will be trained to figure out the hyperparameters that best fit the THCHS-30 corpus. To improve the performance, some normalization and regularization techniques are added into the model. Then, different methods of data augmentation will be applied to the dataset for anti-noise research. Lastly, the two enhanced models will be trained respectively, with only no data augmentation or all augmentation methods, and evaluated by the noise-contaminated test set, to compare the difference between model frameworks.

3.3. Configurations and Evaluation Method

The configurations are consistent in all three trials to minimize the effects of irrelevant factors. As for BiLSTM hyperparameters, there are 3 bidirectional LSTM layers to capture temporal dependencies, 256 hidden units per direction, which leads to a final bidirectional output dimension of 512, which performed best during training. Then there are the training setups. The number of training epochs is 50, which is enough for all the models to converge. The loss function used here is Connectionist Temporal Classification (CTC) loss, which is ideal for sequence-to-sequence tasks where input-output alignment is unknown. The optimizer is Adam optimizer with L2 regularization and the learning rate scheduler used here is OneCycle Learning Rate Policy with 40% warm up phase and appropriate learning rate, which can dynamically adjust learning rates to accelerate convergence

DOI: 10.54254/2755-2721/145/2025.21896

via high learning rates during the cycle and stabilize training via gradual warmup and cooldown phases.

The CNN model combined with the BiLSTM has two 1D convolutional layers with "same" padding to maintain temporal resolution, and increase the number of channels from 40 (dimension of MFCC) to 64 and from 64 to 128. The attention mechanism processes BiLSTM outputs through a linear layer with tanh activation, computes sequence-masked softmax scores using a learnable vector v, generates weighted context vectors across time steps, and concatenates them with original features (2H→4H) to enhance temporal context awareness of the model.

3.4. Evaluation

The models are evaluated using Character Error Rate (CER), a common evaluation method in speech recognition. In this study, CER is computed via the Levenshtein distance, a function that measures the minimum number of single-character edits (insertions, deletions, or substitutions) required to change one sequence into another. After epochs of training, the models will generate their predictions of the test audio without noise to compute CER with the corresponding label, and finally be tested by the audio mixed with noise.

4. Results and Discussion

4.1. Baseline Model

Table 1: Testing results of CER of BiLSTM model

Data Augmentation	Test without noise	Test with background noise		
		Car noise	Café noise	White noise
No Augmentation	9.05%	18.75%	47.05%	61.92%
Random Speed	9.17%	16.55%	45.79%	62.98%
Random Instantaneous Silence	9.09%	19.12%	46.52%	62.01%
Random Increased/Decreased volume	9.02%	16.15%	42.25%	61.24%
Time-Frequency Masking	9.43%	19.90%	47.00%	62.84%
All Augmentation	9.66%	15.56%	39.61%	62.04%

Table 1 shows the testing results of the standalone BiLSTM model. The BiLSTM model achieves a CER of 9.05% in the test of clean audio without data augmentation. With different kinds of data augmentation, the CER results fluctuate about 9.10% except the results of the Time and Frequency Masking method and all augmentation. The probable reason for the increase in CER of time-frequency masking is that the process disrupts critical tonal features in Mandarin, leading to obscure phoneme transitions and consonant-vowel structures.

Similar to the results in Wang & Zhang, the CER of noise-mixed test increased greatly especially in the white noise test [5]. Among all the 4 methods of data augmentation, the method of increasing or decreasing volume performs best in lowering CER of car and café noise tests, and with all kinds of augmentation, the model got the lowest CER in car and café noise tests, which shows that certain data augmentation methods are quite effective to improve model's robustness against noise.

However, there are no significant decreases in CER of the white noise test, no matter the kind of data augmentation applied to the audio data. White noise has uniform distribution characteristics in the time-frequency domain. Conventional enhancement methods (such as spectral masking) mainly

DOI: 10.54254/2755-2721/145/2025.21896

disrupt local features, whereas white noise affects the global signal-to-noise ratio, leading to inadequate targeting of enhancement strategies.

