Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.037380
Article

Improving Speech Enhancement Framework via Deep Learning

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

[1]Department of Information Technology, Takming University of Science and Technology, Taipei City, 11451, Taiwan
[2]Department of Electrical Engineering, National Chin-Yi University of Technology, Taichung, 411030, Taiwan
*Corresponding Author: Wen-Tsai Sung. Email: songchen@ncut.edu.tw
Received: 01 November 2022; Accepted: 06 January 2023

Abstract: Speech plays an extremely important role in social activities. Many individuals suffer from a "speech barrier," which limits their communication with others. In this study, an improved speech recognition method is proposed that addresses the needs of speech-impaired and deaf individuals. A basic improved connectionist temporal classification convolutional neural network (CTC-CNN) architecture acoustic model was constructed by combining a speech database with a deep neural network. Acoustic sensors were used to convert the collected voice signals into text or corresponding voice signals to improve communication. The method can be extended to modern artificial intelligence techniques, with multiple applications such as meeting minutes, medical reports, and verbatim records for cars, sales, etc. For experiments, a modified CTC-CNN was used to train an acoustic model, which showed better performance than the earlier common algorithms. Thus a CTC-CNN baseline acoustic model was constructed and optimized, which reduced the error rate to about 18% and improved the accuracy rate.

Keywords: Artificial intelligence; speech recognition; speech to text; CTC-CNN

1 Introduction

In terms of application concept, the purpose of speech recognition is to train an artificial intelligence (AI) model to detect sounds and send a text message or semantic understanding and convert it into voice commands to control smart devices, making human life more convenient with the aid of smart technology. Based on the industry's demand for AI speech recognition, the speech recognition cloud service can provide the industry with a customized speech recognition model to achieve intelligent control or operation mode to assist related industries in innovative applications of human-computer interaction. This study discussed smart homes, voice assistants, the internet of vehicles, and mobile vehicles that operate in the form of cloud services and accept users' requests at any time through the application programming interface (API).

Many start-up information technology (IT) companies and large enterprises are optimistic about the application of speech recognition and have invested in the field of AI speech recognition and

continue to improve the accuracy of its recognition. The current AI speech recognition function has been introduced to the market in the period when the states are at war; the technology can be used in the industry, smart speakers and other smart home appliances, first aid, insurance market, customer service market, as well as meeting record production, paperwork operations, and corporate interviews to assist talent selection, etc. Different fields meet the technical level required by the individual consumer market and the industry. Concurrently, speech recognition also penetrates and extends to the in-vehicle field to seize the business opportunities of in-vehicle voice applications and helps people with language barriers to be understood better, thus becoming an accurate microphone for people with language barriers.

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, this study proposes improving the connectionist temporal classification convolutional neural network (CTC-CNN) architecture to help these people communicate normally. Deep learning has been predominantly used in visual recognition, speech recognition, natural language processing, biomedicine, and other fields, where it has achieved excellent results. However, 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 employed 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. This paper used state-of-the-art techniques to verify this study method. Experimental results indicate an outstanding effect. Fig. 1 illustrates the historical evolution of automatic speech recognition [2].

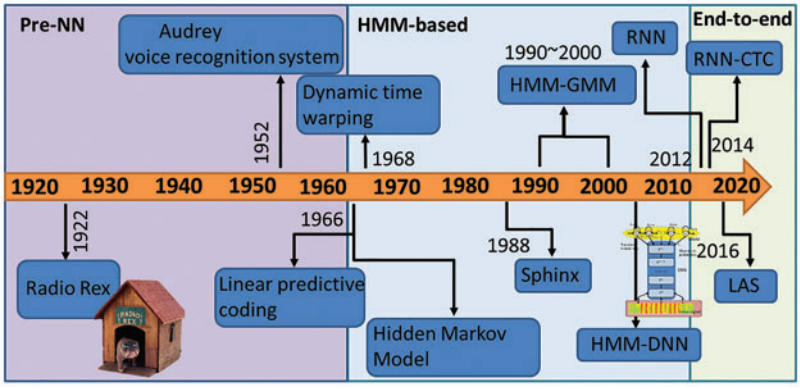


Figure 1: Historical evolution of automatic speech recognition

The sound of Parrotron launched by Google is very close to the original meaning of the input voice. AI speech recognition can help people with language barriers to understand better, and it can be called an accurate "microphone" for such people. This research combined AI speech recognition technology with the improved method of deep learning to solve the current difficulties in speech recognition, such as regional accents or recognition of speech with accents. The accuracy of the recognition is still a big challenge to be solved. At the same time, speech is not just about expressing language. If the emotion of the speaker and other parameters are included in the recognition mechanism of speech emotion during the recognition process, and the automatic response service is moderately added with empathetic intonation and words, it will be more helpful in improving customer service in perceptual appeal. In recent years, although many technology companies have used the AI speech recognition technology

of deep learning multi-layer artificial neural networks to improve their recognition accuracy, a good neural network must rely on a large amount of data. To overcome the complex and harsh acoustic environment, AI speech recognition Technology is bound to be combined with big data, the Internet of Things, and other technologies to break through the development of AI acoustics and achieve the purpose of improving the quality of human life [3].

2 Literature Survey

Speech recognition technology has been applied to many fields, mainly intelligent computers, communications and consumer electronics (3C), and driving navigation products. This research is necessary to understand basic language theory to help people with speech disabilities. The results of this study will help people who suffer from language barriers and communication difficulties due to diseases or various disabilities and may bring more convenience to their lives. In the experimental framework, the study model takes into account linguistics, speech recognition applications, and deep learning techniques [4].

