Intelligent Automation & Soft Computing
DOI:10.32604/iasc.2021.016457
Article

Tech Science Press

Oral English Speech Recognition Based on Enhanced Temporal Convolutional Network

Hao Wu[1,*] and Arun Kumar Sangaiah[2]

[1]Hunan Radio and TV University, Changsha, 410004, China
[2]School of Computing Science and Engineering, Vellore Institute of Technology, Tamil Nadu, 632014, India
[*]Corresponding Author: Hao Wu. Email: whwendy@163.com
Received: 02 January 2021; Accepted: 02 February 2021

Abstract: In oral English teaching in China, teachers usually improve students' pronunciation by their subjective judgment. Even to the same student, the teacher gives different suggestions at different times. Students' oral pronunciation features can be obtained from the reconstructed acoustic and natural language features of speech audio, but the task is complicated due to the embedding of multimodal sentences. To solve this problem, this paper proposes an English speech recognition based on enhanced temporal convolution network. Firstly, a suitable UNet network model is designed to extract the noise of speech signal and achieve the purpose of speech enhancement. Secondly, a network model with stable parameters is obtained by pre training, which is helpful to distinguish the spoken speech signals. Thirdly, a temporal convolution network with residual connection is designed to infer the meaning of pronunciation. Finally, the speech is graded according to the difference between the output value and the real result, according to the details of students' oral pronunciation, the intelligent guidance of students' oral pronunciation can be realized. The experimental results show that the model file obtained after training is improved under the controlling of file size. From the test results of LibriSpeech ASR *corpus*, it demonstrates the effectiveness and advantage of this approach.

Keywords: Temporal convolutional network; college English teaching; speech recognition; teaching model

1 Introduction

With the fast-growing of computer application, computer-aided instruction (CAI) has been widely used in English teaching. The combination of machine learning and ASR can effectively correct students' pronunciation and promote their oral English learning.

The auxiliary oral pronunciation mainly involves three aspects, namely, speech preprocessing, speech recognition, and speech matching and guidance. At present, the bottleneck of speech recognition mainly focuses on speech recognition, while the methods of speech preprocessing have been quite mature. With the achievements of using deep learning in speech recognition, ASR is also used speech recognition to

check the entrance guard and so on. Since there are few corpora suitable for oral English teaching in China, the application of speech recognition in English teaching is only a preliminary exploration.

2 Related Researches

This paper will introduce three aspects of work related to this study, mainly including the development of oral English teaching research, speech preprocessing technology, and ASR.

2.1 The Development of Oral English Teaching

The development of oral English in China began several decades ago. Before that, there were few researches on oral English in China. The empirical research based on data was very rare. In recent years, with the development of linguistics, cognitive linguistics, pedagogy, other related disciplines, and the improvement of domestic and foreign researchers' teaching ability, the quality of oral English teaching has been enhanced. Because of the breakthrough change of the content and method of English teaching and the improvement of oral English teaching [1,2], English examination gradually becomes social, and more emphasis is paid to the cultivation of practical application ability. Compared with the social demand for the ability of listening and speaking, foreign language teaching is still obviously insufficient in the cultivation of students' ability of "speaking". General speaking, there is still much room for improvement in oral English teaching in China.

In the traditional oral English teaching, the teaching concept, content, and method are backward. Traditional oral English teaching is mainly based on oral expression and indoctrination, and the teaching method is single. For a long time, grammar and translation have been dominant in English teaching in China. The mode of translation method is teacher-centered, students only passively input and store information, rarely could output language. Teachers themselves grow up in the traditional English education mode, lack of oral teaching thinking and methods, accuracy, fluency is not strong. Teachers' teaching follows the concept of exam-oriented education [3,4], and English class is still a single training mode dominated by "cramming". In the classroom, teachers still interpret English in the way of Chinese thinking, which is not to let students integrate into the communication scene to deeply understand [5]. Oral teaching and oral training time are insufficient, which is not conducive to the cultivation of students' oral ability, cannot arouse enough students' attention, the classroom atmosphere is dull, and cannot mobilize the enthusiasm of students. Teachers mainly explain words and grammar. Students have mastered a large number of English words and grammar knowledge. They can pass various examinations, but once they encounter English native speakers, they are a bit confused. Students' oral communication ability is generally low. It is very likely that it is difficult to express their ideas and communicate directly in proper language except for the simplest responses and greetings. At this situation, people call it as "dumb English". Therefore, in English teaching, the quality of information is as important as the fluency of information expression [6,7].

