An End-to-End Speech Recognition Model for the Northern Shaanxi Dialect: Design and Evaluation

Yi Qin * and Feifan Yu

Posted Date: 5 December 2024

doi: 10.20944/preprints202412.0477.v1

Keywords: dialect speech recognition;coal mining industry; end-to-end; Conformer model; Transformer model; Connectionist Temporal Classification(CTC)

 doi:10.20944/preprints202412.0477.v1

Article

An End-to-End Speech Recognition Model for the Northern Shaanxi Dialect: Design and Evaluation

YiQin [1,*] and Feifan Yu [2]

[1] College of Computer Science & Technology, Xi'an University of Science and Technology, Xi'an 710054, China 1; qiny@xust.edu.cn
[2] SHCCIG Yubei Coal industry Co.,Ltd., Xi'an, Shaanxi 710900; 1312250556@qq.com
* Correspondence: qiny@xust.edu.cn;

Abstract: The coal mining industry in Northern Shaanxi is robust, with a prevalent use of the local dialect, known as "Shapu," characterized by a distinct Northern Shaanxi accent. This study addresses the practical need for speech recognition in this dialect. We propose an end-to-end speech recognition model for the Northern Shaanxi dialect, leveraging the Conformer architecture. To tailor the model to the coal mining context, we developed a specialized corpus reflecting the phonetic characteristics of the dialect and its usage in the industry. We investigated feature extraction techniques suitable for the Northern Shaanxi dialect, focusing on the unique pronunciation of initial consonants and vowels. A preprocessing module was designed to accommodate the dialect's rapid speech tempo and polyphonic nature, enhancing recognition performance. To enhance the decoder's text generation capability, we replaced the Conformer decoder with a Transformer architecture. Additionally, to mitigate the computational demands of the model, we incorporated CTC joint training for optimization. Experimental results on our self-established voice dataset for the Northern Shaanxi coal mining industry demonstrate that the proposed Conformer-Transformer-CTC model achieves a 9.2% and 10.3% reduction in word error rate compared to the standalone Conformer and Transformer models, respectively, confirming the advancement of our method.

Keywords: dialect speech recognition;coal mining industry; end-to-end; Conformer model; Transformer model; Connectionist Temporal Classification(CTC);

1. Introduction

In recent years, the rapid advancement of artificial intelligence technology has spurred the further development of intelligent coal mine construction [1]. The introduction of policies such as the "Intelligent Coal Mine Guide (2021 Edition)" and the "Trial Measures for the Acceptance Management of Intelligent Demonstration Coal Mines" has underscored the growing necessity of establishing intelligent coal mines that leverage related artificial intelligence technologies [2-3]. Given the unique characteristics of the coal industry, developing an intelligent coal mine that includes a voice interaction system tailored to the sector is crucial for ensuring the safety of coal mine production.

Northern Shaanxi, one of the most coal-rich regions in China, is at the forefront of producing and managing the production environment, conferences, dispatching, and command operations within the coal industry. The integration of a voice interaction system that accommodates the local dialect is vital for enhancing communication efficiency and safety in these contexts.

Management personnel in the coal mining industry predominantly communicate using the Northern Shaanxi dialect, commonly referred to as "Shaanxi Pu," which is characterized by a distinct Northern Shaanxi accent. The dialect serves not only as a cultural emblem but also as a carrier of traditional cultural heritage. Consequently, it is imperative to compile a corpus of the Northern Shaanxi dialect and to develop speech recognition capabilities tailored to the coal mining industry.

 doi:10.20944/preprints202412.0477.v1

This initiative is vital for both preserving the regional culture and enhancing operational efficiency within the sector.

To address the challenge of dialect recognition within the field of speech recognition, initial research efforts by scholars involved adapting traditional speech recognition models for dialect recognition purposes. This included the application of linear predictive coding (LPC), dynamic time warping (DTW), and hidden Markov models (HMMs) as technical frameworks for dialect identification. Furthermore, the integration of Gaussian Mixture Model (GMM) technology in speech modeling has notably enhanced recognition rates. For instance, studies [4-6] have utilized the GMM to develop dialect recognition systems for the Mongolian dialect, Chongqing dialect, and the Shuozhou dialect in Shanxi Province.