CTC-CNN used in this study is a sequence prediction method combined with deep learning. Its dynamic programming concept is similar to Hidden Markov Models (HMM). In the field of speech recognition, CTC-CNN has lower complexity and better performance than the performance of HMM. Previous studies have found that the performance of CTC is better than that of HMM; the learning process of CTC and the distribution of errors were not explored. In this study, the sequence prediction results of the basic Deep Neural Network (DNN) were combined with CTC-CNN and observed to analyze the phoneme accuracy rate of the whole domain and a specific position and the relationship between the phoneme substitution error, and the number of training samples was established. The improved CTC method was then integrated into the learning process of speech recognition [5].

2.1 Convolutional Neural Networks Applied in This Study

2.1.1 The Connectionist Temporal Classification-Convolutional Neural Network Acoustic Model

With the development of deep learning, speech recognition systems began using deep learning-based acoustic models and achieved remarkable results. In the speech recognition system, the acoustic model is an important underlying model, and its accuracy directly affects the performance of the entire system. This study mainly used CNN to construct the acoustic model and combined it with the connection sequence classification algorithm, which significantly improved the accuracy and performance of the speech recognition system. The latest end-to-end speech recognition framework abandons the more restrictive HMM model and directly optimizes the input and output sequences likelihood, which significantly simplifies the training process [6]. As the output of the acoustic model, the choice of modeling unit is also one of the factors affecting the performance of the acoustic model. When choosing a modeling unit, it is necessary to consider whether the modeling unit fully represents the context information and whether it can describe the generalization of acoustic features. Based on the establishment of the baseline acoustic model, the error rate of the speech-to-pinyin sequence was significantly reduced in this study by continuously optimizing the acoustic model [7].

When the CTC-based acoustic model recognizes speech, the acoustic feature parameters are further extracted through the CNN, and then the posterior probability matrix is output through the fully connected network and the SoftMax layer. This study adopted a non-complete end-to-end speech recognition framework when constructing the speech recognition system; that is, the acoustic model uses the end-to-end recognition framework to convert speech into pinyin sequences and then uses a language model that converts pinyin sequences into text. The schematic diagram of the CTC-CNN

acoustic model is shown in Fig. 2 [8]. The CNN was used to construct the acoustic model and the CTC-CNN algorithm was used to realize the conversion from pinyin sequence to pinyin sequence. The maximum probability label for each node was thus used as the output sequence. Finally, the optimization of the CTC-CNN decoding algorithm provides label sequence marker recognition results as outputs.

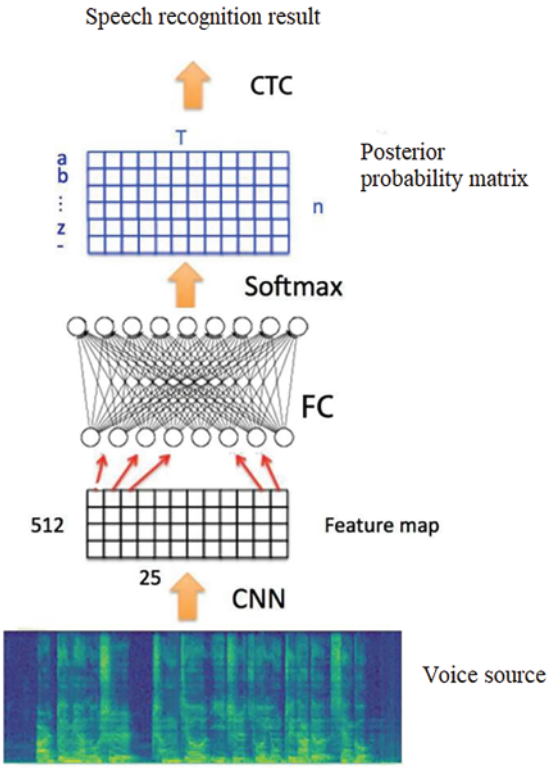


Figure 2: Schematic diagram of the connectionist temporal classification-convolutional neural network acoustic model

2.1.2 The Connectionist Temporal Classification-Convolutional Neural Network Technical Architecture

The technical architecture of CTC mainly includes the following:

(1) Use the forward algorithm idea of HMM and effectively calculate the CTC loss function and its derivatives according to the dynamic programming algorithm to solve the CNN end-to-end training problem [9].
(2) Extend the output layer of CNN, add a many-to-one spatial mapping between the output sequence and the recognition result (label sequence) and accordingly define the CTC loss function.
(3) Integrates the CTC decoding algorithm to effectively realize the end-to-end prediction of sequence data [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 the speech input, the output sequence with the highest probability is selected, and after CTC decoding

optimization, the final recognition result (x) is provided as 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, first, S is assumed to be the training data set, X is the input space, Z is the target space (the set of labeled sequences), and L is defined as the sum of all output labels (modeling units). CTC is set to extend L to $L' = \text{L} \cup \text{Blank}$. Under given conditions, the probability of the label k output 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).

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

This study defined the mapping relationship B: $L'^T \to L^{\leq T}$ from the path π to the label sequence l. This mapping relationship will keep only one consecutive and identical label in the output sequence in the path π and remove the Blank label [11]. Then, to calculate the probability of label sequence $l \in L^{\leq T}$, all path probabilities belonging to l were accumulate using the formula in Eq. (4).

