International Journal of Computer Applications Technology and Research
Volume 14–Issue 08, 33 – 37, 2025, ISSN:-2319–8656
DOI:10.7753/IJCATR1408.1003

End-to-End Chinese Speech Recognition Based on CNN and CTC

Chao Tang
Chengdu University of Information Technology
College of Communication Engineering
ChengDu, China

Zehua Lv
Chengdu University of Information Technology
College of Communication Engineering
ChengDu, China

Ximing Yuan*
Chengdu University of Information Technology
College of Communication Engineering
ChengDu, China

Abstract: Traditional acoustic models have limitations such as complex components, difficulty in joint training, and the need for data pre-alignment. To this end, this paper proposes an end-to-end Chinese speech recognition model that combines a 1D gated convolutional neural network with Connectionist Temporal Classification(CTC). the core of the model consists of stacked multi-layer 1D convolutional networks to extract contextual high-level features, gated linear units (GLU) to suppress gradient dispersion, and CTC to achieve end-to-end training and decoding of Chinese characters at the character level. Experiments demonstrate that the model significantly improves the performance over the baseline model on the public dataset, with a CER reduction above of 2.5%.

Keywords: Speech Recognition; End-to-End; Convolutional Neural Networks; Gated Linear Units; Connectionist Temporal Classification;

1. INTRODUCTION

Traditional automatic speech recognition (ASR) systems are commonly based on GMM-HMM[1] acoustic models and generate sentences by reordering word sequences with the help of external language models. The core problem lies in the modular design: the system consists of independent modules such as acoustic models, pronunciation dictionaries, and language models. This not only requires the integration of multi-disciplinary knowledge in phonetics and linguistics, but also leads to the fact that each module only optimizes its own goal during training, which leads to the accumulation of errors in the inference stage. In addition, such systems rely heavily on high-quality training data that must be forced to be pre-aligned (i.e., ensuring that each frame of input corresponds to a specific label), which makes data preparation costly. In summary, the design and training of high-performance conventional ASR systems face great complexity and difficulties. Unlike conventional systems, the end-to-end ASR system is a Sequence-to-Sequence model that directly maps acoustic signals to character/word sequences, eliminating the pre-alignment step. At its core, it integrates most of the components in a single deep neural network (DNN) and optimizes them with a unified objective function. Given the advantages of the end-to-end approach, this paper proposes a CNN model with integrated gated linear units (GLU)[2]. The model utilizes a one-dimensional CNN[3-4] to jointly extract contextual information for feature abstraction and enhance expressive capability, and employs Connected Timing Classification (CTC)[5] to achieve end-to-end Chinese speech recognition.

2. RELATED WORK

End-to-end speech recognition systems can directly map input speech to sequences of letters or words and jointly train components such as acoustic models and pronunciation models within a single system. Their implementation methods are primarily divided into two categories: CTC-based methods and attention mechanism-based methods. CTC-based methods address the issue of input sequence lengths exceeding output sequence lengths by introducing the CTC criterion, and can directly achieve end-to-end modeling when combined with deep neural networks (DNN). Attention-based methods typically include encoder and decoder networks. The encoder converts input speech into high-level feature representations; the attention mechanism dynamically determines which encoder features to focus on, generating context vectors; the decoder then combines these context vectors with the embedding information of the previous predicted symbol to predict the next output symbol. Yao Yu et al[6]. constructed an acoustic model based on a bidirectional long short-term memory neural network (BLSTM) and trained it using the CTC criterion, thereby establishing an end-to-end Chinese speech recognition system based on BLSTM-CTC. Wang et al[18]. proposed an end-to-end Mandarin speech recognition model combining CNN+BLSTM+CTC, utilizing a convolutional neural network (CNN) to learn local speech features, BLSTM to capture contextual information, and CTC for decoding. Chan et al. proposed the Listen-Attend-Spell (LAS) model, which consists of a listener (a pyramid-shaped BLSTM network) and a speller (an attention-based recurrent network decoder). The speller predicts the next character based on all generated characters and the entire sound sequence. Chiu et al[7]. optimized the LAS model by adopting a multi-head attention mechanism, significantly improving performance. Zhou et al[8]. introduced the Transformer architecture into the field of speech recognition, proposing a greedy cascaded decoder Transformer model that performs well in Mandarin speech recognition tasks.