4.2. Comparative Experiments

Table 2: Results of CER of CNN-BiLSTM model

Data Augmentation	Test without noise	Test with background noise		
		Car noise	Café noise	White noise
No Augmentation	10.99%	14.65%	41.97%	55.43%
All Augmentation	9.66%	14.72%	38.14%	61.42%

Table 3: Results of CER of BiLSTM model

Data Augmentation	Test without noise	Test with background noise		
		Car noise	Café noise	White noise
No Augmentation	11.64%	19.76%	47.55%	60.35%
All Augmentation	12.66%	18.92%	44.56%	63.41%

Table 2 and Table 3 show the results of CNN-BiLSTM model and BiLSTM model combined with attention mechanism. The CER result (without noise) of BiLSTM model combined with CNN is 10.99% when the model converged. The possible reason for this little increase in CER is that the CNN's limited receptive fields and improper feature fusion may degrade discriminative acoustic patterns crucial for Mandarin tones, offsetting BiLSTM's sequential modeling advantages to some extent. However, there are significant decreases in CER of noise tests. Even with no data augmentation, the BiLSTM model combined with CNN performs better compared to the pure BiLSTM model trained with augmented data, demonstrating the great strength of CNN against noise including white noise.

While data augmentation improves CNN-BiLSTM model's robustness to café noise (3.83% decrease), it unexpectedly degrades performance in car and white noise tests. This phenomenon suggests that spectral/time masking may conflict with Mandarin's tone-dependent phoneme structures, particularly when critical pitch contours are obscured.

As for the BiLSTM combined with attention mechanism, the results didn't show much improvement in the accuracy of noise tests. The attention mechanism may fail to align critical tonal transitions in Mandarin speech, introducing noise from redundant context while BiLSTM's inherent bidirectional modeling already captures sufficient phonetic dependencies.

5. Conclusion

The baseline BiLSTM model achieved 9.05% CER on clean speech and performed badly in noisy conditions (e.g., 61.4% CER under white noise). While appropriate data augmentation improved robustness against noise, time-frequency masking increased CER by disrupting Mandarin's tonal transitions. Combining CNN method with BiLSTM model substantially enhanced noise robustness, leading to lower CERs compared to the BiLSTM model with augmented data. Attention mechanisms failed to improve performance, likely due to misalignment between attention weights and tonal boundaries.

This research confirms that combining CNN method with BiLSTM model and appropriate data augmentation methods can significantly improve the anti-noise ability of Mandarin speech recognition models, which achieved great reduction in CER in all the three kinds of noise tests compared to the standalone BiLSTM model. However, a key limitation of this study is that the

DOI: 10.54254/2755-2721/145/2025.21896

experiments treat Mandarin ASR as a black-box process, which means that important linguistic features of Chinese like tonal contours and syllable-timed rhythm, thereby limiting the final accuracy.

Mandarin ASR has already been implemented in many fields including voice assistants, car voice controls, factory command systems and so on. Future works will focus on building lightweight models for edge devices, real-time noise handling and dialect-related tonal analysis, to cope with challenges of the real world like noisy public spaces, and pave the way for human-machine interaction in dynamic environments.

References

[1] *Fan, H. H. (2017). Analysis and application of speech recognition technology. Electronics World, (17), 140.*

[2] *Zhang, J. (2024). Application of artificial intelligence in speech recognition. Computer Knowledge and Technology, 20(17), 46–48.*

[3] *Zhang, X. D., Zhang, Y. Y., Hu, H. D., et al. (2024). Research on speech recognition technology based on deep learning. Electronic Production, 32(16), 63–65.*

[4] *Liu, Y., & Lian, M. M. (2024). Construction method of speech recognition system based on 1D convolutional neural network. Audio Engineering, 48(10), 77–79.*

[5] *Wang, D., & Zhang, X. (2015). THCHS-30: A free Chinese speech corpus. arXiv.*

[6] *Abdel-Hamid, O., Mohamed, A. R., Jiang, H., Deng, L., Penn, G., & Yu, D. (2014). Convolutional neural networks for speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22(10), 1533–1545.*

[7] *Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. Proceedings of the International Conference on Learning Representations (ICLR).*

[8] *Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). SpecAugment: A simple data augmentation method for automatic speech recognition. Proceedings of the Interspeech 2019, 2613–2617.*