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

For any moment t, the posterior probability of the label sequence was calculated using the forward and backward probabilities, and the calculation formula is shown in Eq. (5).

$$P(l|x) = \sum_{s=1}^{|l'|} \frac{\alpha(t, s)\beta(t, s)}{\text{y}_{l'_s}^t} \tag{5}$$

With the posterior probability formula $(l|x)$ of the label sequence, the training target could be optimized, and the parameters were 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. (6).

$$-\frac{\partial \ln P(l|x)}{\partial u_k^t} = y_k^t - \frac{1}{P(l|x)\, y_k^t} \sum_{s \in lab(l,k)} \alpha(t, s)\beta(t, s) \tag{6}$$

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 [12]. This study adopted 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 each frame is accepted by the algorithm [13]. The recognition result of the final acoustic model is given by Eq. (6). In essence, the acoustic model of CTC can be directly output end-to-end as Chinese characters. Due to the limitation of the training corpus and the complexity of the model, Pinyin was the output of the acoustic model; the final result of speech recognition was obtained by entering the pinyin sequence as input into the language model [14].

2.1.3 Construction and Training of Baseline Acoustic Model

The structure of convolutional and pooling layers in this study shows that input features with slight deformation and displacement could be accurately recognized. The training mode of CNN parallel computing effectively shortens the training time and utilizes the powerful parallel processing capability of the graphics processing unit (GPU). CTC-CNN illustrates the optimization of the neural network function loss and the optimization of the output sequence. Therefore, this study proposed a CTC-CNN acoustic model based on CNN combined with an improved CTC algorithm [15].

3 System Architecture

3.1 System Design

ReSpeaker Mic Array v2.0.1 was used to record voice data and compares the recorded voice signal with the voice database. Fig. 3 presents a diagram of the hardware architecture used in this experiment. It includes ReSpeaker Mic Array v2.0.1 and display. Algorithms are processed, and the results of calculations are displayed to show words, sentences, and phrases after speech have been converted to text [16].

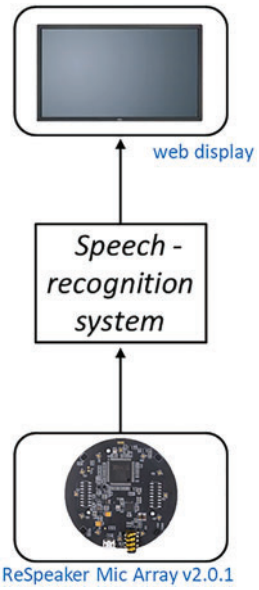


Figure 3: Hardware architecture diagram

ReSpeaker Mic Array v2.0.1 records speech signals of speech-impaired persons and extracts the original recorded speech recording files through the Python algorithm. The overall structure and flow chart of the speech recognition assistance system for language-impaired persons is presented in Fig. 4. These features are repeatedly compared and decoded in acoustic comparison and language decoding until the computed result is very similar to or exactly what the speaker intended, i.e., produces the expected output. The final result is displayed in text form on the machine [17].

Algorithms extract speech features from the extracted raw data and produce extracted features. These vectors are calculated algorithmically by the speech recognition system, which includes acoustic comparison and language decoding. Language models are generated in the same way as acoustic models and are trained and tuned through text corpora, building common words or sentences, and are

even multilingual. The acoustic model uses the language corpus to train and adjust the acoustic model to achieve cross-comparison with the speaker's pronunciation, words, and expressions to improve the recognition accuracy [18].

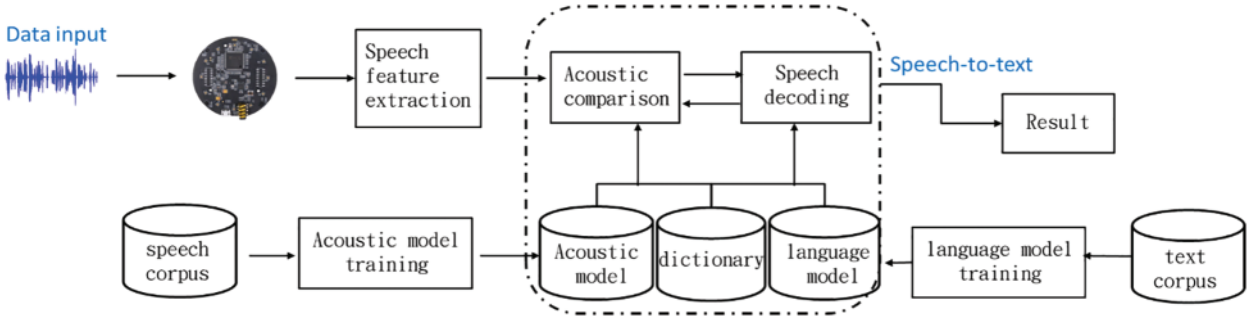


Figure 4: System architecture diagram

3.2 ReSpeaker Mic Array v2.0.1

ReSpeaker Mic Array v2.0.1 was the radio hardware component used in this experiment. This new chipset includes several speech recognition algorithms for improved performance. This array can be stacked (connected) on top of the original ReSpeaker Core to significantly improve voice interaction performance. This system can be used in a variety of occasions, such as smart speakers, smart voice assistant systems, voice conference systems, car voice assistants, etc. The module is equipped with XMOS' XVF-3000 IC, which integrates advanced Digital Signal Processor (DSP) algorithms, including Acoustic Echo Cancellation (AEC), beamforming, demixing, noise suppression, and gain control.

The ReSpeaker Mic Array v2.0.1 module has many voice algorithms and functions. This module contains four high-performance digital microphones (MP34DT01-M), built-in capacitive sensing elements, and an I2C interface, supports far-field voice capture, and far-field voice capture recording, and understands needs within a range of up to 5 m. Figs. 5a and 5b present the diagrams of the ReSpeaker Mic Array v2.0.1 and the module system [19], respectively. The module also improves recording quality, reduces ambient voice echo, and employs AEC to eliminate current audio output. The module features the WM8960, a low-power stereo codec with a Class-D speaker driver capable of delivering 1 W per channel at an 8 W load.