3. DATA PROCESSING

In this paper, we use the open-source AISHELL-1 Chinese Mandarin speech dataset[9] for model training and experimentation. The dataset contains more than 140,000 voice samples recorded by 400 speakers, with a total duration of about 165 hours. The dataset is divided into three parts: training

International Journal of Computer Applications Technology and Research
Volume 14–Issue 08, 33 – 37, 2025, ISSN:-2319–8656
DOI:10.7753/IJCATR1408.1003

set, validation set and test set according to the standard way, and its specific sample distribution is shown in Table 1.

Table 1.Data set segmentation

Datasets	Duration/h	Number-of-recordings/article
Train	131	106091
Dev	17	18753
Test	17	16756

The Character Error Rate (CER), which is commonly used in Mandarin Chinese speech recognition tasks, is calculated as shown in Equation 1.

$$\mathrm{CER} = \frac{I + D + R}{L} \times 100\% \quad (1)$$

Where:I is the number of insertion errors, D is the number of deletion errors, R is the number of replacement errors, and L is the total number of words in the true label.

4. MODEL STRUCTURE

The structure of the deep neural network model based on CNN+GLU+CTC proposed in this paper is shown in Figure. 1. Firstly, the input raw audio sequence is preprocessed and features are extracted. Then, the feature sequence is processed by three groups of eight convolutional blocks (CNN blocks). Each convolutional block consists of three sequential operations, namely 1D convolution, gated linear unit and dropout. Different groups of convolutional blocks have different parameter settings. The output features of a convolutional block are then passed through a one-dimensional convolutional layer with a convolutional kernel size and step size of one. The function of this layer is to map the features of each time step into probability distributions corresponding to different Chinese characters. Finally, the connected temporal classification CTC layer receives the above probability distributions and decodes them to output the final labeled sequence y. The final labeled sequence y is the result of the convolutional block.

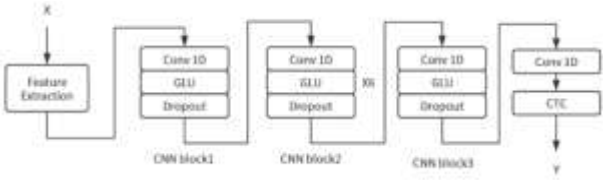

Figure 1.Model structure

4.1 Feature Extraction

The most commonly used acoustic features in end-to-end speech recognition systems are the Mel Frequency Cepstrum Coefficients (MFCC)[10] and the Filter Bank Based Features (Fbank)[11].The MFCC is a cepstrum parameter extracted in the frequency domain of the Mel scale, designed according to the auditory properties of the human ear.The Fbank feature omits the last step of the MFCC extraction process, the Discrete Cosine Transform (DCT)[12], which retains more of the original speech information than the MFCC. The Fbank feature omits the final step in the MFCC extraction process, the Discrete Cosine Transform (DCT), which retains more of the original speech information than the MFCC. In recent years, the direct use of Spectrogram[13] or even the original speech waveform for modeling has also attracted more and more attention. In this paper, we carry out a comparative study on the above features and conduct experiments using Spectrogram, Fbank and MFCC features respectively. The specific flow of feature extraction is shown in Figure 2.

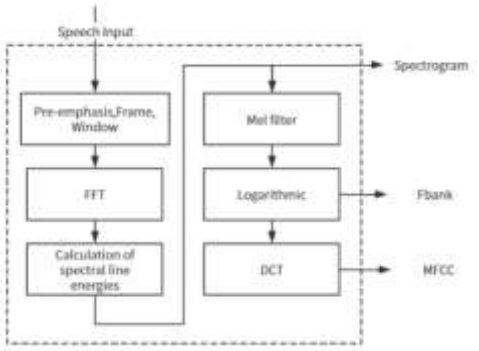

Figure 2.Audio feature extraction process

4.2 Gated Convolutional Neural Network

Convolutional Neural Network (CNN) were first widely used in image recognition[14], and then successfully introduced into natural language processing and speech recognition. The application of CNN to speech recognition has significant advantages: not only can it accurately control the context-dependent length of model perception, but it can also realize hierarchical feature extraction by stacking multiple layers of CNN-neighboring input elements interact with each other in the shallow layer of the network, while distant elements interact with each other in the deeper layer of the network, so as to progressively extract higher-level, more abstract feature representation. In this paper, a one-dimensional convolutional neural network incorporating gated linear units (GLU) is used for training, with the goal of directly mapping the input audio feature sequences to the corresponding Chinese character sequences. The structure of this gated convolutional neural network is shown in Figure. 3.

