Speech Recognition via CTC-CNN Model

Wen-Tsai Sung
National Chin-Yi University of Technology
Hao-Wei Kang
National Chin-Yi University of Technology
Sung-Jung Hsiao (sungjung@gs.takming.edu.tw)
Takming University of Science and Technology https://orcid.org/0000-0002-0723-1632

Research Article

Keywords: artificial intelligence, speech recognition, speech to text, DNN, ASR

Posted Date: November 14th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2226611/v1

Speech Recognition via CTC-CNN Model

Wen-Tsai Sung[1], Hao-Wei Kang[1] and Sung-Jung Hsiao[2,*]

[1]Department of Electrical Engineering, National Chin-Yi University of Technology, Taichung, 411030, Taiwan
[2]Department of Information Technology, Takming University of Science and Technology, Taipei City, 11451, Taiwan
[*]Corresponding Author: Sung-Jung Hsiao. Email: sungjung@gs.takming.edu.tw
Received: XX Month 202X; Accepted: XX Month 202X

Abstract: In modern society, human communication is increasingly frequent, and speech plays an extremely important role in social activities. Words can be used to express emotions and thoughts. However, numerous individuals are troubled by "language barriers", due to which their communication with others is limited. This study proposes a method to address the needs of speech-impaired and deaf-mute individuals. A basic deep neural network (DNN) acoustic model was established through a voice database combined with a deep neural network. A sound sensor was used to convert the collected voice signals and process them into text or corresponding voice signals to improve communication. This method can be extended to modern artificial intelligence technology, with diversified applications, such as verbatim transcripts of meeting minutes, medical reports, automotive, sales, etc. The results obtained in this study demonstrate the efficiency of the proposed method and discuss its significance.

Keywords: artificial intelligence; speech recognition; speech to text; DNN; ASR

1 Introduction

Speech is a linguistic term coined by the Swiss linguist Saussure; it is a concept that is opposite to language. Speech activity is mainly controlled by the individual's free will; it has characteristics of personal pronunciation, word use, expression and emotion, etc. In contrast, language is a social part of speech activity, not dominated by individual will, but shared by members of society, and arises as a social psychological phenomenon. Speech activity, as defined by the linguist Saussure, is used to collectively describe the phenomenon of human speech. Human language is a natural and effective means of communication, and it is required at most levels of life to communicate with and be understood by others. Verbal communication is taken for granted by most people. In contrast, if an individual's pronunciation or expression makes it difficult for others to even understand what they are saying, it is highly inconvenient and frustrating.

Millions of people worldwide are unable to pronounce correctly and fluently due to disorders, such as strokes, amyotrophic lateral sclerosis (ALS), cerebral palsy, traumatic brain injury, or Parkinson's disease. In response to this problem, we propose an end-to-end neural network architecture, as the connectionist temporal classification-convolutional neural network (CTC-CNN) to help these people communicate normally.

Deep learning is predominantly used in visual recognition, speech recognition, natural language processing, biomedicine, and other fields, where it has achieved excellent results. However, in the field of speech recognition, the performance of the acoustic model directly affects the accuracy and stability of the final speech recognition system, such that it is necessary to consider its establishment, optimization, and efficiency in detail [1].

The experiments in this study employ CTC-CNN, which exhibits a better performance than the earlier commonly employed gaussian mixture model-hidden Markov model (GMM-HMM) acoustic model, to train the acoustic model. We use state-of-the-art techniques to verify our method. Experimental results indicate an outstanding effect. Fig. 1 illustrates the historical evolution of automatic speech recognition (ASR).

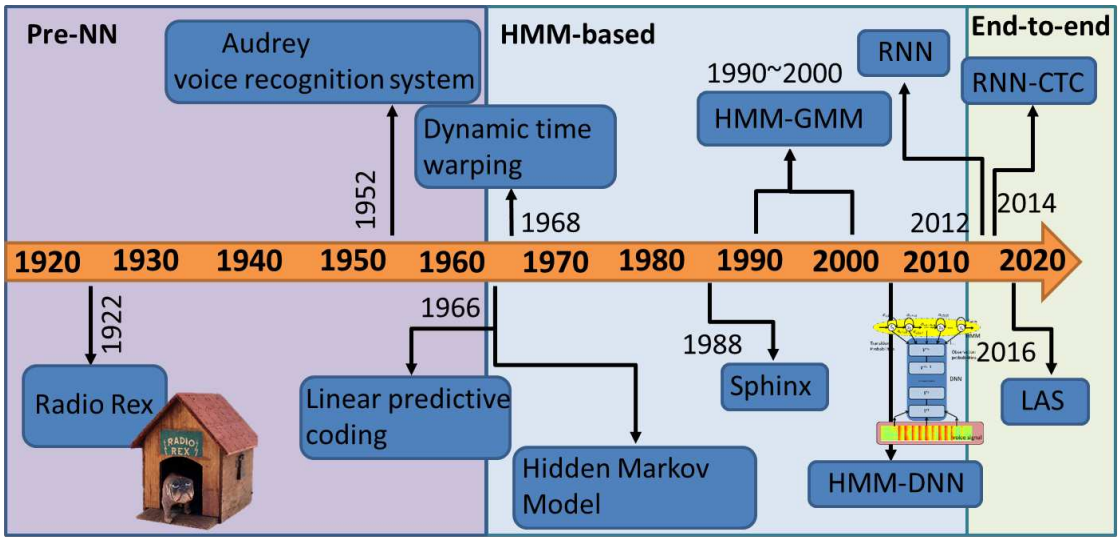


Figure 1: Historical evolution of automatic speech recognition

2 Literature Survey

In this era of technological progress, speech recognition technology has been applied in numerous fields, most of which are mainly based on intelligent electronics and driving navigation products. In addition to helping people troubled by language barriers and unable to communicate normally due to disease or various disorders, this research has the potential to bring more convenience to their lives. In the experimental architecture, this model considers linguistics, speech recognition applications, and deep learning techniques. To provide assistance to individuals with speech recognition and language impairments, it is necessary to understand basic linguistic theory. Because language belongs to human spontaneous speech, it contains numerous irregular variables, such as personal pronunciation, words, expressions, and various factors that lead to a certain degree of complexity in establishing the acoustic model to ensure that it meets the requirements as much as possible. Deep learning serves to improve the efficiency and accuracy of the acoustic model performance.

2.1 Speech Recognition

In recent years, the use of speech recognition has been spreading widely across various fields, and it is no longer limited to intelligent electronics products, but gradually expanding to the healthcare industry and even to product sales and customer services. A good speech recognition system must allow organizations to customize and adapt the technology to their specific needs, ranging from nuances in language to speech to everything else. For example:

1. Language weighting：A discriminative weighted language model is proposed to better distinguish similar languages. Similar utterances or words are weighed to improve the accuracy [2].

2. Speaker markers：Speaker selection, taking turns, elaboration, and digression. After providing definitions of discourse markers, turns, floor control types/turn segments, topic units and actions, a list of verbal and non-verbal discourse markers is specified and grouped into subcategories according to their semantic relationship [3].

3. Acoustic training：Building acoustic models from large databases has been shown to benefit the accuracy of speech recognition systems. Deep learning is employed to train these systems to adapt to various acoustic environments, such as speaker pronunciation, speech rate, and pitch, etc., to cope with a variety of different situations.

4. Indecent content filtering: Filters are used to identify profanity words or nonsense particles, etc., and eliminate this type of speech [4].

2.1.1 Pattern Recognition

Current mainstream large-vocabulary speech recognition systems mostly use statistical pattern recognition technology. A typical speech recognition system based on statistical pattern recognition method consists of the following basic modules:

1. Signal processing and feature extraction modules: The main work of this module is to extract sound features from the input signal and provide them with an acoustic model for processing. It also includes signal processing techniques to minimize the influence of environmental noise, channel, speaker, and other factors on the characteristics.

2. Acoustic model: Typical systems are mostly modeled based on first-order hidden Markov models.

3. Pronunciation dictionary: The pronunciation dictionary contains the vocabulary and pronunciations that can be handled by the system. The pronunciation dictionary actually provides the mapping between the acoustic model modeling unit and the language model modeling unit.

4. Language model: A statistical language model represents a probability distribution over a sequence of words, and a language model mainly provides the context to distinguish two words and phrases that have similar pronunciations but different meanings, as shown in the example in Fig. 2. This model is often used in numerous natural language processing applications, such as speech recognition, machine translation, and part-of-speech tagging, etc. Because words and sentences are of any length in any combination, strings that have not appeared in the training process will appear. This further makes it difficult to estimate the probability of strings in the database.

5. Decoder: The decoder is one of the core aspects of the speech recognition system. It mainly uses the input signal to find the word string that outputs the signal with the greatest probability according to acoustics, language models, and dictionaries.

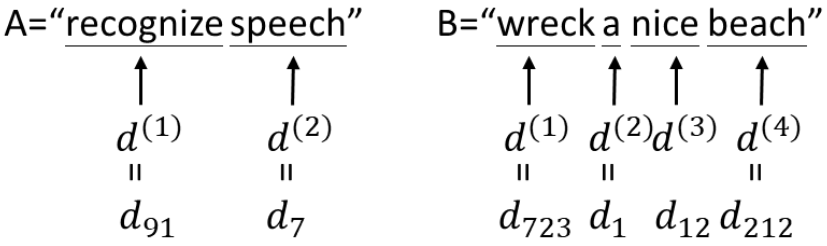


Figure 2: Two homonymous English strings

Fig. 2 illustrates two homonymous English strings. In the language model, in addition to the pronunciation that affects the accuracy of speech recognition, punctuation is likewise an important reason that affects the recognition of the system. Therefore, we discuss several considerations when constructing a post-processing system: (1) Restoring the original requires a high-accuracy model of text punctuation and capitalization. The model must make quick inferences about interim results and catch up on instant captions. (2) Using several resources: Speech recognition is an AND operation-intensive technology, such that punctuation patterns do not need to be so computationally intensive. (3) Ability to handle text not listed in the vocabulary: sometimes, the system must add punctuation or capitalization to text that the model has not seen before.

2.1.2 Speech Recognition Algorithms

Speech recognition is considered one of the most complex fields in modern technology, as it involves linguistics, mathematics, and statistics. At present, the common speech recognition system is mainly composed of several technologies, such as: speech signal input, feature extraction, and acoustic model establishment, feature vector, decoder, and result output. Speech recognition technology is evaluated based on its accuracy, the word error rate (WER), and speed. A variety of factors affect the misspelling rate.

Here are some of the various algorithms and techniques that are currently most commonly used to recognize speech and convert it to text:

1. Natural Language Processing (NLP) belongs to the field of artificial intelligence, which focuses on language interaction between humans and machines through speech and text. Numerous mobile devices currently incorporate speech recognition into their systems to provide more assistance.

2. Hidden Markov Model (HMM) is used as sequence model in speech recognition, assigning labels to each unit in the sequence, i.e., words, syllables, sentences, etc. These labels map between them and the input provided, such that it can determine the most appropriate sequence of labels.

3. N-grams is the simplest type of language model (LM) that assigns probabilities to sentences or phrases. An N-gram is a sequence of N words. For example: “How are you” is a ternary or, and “I'm fine thank you” is a quaternary. Grammar and specific word sequence probabilities are used to enhance recognition and accuracy.