In the present stage, teaching methods and means have changed in oral English teaching. With the popularization and application of multimedia facilities and network platform, oral English teaching uses the advantages of network multimedia technology feedback, unlimited repetition, unlimited by time and place, to mobilize students' autonomous learning ability. Multimedia network technology integrates text, sound, image, and animation, which can provide real language materials and communication scenes. It can create a variety of relaxed and pleasant practice environment, and ensure that students have enough opportunities to train oral English, then the oral English teaching is no longer limited into the classroom through face-to-face teaching [8,9]. Based on massive comprehensible language input and output, learners have a lot of opportunities to contact and use English. They improve oral ability, mobilize the enthusiasm and initiative of learning [10,11]. If conditions permit, some schools can introduce foreign teachers, use

speech recognition software, promote oral teaching tasks with situational teaching method, emphasize the core skills of speaking, that is, the knowledge, skills and strategies used by speakers, so as to help students acquire oral ability subconsciously, adapt to the classroom, complete oral practice, stimulate students' interest in learning, improve their language ability and oral communication ability [12]. Classroom teaching takes "students as the main body, teachers as the leading", the arrangement of teaching content is gradual and the teaching methods are relaxed and diverse. It also needs to create a relaxed and happy classroom atmosphere, so that students can eliminate the anxiety of tension and in the best learning state, students have enough understanding the importance of oral English, then they have the courage to speak, dare to try and express themselves, then actively participate in classroom activities, at last they have a sense of progress and sense of achievement [13].

2.2 Speech Preprocessing

At present, speech recognition is usually carried out online, which is automatic recognition of users' real-time speech. In this identification process, it can be divided into two modules: front-end and back-end. The main functions of the front-end include endpoint detection, noise reduction, feature extraction, etc. Speech signal preprocessing, also known as front-end processing [14], refers to processing the original speech before feature extraction, so that the processed signal can meet the actual needs, which is great significance to improve the processing accuracy.

At present, speech recognition system can achieve high recognition performance in relatively quiet and non-interference scenes. However, due to the complexity of the real scene, the ideal level of speech recognition is not so easy to achieve. On the one hand, voice signals are transmitted from the sound source to the receiver in the form of sound waves in the air and other media. In the process of transmission, various interference factors will be encountered, such as environmental background noise, reverberation, etc., which will greatly reduce the speech quality, and it is also a great challenge for the machine to accurately identify the signals after such complex factors are interfered with [15]. On the other hand, most of the current researches focus on the single target speech recognition, that is, only one speaker speaks at one moment. When the number of speakers increases, the speeches of different speakers are mixed, which will make the recognition accuracy decline sharply. It is still a difficult problem to solve the problem of multi-person conversation speech recognition in complex acoustic scene.

In practice, the audio signal in the real world is only partially damaged in most cases. Therefore, it is also effective for automatic speech recognition to detect the broken input signal fragments by preprocessing technology and then remove them from the next processing stage [16]. Speech signal preprocessing is an indispensable part of speech signal processing. In the actual application environment, speech will be interfered by environmental noise in varying degrees. Before feature extraction, whether it is speech recognition or speech coding synthesis, the input speech signal should be preprocessed first in order to eliminate the human vocal organ itself, or other factors affect the quality of voice signal which are caused by the acquisition of the aliasing, high-order harmonic distortion, high-frequency and so on. Schmid [17] proposed to ensure that the signal obtained by subsequent speech processing is as uniform and smooth as possible, so as to improve the quality of speech processing. Through converting the analog signals into digital signals for computer processing, and then feature extraction is performed. Some parameters reflecting the characteristics of speech signals are used to represent speech. Finally, different processing methods are adopted according to different tasks.

There are different preprocessing algorithms for different types of interference. For example, in high noise environment, noise reduction pretreatment is needed; in high reverberation scene, it is necessary to carry out de reverberation pretreatment; in multi-speaker scene, speaker segmentation clustering preprocessing or speech separation preprocessing are needed. In face-to-face communication, people need to shorten the distance to reduce background noise, and in telephone communication, they need to

increase the volume. Many speech recognitions are not so friendly to the elderly, because the elderly speak slowly and keep silent for a long time. If the existing speech recognition system is modified, there will be additional costs. Therefore, the use of preprocessing can not only provide better speech recognition, but also greatly reduce the extra cost. In the field of SER (Speech Emotion Recognition) [18], preprocessing is the first step after mobile phone data, which will be used to train classifiers in SER system. Some of these preprocessing techniques are used for feature extraction, while others are used for feature normalization so that the changes of loudspeakers and recordings will not affect the recognition process [19].

Hu et al. [20] proposed a scheme based on pre-processing and post-processing, which can reduce the error by eliminating the harmonic component of speech. The post-processing scheme can not only effectively reduce the music noise, but also make the residual noise sound more natural. Considering that the actual speech signal is an analog signal, which should be sampled in a certain period and discretized according to Nyquist sampling theorem before digital processing of speech signal.