Due to the reliance of traditional dialect recognition methods on extensive corpora and manual annotation, these approaches incur high costs and often yield only moderate recognition performance. Moreover, the advent of deep learning has transformed the landscape of speech recognition technology.

Technology, particularly deep learning, not only streamlines the process of speech recognition but also markedly enhances recognition accuracy. For instance, the study in literature [7] developed an end-to-end Listen, Attend and Spell (LAS) model for Tujia speech recognition, incorporating a multi-head attention mechanism to boost the accuracy of Tujia dialect recognition. In document [8], an end-to-end dialect speech recognition method based on transfer learning is introduced, which leverages shared feature extraction to enhance the recognition performance of low-resource dialects. Document [9] presents an end-to-end speech recognition system that integrates a multi-head self-attention mechanism with a residual network (ResNet) and a bidirectional long short-term memory network (Bi-LSTM), significantly improving the recognition of Jiangxi and Hakka dialects. Nonetheless, these models could benefit from further enhancements in incorporating contextual semantic information and capturing positional details.

Currently, end-to-end (E2E) speech recognition technology has yielded substantial research outcomes [10]. Architectures such as Recurrent Neural Networks (RNNs) [11], Convolutional Neural Networks (CNNs) [12-13], self-attention-based Transformer networks [14], and the Conformer [15] have emerged as prominent backbone structures for Automatic Speech Recognition (ASR) models, garnering significant attention. However, the Conformer decoder exhibits limited text generation capabilities, the Transformer model is computationally intensive, incurs substantial memory costs, and has a weaker ability to capture local features.

Addressing these limitations, this study introduces a Conformer-Transformer-CTC (Connectionist Temporal Classification) fusion approach for dialectal speech recognition systems. The proposed method leverages the audio modeling prowess of Conformer as the encoder, utilizes Transformer for text generation as the decoder, and harnesses the flexible alignment capabilities of CTC to construct an end-to-end dialect speech recognition model, thereby enhancing speech recognition accuracy.

2. Related Work

The end-to-end speech recognition process entails the consolidation of acoustic, pronunciation, and linguistic factors within a single deep neural network (DNN) to streamline the modeling process and achieve direct mapping from speech input to text output [16-17]. As deep learning techniques gain widespread application across various domains, end-to-end speech recognition models have demonstrated superior recognition performance compared to traditional speech recognition models, garnering increasing interest and attention [18-19].

Currently, the prevalent end-to-end methods include: a) Connectionist Temporal Classification (CTC) [20], which computes the model's loss and optimizes it using forward and backward algorithms [21]. Study [22] further introduces intermediate CTC loss to regularize CTC training and enhance speech recognition performance. b) Recurrent Neural Network Transducer (RNN-T) [23], which is suitable for streaming speech recognition as it can recall past information. c) The Encoder-Decoder architecture with an attention mechanism [24] has garnered significant attention and

 doi:10.20944/preprints202412.0477.v1

research. In recent years, the diverse modeling approaches in end-to-end Automatic Speech Recognition (ASR) systems have sparked interest in developing hybrid methods to leverage their complementary strengths for speech recognition [25]. For instance, the research in [26] combined CTC with the Transformer decoder to achieve higher recognition accuracy. The study in [27] integrated the self-attention mechanism and the multi-layer perceptron module into dual branches within the model, yielding exceptional recognition performance.

In essence, the integration of the strengths of different architectures has demonstrated potential for enhancing speech recognition performance in recent research [28]. In light of this, this study harnesses the benefits of end-to-end models to construct a speech recognition system for the Northern Shaanxi dialect, aiming to improve recognition performance.

3. Method

3.1. Corpora Establishment

Currently, there is a significant lack of openly accessible corpora concerning the dialect used within the coal mining sector in northern Shaanxi. To facilitate research in the realm of speech recognition for the northern Shaanxi dialect, this study has successfully compiled a specialized dialect corpus for the northern Shaanxi coal mining industry. The methodology employed in constructing this corpus is detailed in Figure 1.