4. Neural networks are mainly used in deep learning algorithms. They learn the mapping function through supervised learning and adjust it according to the loss function during gradient descent.

5. Speaker Discrimination (SD) algorithms identify and separate utterances by speaker identity. This helps the system make better distinctions between individuals in a conversation **[5]**.

2.2 Convolutional Neural Networks Applied in This Study

2.2.1 CTC-CNN Acoustic Model

In the speech recognition system, the acoustic model is an important underlying model, and its accuracy directly affects the performance of the entire system. When acoustic features remain unchanged, the performance of the speech recognition system is mainly improved by optimizing the acoustic model. Early speech recognition systems mainly employ the GMM-HMM acoustic model, which is a shallow model. Thus, it is difficult to accurately describe the state space distribution of features. Furthermore, the frame-by-frame training mode requires mandatory alignment of the training speech, which increases the difficulty of model training. With the development of deep learning, speech recognition systems began to use deep learning-based acoustic models, and achieved remarkable results. The latest end-to-end speech recognition framework abandons the more restrictive model of HMM, and directly optimizes the likelihood of input and output sequences, which significantly simplifies the training process. Deep neural networks, loop neural networks, and convolutional neural networks achieved great results in the field of speech recognition with their own advantages **[6]**.

In this study, the convolutional neural network was mainly used to build the acoustic model combined with the connection sequence classification algorithm, which significantly improves the accuracy and performance of the speech recognition system. Based on the establishment of the baseline acoustic model, this research significantly reduces the error rate of the speech-to-pinyin sequence by continuously optimizing the acoustic model. As the output of the acoustic model, the choice of the modeling unit is also one of the factors affecting the performance of the acoustic model. When selecting a modeling unit, it is necessary to consider: (1) whether the modeling unit fully represents the context information, i.e., the accuracy of the modeling unit; (2) whether it can describe the acoustic features for generalizability; (3) whether there is sufficient language material that can satisfy the modelling unit for

model training and trainability. When building the speech recognition system in this study, a non-complete end-to-end speech recognition framework is employed, i.e., the acoustic model uses the end-to-end recognition framework to convert speech into pinyin sequences, and then uses the language model to convert the pinyin sequences into text. In this study, a convolutional neural network is used to build the acoustic model, which is combined with the connectionist temporal classification (CTC) algorithm to realize the conversion of phonetic to pinyin sequences. Traditional classification methods face problems, such as unequal input and output lengths, and frame-by-frame training is required. CTC can directly map the input speech sequence into a string of text sequences, such that it can optimize the likelihood of the input and output sequences, which significantly simplifies the training process. The essence of the acoustic model based on CTC remains a sequence classification problem, meaning that the output of each node in the output layer of the neural network selects a generation path with the highest probability. Therefore, the input and output of CTC are often in a many-to-one relationship [7]. When the CTC-based acoustic model recognizes speech, the acoustic feature parameters are further extracted through the convolutional neural network, and then the posterior probability matrix is output through the fully connected network and the SoftMax layer. The maximum probability label of each node is thus used as the output sequence. Finally, the optimized output label sequence of the CTC decoding algorithm marks the recognition result. The schematic diagram of the CTC-CNN acoustic model is shown in **Fig. 3 [8]**.

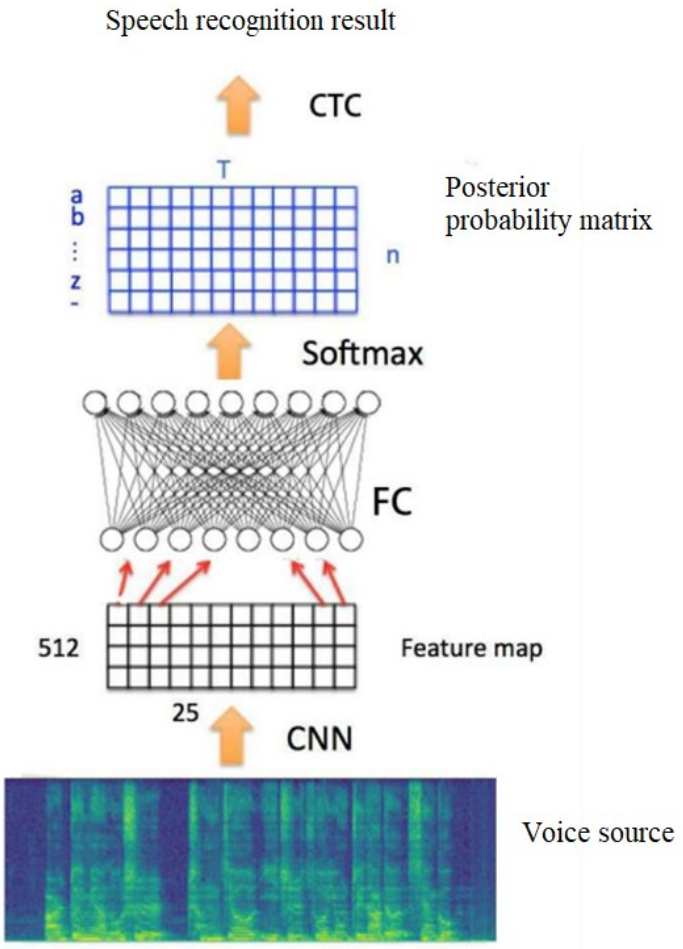


Figure 3: Schematic diagram of CTC-CNN acoustic model

2.2.2 Core Idea of CTC

The core ideas of CTC mainly include the following parts:

(1) Expanding the output layer of CNN, adding a many-to-one spatial mapping between the output sequence and the recognition result (label sequence), and defining the CTC loss function on this basis.

(2) Drawing on the idea of the forward algorithm of HMM, the dynamic programming algorithm is used to effectively calculate the CTC loss function and its derivative, thus solving the problem of end-to-end training of CNN [9].

(3) Combined with the CTC decoding algorithm, the end-to-end prediction of sequence data is effectively realized [10].

Assuming that the speech signal is x, and the label sequence is l, the neural network obtains the probability distribution of the label sequence $(l|x)$ during the training process. Therefore, after inputting the speech, the output sequence is selected with the highest probability, and after CTC decoding optimization, the final recognition result $O(x)$ can be output, where the operation formula is shown in Eq.(1).

$$O(x) = \text{argmax}P(l|x) \tag{1}$$

Given a CNN acoustic model for CTC derivation training, we first assume that S is the training data set, X is the input space, Z is the target space (the set of labelled sequences), and L is defined as the sum of all output labels (modeling units). Set, CTC extends L to $L' = \text{L} \cup \text{Blank}$. Under given conditions, the probability of outputting label k at time t can be expressed as Eq.(2).

$$y_k^t = P(\text{O}_t = k|x_1, x_2, \ldots, x_t) \tag{2}$$

Assuming that under the condition of a given input sequence x, the output label probability is independent at time t, and L'^T is defined as the set of output sequences of length T composed of L', then the conditional probability formula of a path $\pi \in L'^T$ is given by Eq. (3).

$$P(\pi|x) = \prod_{t=1}^{T} \text{y}_{\pi_t}^t \tag{3}$$

We define the mapping relationship B: $L'^T \rightarrow L^{\leq T}$ from the path π to the label sequence l. Using this mapping relationship will keep only one consecutive and identical label in the output sequence contained in the path π, and remove the Blank label. Then, to calculate the probability of label sequence $l \in L^{\leq T}$, it is necessary to accumulate all path probabilities belonging to l, and the calculation formula is shown in Eq. (4).

$$P(l|x) = \sum_{\pi \in B^{-1}(l)} P(\pi|x) \tag{4}$$

The mapping of path π to label sequence l is shown in Fig. 4.

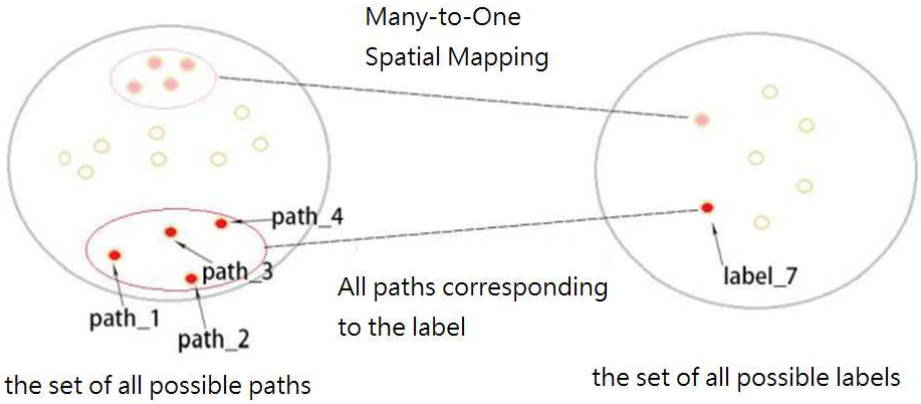


Figure 4: Mapping output to label sequence

Fig. 4 shows that the probability of label sequence $label\_7$ is equal to the total probability of its

entire path, that is, $P\ (label\_7) = P\ (path\_1) + P\ (path\_2) + P\ (path\_3) + P\ (path\_4)$. It is impractical to directly violently calculate $(l|x)$, which will increase the training time of the model and occupy computing power. Borrowing the forward and backward algorithm in HMM effectively solves $(l|x)$, and it is assumed that under the condition of a given input sequence x, the output label probability at time t is independent, such the transition probability between states does not need to be considered. The derivation diagram of the forward and backward algorithm is shown in **Fig. 5 [11]**.

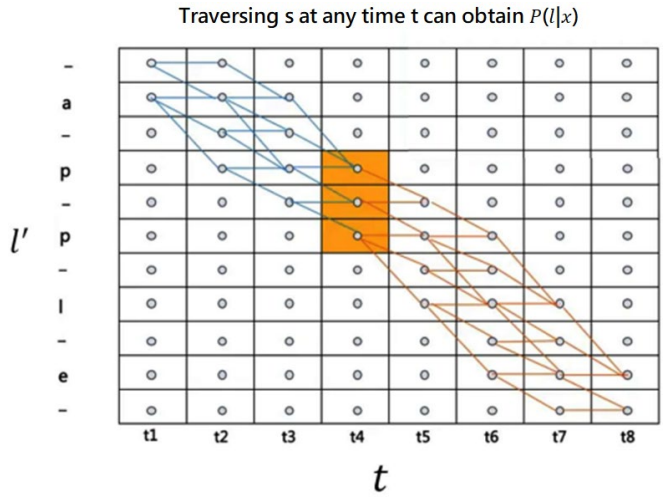


Figure 5: Derivation of forward and backward algorithm

The calculation of the forward-backward algorithm is as follows: For the input sequence x and the label sequence l with the time sequence length T, the extended label sequence is l', and the length of the extended label sequence is $|l'|=2|l|+1$, defining the first t. The forward probability of outputting the extended label at the sth position at the moment is $\alpha(t,s)$, and the posterior probability calculation formula of the label sequence is shown in **Eq.5 [12].**

$$P(l|x) = \alpha(T,|l'|) + \alpha(t,|l'|-1) \tag{5}$$

Before calculating the forward probability, the parameters must be initialized first, and the Blank label is abbreviated as b, such that the calculation formula is **Eq.6:**

$$\alpha(1,1) = y_b^1;\ \alpha(1,2) = y_{l_2'}^1;\ \alpha(1,s) = 0(\forall s > 2) \tag{6}$$