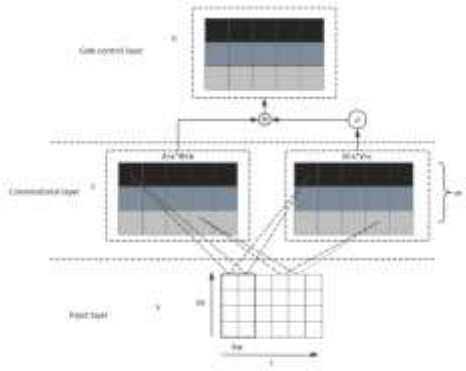

Figure 3.One-dimensional gated convolutional neural network structure

Let the sequence received by the input layer be denoted as $\mathrm{X} = (X_1, X_2, \ldots, X_T)$, where $X_i \in R^{dx}$ (each frame of input is a dx-dimensional feature vector). The sequence is computed by a convolutional layer with kernel width kw, stride s, and depth / number of filters m. The formula is shown in Equation 2.

$$c_t^i = \sum_{j=1}^{dx} \sum_{k=1}^{kw} w_{i,j,k} x_{s\times(t-1)+k}^j + b_i \quad \forall 1 \le i \le m \quad (2)$$

International Journal of Computer Applications Technology and Research
Volume 14–Issue 08, 33 – 37, 2025, ISSN:-2319–8656
DOI:10.7753/IJCATR1408.1003

where: $w \in R^{m \times dx \times kw}$ and $b \in R^{m}$ are the parameters to be learned by the model. In Figure. 3 we can see that the calculated output tensor size is size=(3,5) whenT=6,dx=4,kw=2,s=1,m=3. For the gated linear unit, the calculation formula is shown in Equation 3.

$$h(x)=(x^{*}\mathrm{W}+\mathrm{b}) \otimes \sigma(x^{*}\mathrm{V}+\mathrm{c}) \tag{3}$$

where:* is the convolution operation,⊗ is the multiplication of the corresponding elements of the matrix,σ is the sigmoid function. By introducing the gating mechanism, the network is able to dynamically regulate the flow of information in the hierarchical structure: it selectively enhances the delivery of valuable information while suppressing irrelevant information, thus improving the feature extraction capability of the network. In addition, the gated convolutional structure provides a linear pathway for gradient back propagation while maintaining a strong nonlinear modeling capability, which effectively alleviates the gradient dispersion problem in deep networks.

4.3 Connectionist Temporal Classification

Connectionist Temporal Classification (CTC) is a technique for mapping an input sequence x of mismatched length to an output sequence y (where |x| > |y|). It is able to map audio data directly to text sequences, skipping intermediate phonetic representations (e.g., phonemes). In addition, CTC does not require any pre-alignment between the input sequence and the target sequence. The Connectionist Timing Classification (CTC) algorithm consists of two main steps: path probability computation and path aggregation. The core mechanism is the introduction of a special blank tag (denoted by the symbol "-" in this paper) to characterize the separation between silent frames or characters. Define L as a dictionary containing all labeled text characters with total number of characters N, then the extended dictionary L′=L∪{"-"} with blank label "-" added, has length (N+1). For an input sequence $\mathrm{X}=(X_1, X_2, \ldots, X_T)$ of length T frames and dimension dx for each frame, an (N+1)-dimensional vector is computed by the CTC at each time step. the CTC converts the output vector into a probability distribution matrix $y=(y^1, y^2, \ldots, y^T)$ by means of the Softmax function. , where each frame is $y^t=(y_1^t, y_2^t, \ldots, y_k^t, \ldots, y_{N+1}^t)$, and y_k^t represents the probability of outputting the kth character in the dictionary L′ at the moment t, then we have $\sum_k y_k^t = 1$. After the calculation, the mapping relationship realized by CTC is shown in Equation 4.