2.3 Speech Recognition Technology

The training of speech recognition model belongs to the learning of an undivided sequence of tags, that is, there is speech and corresponding text, but the corresponding relationship between which speech segment and which word or which word is not expressed. RNN (Recurrent Neural Network) is a powerful sequential learning model, but it requires pre-segmented training data to transform the model output into label sequence through post-processing, so its application is limited. Although the position of each character in the audio can be manually mapped, it is time-consuming. Therefore Gomez et al. proposed a new RNN training method CTC [21], which supports the direct prediction of tags on undivided sequences. CTC transforms the output of the network into a conditional probability distribution on the tag sequence. For a given input, the network completes the classification by selecting the most likely tag and outputs the result. However, CTC cannot establish the dependency relationship between the model outputs, and lacks the ability of language modeling, meanwhile it cannot be integrated language models for joint optimization. To solve this problem, Graves proposed the RNN-Transducer [22]. Based on the CTC encoder, a RNN called Prediction Network is added. The Prediction Network takes the output of the previous frame as the input, while the output of the prediction network and the output of the encoder are sent to a Joint Network. Finally, the output of the Joint Network is input to Softmax to obtain the classification probability. In this way, RNN Transformer can learn the sequence relationship between before and after the sequence.

Attention based models are more and more widely used in machine learning [23] and deep learning. Chan et al. introduced attention mechanism and proposed LAS [24], which is different from CTC and RNN-Transformer. It uses attention mechanism to effectively align. LAS consists of two main parts: Listener, Attach and Spell. The former uses the encoder to extract hidden features from the input sequence, while the latter uses the attention module to get the context vector, and then the decoder uses the context vector and the output of previous frames to generate the final output. The accuracy of LAS model is slightly higher than other models because it takes all the information of context into account. However, because it needs the information of context, it cannot carry out ASR (Automatic Speech Recognition) in streaming form. In addition, the length of input voice has a great impact on the accuracy of the model. The above three methods belong to the most basic methods, and other methods are mostly based on them.

There are many effective end-to-end methods based on deep learning [25]. These methods use RNN layer instead of acoustic model and train with standard sequence. However, these models are complex in calculation and need to take a long time to train. Collobert et al. [26] proposed Wav2Latter, which does not sacrifice accuracy to speed up training. It is completely dependent on its loss function to handle the alignment of audio and transcriptional sequences, and the network itself is only composed of convolution units, which is a full convolution network, so it can have shorter training time and lower hardware requirements.

The end-to-end training of ASR system usually not only consumes a lot of computing resources, but also needs a lot of data sets for training, which leads to the increase of training cost. Kunze et al. [27] explored a model-based adaptive transfer learning method based on Wav2Latter method. In the experiment, they found that it was feasible to transfer the network trained on English *corpus* to German *corpus*. Training from scratch, using the parameters trained in English *corpus* to train on German can effectively shorten the time, and the model of transfer learning training can get the same score as the model only using German with less training data. Sequence to sequence model is simpler in training than other types of models. However, there is no sequence-to-sequence model that can reach the most advanced level in a large number of vocabulary continuous language recognition tasks, until Chiu et al. [28] proposed an advanced level sequence to sequence language recognition model, which was improved based on LAS. The main parts of the improvement are word level modeling instead of ordinary phoneme or letter level [29], adding multiple attention mechanisms, minimizing word error rate, using Scheduled Sampling to reduce the difference between training and testing [30], and using Label Smoothing alleviates the possible over fitting problem and asynchronous training speeds up the training. The final performance is greatly improved compared with the original ASR [31].

3 Temporal Convolutional Network (TCN)

TCN is an innovative structural of CNN applied to timing problems [32,33]. The convolution relationship in TCN is causality [34], which is mainly based on Causal Convolution and Dilated Convolution, it shows that there is no information from the future to the past. The input and output of TCN can be made equal length by proper mapping, and the effective representation information can be constructed by residual layer and hole convolution.

3.1 Sequence Modeling

Given a spoken speech input sequence $(x_0, x_1, \cdots, x_T)$, the constraint condition is that only the observation data before the current can be used, and predict the corresponding results is realized by calculating the network mapping function f of the sequence model. The objective function is to minimize the loss between the predicted output and the real output, where f can be represented as:

$$\hat{y}_0, \hat{y}_1, \cdots, \hat{y}_T = f(x_0, x_1, \cdots, x_T) \tag{1}$$

3.2 Causal Convolution

Each layer adopts one-dimensional full convolution network architecture, and each hidden layer and input layer are equal in length. By introducing padding operation, the length of subsequent layers is maintained, and there is no information from the future to the present. The layer structure of causal convolution is shown in Fig. 1.