The recursive calculation formula of forward probability obtained by recursion is shown in **Eq.7.**

$$\alpha(t,s) = \begin{cases} \left(\alpha(t-1,s) + \alpha(t-1,s-1)\right)y_{l_s'}^t & if\ \ l_s' = b\ \ or\ \ l_{s-2}' = l_s' \\ \left(\alpha(t-1,s) + \alpha(t-1,s-1) + \alpha(t-1,s-2)\right)y_{l_s'}^t & \text{others} \end{cases} \tag{7}$$

The backward algorithm is similar to the forward algorithm. The backward probability of outputting the extended label at the sth position at time t is defined as $(t,)$, and the posterior probability calculation formula of the label sequence is shown in **Eq.8.**

$$P(l|x) = \beta(1,1) + \beta(1,2) \tag{8}$$

Before calculating the backward probability $(t,)$, we initialize the parameters, as shown in **Eq.9.**

$$\beta(T,|l'|) = y_B^T, \beta(T,|l'|-1) = y_{l'_{|l'|-1}}^T, \beta(T,s) = 0(\forall s < |l'|-1) \tag{9}$$

The recursive calculation formula of the backward probability obtained by recursion is shown in **Eq.10.**

$$\beta(t,s) = \begin{cases} \left(\beta(t+1,s) + \beta(t+1,s+1)\right)y_{l_s'}^t & if\ \ l_s' = b\ \ or\ \ l_{s+2}' = l_s' \\ \left(\beta(t+1,s) + \beta(t+1,s+1) + \beta(t+1,s+2)\right)y_{l_s'}^t & \text{others} \end{cases} \tag{10}$$

For any moment t, the posterior probability of the label sequence is calculated using the forward and backward probabilities, and the calculation formula is shown as **Eq.11.**

$$P(l|x) = \sum_{s=1}^{|l'|} \frac{\alpha(t,s)\beta(t,s)}{y_{l'_s}^t} \tag{11}$$

With the posterior probability calculation formula $(l|x)$ of the label sequence, the training target can be optimized, and the parameters can be updated. The loss function of CTC is defined as the negative log probability of the label sequence on the training set S. Then, the loss function (x) output of each sample is given by **Eq.12.**

$$L(x,l) = -lnP(l|x) \tag{12}$$

The loss function L_S of the entire training set is given by **Eq.13.**

$$L_S = -\sum_{(x,l)\in S} lnP(l|x) \tag{13}$$

The loss function L takes the derivative of the network output parameter y_k^t, and its operation formula is shown in **Eq.14.**

$$\frac{\partial \mathrm{L}}{\partial y_k^t} = -\frac{\partial}{\partial y_k^t}\ln P(l|x) = -\frac{1}{\ln P(l|x)}\frac{\partial}{\partial y_k^t}\ln P(l|x) \tag{14}$$

The chain rule yields the partial derivative of the loss function to the network output u_k^t without the SoftMax layer. Because a character k may appear multiple times in a label sequence, a set is defined to represent the position where k appears: $(l) = \{s{:}l_s{=}k\}$. It is obtained as **Eq.15.**

$$-\frac{\partial \ln P(l|x)}{\partial u_k^t} = y_k^t - \frac{1}{P(l|x)y_k^t}\sum_{s\in lab(l,k)} \alpha(t,s)\beta(t,s) \tag{15}$$

The parameters of the neural network part are updated layer by layer and frame by frame according to the back-propagation algorithm. When CTC decodes the output, the output sequence must be optimized to obtain the final label sequence. This study adopts the Best path decoding algorithm, assuming that the probability maximum path π and the probability maximum label l^* have a one-to-one correspondence, meaning that the many-to-one mapping B is degenerated into a one-to-one mapping relationship, and the algorithm accepts each frame. The label sequence corresponding to the output sequence generated by the maximum probability label is used as the final recognition result. First, we must calculate the maximum probability path π^* output by the network, and the operation formula is shown in **Eq.16 [13].**

$$\pi^* = argmax\, P(\pi|x) \tag{16}$$

Then, we calculate the label sequence output by the network, and define $l^* = (\pi^*)$. The formula of l^* is given by **Eq.17.**

$$l^* = \mathrm{B}\left(argmax_\pi\, P(\pi|x)\right) \tag{17}$$

The recognition result of the final acoustic model is given by **Eq.17** In essence, the acoustic model of CTC can be directly output to Chinese characters end-to-end. Due to the limitation of the training corpus and the complexity of the model, the output of the acoustic model in this study is Pinyin; the final result of speech recognition is obtained by inputting the pinyin sequence into the language model **[14-16].**

2.2.3 Construction and Training of Baseline Acoustic Model

In the convolutional neural network, the structure of the convolutional and pooling layers indicate that the input features with slight deformation and displacement are accurately recognized. This translation invariance characteristic is beneficial to the recognition of spectrogram features. The training mode of parallel computing of the convolutional neural network effectively shortens the training time and utilizes the powerful parallel processing capability of GPU. CTC illustrates the optimization of the loss function of the neural network and the optimization of the output sequence. Therefore, this study proposes a CTC-CNN acoustic model based on CNN combined with the CTC algorithm. The overall structure of

the CTC-CNN acoustic model is shown in Fig. 6 [17][18].

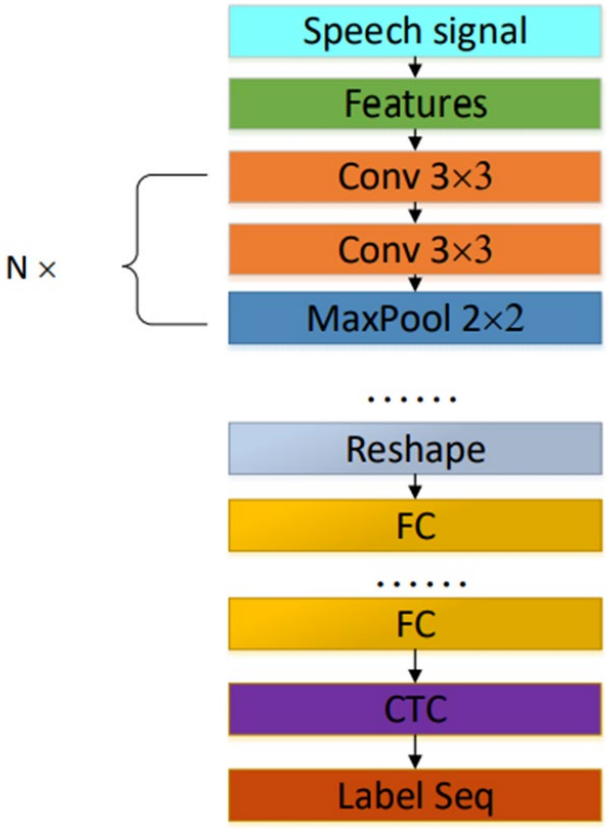


Figure 6: Structure of CTC-CNN acoustic model

3 System Architecture

3.1 System Design

Fig. 7 portrays the architectural diagram of the hardware employed this experiment. It comprises the ReSpeaker Mic Array v2.0.1 and display screen. The ReSpeaker Mic Array v2.0.1 is used to record voice data, and the recorded voice signals are compared with the voice database. The algorithm is processed, and calculated results are displayed on the display screen to display the words, sentences, or phrases after the speech is converted into text.

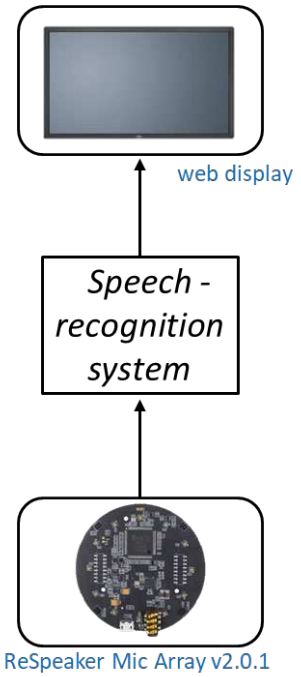


Figure 7: Hardware architecture diagram

Fig. 8 shows the overall structure and flow chart of the speech recognition assistance system for language-impaired individuals. ReSpeaker Mic Array v2.0.1 records the speech signals of individuals with language impairments, and extracts the recorded original voice recording files through a Python algorithm. Then, the algorithms extract voice features from the extracted raw data, and yield the extracted features. The vectors are calculated algorithmically by the speech recognition system, which includes acoustic comparison and language decoding. The features are repeatedly compared and decoded in acoustic comparison and language decoding, until the calculated result is very similar or correct to the original intention of the speaker, i.e., it yields the intended output. The final result is presented in the form of text to be displayed on the vehicle.

The upper layer of acoustic comparison and language decoding is mainly divided into three parts, namely the acoustic model, pronunciation dictionary, and language model. Among them, the acoustic model uses the language corpus to train and adjust the acoustic model, enabling cross-comparison with the speaker's pronunciation, words, and expressions to improve the accuracy of identification. The language model is generated in the same manner as the acoustic model. The language model is trained and adjusted through the text corpus, to establish common words or sentences, and even multi-languages.

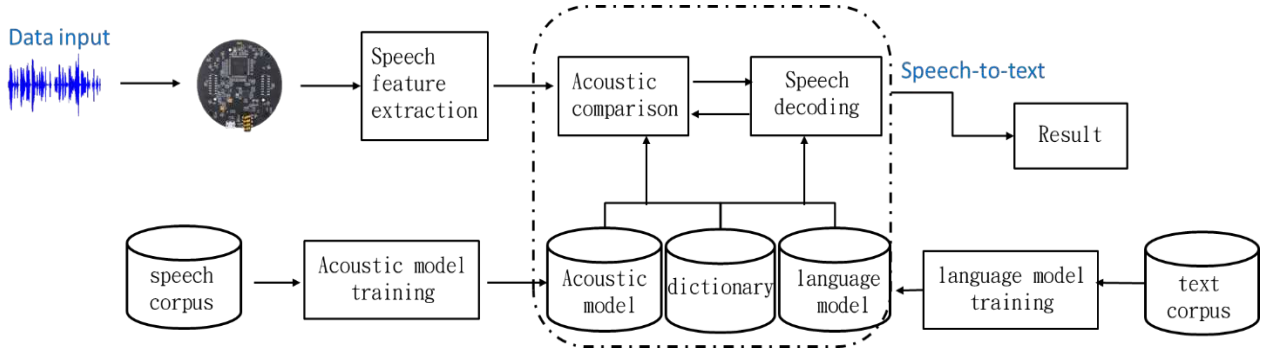
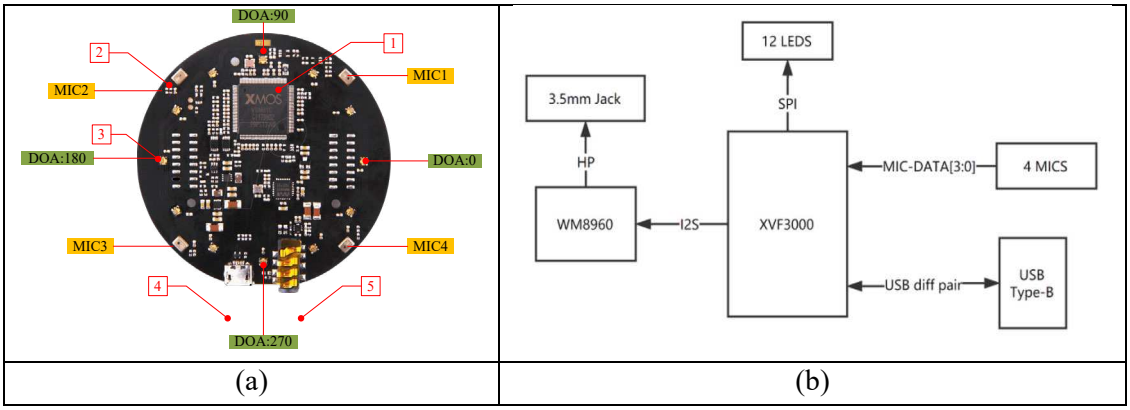