3.3 System Technology Description

An improved end-to-end speech enhancement architecture was proposed herein. The end-to-end model framework UNet [20] contains the main structure of the framework. UNet neural network was initially applied to process medical images, and good results were achieved. By combining the evaluation index with the loss function, the commonality and difference between different evaluation indexes were used to improve the perception ability of the model and obtain clearer speech.

The structure of the model proposed is shown in Fig. 6. The architecture consists of three parts, namely, the preprocessing of the original audio signal, the codec module based on the UNet architecture, and the post-processing of enhanced speech synthesis. The defects and problems in time-frequency transformation are avoided by directly modeling the time-domain speech signal. The problem is to transform the one-dimensional signal into a two-dimensional signal through convolution

operation so that the neural network can mine the speech signal in high-dimensional space and deep representation.

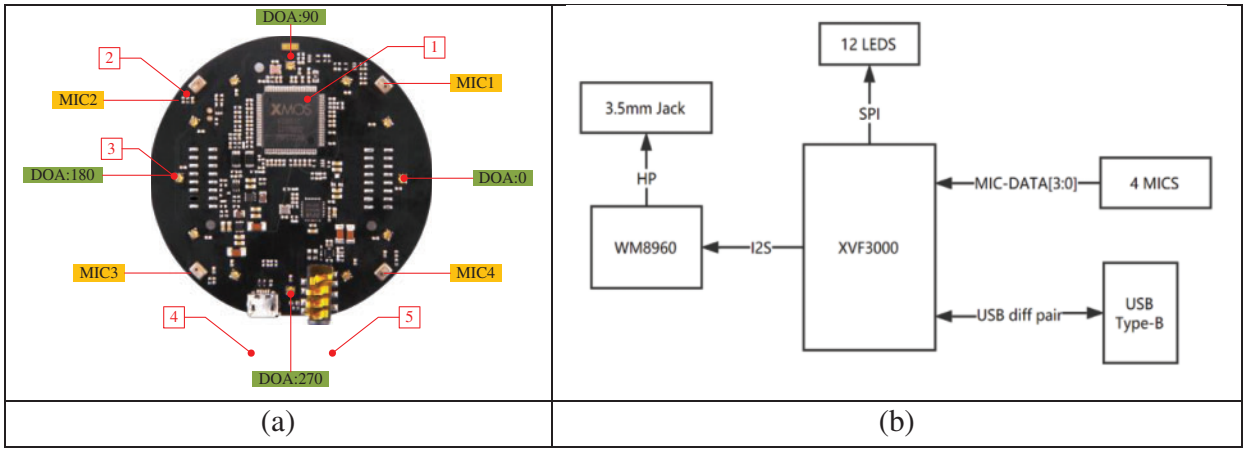


Figure 5: (a) ReSpeaker mic array v2.0.1 (b) ReSpeaker mic array v2.0.1 system diagram

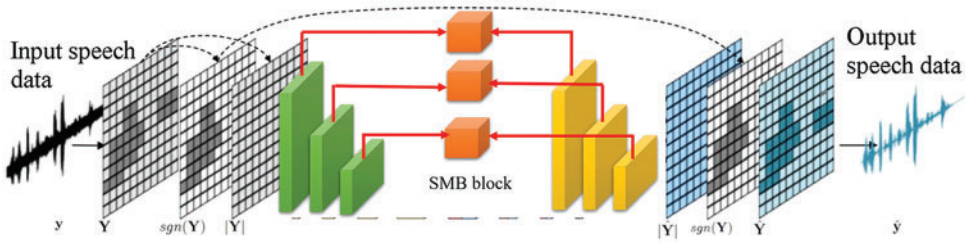


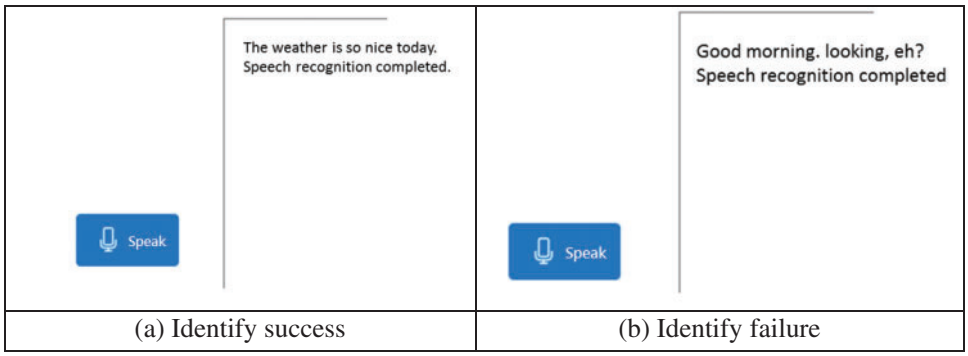
Figure 6: End-to-end speech enhancement framework

The main structure of UNet consists of an encoding stage (the left half of UNet) and a decoding stage (the right half of UNet). Between each corresponding encoding stage and decoding stage, skip connections are used. Here, the skip connections are not residuals but splicing methods. To reduce the number of parameters and the complexity of the model, the up-sampling operation of the decoding part of UNet here does not involve deconvolution, but a bilinear interpolation method [21], which is the improvement of this research.

4 Analyses of Experimental Results

4.1 Basics Experimental Results

First, the system records on the ReSpeaker Mic Array v2.0.1. Subsequently, the algorithm quickly performs speech recognition and displays the speaker's incomplete or broken sentences on the vehicle to help people with speech disabilities to communicate smoothly and quickly [22]. Figs. 7a and 7b are screenshots of the experimental results of the use of a Web Graphical User Interface (GUI). Fig. 7a shows that the speaker said, "The weather is nice today", and the system successfully displayed the speaker's complete sentence. Fig. 7b shows the situation where the speaker said "good morning" twice in a row, but the recognition result showed homophones.