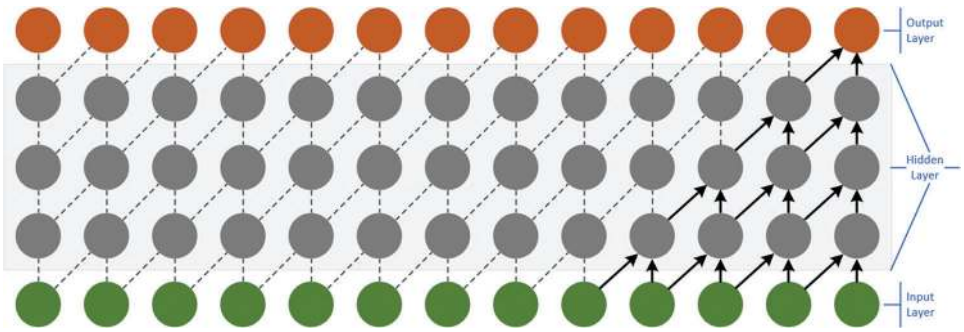

Figure 1: Causal convolutional layers

In the causal convolution structure, the lower layer only depends on the value t of the next layer and its previous value. Each layer of causal convolution cannot see the future data, which is a unidirectional connection structure and is a strict time constraint model, by which it is called causal convolution.

3.3 Dilated Convolution

However, there is a problem in this model. If the output record is longer, a deep network must be built, and the calculation complexity will increase exponentially. Considering that only simple causal convolution can only see the past data with a linear relationship with the network depth. Therefore, it is very important to expand the Dilated Convolution of the sensing field. For the one-dimensional sequence and filter, the operation of dilated convolution on each element s in the sequence is defined as:

$$F(s) = (x *_d f)(s) = \sum_{i=0}^{k-1} f(i) * x_{s-d*i} \tag{2}$$

Where d is the dilated factor and k is the filter size. s-d*i indicates the past direction. The characteristic sense field is improved by increasing the dilated factor d, to realize the spanning of long and effective historical data. The layer structure of dilated convolution is shown in Fig. 2.

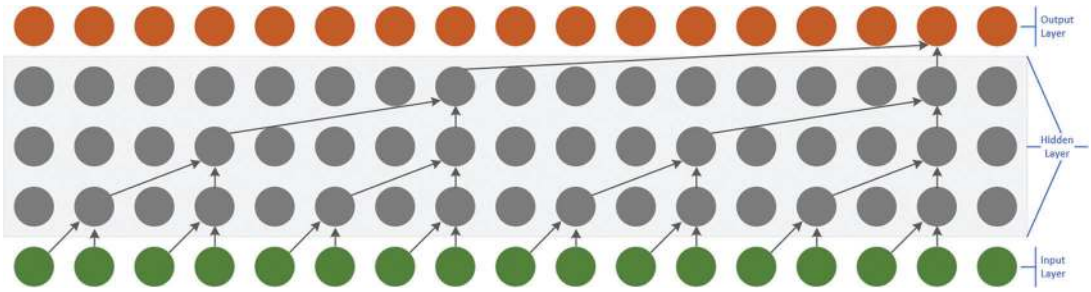

Figure 2: Dilated convolution

Unlike traditional convolution, the input of the dilated convolution is allowed to have interval sampling, in which the sampling rate is controlled by d in the graph. $d = 1$ of the bottom layer means that every point is sampled when the input and the second layer is $d = 2$ upward, which means that when the input, one sampling is taken as input for every 2 points. The higher the level, the larger the size of d. Therefore, the Dilated Convolution makes the effective window size increase exponentially with the increase of the number of layers, so that the convolution network can obtain a large sense field with fewer layers.

3.4 Residual Connection

The residual connection has been proved to be an effective mean to train deep network, which makes the network transmit information in a cross-layer way. In this paper, a residual block is constructed to replace the convolution of one layer. As shown in Fig. 3, a residual block contains two levels of convolution and nonlinear mapping. In each layer, the Weighted Norm and Dropout are added to regularize the network.

A residual block contains a branch F pointing to the transformation operation whose output is added to the original input:

$$O = Activation(x + F(x)) \tag{3}$$

The using of the residual structure is very effective for the calculation of deep network, and the previous input will be helpful to the subsequent learning process. The receptive field of TCN is determined by the

depth of network n, the size of core k and the cavity factor d, which is very important for the deeper and larger TCN to be stable. When the prediction results depend on the 2^{10} data and a high-dimensional input sequence, then a 10 layers network is needed. Each layer has multiple filters to extract features. Since the input and output sizes in TCN are different, 1×1 convolution is used to make them have the same size.