Figure 8: System architecture diagram

3.2 ReSpeaker Mic Array v2.0.1

The radio hardware component employed in this experiment is the ReSpeaker Mic Array v2.0.1, which is launched by SeeedStudio. This is an upgrade to the original ReSpeaker Mic Array v1.0. This upgraded version is based on XMOS' XVF-3000, a chipset with significantly higher performance than the previously used XVSM-2000. This new chipset includes several speech recognition algorithms to improve its performance. This array can be stacked (connected) on top of the original ReSpeaker Core to significantly improve the voice interaction performance. The microphones in this release are likewise improved, providing a significant boost in performance compared to the first-generation microphone array with only four microphones. This can be used in numerous occasions, such as: smart speakers, intelligent voice assistant systems, voice conference systems, and car voice assistants. The ReSpeaker Mic Array v2.0.1 module has numerous voice algorithms and features, and the maximum sampling rate is 16 kHz. This small chip has the benefits of numerous functions, as the module is equipped with XMOS's XVF-3000 IC, which integrates advanced DSP algorithms, including acoustic echo cancellation (AEC), beamforming, demixing, noise suppression, and gain control. Furthermore, it also performs on-chip speech algorithms. This module contains four high-performance digital microphones (MP34DT01-M) The microphone is an ultra-compact, low-power, omnidirectional, digital MEMS microphone with built-in capacitive sensing element and I2C interface, and supports far field voice capture, far-field voice capture recording, and understands requirements up to five meters away. By focusing on the correct sound, as DoA reveals the direction of the source to the device, BF allows the device to focus only on the sound coming from the target direction, ignoring background noise and chatting through NS. The module also improves recording quality, reduces ambient voice echo, and employs AEC to remove the current audio output. The module also features the WM8960, a low-power stereo codec with Class D speaker drivers capable of delivering 1 W per channel at an 8 W load. It further supports USB Audio Class 1.0 (UAC 1.0) and has twelve programmable RGB LED indicators for user freedom. **Figs. 9(a)** and **(b)** show the ReSpeaker Mic Array v2.0.1 and the module system diagram, respectively **[19]**.



(a) (b)

Figure 9: (a) ReSpeaker Mic Array v2.0.1(b) ReSpeaker Mic Array v2.0.1 System Diagram

3.3 System Technology Description

The study proposes an end-to-end speech enhancement architecture that (1) models the original speech waveform time domain signal, bypassing the phase processing operation in the traditional time-frequency conversion, and avoiding phase pollution, (2) transforms the one-dimensional time-domain speech signal, mapping it to a two-dimensional representation. More sufficient information is mined from the high-dimensional representation of the speech signal, and the codec network is subsequently used to learn the mapping from noisy to clean speech. This represents the dimensionality reduction and reconstruction into a time-domain waveform signal, (3) by combining the evaluation index with the loss function, the commonality and difference between different evaluation indexes are used to improve the perceptual ability of the model and obtain clearer speech.

The end-to-end model framework UNet [20] comprises the main structure of the framework, as shown in Fig. 10. The UNet neural network was initially applied for medical image processing and achieved good results. The main structure of UNet is composed of an encoding stage (left half of UNet) and a decoding stage (right half of UNet). Between each corresponding encoding stage and decoding stage, skip connections are used. The skip connections herein are not residual, as it is not the calculation method of the residual, but the method of splicing.

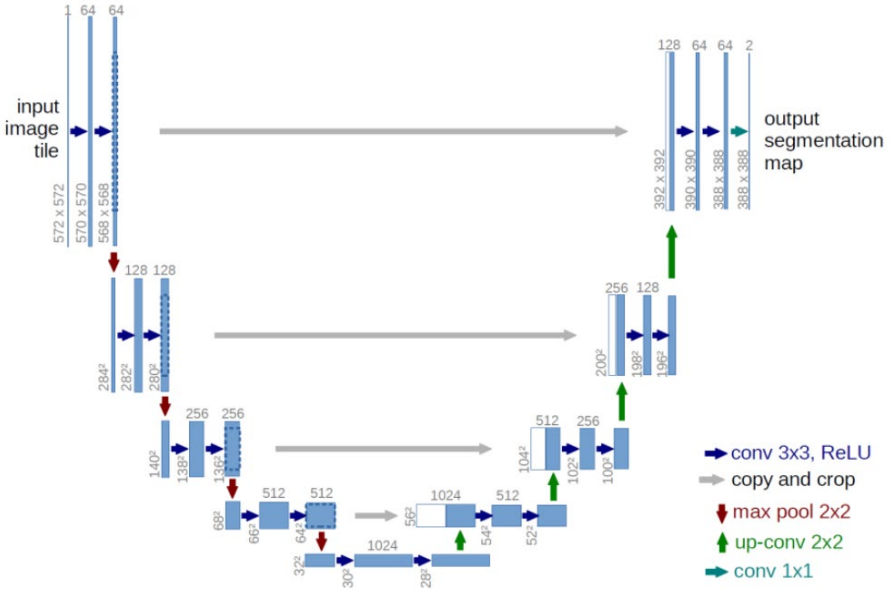


Figure 10: End-to-end model framework UNet

The structure of the model proposed in this study is shown in Fig. 11. The architecture consists of three parts, namely, preprocessing of the original audio signal, encoding and decoding module based on the UNet architecture, and post-processing of enhanced speech synthesis. By directly modeling the time domain speech signal, we avoid the defects and problems in the time-frequency transformation, and convert the one-dimensional signal into a two-dimensional signal through the convolution operation, such that the neural network can mine the speech signal in the high-dimensional space and deep representation. To reduce the amount of parameters and the complexity of the model, the up-sampling operation in the decoding part of UNet here is not deconvolution, but bilinear interpolation [21].

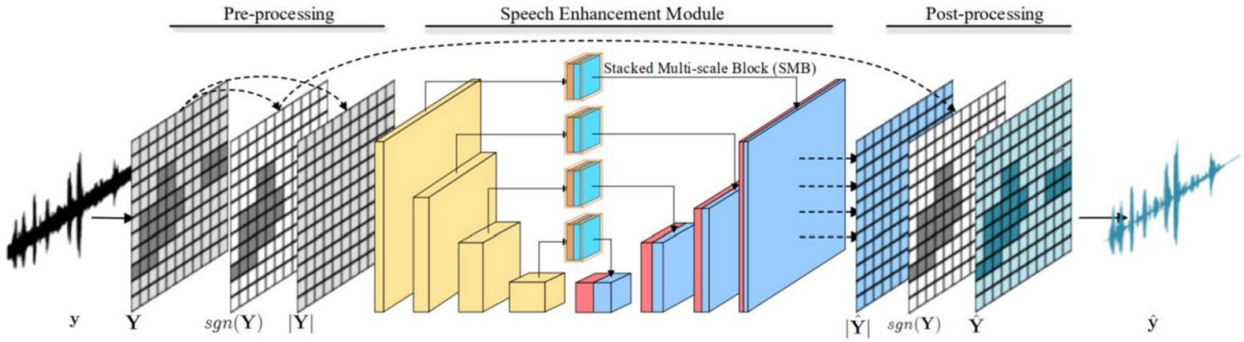
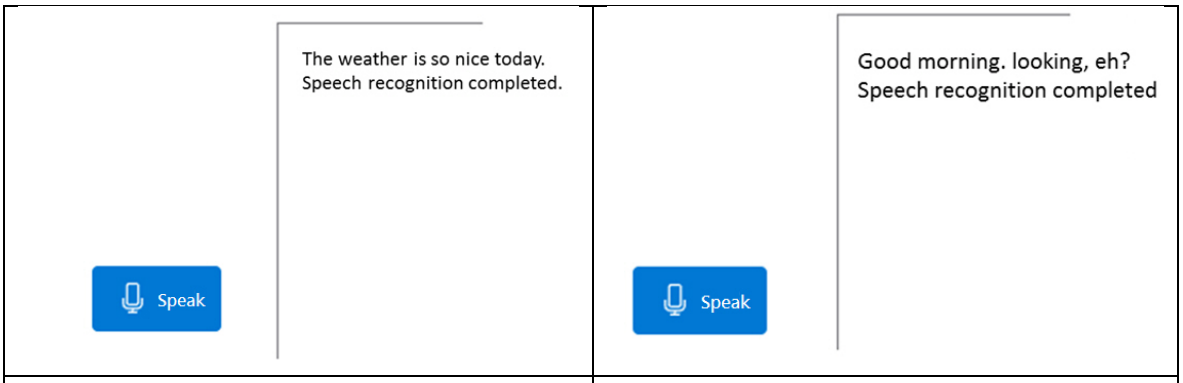


Figure 11: End-to-end speech enhancement framework

4 Analyses of Experimental Results

4.1 Basics Experimental Results

Experimental results are presented in **Figs. 12(a), (b)** as screenshots of the web GUI interface. **Fig.12 (a)** shows the speaker saying "The weather is so nice today", and the system successfully displays the speaker's complete sentence. **Fig.12(b)** shows the speaker saying "Good morning" twice in a row, but the recognition result is only successful one time; the result of the second time presents the situation of homophones. First, a voice recording is made on the ReSpeaker Mic Array v2.0.1. Subsequently, the algorithm rapidly performs voice recognition and displays the speaker's incomplete or intermittent sentences on the vehicle, helping the language-impaired person to communicate smoothly and quickly with others **[22]**.