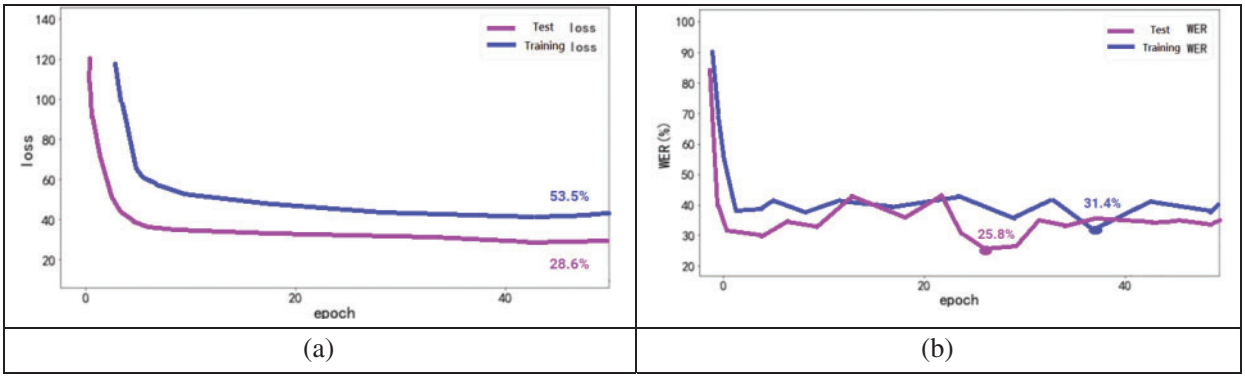


(a) Identify success (b) Identify failure

Figure 7: Experimental results using the acoustic model developed in this study

4.2 Experimental Data Analysis

To validate and train the CTC-CNN baseline acoustic model, THCHS-40 (Technology Human Codes High Speed-40) and STA-CMDS (Speech Technology Analysis-Chinese Mandarin Database Scripts) speech datasets were used as training datasets, and the dataset was categorized into the training and test sets. The training results are shown in Figs. 8a and 8b. After 62 epochs of training, the word error rate of the acoustic model training set was about 53.5%, and the word error rate of the test set was stable at about 28.6%. A certain overfitting phenomenon was observed. This experiment proves that it was difficult to practically apply the word error rate. Thus, it is necessary to optimize and adjust the network structure parameters to further improve the accuracy of the acoustic model.



(a) (b)

Figure 8: (a) Loss variation of baseline acoustic model (b) Word error rate change of baseline acoustic model

Table 1 lists the recognition accuracy for multiple consecutive words. The results are a representative of 2500 test datasets. When the speaker spoke only one word, the recognition accuracy was the highest, with an accuracy rate of 98.11%. In contrast, when the speaker uttered sentences of more than five words, the recognition accuracy dropped to only 93.77%. Sentences longer than five words may cause the system to misjudge words due to speakers' punctuation marks or words that sound too similar. For example, "recognize speech" and "wreck a nice beach" have similar pronunciations in English, and "factors" and "Sound Wave" have similar pronunciations in Chinese. Fig. 9 presents a forecast trend chart of recognition accuracy rates of various word counts. In addition to the above-mentioned conditions, environmental noise factors may also lead to a decline in recognition accuracy.

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

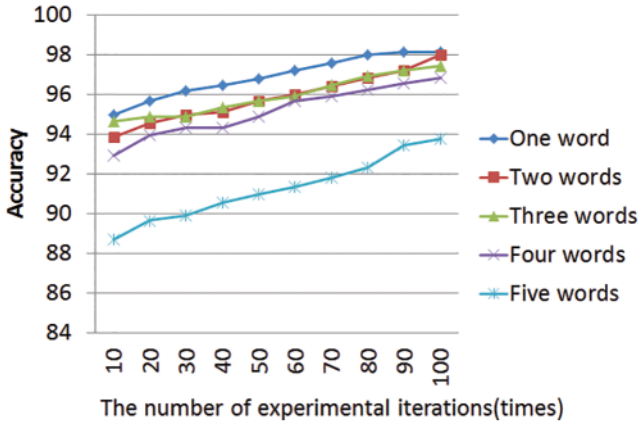


Figure 9: Prediction chart of the trend of various word count recognition accuracy rates

The surrounding environmental noise was also considered in this study; 80 tests were performed at each decibel level, and the data are shown in Table 2. Taking the noise of 80–90 dB as an example, the accuracy rate is 88.18%, which is the worst performance among all levels. Fig. 10 shows the predicted trend of the effect of environmental noise on recognition accuracy. At 40–60 dB, the accuracy was as high as 97.33% due to less noise pollution.

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

For this reason, this experiment was designed to verify their time-frequency diagrams according to the characteristics of Chinese initials and vowels. Due to the similarity of Chinese pronunciation, the recognition error rate of the system is expected to increase significantly. Chinese Pinyin has 21 initials and 16 finals. The vowels are formed mainly through changes in the shape of the mouth, while consonants are formed by controlling the airflow through certain parts of the mouth or nasal cavity.