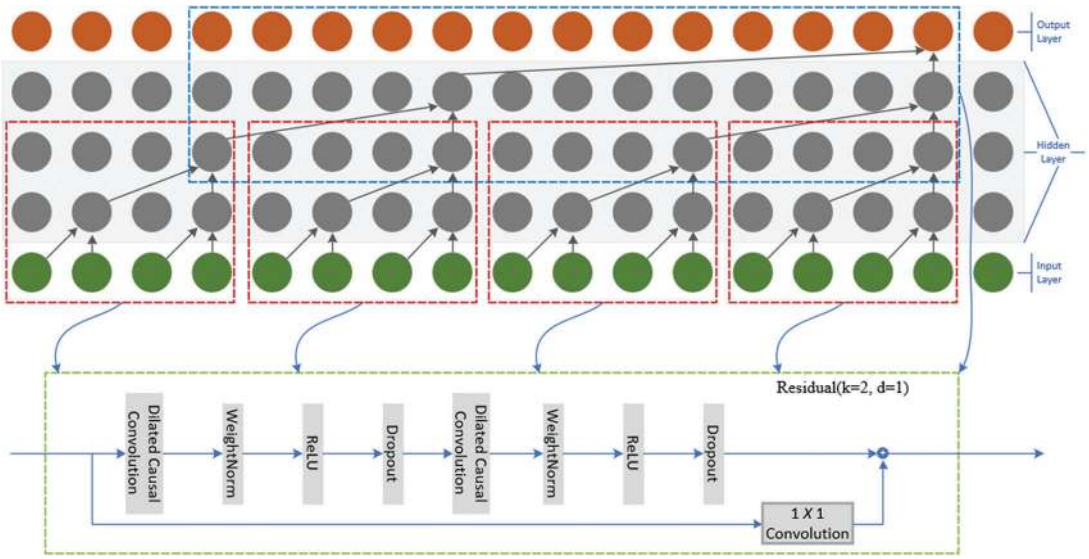

Figure 3: Dilated causal convolution and residual connection

3.5 Speech Enhancement

The speech enhancement is realized by UNet network structure, which is suitable for the separation of human sound and noise. UNet was first used in biomedical image segmentation to improve the accuracy and location of the microimaging of neuron structure. UNet is based on the complete convolution network, which is mainly composed of a combination of a network structure of a sequence of down sampling and upper sampling. The image size of each layer of the lower sampling convolution layer is halved, the number of channels is doubled, and the image is encoded as a small and deep representation. The reverse convolution layer of the upper sampling is just the opposite, especially in the upper sampling process, which combines the lower sampling information and the upper sampling input information. If the details are restored, the original size image precision can be restored step by step.

The offset of a pixel is not considered as the main distortion for the restoration of natural images. However, such a small linear offset can also cause disaster impact in frequency domain processing. In speech recognition, even the slight deviation in time dimension may produce jitter and be heard. UNet establishes an additional direct connection between the same layer structure of encoder and decoder, which makes the low-level information directly from high-resolution input stream to high-resolution output.

3.5.1 Speech Enhancement Structure

The objective of the speech enhancement neural network is to predict the input human sound and noise indirectly. The output of the decoder layer is a soft mask, which is multiplied by elements of the mixed spectrum to obtain the final estimate. Fig. 4 outlines the network structure. In this paper, the authors choose to train two independent models to extract the human sound and noise components of the signal, to provide more different training schemes for the two models in the future.

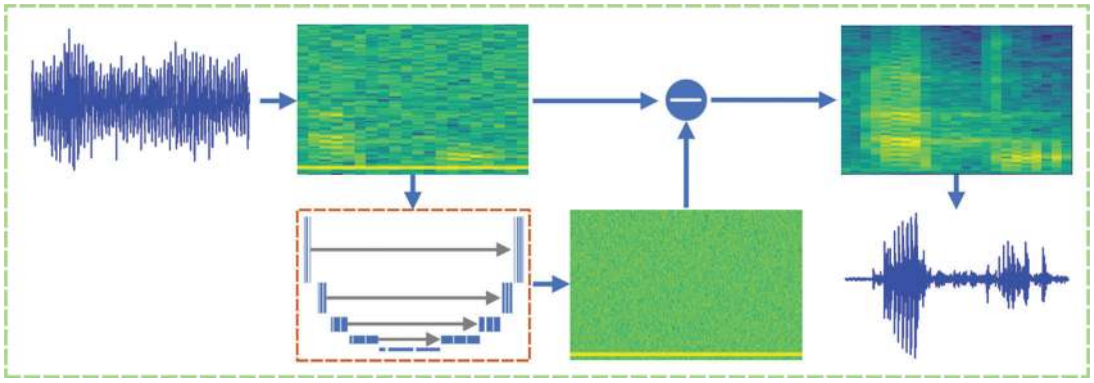

Figure 4: Speech enhancement structure

3.5.2 Model Training

Suppose X represents the data of the spectrum of the original mixed-signal, that is, the audio containing human voice and external noise components, and Y represents the information of the spectrum map of the target audio. The latter refers to the human voice (Y_v) or noise (Y_i) component of the input signal. The loss function used for training model is L1, that is, the norm of the difference between the target spectrum and the mask input spectrum

$$L(X, Y; \Theta) = \|f(X, \Theta) \odot X - Y\|_{1,1} \tag{4}$$

where $f(X, \Theta)$ is the output mask of the network model, which is applied to the output of the network model with input X. The parameter Θ is the mask generated by the model. Two UNet networks, Θ_V and Θ_I are trained to predict the spectrum mask of the human voice and external noise.