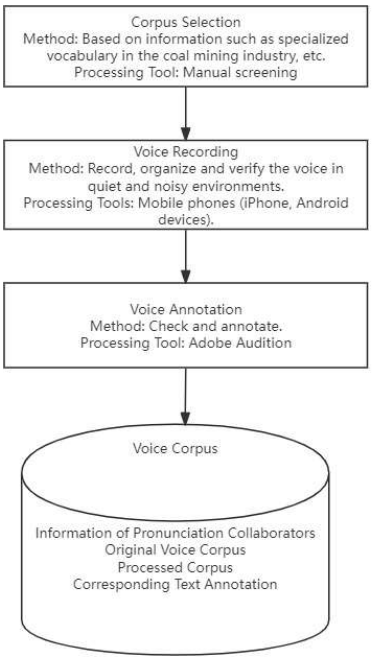

Figure 1. Construction process of dialect corpora.

The initial phase involves the careful selection of speech materials from the region. By analyzing the distinctive features of the northern Shaanxi dialect, which are summarized in Table 1, the corpus was compiled. The selection process was guided by the textual content of coal mine dispatch logs, industry-specific terminology from the coal mining sector, and relevant industrial texts. Subsequently, recorded transcripts were created based on this selected material.

 doi:10.20944/preprints202412.0477.v1

Table 1. Examples of the characteristics of northern Shaanxi dialect.

	Mandarin example	Example of northern Shaanxi dialect	remarks
Vocabulary characteristics	rainbow	pass through	Keep ancient words
taxeme	cause to trip	Not bad	There are a lot of Consonant words
	Relatives and friends	Friends	There are a lot of Inverse order word
	basket	Basket basket	The overlap of nouns
	rub	Wipe	The overlap of verbs

The recording protocol adopted in this study involves the collection of dialect data by 20 volunteers. The recordings are primarily based on the text from the northern Shaanxi coal mine scene-specific dialect dataset. Among the 20 volunteers, there are 13 males and 7 females, with ages ranging from 18 to 40 years. All participants are native to northern Shaanxi, and the dialects recorded are exclusively of the northern Shaanxi variety. The resultant dataset is detailed in Table 2.

Table 2. Self-built corpus of northern Shaanxi pronunciation.

data set	The number of people	duration	The number of sentences
training set	15	21h	8010
test set	5	2h	760
development set	5	2h	760

To guarantee the quantity and integrity of the data, this study employs professional recording equipment to capture the audio. The recordings are saved in WAV format with a 16 kHz sampling rate. Subsequent to recording, the audio files are meticulously checked and annotated using Adobe Audition.

In the final phase of data preparation, all recorded data, including participant information, the original phonetic corpus, the annotated and processed corpus, along with the corresponding text annotations, are systematically organized and stored within a unified corpus repository. This structured approach ensures that the dataset is both comprehensive and accessible for further analysis and use in speech recognition model training.

3.2. Conformer-Transformer-CTC Model Structure

To achieve a dialect speech recognition model with enhanced dialect recognition accuracy, this paper introduces a Conformer-Transformer-CC based speech recognition system. The system comprises several key modules: a prepossessing module, which includes both speech and text prepossessing, and a code module where the encoder utilizes the Conformer architecture, while the decoder combines Transformer and CTC for joint decoding. Figure 2 illustrates the end-to-end architecture of the dialect Conformer-Transformer-CC speech recognition system. This integrated approach is designed to leverage the strengths of both Conformer and Transformer networks, along with the benefits of CTC for robust and accurate dialect recognition.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 December 2024 doi:10.20944/preprints202412.0477.v1