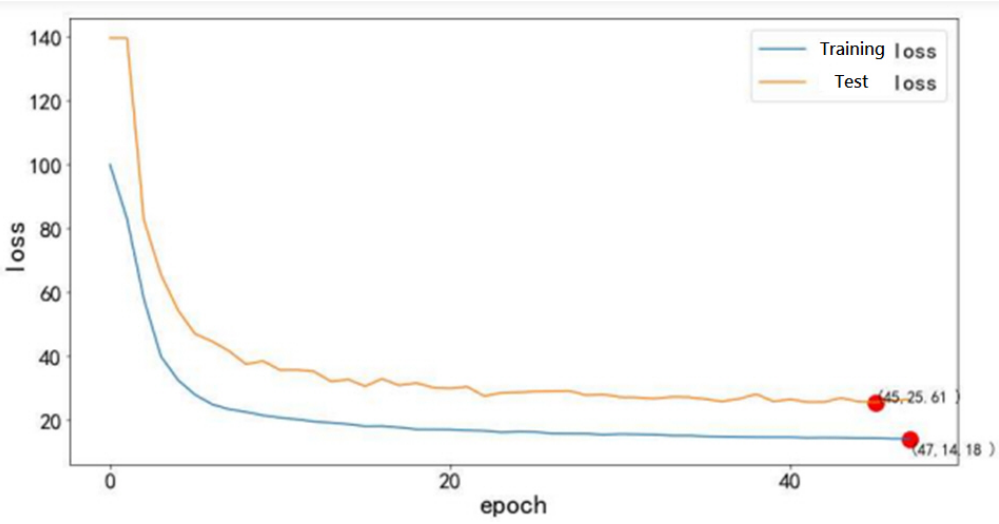


(a) Identify success | (b) Identify failure

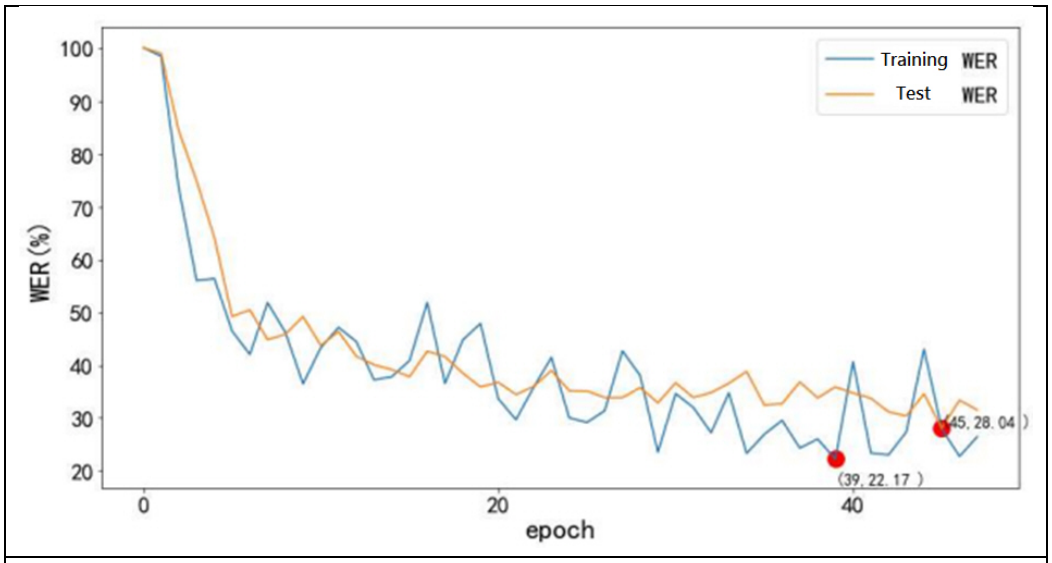
Figure 12: Experimental results

4.2 Experimental Data Analysis

To train and verify the CTC-CNN baseline acoustic model, THCHS-30 and ST-CMDS speech datasets were used as training data sets, and the data sets are divided into training and test sets. The training results are shown in **Figs. 13(a)** and **(b)**, which show that after 54 epochs of training, the word error rate of the acoustic model training set is about 31 %, and the word error rate of the test set is stable at about 43 %. There is a certain overfitting phenomenon. Namely, the 43 %-word error rate is difficult to put into practical application, such that it is necessary to optimize and adjust the network structure parameters to further improve the accuracy of the acoustic model.



(a)



(b)

Figure 13: (a) Loss variation of baseline acoustic model (b) Word error rate change of baseline acoustic model

Tab. 1 lists the recognition accuracies of a number of consecutive words. The results are obtained from 100 test datasets. When the speaker only speaks one word, the recognition accuracy is the highest, and the accuracy rate can reach 98.11 %. In contrast, when the speaker utters sentences with more than five words, the recognition accuracy rate falls to only 93.77 %. Sentences with more than five characters may cause the system to misdiagnose the words due to the speaker's punctuation or if the pronunciation of the words is too similar, for example: "recognize speech" and "wreck a nice beach" have similar pronunciations in English, and "factors" with similar pronunciation in Chinese and "Sonic" etc. In addition to the above-mentioned situations that affect the accuracy of identification, the environmental noise factor may also lead to a decrease in the accuracy of identification. Fig. 14 shows prediction trend chart of various word count recognition accuracy rates.

Table 1: Recognition accuracy of each character number

Word count	*Accuracy [%]*
One	98.11
Two	97.99
Three	97.45
Four	96.84
Five	93.77

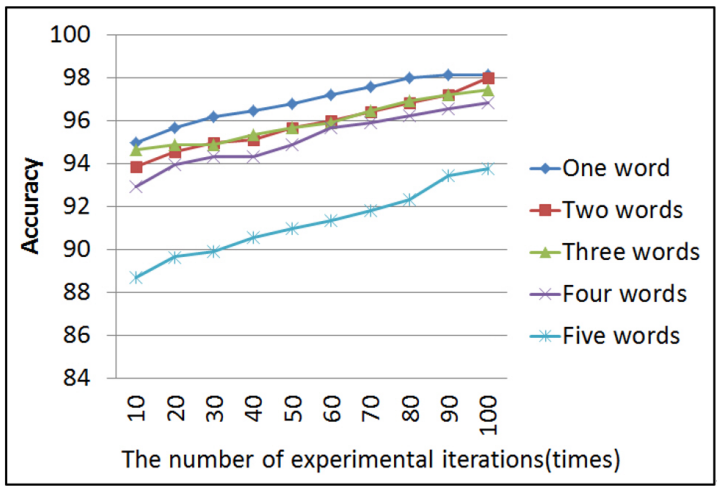


Figure 14: Prediction trend chart of various word count recognition accuracy rates

Therefore, this experiment also considers the surrounding ambient noise, and conducts 80 tests at each level of decibels, as shown in the following **Tab. 2**. Taking the noise of 80–90 dB as an example, at this level of noise, the accuracy rate is 88.18 %, which represents the poorest performance among all levels. In contrast, at 40–60 dB, owing to less noise pollution, the accuracy rate is as high as 97.33 %. **Fig. 15** shows the prediction trend of environmental noise impact identification accuracy.

Table 2: Environmental noise affects the recognition accuracy

Noise [dB]	*Accuracy [%]*
40 – 50	97.33
50 – 60	95.69
60–70	92.45
70–80	90.32
80–90	88.18

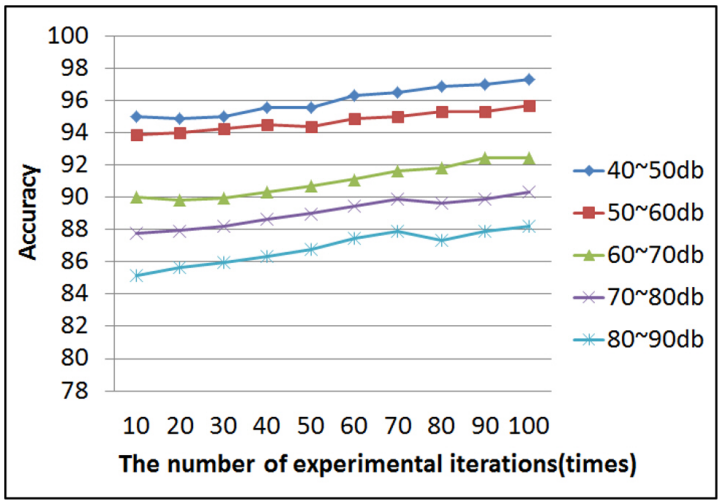


Figure 15: Prediction trend of environmental noise impact identification accuracy

Because of the similarities in Chinese pronunciation, the recognition error rate of the system is expected to increase significantly. To this end, we designed this experiment based on the characteristics of Chinese consonants and vowels to verify their time-frequency map. There are 21 consonants and 16 vowels, respectively, in the Chinese phonetic alphabet. The forming of vowels mainly occurs by the change of mouth shape, while the consonants are formed by controlling the flow of air through certain parts of the oral cavity or nasal cavity. Therefore, the energy of consonants is small, their frequency is high, and the time is short, and most of them appear before vowels. Conversely, vowels have higher energy, lower frequency, and longer duration, and usually appear after consonants or independently. The energy and frequency difference of vowels can be verified through time-frequency graph experiments, and this difference can be used to perform simple vowel identification [23].

The experimental results in Fig. 16 indicate that the amplitudes of consonants are small, while the amplitudes of vowels are relatively large. Taking the character "jin" as an example, the amplitude of the first consonant "ji" is relatively small, and the amplitude does not increase significantly until the vowel "yī" appears. However, if there are words with double vowels (for example: wāng), the amplitude will always be very large, resulting in blurred boundaries between sounds. In this situation, it is more difficult to use the amplitude to determine the change of vowel sounds.

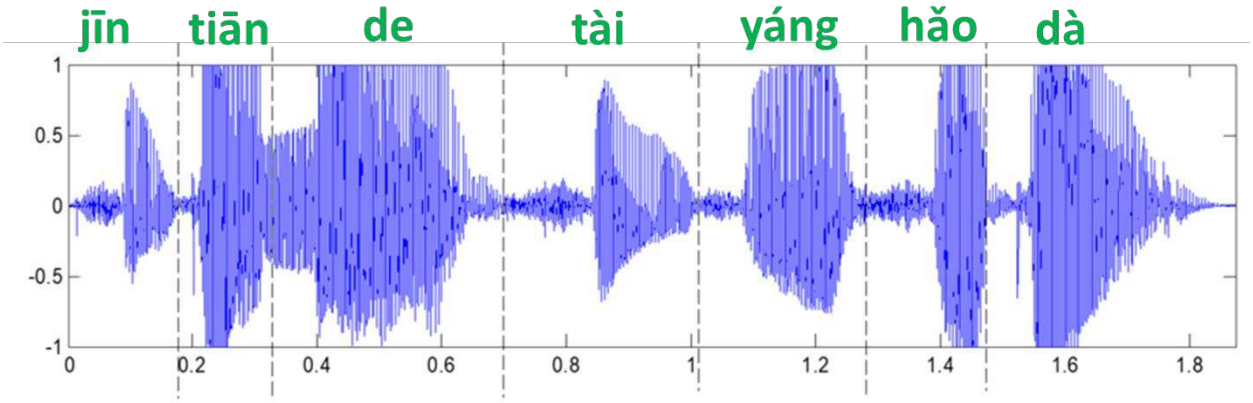


Figure 16: Consonant sound time-frequency diagram

4.3 Tuning and Optimization of Acoustic Models

The models are trained through continuous design and improvement of the relevant parameters of the acoustic model, and finally the model with excellent performance is selected according to the evaluation index. The baseline acoustic model in this study faces challenges such as long training time, high error rate, and a certain degree of overfitting. Common optimization strategies for neural networks include the dropout, normalization, and residual modules. Dropout was first proposed by Srivastava et al. in 2018, which can effectively solve the problem of overfitting. Normalization was first proposed by Segey Loffe and Christian Szegedy in 2020, which can speed up the model convergence and alleviate the overfitting problem to a certain extent. The residual module was proposed by Kaiming He et al. in 2022 [24], which solves the problem of gradient disappearance caused by the deepening of network layers.