$$y=F(x), F:(R^{dx})^T \rightarrow (R^{N+1})^T \tag{4}$$

Taking one element of the dictionary L′ at each time step t and arranging it in temporal order, an output sequence π, which we call a path, is obtained. Under the condition that the input is x, the probability that the output path is π is shown in Equation 5.

$$p(\pi \mid x)=\prod_{t=1}^{T} y_{\pi_t}^t, \forall \pi \in L'^T \tag{5}$$

The path probability is computed by multiplying the probabilities of the corresponding labeled characters output at each time step on the path π. The path probability is calculated by multiplying the probabilities of the corresponding labeled characters output at each time step on the path π. However, the path length is equal to the input sequence length T, while the actual labeled text is often much shorter than T. To resolve this length discrepancy, we should merge related paths that point to the same shorter labeled sequence. This merging process is generally divided into two steps:

- Need to delete consecutive duplicate labels: that is, when the same label is output in consecutive time steps, only the first one is retained. For example, the paths "ss-aay" and "s-ayyy" (both of which are 6 time-steps) are changed to "s-ay" after this operation.
- Remove the blank label "-" from the path. The "-" means that there is no output for this frame and it should be removed to get the final label sequence. The label "s-ay" obtained in the first step, after removing the "-", the final output is "say".

Usually, a very short output sequence can be obtained by combining multiple paths π. The fence diagram[15] in Figure. 4 gives all the legal paths of the labeling sequence "say" when the path length is 6.

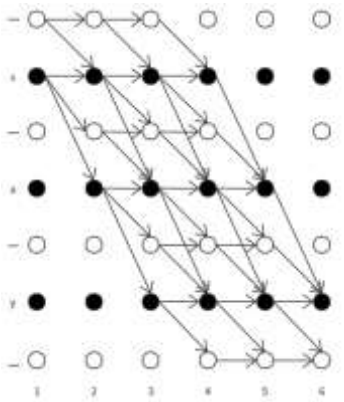

Figure 4.CTC Fence diagram

The purpose of path merging is twofold: to extract the final labeled sequence from the path, and to compute the probability of that sequence. We define the transformation B whose function is to remove blanks and consecutively repeated characters from the sequence (e.g. B(ss-aay) = B(s-ayyy) = "say"). Thus, the probability of outputting the sequence of labels l given input x is given by Equation 6.

$$P(l \mid x)=\sum_{\pi \in B^{-1}(l)} P(\pi \mid x) \tag{6}$$

The above calculations show that the probability of a label sequence is differentiable. This allows us to apply the backpropagation algorithm[16] to train the model to maximize the probability of the true label sequence. After the model training is complete, the label sequence with the highest probability is used as the final output of speech recognition during prediction.

5. EXPERIMENTS AND ANALYSIS

5.1 Comparison of different input features

To compare the three types of input features of Spectrogram, Fbank and MFCC, the experiment used the original 16 kHz sampling rate of speech data. The feature extraction process is as follows: Firstly, frame splitting is carried out with a window width of 20 ms and a frame shift of 10 ms, and a Hamming window is applied. Then calculate the FFT and spectral line energy to obtain 161-dimensional Spectrogram features; Then, the Mel filter bank is applied to obtain 40-dimensional Fbank features; Finally, calculate the MFCC of Fbank and its first-

order and second-order differences to form the common 39-dimensional MFCC features.

As shown in Table 2, comparing the optimal CER performance of the system with different input features, Fbank is optimal, Spectrogram is second, and MFCC is the worst. The reason is that MFCC loses part of the information due to the additional manual feature extraction step; Spectrogram retains complete information but has high redundancy. Under the limited dataset, it is difficult for the model to effectively filter the key information in Spectrogram, so its performance is not as good as that of Fbank.

Table 2. Comparison of different input features

Input features	CER/%
Spectrogram	17.1
Fbank	16.5
MFCC	20.2

5.2 Comparison of performance with baseline models

The model from the literature is chosen as the baseline, the input features are all Fbank, and none of the external language models are used. Table 3 shows the performance comparison results between this paper's model and the baseline model on the test set. On the AISHELL-1 dataset, the CNN+GLU+CTC model in this paper achieves a lower error rate with a CER reduction of 3.1% and 2.5% compared to the BLSTM+CTC[17] and CNN+BLSTM+CTC[18] baseline models, respectively.