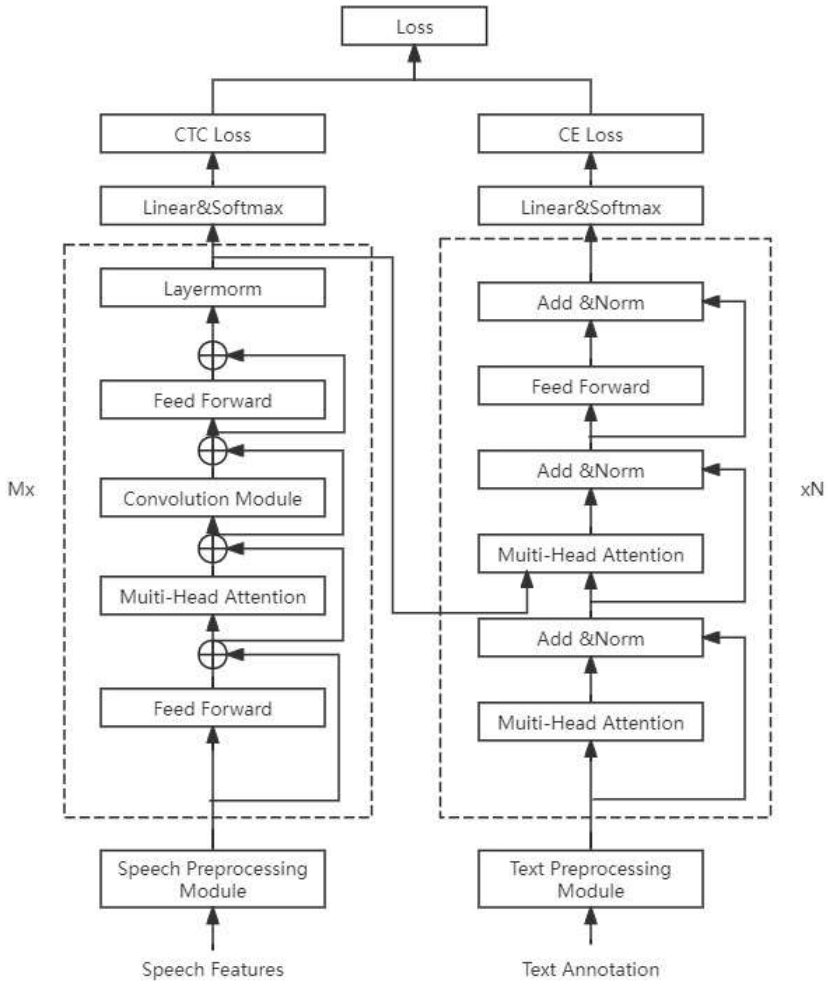

Figure 2. End-to-end dialect Conformer-Transformer-CC speech recognition system.

3.2.1. Preprocessing Module

The preprocessing module includes a speech preprocessing module and a text preprocessing module, as shown in Figure 3.

 doi:10.20944/preprints202412.0477.v1

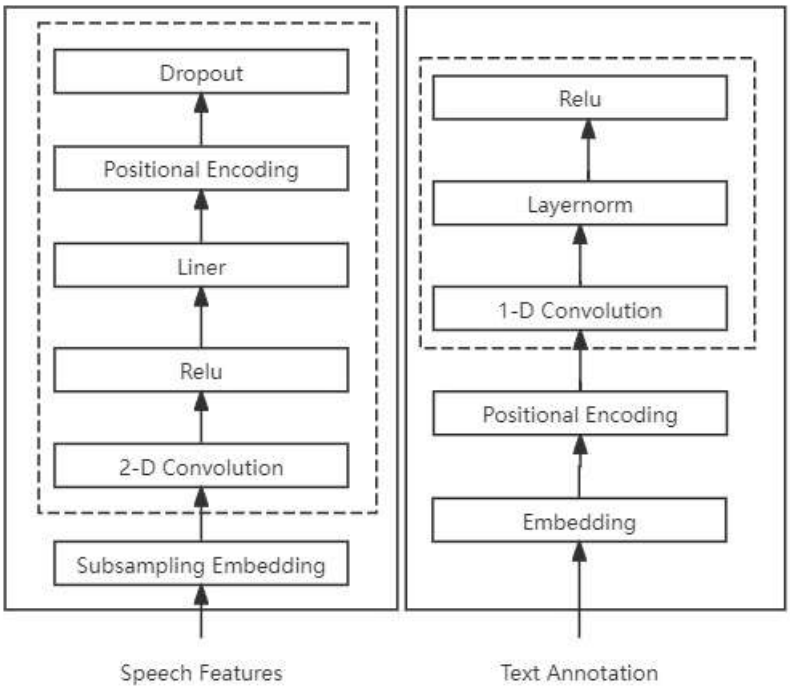

Figure 3. Preprocessing module.