The features of the neural network input generally follow the standard normal distribution, and generally perform well for shallow models. However, as the depth of the network increases, the nonlinear layer of the network will make the output results interdependent, and no longer meet a standard normal distribution. The problem of the output center offset will occur, which brings difficulties to the training of the network model. The training of deep models is particularly difficult. To solve the problem of model convergence, a normalization operation is added to the middle layer, i.e., a normalization process is performed on the output of each layer to make it conform to the standard normal distribution. Through this processing, the network input conforms to the standard normal distribution, which can be well trained, thus speeding up the convergence speed. The data dimension processed by the convolutional neural

network is a four-dimensional tensor, such that there are numerous normalization methods: layer normalization (LN), instance normalization (IN), group normalization (GN), batch normalization (BN), etc. [25].

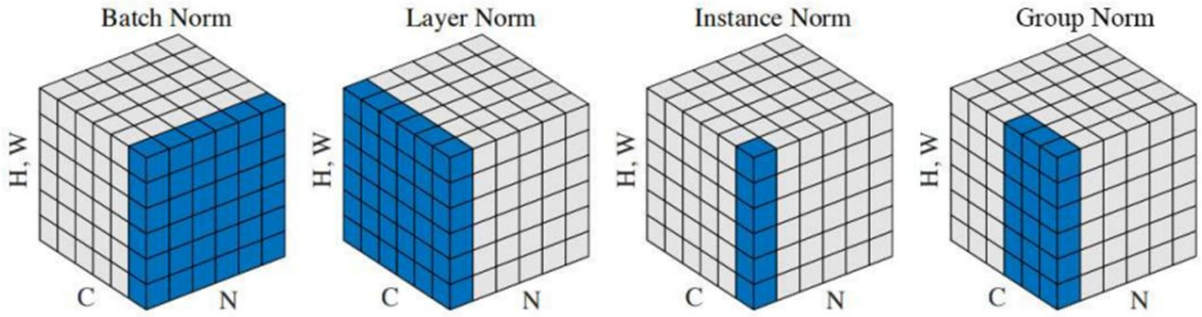


Figure 17: Comparison diagrams of applied normalization.

Fig. 17 illustrates schematic diagrams of the normalizations for comparison. Taking a piece of voice data as an example, as the voice frequency range is roughly 250–3400 Hz, and the high frequency is 2500–3400 Hz, four intrinsic mode functions (IMF) component frequency diagrams are decomposed by the normalized comparison method, as shown in Figs. 18(a),(b),(c), and (d). From the density of the normalized amplitude value of each IMF component, the high frequency region of speech is mainly concentrated in the first IMF component. Figs. 18(a),(b),(c) and (d) indicate that the high-frequency region of the speech signal can be effectively extracted by empirical mode decomposition (EMD) decomposition. However, for the feature parameter extraction method of the high-frequency region, the traditional extraction algorithm is not suitable, and one must seek the high-frequency feature parameter extraction algorithm [26].

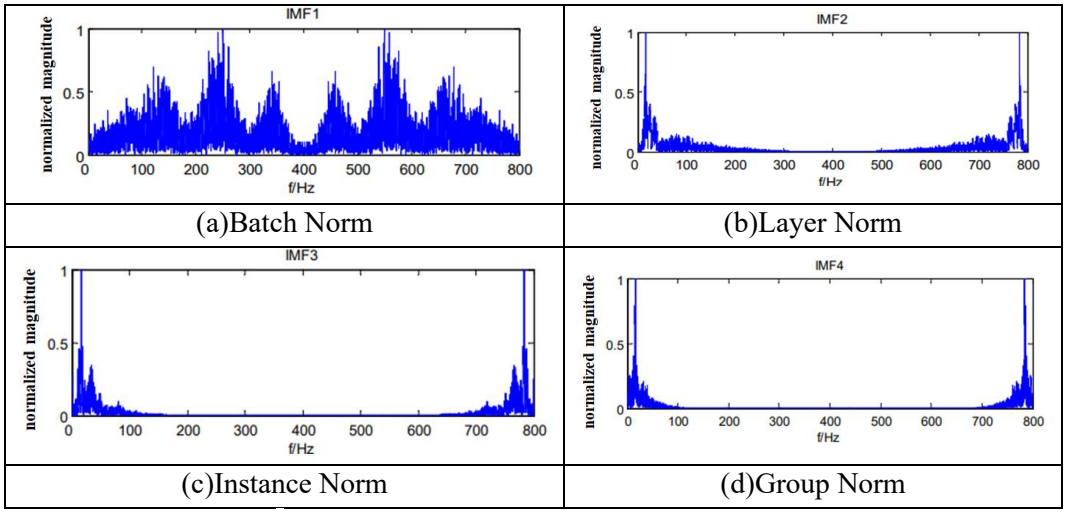


Figure 18: Comparison diagram of normalizations based on IMF.

Fig. 19(a) shows a schematic diagram of the Residual module, which transmits original input information to the output layer through a new channel opened on the network side. The Residual module directly transfers the input of the previous layer to a later layer by adding a congruent mapping layer. The principle of Dropout to suppress overfitting is to temporarily set some neurons to zero during network training, and ignore these neurons for parameter optimization, such that the network structure of each repeated operation training is different, so as to avoid network reliance on a single feature for classification and prediction. Dropout, a method of training multiple neural networks and then averaging the results of the entire set, instead of training a single neural network, increases the sparsity of the network model and improves its generalization. The schematic diagram of Dropout is shown in Fig. 19(b).

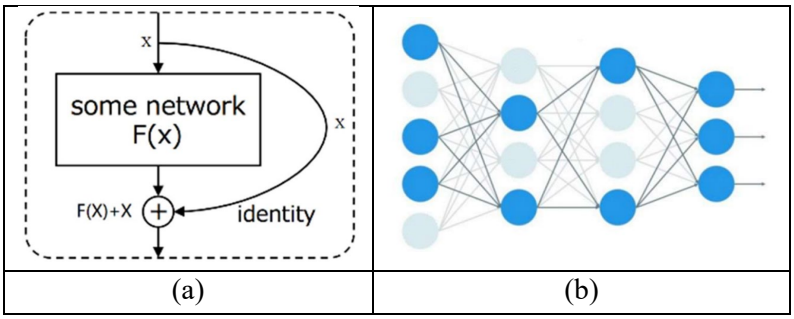


Figure 19: (a) Schematic diagram of Residual module; (b) Schematic diagram of Dropout.

Figs. 20(a) and **(b)** show the training comparison of the baseline acoustic model and improved acoustic model, respectively **[27]**.

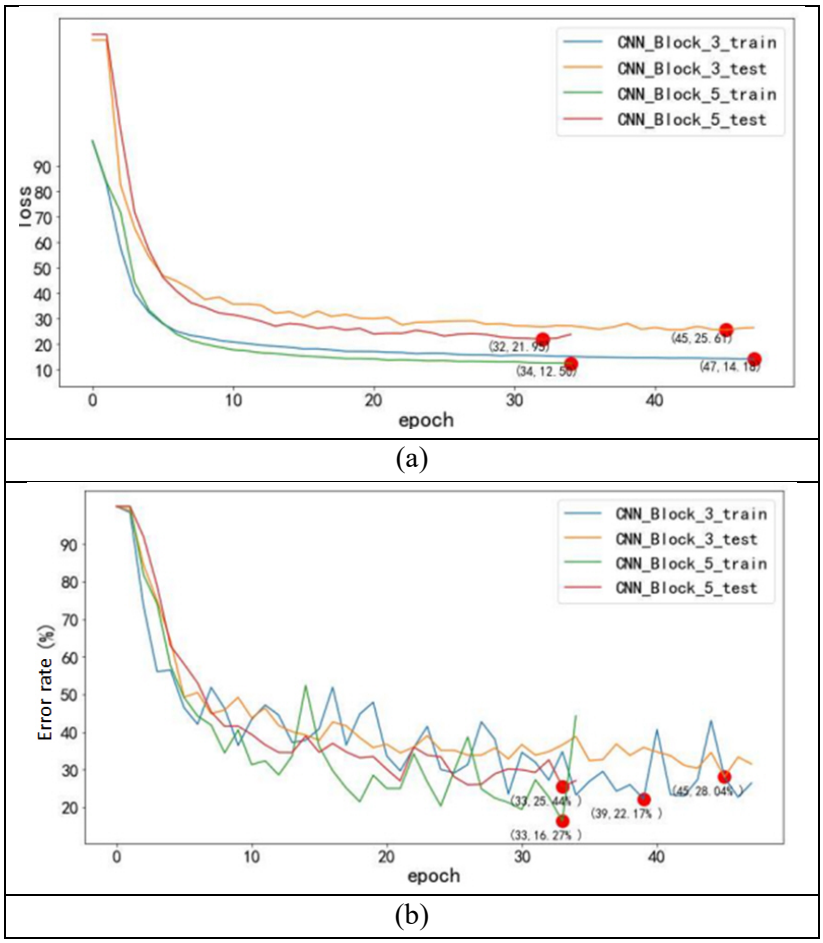


Figure 20: (a) Loss comparison of acoustic model; (b) Comparison of WER of acoustic model.

Fig. 20(b) shows that by increasing the depth of the network model, the improved acoustic model reduces the WER by 3.5 % on the test set compared to the baseline model. Although the error rate drops, the effect is still unsatisfactory. **Fig. 20(a)** shows that the improved acoustic model still faces the overfitting problem. Therefore, further optimization of this improved acoustic model is required. In the improved acoustic model, the number of network layers has reached 25. If the network layer continues to be deepened, the training time becomes too long, which likewise affects the decoding performance. To solve the overfitting problem, Dropout and BN layers are employed in the network model. The network model structure is shown in **Fig. 21 [28]**.

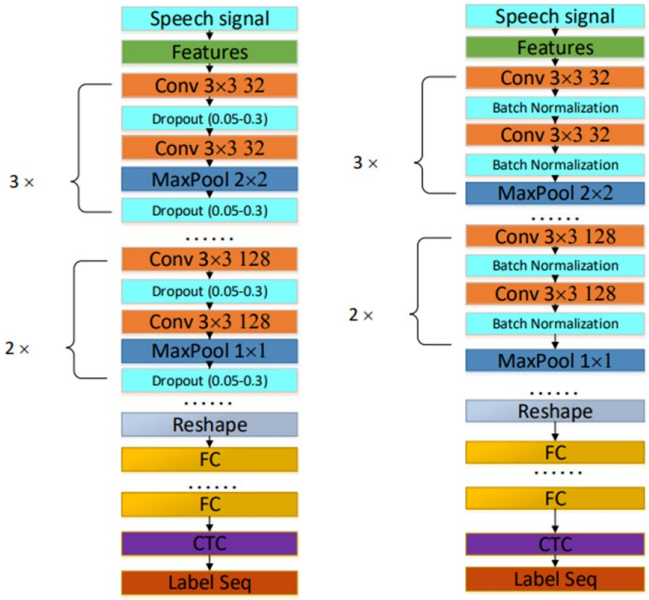


Figure 21: Comparison of structures between dropout and BN acoustic models

Figs. 22(a) and **(b)** show the training comparison diagrams of the Dropout and BN acoustic models.