3.5.3 Network Architecture Details

Each encoder layer of UNet is composed of step-by-step 2D convolution with a step of 2 and core of 5×5, batch normalization and linear unit (Rectified Linear Unit, ReLU) with a correction of 0.2. In the decoder, step deconvolution is used, the step size is 2, core size is 5×5, batch processing is standardized, common ReLU, and 50% loss are used in the first three layers. In the last layer, a sigmoid activation function is used and Adam optimizer is used to train the model.

Due to a large amount of calculation during model training, to speed up the processing, the input audio frequency is reduced to 8,192 hz, and the short-time Fourier transform with the size of 1,024 frames and the hop length of 768 frames is input into the network as the target, and the spoken English speech spectrogram is standardized to the range of value [0~1].

3.5.4 Audio Signal Reconstruction

Generally, environmental noise keeps stable in the sampling process. If the spectrum of the noise signal can be reconstructed effectively, the enhancement process will be easy to implement. Through the UNet neural network model to operate the size of the audio spectrum, including spoken speech and external noise of the audio signal after the UNet network prediction reconstruction of the external noise spectrum, and then the original spectrum amplitude minus the predicted external noise spectrum, the enhanced real spoken speech spectrum can be obtained.

3.5.5 Training Data Construction

The training data described by the model structure is expressed in the form of two tuples (voice components and external noise components). It is unrealistic to assume that the acquisition of spoken English signals can

obtain unmixed information through multi-channel. It is a good strategy to reconstruct spoken speech through UNet. A good model is trained through the collected oral English speech *corpus* in the early stage, to effectively separate the external noise components and enhance the spoken English speech signal.

4 Oral English Experiment and Result Analysis

In this part, the method is evaluated on LibriSpeech English *corpus*, and the results are compared with other popular models.

4.1 Experiment Setup

This paper trains and designs the model on the LibriSpeech ASR *corpus*, which is a speech database published in 2015, including 1,000 h of English pronunciation and corresponding words. The training *corpus* consists of 460 hours of 16 kHz English speech. For speech recognition, the LibriSpeech *corpus* is used to report the results. LibriSpeech is the most commonly used English *Corpus* in public. After being cut and sorted into 10 s audio files with text annotation, LibriSpeech is very suitable for testing spoken English speech recognition.

Training set: the training set is used to train the model, following the characteristics of a large training set, small development and test set, accounting for the vast majority of all data. **Development set:** Development set is used to test the model trained by the training set, and optimized the model continuously through the test results. **Test set:** the *corpus* is used for a final evaluation of the trained model after the training. Speech recognition performance is evaluated by the phoneme error rate (PhonemeErrorRate, PER) and word error rate (WordErrorRate, WER).

4.2 Sentence Embedding Experiment

Sentence embedding should consider determining which kind of embedding to use and how to combine them into a single sentence embedding. Intermediate embedding refers to the selection of two baseline embeddings as the representation of learning intermediate speech. The reason why they are chosen is that they are easy to use and suitable for language tasks. Although this method needs word alignment, it can be trained unsupervised. Different pronunciation of the same word has different embedding, which is the same as Word2Vec, but it is suitable for oral English. Phoneme2Vec is the same as Speech2Vec, which applies to phonemes. Each pronunciation of phoneme is encoded as an embedding, which is very effective for fine-grained tasks. Speech2Vec and Phoneme2Vec are trained on LibriSpeech ASR *corpus*, and some special models can be used to calculate word and phoneme alignment.

By embedding and fusing each word or phoneme, they are combined into a sentence embedding. By calculating the sum of the elements embedded in the middle of each word or phoneme position and dividing by the number of words in the sentence, the middle embedding is converted into sentence embedding, and then the vector value is input into the deep neural network to generate the final sentence embedding. The comparison of phoneme, word and sentence embedding is shown in Tab. 1.

Table 1: Comparison of phoneme, word, and sentence embeddings

Embedding Level	LibriSpeech WER
Phoneme	76 ± 9
Word	30 ± 6
Sentence	14 ± 2

4.3 A Comparative Experiment of Different Corpuses

In the 300 h switchboard *corpus*, the WER of the proposed scheme is 10.4%, and that of the 1,000 h LibriSpeech *corpus* is 3.8%. The size of network models obtained by TCN training is 29.4 M, while the size of ETCN (Enhanced Temporal Convolutional Network) model models is unchanged, but the performance is improved. The comparison is shown in Tab. 2.