The speech processioning module developed in this study comprises a down-sampling module (Sub-sampling Embedding), a convolution module, and position coding (Position Encoding).

Initially, the speech features undergo down-sampling. Dialectal speech may exhibit significant variations in temporal length or rapid speaking rates. Down-sampling reduces the time dimension's resolution, enabling the model to more effectively handle rapidly changing features.

Subsequently, the convolution module includes a 2D-Convolution layer followed by a Rel U activation layer to perform the convolution operation. Dialectal speech acoustic features often exhibit extemporization correlations. Two-dimensional convolution allows for the capture of these acoustic feature patterns across both time and frequency dimensions, facilitating the learning of local features within dialectal speech. Furthermore, the ReLU activation function enhances the model's expressive capacity, particularly for acoustic features that deviate from standard language norms.

Following this, a linear layer is employed to extract dialect-specific acoustic features and patterns, yielding a more tailored feature representation. Additionally, fixed position coding, utilizing sine and cosine functions, is applied to better comprehend the sequential distribution and structure of dialectal speech features, as detailed in formulas (1-2).

Conclusively, the application of Dropout operations involves the random suppression of certain features, which decreases the model's reliance on individual features and, in turn, strengthens its robustness and generalization capabilities.

$$PE_{(pos,2i)} = \sin(pos / 10000^{2i/d_{model}}) \quad (1)$$

$$PE_{(pos,2i+1)} = \cos(pos / 10000^{2i/d_{model}}) \quad (2)$$

The position encoding matrix, which indicates the specific location of the current word within the sequence. It represents the i-th dimension of the word vector and the dimension size of the word vector.

Through the aforementioned processes, the speech processioning module facilitates dimensional reduction, feature transformation, and position modeling for dialectal speech. This results in a feature representation with enhanced discriminating power, thereby improving the

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 December 2024 doi:10.20944/preprints202412.0477.v1

performance and accuracy of dialect recognition. Utilizing this speech processioning module enables the model to more effectively adapt to the characteristics of dialectal speech and to extract dialect-specific acoustic features and patterns.

The text processioning module developed in this paper includes an embedding layer, position coding, and convolution modules.

Initially, the embedding layer is employed to convert text labels into dense vector representations, capturing the semantic relationships between words. Subsequently, the same position coding technique used in the speech preprocessing module is applied to encode the positional information of the words. Following this, a 1-D convolutional layer is utilized to extract implicit positional information and to capture more nuanced local semantic details. Layer normalization is applied to normalize the model's output at this stage. Finally, the ReLU activation function is introduced to incorporate non-linearities, thereby enhancing the model's representational capacity.

By undergoing this text processing pipeline, the speech recognition system can enhance its comprehension and expression of textual information, thereby improving the overall performance of speech recognition.

3.2.2. Codec

The advantage of Conformer architecture as an encoder lies in its ability to process both time-domain and frequency-domain features of the audio signal. This dual processing capability allows the Conformer to yield a rich audio representation, enhancing its understanding of the input audio signal and providing more informative features for the subsequent decoder stage. Nevertheless, the Conformer structure exhibits limited text generation capabilities within its decoder. To address this, the present study employs a Transformer-based decoder. The self-attention mechanism inherent in the Transformer is adept at managing long-range dependencies, which allows the decoder to take into account the global context when generating text. This results in the production of accurate and coherent textual outputs. The proposed architecture is depicted in Figure 4.