Table 3. Comparison of performance with baseline models

Models	CER/%
BLSTM+CTC	22.3
CNN+BLSTM+CTC	21.7
1D-CNN+GLU+CTC	19.2

6. CONCLUSION

In this paper, we propose a Chinese Mandarin speech recognition system that directly outputs Chinese character labels with a one-dimensional gated convolutional neural network combined with CTC at its core.This design simplifies the training and decoding process of the system. Experiments were conducted to explore the effects of multiple input features on the performance. The results show that the system significantly reduces the error rate compared to the baseline on the AISHELL-1 dataset. In the future, further optimization of the model structure and use of data augmentation to compensate for data deficiencies are planned with the aim of continuing to improve the performance of the system.

REFERENCES

[1] K.-M. Y ,W. K .A study on user defined spoken wake-up word recognition system using deep neural network-hidden Markov model hybrid model[J].Journal of the Acoustical Society of Korea,2020,39(2):131-136.

[2] KimM S .End-to-end speech-denoising deep neural network based on residual-attention gated linear units[J].Electronics Letters,2024,60(20):e70020-e70020.

[3] Nanmalar M ,Joysingh J S ,Vijayalakshmi P , et al.A feature engineering approach for literary and colloquial Tamil speech classification using 1D-CNN[J].Speech Communication,2025,173103254-103254.

[4] Anver R S ,Deepambika A V ,Rahiman A M , et al.Emotional Speech Generation: An Approach Using Convolutional Neural Networks (CNN) Based Generative Adversarial Network[J].Circuits, Systems, and Signal Processing,2025,(prepublish):1-23.

[5] Ning J ,Dai Y ,Li G , et al.Semi-supervised End-to-end Speech Recognition[J].Advances in Computer, Signals and Systems,2023,7(3):.

[6] YAO Yu,RYAD Chellali.End-to-end Chinese speech recognition system using bidirectional long short-term memory networks and weighted finite-state transducers[J].Journal of Computer Applications

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

[8] Zhou Shiyu,Dong Linhao,Xu Shuang,et al.A comparison of modeling units in sequence-to-sequence speech recognition with the transformer on mandarin Chinese[C]//International Conference on Neural Information Processing

[9] Bu H,Du J,Na X,et al.Aishell-1:An open-source mandarin speech corpus and a speech recognition baseline[C]//20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment.IEEE,2017:1-5.

[10] Vajrobol V ,Aggarwal N ,Saxena J G , et al.Enhancing speaker identification in low-resource multilingual languages using hybrid MFCC-Chroma STFT and transformer encoder[J].Multimedia Tools and Applications,2025,(prepublish):1-35.

[11] Ke W .Study on recognition and classification of English accents using deep learning algorithms[J].Journal of Intelligent Systems,2023,32(1):.

[12] Shahrzad B A ,Mansour V ,Mohammadreza M .Noise Reduction of Lung Sounds based on Singular Spectrum Analysis combined with Discrete Cosine Transform[J].Applied Acoustics,2022,199.

[13] Tao H ,Li S ,Wang X , et al.Analysis and Research on Spectrogram-Based Emotional Speech Signal Augmentation Algorithm[J].Entropy,2025,27(6):640-640.

[14] Wan G ,He Q ,Zhang Q , et al.A Novel Lightweight Algorithm for Sonar Image Recognition[J].Sensors,2025,25(11):3329-3329.

[15] Wu C ,Sun H ,Huang K , et al.MPSA-Conformer-CTC/Attention: A High-Accuracy, Low-Complexity End-to-End Approach for Tibetan Speech Recognition[J].Sensors,2024,24(21):6824-6824.

[16] A. H A ,A. A A ,Islam H , et al.Arabic speech recognition using end-to-end deep learning[J].IET Signal Processing,2021,15(8):521-534.

[17] ZHANG Limin,WANG Yanzhe,ZHANG Bingqiang,et al.Mandarin recognition and improvement based on CTC criterion[J].Computer Engineering,2019,45(6):

[18] Wang Dong,Wang Xiaodong,Lv S.End-to-end mandarin speech recognition combining CNN and BLSTM[J].Symmetry,2019,11(5):1-19.