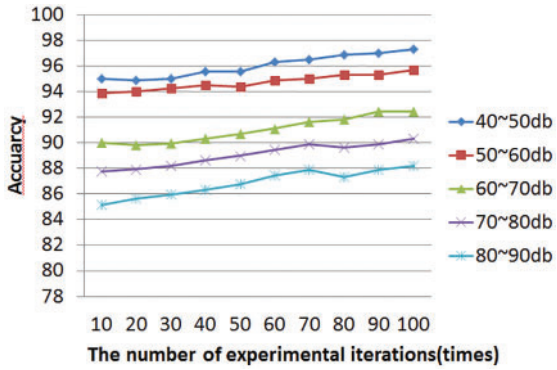


Figure 10: Prediction trend of the impact of environmental noise on sound identification accuracy

The difference in energy and frequency of Chinese vowels can be verified by time-frequency graph experiments, and this difference can be used for simple vowel recognition [23]. Therefore, consonants have low energy, high frequency, and short duration and often appear before vowels; conversely, vowels have higher energy, lower frequency, and longer duration and usually appear after consonants or independently. In the case of words with diphthongs (e.g., wāng), the amplitude is always very large, thus blurring the boundaries between sounds. In this case, it is difficult to use the amplitude to judge the change of the vowel.

The results in Fig. 11 show that consonants have small amplitudes, while vowels have relatively larger amplitudes. Taking the word "Jin" as an example, the amplitude of the initial consonant "ji" is relatively small, and the amplitude does not increase significantly until the appearance of the vowel "yī".

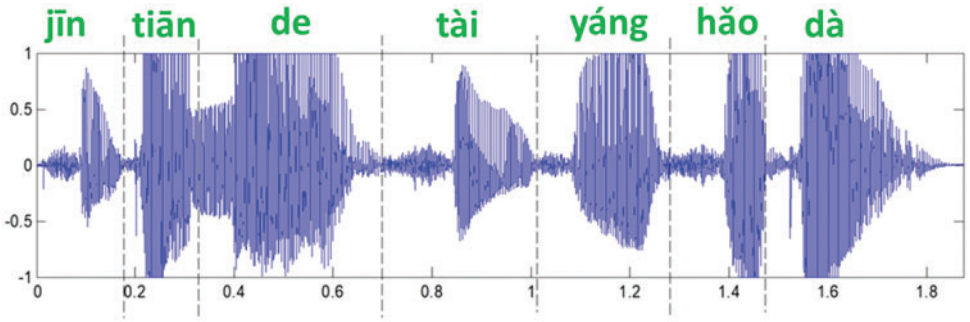


Figure 11: Consonant sound time-frequency diagram

4.3 Optimization of Acoustic Models

The baseline acoustic model in this study has certain challenges, such as a long training time, a high error rate, and a certain degree of overfitting. The model was trained through the continuous design and improvement of the relevant parameters of the acoustic model, and finally, the model with excellent performance was selected according to the evaluation index. Common optimization strategies for neural networks include dropout, normalization, and residual modules. Dropout was first proposed by Srivastava et al. (2018), and the overfitting problem could be effectively solved. The residual module, proposed by Zhou et al. in 2021 [24], could solve the problem of gradient

disappearance caused by the deepening of network layers. Normalization was first proposed by Segey Loffe and Christian Szegedy in 2020, which can speed up model convergence and alleviate the overfitting problem to a certain extent. However, as the depth of the network increases, the nonlinear layer of the network renders the output results to depend on each other and no longer fulfills the standard normal distribution. A problem of output center offset occurs, which brings difficulties to the training of the network model [25].

The characteristics of neural network input generally obey the standard normal distribution and generally perform well for shallow models. The training of deep models is particularly difficult. Through current processing, the network input conforms to the standard normal distribution and can be trained well, thus speeding up the convergence [26]. The data dimension processed by the CNN is a four-dimensional tensor, so it involves many normalization methods, including layer normalization (LN), instance normalization (IN), group normalization (GN), batch normalization (BN), etc. [27]. To solve the problem of model convergence, a normalization operation was added in the middle layer; that is, the output of each layer was normalized to make it conform to the standard normal distribution. Fig. 12 shows a schematic diagram of normalization for comparison.