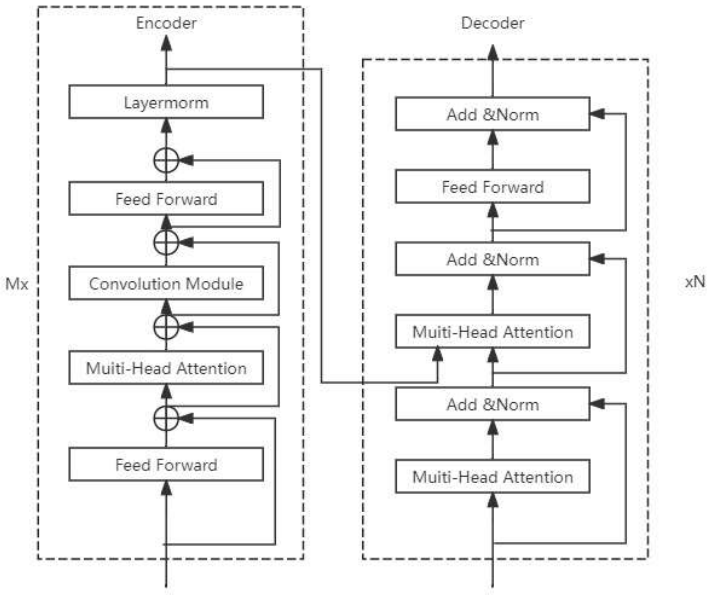

Figure 4. Codec structure.

- encoder

The Conformer model primarily consists of four key modules: the first feedforward module (Feed Forward), the multi-head attention module (Multi-Head Self-Attention), the convolutional module (Convolution Module), and the second feedforward module. For the Conformer, the process

 doi:10.20944/preprints202412.0477.v1

of computing the output for the input vector at the first encoder stage can be described by the following formula:

$$\tilde{x}_i = x_i + \frac{1}{2} FFN(x_i) \quad (3)$$

$$x'_i = \tilde{x}_i + MHSA(\tilde{x}_i) \quad (4)$$

$$x''_i = x'_i + Conv(x'_i) \quad (5)$$

$$y_i = Layernorm(x''_i + \frac{1}{2} FFN(x''_i)) \quad (6)$$

Among the components, FFN denotes the Feedforward module. 'First' signifies the initial feedforward module, succeeded by the 'Second' feedforward module. MHSA stands for the Multi-Head Self-Attention module, while 'Conv' is an abbreviation for the Convolution module. 'Layernorm' represents Layer Normalization. Each of these modules incorporates a residual connection to enhance the flow of gradients and stabilize training.

- decoder

Order represents the input sequence, represents the advanced sequence, is the output sequence, the probability of encoding and decoding is as follows:

$$x = (x_1, x_2, \ldots\ldots x_n)\ h = (h_1, h_2, \ldots\ldots h_t)\ y = (y_1, y_2, \ldots\ldots y_i)$$

$$P(y \mid x) = \prod_{t=1}^{T} P(y_t \mid h, y < t) \quad (7)$$

At each time, the conditional dependence of the output on the encoder features is calculated through the attention mechanism. The attention mechanism is a function of the hidden state of the current decoder and the encoder output features, which is compressed into a context vector. Is the learning parameter. The attention distribution is obtained by performing the normalization:

$$t\ h\ v^t\ b\ W_h\ W_d\ softmax$$

$$\alpha_t = soft\max(v^t \tanh(W_h h_i + W_d d_t + b) \quad (8)$$

Using and hiding states, the corresponding context direction is obtained by using weighted sums:

$$\alpha_t\ h_i$$

$$c_t = \sum_{i=1}^{K} \alpha_t h_i \quad (9)$$

The final model decoder training loss function is defined as:

$$Att_{loss} = -\ln(P(y \mid x)) \quad (10)$$

3.2.3. CTC Auxiliary Training

Methods for improving speech recognition assistance tasks were investigated by assuming different levels of learning representations at different levels[29]. In this paper, the CTC objective function is integrated into the Conformer-Transformer fusion model. End-to-end training through CTC learning sequence-to-sequence mapping without explicit alignment, which to reduce irregular alignment of Attention-based Encoder-Decoder (AED) models for better performance.

During training, the model consists of three parts, the Conformer encoder, the Transformer decoder, and the CTC decoder. The end-to-end dialect speech recognition training is as follows:

 doi:10.20944/preprints202412.0477.v1

$$h = Encoder(x) \tag{11}$$

$$P(y \mid x) = Decoder(h) \tag{12}$$

The output of the encoder was used to calculate the CTC loss, setting the training set as, Then the CTC loss function is:

$$CTC_{loss} = -\sum_{(X,Y)\in S} \ln P(y \mid x) \tag{13}$$

Combining the CTC loss and the decoder loss facilitates the convergence of the decoder, while enabling the hybrid model to exploit the label dependence. Since the CTC is used to assist the decoder alignment, the CTC is less weighted in the fusion. The total loss function is defined as the weighted sum of the CTC and the decoder loss:

$$T_{loss}(x,y) = \lambda CTC_{loss}(x,y) + (1-\lambda)\ Att_{loss}(x,y) \tag{14}$$

where,, is used to measure the importance of CTC loss and decoder loss. It represents speech features and represents text annotation.