Table 2: Comparison of different methods on librispeech and SWBD-300 datasets

Models	Librispeech				Switchboard
	Test-clean	Dev-clean	Test-other	Dev-other	Train_dev
IBM Deep CNN	–	–	–	–	16.9
Microsoft Spatial smoothing+ Lattice-free MMI	8.46	–	13.62	–	15.9
Google Network-in-Network+ Batch Normalization+ ConvLSTM	7.28	7.05	12.23	12.1	14.9
pFSMN-Chain	3.62	3.28	8.45	8.37	10.89
TCN	4.6	4.5	10.62	10.21	13.91
ETCN*	3.8	2.56	7.6	7.5	10.4

5 Summary

In this paper, the subsequent recognition is proved to be effective after the speech enhancement processing. The time additional enhancement processing is very small, which can be ignored, but the effect is obvious. The introduction of residual connection based on TCN model is effective for the training of the deep network, otherwise, the effect will be worse with the deepening of the network, but the amount of calculation may increase slightly with the addition of residual connection.

Although this paper has made some progress in spoken English speech recognition, there are still some improvements. How to further improve the ability of ETCN in parallel processing sentences; whether different computing units can be activated according to different tasks, including convolution kernel size and expansion coefficient; how to effectively improve the adaptability of ETCN transfer learning, the authors hope it can be used for further study.

Acknowledgement: We would like to thank all the parties involved in this research work.

Funding Statement: This work was financially supported by the Scientific Research Fund of Hunan Provincial Education Department of China (Grant No. 20C1246); and the 2021 Project of Hunan Province Social Science Achievement Evaluation Committee (Grant No. XSP21YBC202).

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

References

[1] Z. Y. Wan and X. S. Gao, "English or Chinese as medium of instruction? International students' perceptions and practices in Chinese universities: Exploring international students' perceptions and reported practices with regard to English and Chinese as mediums of instruction in parallel programs in Chinese universities," *English Today*, vol. 36, no. 1, pp. 37–44, 2020.

[2] Y. Zhao, J. Yue, W. Song, X. Xu, L. Li *et al.,* "Tibetan multi-dialect speech recognition using latent regression bayesian network and end-to-end mode," *Journal on Internet of Things*, vol. 1, no. 1, pp. 17–23, 2019.

[3] M. A. Pasban and M. H. Narafshan, "The relationship between learners' academic goal motives and L2 (second language) Willingness to communicate in English language classes: A look at academic goal motives' orientations," *Cogent Psychology*, vol. 7, no. 1, pp. 1824307, 2020.

[4] L. Zuo, "Computer network assisted test of spoken English," *Computer Systems Science and Engineering*, vol. 34, no. 6, pp. 319–323, 2019.

[5] Y. G. Butler, J. Lee and X. L. Peng, "Failed policy attempts for measuring English speaking abilities in college entrance exams: Cases from China, Japan, and South Korea," *English Today*, pp. 1–7, 2020.

[6] C. Y. Liu, "Application of speech recognition technology in pronunciation correction of college oral English teaching," In: V. Sugumaran, Z. Xu, H. Zhou (eds.), *Proc. of Int. Conf. on Application of Intelligent Systems in Multi-modal Information Analytics (MMIA)*, vol. 1234, pp. 525–530, 2020.

[7] Z. Zeng, "Implementation of embedded technology-based english speech identification and translation system," *Computer Systems Science and Engineering*, vol. 35, no. 5, pp. 377–383, 2020.

[8] X. G. Li, J. H. Chen, Z. Chen and Y. N. Chen, "An approach to evaluation index and model of undergraduates' spoken English pronunciation," in *Proc. of Int. Conf. on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, Changsha, China, pp. 2189–2193, 2016.

[9] S. Zhou, L. Chen and V. Sugumaran, "Hidden two-stream collaborative learning network for action recognition," *Computers, Materials & Continua*, vol. 63, no. 3, pp. 1545–1561, 2020.

[10] J. C. Mora and M. Levkina, "Task-based pronunciation teaching and research: Key issues and future directions," *Studies in Second Language Acquisition*, vol. 39, no. 2, pp. 381–399, 2017.

[11] J. Y. M. Zhang, J. Liu and X. Y. Lin, "Improve neural machine translation by building word vector with part of speech," *Journal on Artificial Intelligence*, vol. 2, no. 2, pp. 79–88, 2020.

[12] L. Wang and F. Fang, "Native-speakerism policy in English language teaching revisited: Chinese university teachers' and students' attitudes towards native and non-native English-speaking teachers," *Cogent Education*, vol. 7, no. 1, pp. 1778374, 2020.