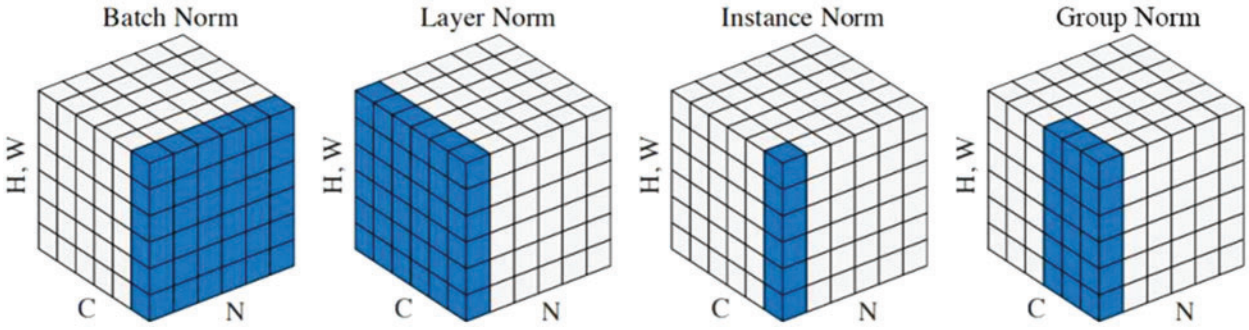


Figure 12: Comparison diagrams of applied normalization

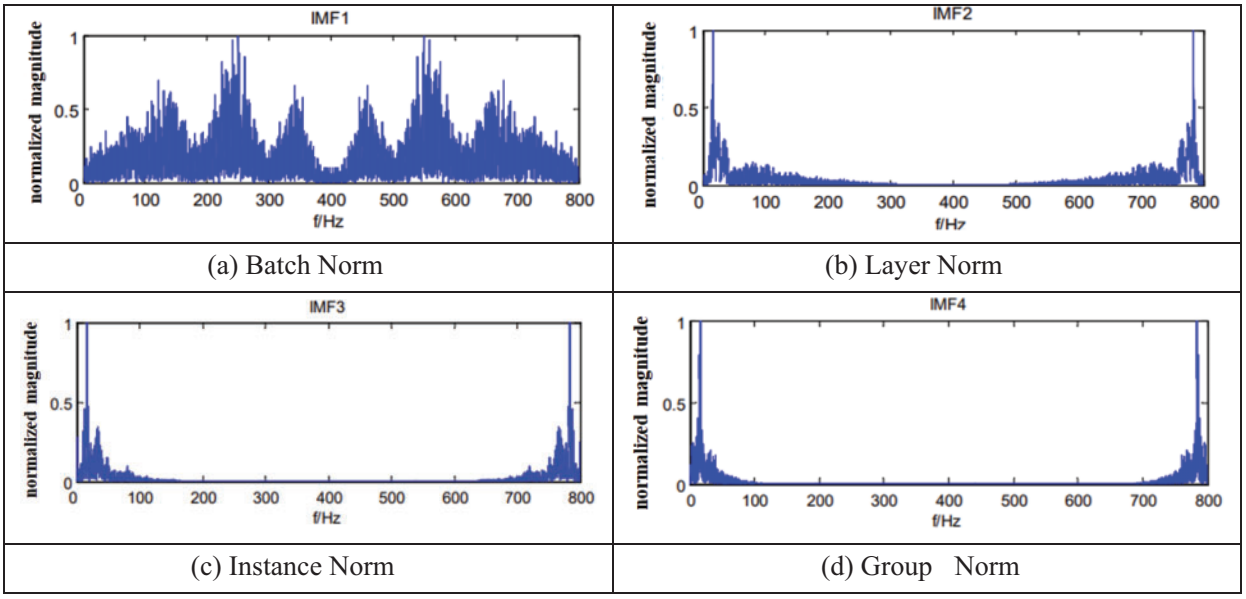
Taking a piece of speech data as an example, since the speech frequency range was roughly 250∼3400 Hz and the high frequency was 2500∼3400 Hz, four Intrinsic Mode Function (IMF) component frequency diagrams are decomposed by the normalized comparison method, as shown in Figs. 13a–13d. The high-frequency region of speech is mainly concentrated in the first IMF component, which can be seen from the density of normalized amplitude values of each IMF component. For the feature parameter extraction of the high-frequency region, the traditional extraction algorithm is not applicable, and a high-frequency feature parameter extraction algorithm must be sought [28].

Figs. 13a–13d show that the high-frequency region of the speech signal can be effectively extracted by Empirical Mode Decomposition (EMD) decomposition.

The original input information is transmitted to the output layer through the newly opened channel on the network side. Fig. 14a presents a schematic diagram of the residual module. 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 so that the network structure is different for repeating operation training, thus avoiding the dependence of the network on a single feature for classification and prediction.

The residual module directly transfers the input of the previous layer to the next layer by adding a congruent mapping layer. The schematic diagram of Dropout is presented in Fig. 14b. Dropout is a method that involves training multiple neural networks and then averaging the results of the entire

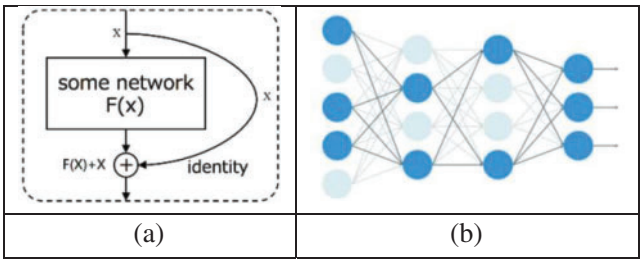
set instead of training a single neural network, which increases the sparsity of the network model and improves the generalization ability.



(a) Batch Norm (b) Layer Norm

(c) Instance Norm (d) Group Norm

Figure 13: Comparison diagram of normalizations based on intrinsic mode function



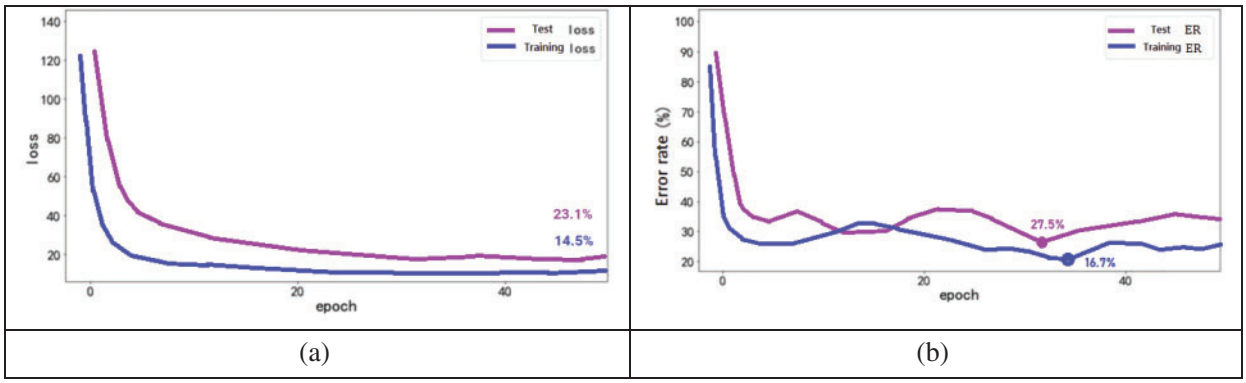
(a) (b)

Figure 14: (a) Schematic diagram of residual module; (b) Schematic diagram of dropout

Figs. 15a and 15b show the training comparison of the baseline acoustic model and improved acoustic model, respectively [29].