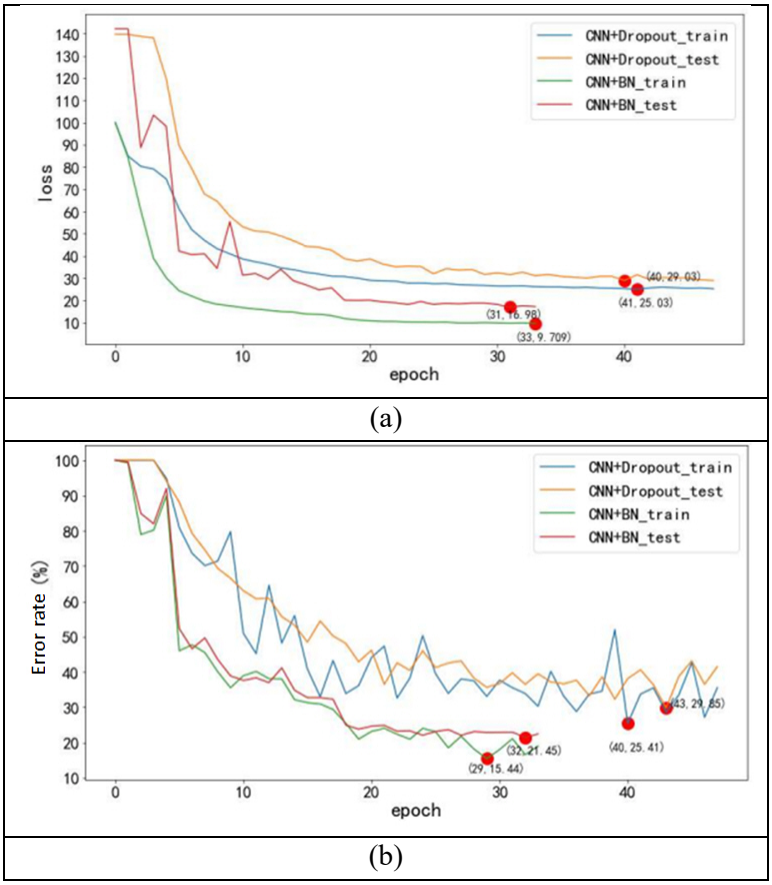


Figure 22: (a) Comparison of loss between dropout and BN acoustic models; (b) Comparison of WER between dropout and BN acoustic models.

Fig. 22(a) shows that both the Dropout and the BN acoustic models play a role in suppressing overfitting. However, as indicated in **Fig. 22(b)**, the error rate of the acoustic model using Dropout does

not drop but rises instead, revealing the opposite effect. The acoustic model using BN effectively reduces the error rate, and at the same time accelerates the convergence, such that the training speed of the model is accelerated. The error rate of the BN acoustic model drops to 23.67 %, indicating an 8 % improvement over the baseline acoustic model. Considering the gradient vanishing problem that may be imposed by the deep convolutional neural network, the residual module is added on the basis of the BN acoustic model, which is expected to further reduce the error rate. Fig. 23 shows the acoustic model with the added Residual module.

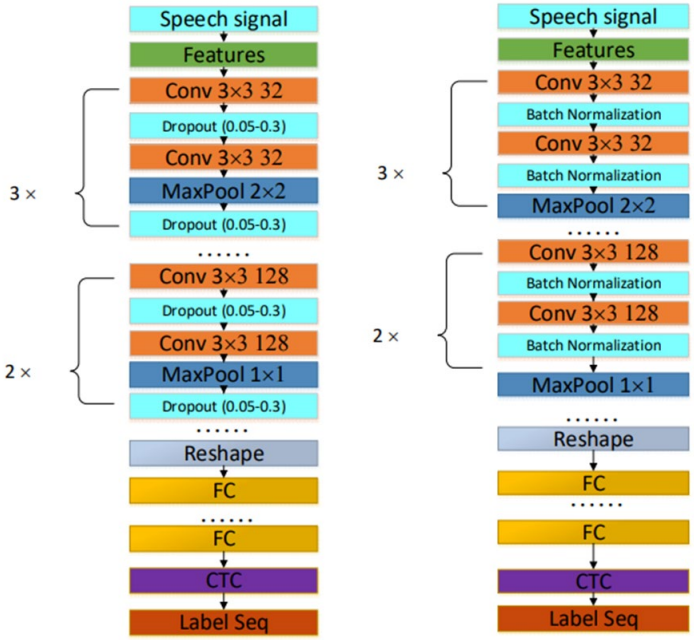
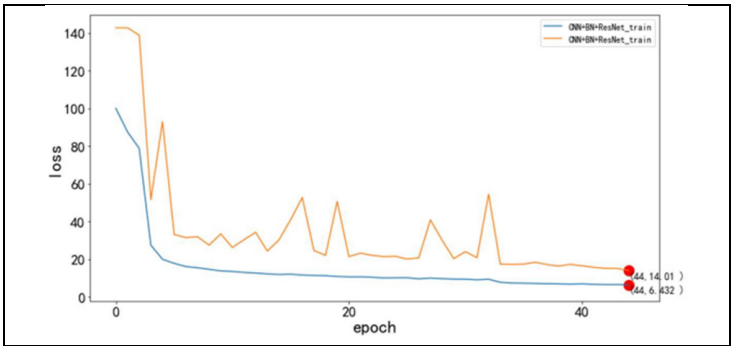
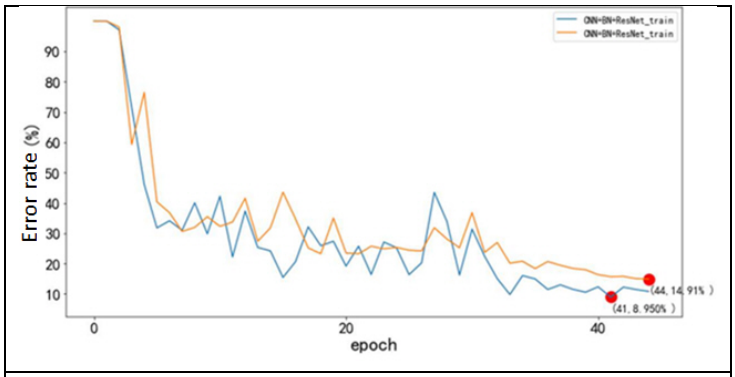


Fig. 23: Acoustic model of Residual plus BN

Fig. 24(a) shows that the Residual plus BN acoustic model has the fastest convergence speed among all models, i.e., the Residual module effectively alleviates the problem of gradient disappearance and speeds up the training speed of the model. As observed in Fig. 24(b), the error rate of the model on the test set is reduced to 12.45 %, which is 17 % higher than the initial baseline acoustic model. An error rate of 13.52 % is already an excellent result on the current scale of the dataset.



(a)



(b)

Figure 24: (a) Training loss diagram of Residual plus BN acoustic model; (b) WER change of Residual plus BN acoustic model.

5 Conclusions and Future Directions

In the speech recognition system, the acoustic model is an important underlying model, whose accuracy directly affects the performance of the entire system. This chapter introduces the construction and training process of the acoustic model in detail, and studies the CTC algorithm that plays an important role in the end-to-end framework. We constructed the CTC-CNN baseline acoustic model, and on this basis, carried out the optimization, reducing the error rate to about 18 %, hence improving the accuracy. Finally, the selection of acoustic feature parameters as well as the selection of modeling units, the speaker's speech speed, and other aspects were compared and verified, and the excellent performance of the CTCCNN_5 plus BN plus Residual model structure is further verified.

This study briefly introduces the historical development of deep learning and the most widely used deep learning models, and presents the development and current situation of these deep learning models in the field of speech recognition. Deep learning research is still in its developmental stage, and the main problems are: (1) training usually must solve a highly nonlinear optimization problem, which easily leads to many local minima in the process of training the network; (2) if the training takes too long, it will cause overfitting of the results. Thus, the use of deep neural networks to solve the robustness problem is currently the hottest topic in the field of speech recognition. In practical applications, the recognition rate of noisy speech is only about 85 %, such that there is no stable, efficient, and universal system that can achieve a recognition rate of more than 95 % for noisy speech. For future research on speech recognition, we believe that the best direction of development is brain-like computing. Only by continuously conforming to the characteristics of speech recognition of the human brain, can the recognition rate of speech be improved to a satisfactory level. However, the existing deep learning technology is far from sufficient to meet this requirement. How to better apply deep learning and meet the market demand for efficient speech recognition systems is a problem worthy of continued focus.

Acknowledgement: This research was supported by the Department of Electrical Engineering at National Chin-Yi University of Technology. The authors would like to thank the National Chin-Yi University of Technology, Takming University of Science and Technology, Taiwan, for supporting this research.

Availability of data and materials

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Competing Interest

The authors declare that they have no conflicts of interest to report regarding the present study.

Funding Statement

The authors received no specific funding for this study.

Authors Contribution

W-T S. is responsible for research planning and providing improvement methods. H-W K. and S-J H. is responsible for thesis writing and experimental verification.

Abbreviations

DNN: Deep neural network
CTC: Connectionist temporal classification
CTC-CNN: Connectionist temporal classification-convolutional neural network
ALS: Amyotrophic lateral sclerosis
HMM: Hidden Markov model
GMM-HMM: Gaussian mixture model-hidden Markov model
ASR: Automatic speech recognition
LN: Layer normalization
IN: Instance normalization
GN: Group normalization
BN: Batch normalization
IMF: Intrinsic mode functions
EMD: Empirical mode decomposition
WER: Word error rate
NLP: Natural language processing
SD: Speaker discrimination
LM: Language model
AEC: Acoustic echo cancellation
UAC 1.0: USB audio class 1.0
UNet: U-shaped network

References

[1] K. Khysru, J. Wei and J. Dang," Research on Tibetan speech recognition based on the Am-do Dialect," *CMC-Computers, Materials & Continua*, vol.73, no.3, pp. 4897-4907, 2022

[2] J. Tang, X. Chen and W. Liu, " Efficient language identification for all-language internet news," *in Proc 2021 International Conference on Asian Language Processing (IALP)*, Singapore, Singapore, pp. 165-169, 2021.

[3] Z. Wang, Y. Zhao, L. Wu, X. Bi, Z. Dawa *et al.*, " Cross-language transfer learning-based Lhasa-Tibetan speech recognition," *CMC-Computers, Materials & Continua*, vol.73, no.1, pp. 629-639, 2022.

[4] S. Yasin, U. Draz, T. Ali, K. Shahid, A. Abid *et al*.," Automated speech recognition system to detect babies' feelings through feature analysis," *CMC-Computers, Materials & Continua*, vol.73, no.3, pp. 4349-4367, 2022

[5] K. Jambi, H. Al-Barhamtoshy, W. Al-Jedaibi, M. Rashwan and S. Abdou," Speak-correct: A computerized interface for the analysis of mispronounced errors," *Computer Systems Science and Engineering*, vol.43, no.3, pp. 1155-1173, 2022.

[6] A. Das, J. Li, G. Ye, R. Zhao and Y. Gong, "Advancing acoustic-to-word CTC model with attention and mixed-units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. vol.27, no.12, pp. 1880 –1892, 2019.

[7] R. P. Bachate, A. Sharma, A. Singh, A. A. Aly, A. H. Alghtani *et al.*," Enhanced marathi speech recognition facilitated by grasshopper optimisation-based recurrent neural network," *Computer Systems Science and Engineering*, vol.43, no.2, pp. 439-454, 2022.

[8] S. Lu, J. Lu, J. Lin and Z. Wang, "A hardware-oriented and memory-efficient method for CTC decoding," *IEEE Access*. vol.7, pp. 120681 – 120694, 2019.