[13] Q. Xie, "Investigating the target language usage in and outside business English classrooms for non-English major undergraduates at a Chinese university," *Cogent Education*, vol. 4, no. 1, pp. 1415629, 2017.

[14] Y. B. Gao, D. Q. Xin, H. G. Tian and X. M. Xu, "Improvement of speech preprocessing model based on linear prediction," in *Proc. of IEEE Conf. on Chinese Automation Congress (CAC)*, Hangzhou, China, pp. 5581–5586, 2020.

[15] T. Dau, C. Christiansen and M. S. Pedersen, "Prediction of speech intelligibility based on an auditory preprocessing model," *Speech Communication*, vol. 52, no. 7–8, pp. 678–692, 2010.

[16] K. Simonchik, S. Aleinik, D. Ivanko and G. Lavrentyeva, "Automatic preprocessing technique for detection of corrupted speech signal fragments for the purpose of speaker recognition," In: A. Ronzhin, R. Potapova, N. Fakotakis (eds.), *Proc. of Int. Conf. on Speech and Computer (SPECOM)*, pp. 121–128, 2015.

[17] D. Schmid, P. Thuene, D. Kolossa and G. Enzner, "Dereverberation preprocessing and training data adjustments for robust speech recognition in reverberant environments," in *Proc. of ITG Sym. on Speech Communication*, Braunschweig, Germany, pp. 1–4, 2012.

[18] S. Kwon, S. J. Kim and J. Y. Choeh, "Preprocessing for elderly speech recognition of smart devices," *Computer Speech & Language*, vol. 36, no. 3, pp. 110–121, 2016.

[19] M. B. Akcay and K. Oguz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, no. 12, pp. 56–76, 2020.

[20] X. H. Hu, S. W. Wang, C. S. Zheng and X. D. Li, "A cepstrum-based preprocessing and postprocessing for speech enhancement in adverse environments," *Applied Acoustics*, vol. 74, no. 12, pp. 1458–1462, 2013.

[21] A. Graves, S. Fernández, F. Gomez and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. of Int. Conf. on Machine Learning (ICML)*, Pittsburgh, Pennsylvania, USA, pp. 369–376, 2006.

[22] A. Graves, "Sequence transduction with recurrent neural networks," arXiv:1211.3711, 2012.

[23] D. Zhang, G. Yang, F. Li, J. Wang and A. K. Sangaiah, "Detecting seam carved images using uniform local binary patterns," *Multimedia Tools and Applications*, vol. 79, no. 13–14, pp. 8415–8430, 2020.

[24] W. Chan, N. Jaitly, Q. Le and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. of Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Shanghai, China, pp. 4960–4964, 2015.

[25] S. R. Zhou, J. Wu, F. Zhang and P. Sehdev, "Depth occlusion perception feature analysis for person re-identification," *Pattern Recognition Letters*, vol. 138, no. 3, pp. 617–623, 2020.

[26] R. Collobert, C. Puhrsch and G. Synnaeve, "Wav2Letter: An end-to-end convnet-based speech recognition system," arXiv:1609.03193, 2016.

[27] J. Kunze, L. Kirsch, I. Kurenkov, A. Krug and J. Johannsmeier, "Transfer learning for speech recognition on a budget," arXiv:1706.00290, 2017.

[28] C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen *et al.,* "State-of-the-art speech recognition with sequence-to-sequence models," in *Proc. of Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Calgary, AB, Canada, pp. 4774–4778, 2018.

[29] Y. He, T. N. Sainath, R. Prabhavalkar, I. McGraw, R. Alvarez *et al.,* "Streaming end-to-end speech recognition for mobile devices," in *Proc. of Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Brighton, United Kingdom, pp. 6381–6385, 2019.

[30] S. Schneider, A. Baevski, R. Collobert and M. Auli, "WAV2vec: Unsupervised pre-training for speech recognition," in *Proc. of Conf. of the Int. Speech Communication Association (INTERSPEECH)*, Graz, Austria, pp. 3465–3469, 2019.

[31] Z. Gao, S. Zhang, M. Lei and I. McLoughlin, "SAN-M: Memory equipped self-attention for end-to-end speech recognition," arXiv:2006.01713, 2020.

[32] S. Zhou and B. Tan, "Electrocardiogram soft computing using hybrid deep learning CNN-ELM," *Applied Soft Computing*, vol. 86, no. 4, pp. 105778, 2020.

[33] D. J. Zeng, Y. Dai, F. Li, J. Wang and A. K. Sangaiah, "Aspect based sentiment analysis by a linguistically regularized CNN with gated mechanism," *Journal of Intelligent & Fuzzy Systems*, vol. 36, no. 5, pp. 3971–3980, 2019.

[34] S. Bai, J. Z. Kolter and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," arXiv:1803.01271, 2018.