Fig. 15a shows that the improved acoustic model still faces the problem of overfitting. Therefore, this improved acoustic model requires further optimization. In the improved acoustic model, the number of network layers reaches 25 layers. Fig. 15b shows that by increasing the depth of the network model, the Windows Error Reporting (WER) of the improved acoustic model on the test set is reduced by 4.3% compared with the baseline model.

After the verification results, Dropout and Back-Propagation Network (BPN) layers are used in the network model to solve the overfitting problem. The network model structure is shown in Fig. 16, which solves the problem of shortening the training time and leading to more accurate decoding performance [30].



(a) (b)

Figure 15: (a) Loss comparison of the acoustic model; (b) Comparison of windows error reporting of the acoustic model

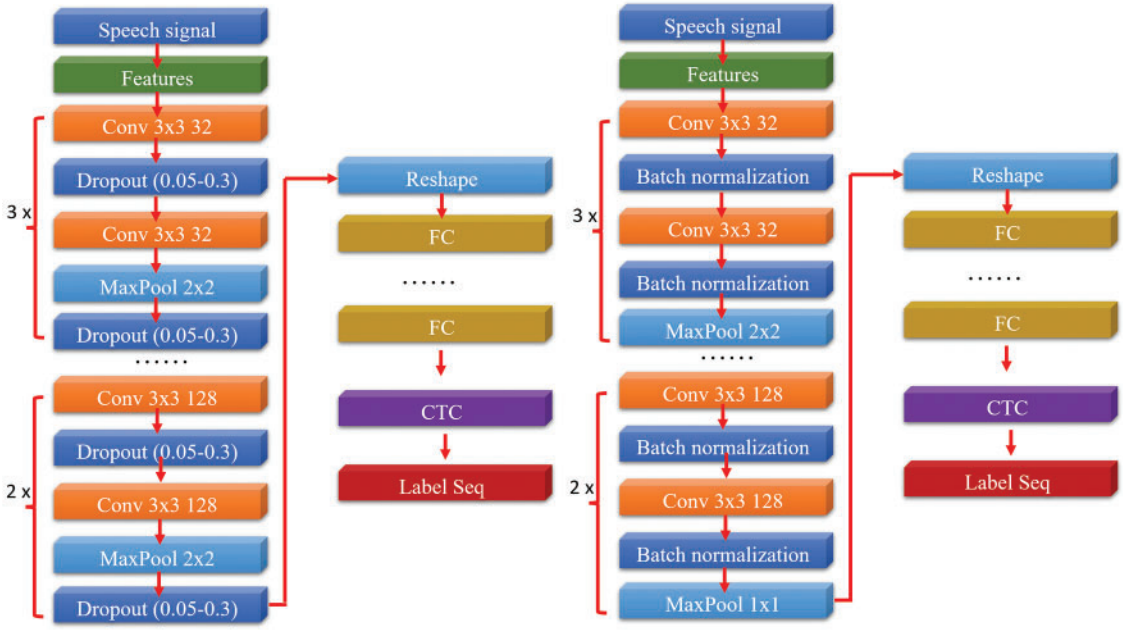


Figure 16: Comparison of structures between dropout and back-propagation network acoustic models

5 Conclusions and Future Directions

This study briefly introduces the historical development of deep learning, the most widely used deep learning models, and the development and current status of these models in the field of speech recognition. In this study, the CTC-CNN algorithm had a key role in the end-to-end framework. The CTC-CNN baseline acoustic model was constructed, improved, and optimized to reduce the error rate to about 16.3%, thus improving the accuracy rate. Finally, the selection of acoustic feature parameters, the selection of modeling units, and the speaker's speech rate were compared and verified, which further established the excellent performance of the CTC-CNN model structure.

In this study, deep learning was integrated into the development stage of the CTC-CNN model. The main problems are: (1) training usually must solve a highly nonlinear optimization problem, which

may easily lead to many local minima during the training process of the network, and (2) a too-long training time may lead to overfitting results. In practical application, the system is stable, efficient, and general-purpose, and more than 97.5% of the recognition rate of noisy speech can be achieved. For future speech recognition research, the best development direction is brain-like computing. The speech recognition rate can be raised to a satisfactory level only by continuously conforming to the characteristics of human brain speech recognition. How to better apply deep learning to meet the market demand for efficient speech recognition systems is a problem worthy of continuous attention.

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.

Funding Statement: The authors received no specific funding for this study.

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

References

[1] K. Khysru, J. Wei and J. Dang, "Research on Tibetan speech recognition based on the Am-do Dialect," *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 Int. Conf. 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," *Computers, Materials & Continua*, vol. 73, no. 1, pp. 629–639, 2022.

[4] M. Sakurai and T. Kosaka, "Emotion recognition combining acoustic and linguistic features based on speech recognition results," in *Proc. 2021 IEEE 10th Global Conf. on Consumer Electronics (GCCE)*, Kyoto, Japan, pp. 824–827, 2021.

[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. B. 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 Int. Conf. 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 Conf. 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 Int. Conf. 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 Int. Conf. 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.
[29] K. Sahu and R. K. Srivastava, "Predicting software bugs of newly and large datasets through a unified neuro-fuzzy approach: Reliability perspective," *Advances in Mathematics: Scientific Journal*, vol. 10, no. 1, pp. 543–555, 2021.
[30] K. Sahu, R. S. Pandey and R. Kumar, "Risk management perspective in SDLC," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 4, no. 3, pp. 1247–1251, 2014.