[9] M. H. Changrampadi, A. Shahina, M. Badri. Narayanan and A. N. Khan," End-to-end speech recognition of Tamil language," *Intelligent Automation & Soft Computing*, vol.32, no.2, pp. 1309-1323, 2022.

[10] Z. Zhao and P. Bell, " Investigating sequence-level normalisation for CTC-Like End-to-End ASR," *In Proc 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, pp. 7792-7796, 2022.

[11] A. Nakamura, K. Ohta, T. Saito, H. Mineno, D. Ikeda *et al*., " Automatic detection of chewing and swallowing using hybrid CTC/Attention," *In Proc 2020 IEEE 9th Global Conference on Consumer Electronics (GCCE)*, Kobe, Japan, pp.810-812, 2020.

[12] H. Wu and A. K. Sangaiah," Oral English speech recognition based on enhanced temporal convolutional network," *Intelligent Automation & Soft Computing*, vol.28, no.1, pp. 121-132, 2021.

[13] E. Yavuz and V. Topuz," A phoneme-based approach for eliminating out-of-vocabulary problem Turkish speech recognition using hidden markov model," *Computer Systems Science and Engineering*, vol.33, no.6, pp. 429-445, 2018.

[14] T. Moriya, H. Sato, T. Tanaka, T. Ashihara, R. Masumura *et al*., "Distilling attention weights for CTC-Based ASR systems," *In Proc 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, pp. 6894-6898, 2020.

[15] L. Ren, J. Fei, W. K. Zhang, Z. G. Fang, Z.Y. Hu *et al*., "A microfluidic chip for CTC whole genome sequencing," *In Proc 2019 IEEE 32nd International Conference on Micro Electro Mechanical Systems (MEMS)*, Seoul, Korea (South), pp. 412-415, 2019.

[16] L. H. Juang and Y. H. Zhao, "Intelligent speech communication using double humanoid robots," *Intelligent Automation & Soft Computing*, vol.26, no.2, pp. 291-301, 2020.

[17] T. A. M. Celin, G. A. Rachel, T. Nagarajan and P. Vijayalakshmi, " A weighted speaker-specific confusion transducer-based augmentative and alternative speech communication aid for dysarthric speakers," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*,vol.27,no.2,pp. 187 - 197, 2019.

[18] Y. Takashima,R. Takashima,T. Takiguchi and Y. Ariki," Knowledge transferability between the speech data of persons with dysarthria speaking different languages for Dysarthric speech recognition," *IEEE Access*. vol.7, pp. 164320 - 164326, 2019.

[19] Y. Yang, Y. Wang, C. Zhu, M. Zhu, H. Sun *et al*., " Mixed-scale Unet based on dense atrous pyramid for monocular depth estimation," *IEEE Access*, vol. 9, pp. 114070 -114084, 2021.

[20] N. Y.H. Wang, H.L. S. Wang, T.W. Wang,S.W. Fu. X. Lu *et al*.," Improving the intelligibility of speech for simulated electric and acoustic stimulation using fully convolutional neural networks," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*,vol.29,pp. 184 – 195,2021.

[21] R. E. Jurdi, C. Petitjean, P. Honeine and F. Abdallah, "BB-UNet: U-Net with bounding box prior," *IEEE Journal of Selected Topics in Signal Processing*. vol.14, no.6, pp. 1189–1198, 2020.

[22] R. Haeb-Umbach, J. Heymann, L. Drude, S. Watanabe, M. Delcroix *et al.*," Far-Field automatic speech recognition," *Proceedings of the IEEE*,vol.109,no.2,pp. 124 – 148,2021.

[23] S. Latif, J. Qadir, A. Qayyum, M. Usama and S. Younis," Speech technology for healthcare: Opportunities, challenges, and state of the art," *IEEE Reviews in Biomedical Engineering*,vol.14,pp. 342 – 356,2021.

[24] H. Zhou, J. Du, Y. Zhang, Q. Wang, Q. F. Liu *et al*., "Information fusion in attention networks using adaptive and multi-level factorized bilinear pooling for audio-visual emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,*vol. 29, pp. 2617– 2629, 2021.

[25] Y. Cai, L. Li, A. Abel, X. Zhu, D. Wang *et al*., "Deep normalization for speaker vectors," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 733 –744, 2020.

[26] G. Kim, H. Lee, B.K. Kim, S.H. Oh and S.Y. Lee, "Unpaired Speech enhancement by acoustic and adversarial supervision for speech recognition," *IEEE Signal Processing Letters*, vol.26, no.1,pp 159 – 163,2019.

[27] Y. Lin, D. Guo, J. Zhang, Z. Chen and B. Yang," A unified framework for multilingual speech recognition in air traffic control systems," *IEEE Transactions on Neural Networks and Learning Systems*,vol.32, no.8,pp. 3608 – 3620,2021.

[28] T. Kawase, M. Okamoto, T. Fukutomi and Y. Takahashi, "Speech enhancement parameter adjustment to maximize accuracy of automatic speech recognition," *IEEE Transactions on Consumer Electronics*, vol.66, no.2, pp. 125 – 133, 2020.

Figure 1: Historical evolution of automatic speech recognition

We use state-of-the-art techniques to verify our method. Experimental results indicate an outstanding effect. The figure illustrates the historical evolution of automatic speech recognition (ASR).

Figure 2: Two homonymous English strings

The figure illustrates two homonymous English strings. In the language model, in addition to the pronunciation that affects the accuracy of speech recognition, punctuation is likewise an important reason that affects the recognition of the system.

Figure 3: Schematic diagram of CTC-CNN acoustic model

Finally, the optimized output label sequence of the CTC decoding algorithm marks the recognition result. The schematic diagram of the CTC-CNN acoustic model is shown in the figure.

Figure 4: Mapping output to label sequence

The figure shows that the probability of label sequence $label\_7$ is equal to the total probability of its entire path, that is, $P\ (label\_7) = P\ (path\_1) + P\ (path\_2) + P\ (path\_3) + P\ (path\_4)$. It is impractical to calculate $(l|x)$ directly violently, which will increase the training time of the model and occupy computing power.

Figure 5: Derivation of forward and backward algorithm

Borrowing the forward and backward algorithm in HMM effectively solves $(l|x)$, and it is assumed that under the condition of a given input sequence x, the output label probability at time t is independent,

such the transition probability between states does not need to be considered. The derivation diagram of the forward and backward algorithm is shown in the figure.

Figure 6: Structure of CTC-CNN acoustic model

CTC illustrates the optimization of the loss function of the neural network and the optimization of the output sequence. Therefore, this study proposes a CTC-CNN acoustic model based on CNN combined with the CTC algorithm. The overall structure of the CTC-CNN acoustic model is shown in the figure.

Figure 7: Hardware architecture diagram

The figure portrays the architectural diagram of the hardware employed this experiment. It comprises the ReSpeaker Mic Array v2.0.1 and display screen.

Figure 8: System architecture diagram

The figure shows the overall structure and flow chart of the speech recognition assistance system for language-impaired individuals. ReSpeaker Mic Array v2.0.1 records the speech signals of individuals with language impairments.

Figure 9: (a) ReSpeaker Mic Array v2.0.1(b) ReSpeaker Mic Array v2.0.1 System Diagram

These figures show the ReSpeaker Mic Array v2.0.1 and the module system diagram, respectively. The module also features the WM8960, a low-power stereo codec with Class D speaker drivers capable of delivering 1 W per channel at an 8 W load. It further supports USB Audio Class 1.0 (UAC 1.0) and has twelve programmable RGB LED indicators for user freedom.

Figure 10: End-to-end model framework UNet

The end-to-end model framework UNet comprises the main structure of the framework, as shown in the figure. The UNet neural network was initially applied for medical image processing and achieved good results.

Figure 11: End-to-end speech enhancement framework

The structure of the model proposed in this study is shown in the figure. The architecture consists of three parts, namely, preprocessing of the original audio signal, encoding, and decoding module based on the UNet architecture, and post-processing of enhanced speech synthesis.

Figure 12: Experimental results

Experimental results are presented in the figure(a)and figure(b) as screenshots of the web GUI interface. The figure(a) shows the speaker saying, "The weather is so nice today", and the system successfully displays the speaker's complete sentence. The figure(b) shows the speaker saying "Good morning" twice in a row, but the recognition result is only successful one time; the result of the second time presents the situation of homophones.

Figure 13: (a) Loss variation of baseline acoustic model (b) Word error rate change of baseline acoustic model

To train and verify the CTC-CNN baseline acoustic model, THCHS-30 and ST-CMDS speech datasets were used as training data sets, and the data sets are divided into training and test sets. The training results are shown in the figure.

Figure 14: Prediction trend chart of various word count recognition accuracy rates

In addition to the above-mentioned situations that affect the accuracy of identification, the environmental noise factor may also lead to a decrease in the accuracy of identification. The figure shows prediction trend chart of various word count recognition accuracy rates.

Figure 15: Prediction trend of environmental noise impact identification accuracy

In contrast, at 40–60 dB, owing to less noise pollution, the accuracy rate is as high as 97.33 %. The figure shows the prediction trend of environmental noise impact identification accuracy.

Figure 16: Consonant sound time-frequency diagram

The experimental results in the figure indicate that the amplitudes of consonants are small, while the amplitudes of vowels are relatively large.

Figure 17: Comparison diagrams of applied normalization.

The figure illustrates schematic diagrams of the normalizations for comparison. Taking a piece of voice data as an example, as the voice frequency range is roughly 250–3400 Hz, and the high frequency is 2500–3400 Hz.

Figure 18: Comparison diagram of normalizations based on IMF

These figures indicate that the high-frequency region of the speech signal can be effectively extracted by empirical mode decomposition (EMD) decomposition.

Figure 19: (a) Schematic diagram of Residual module; (b) Schematic diagram of Dropout.

The figure shows a schematic diagram of the Residual module, which transmits original input information to the output layer through a new channel opened on the network side

Figure 20: (a) Loss comparison of acoustic model; (b) Comparison of WER of acoustic model.

The figure shows the training comparison of the baseline acoustic model and improved acoustic model, respectively

Figure 21: Comparison of structures between dropout and BN acoustic models

If the network layer continues to be deepened, the training time becomes too long, which likewise affects the decoding performance. To solve the overfitting problem, Dropout and BN layers are employed in the network model. The network model structure is shown in the figure.

Figure 22: (a) Comparison of loss between dropout and BN acoustic models; (b) Comparison of WER between dropout and BN acoustic models.

The figure(a) shows that both the Dropout and the BN acoustic models play a role in suppressing overfitting. However, as indicated in the figure(b), the error rate of the acoustic model using Dropout does not drop but rises instead, revealing the opposite effect.

Fig. 23: Acoustic model of Residual plus BN

Considering the gradient vanishing problem that may be imposed by the deep convolutional neural network, the residual module is added based on the BN acoustic model, which is expected to further reduce the error rate. The figure shows the acoustic model with the added Residual module.

Figure 24: (a) Training loss diagram of Residual plus BN acoustic model; (b) WER change of Residual plus BN acoustic model.

The figure shows that the Residual plus BN acoustic model has the fastest convergence speed among all models, i.e., the Residual module effectively alleviates the problem of gradient disappearance and speeds up the training speed of the model.