$$\lambda \in [0,1]\ x\ \ y$$

4. For Experimental Validation

4.1. Experimental Indicators

In this paper, the experimental results are evaluated on the self-built dialect speech data set, and the evaluation algorithm index is word error rate WER, as the evaluation index. In order to align the identified word sequence and the standard word sequence, it needs to replace, delete, or insert certain words. The total number of words inserted, replaced, and deleted is divided by the percentage of the total number of words in the standard word sequence, which is as follows:

$$WER = \frac{I + D + S}{T} \times 100\% \tag{15}$$

This I represents the number of miswords added, D refers to the excluded words, T refers to the total word of the whole sentence, but S refers to the number of words that are replaced. A smaller WER value indicates a better identification effect.

4.2. Experimental Configuration

The hardware configuration used for the experiment is an Intel (R) Xeon (R) Gold 6330 processor, with 32 GB of running memory, and a GPU of NVIDIA GeForce RTX 3090. The software environment used is Anaconda 3 and Python 3.8 environments based on PaddlePaddle 2.4.1 deep learning framework under Ubuntu 20.04.2 LTS operating system. The model was trained simultaneously with the Adam optimizer and the learning rate adaptive change strategy[30].

4.3. Experimental Data

The corpus of northern Shaanxi dialect used in the model includes the prescribed exclusive vocabulary of coal mine industry, self-selected vocabulary, common sayings, coal mine dispatching call, coal mine report and other related corpora, with a total of 9770 sentences, with a total length of 25 hours. The types of the corpus include: vocabulary and grammar, oral culture, dialect dialogue and dialect narration, as shown in Table 3, which are part of the data content of the corpus of northern Shaanxi dialect.

 doi:10.20944/preprints202412.0477.v1

Table 3. Part of the northern Shaanxi dialect data set.

The corpus type	Chapter name	duration
lexicon grammar	A proprietary vocabulary of the coal mining industry (Specified vocabulary)	For 180 minutes and 32 seconds
	Choose a vocabulary	66 Minutes and 13 seconds
	Grammar example sentence	70 Minutes and 32 seconds
Oral culture	common saying	30 Minutes and 11 seconds
	phrase	28 Minutes and 36 seconds
Dialect dialogue	Daily dialogue	70 Minutes and 32 seconds
	Coal mine scheduling call	123 Minutes 2 seconds
Dialect	Coal mine report	60 Minutes and 45 seconds

4.4. Experimental Results and Analysis

In this paper, different experiments are set up from the following three aspects: feature extraction, parameter tuning and recognition rate (comparison experiment), which verify the influence of feature extraction technology, model parameters on the recognition rate and the recognition rate of this model is better than the mainstream model. The details are as follows:

4.4.1. Feature Extraction Experiment

Due to the phonetic phonetic characteristics of dialect, it is difficult to achieve the best performance in the corpus of northern Shaanxi dialect by using a single phonetic feature extraction technology. Therefore, according to the phonetic rhyme characteristics of northern Shaanxi dialect, multiple speech features are extracted, and different feature extraction techniques have an influence on the speech recognition performance of dialects. The results are shown in Table 4.

Table 4. Tyword rates under different feature extraction techniques.

phonetic feature	Error word rate (WER%)
MFCC feature	32.3
FBank feature	29.5
Log-Mels feature	30.2
MFCC+FBank feature	33.2
MFCC+Log-Mels feature	27.6
Fbank+Log-Mels feature	28.8
MFCC + FBank + Log-Mels features	29.4

Considering the above analysis, the root cause can be attributed to the differences in the ability of different characteristics and the information expression mode. The MFCC features are relatively weak, but the combination with other features can provide some complementary information. The combination of different features can provide a more comprehensive and mainly rich audio feature representation, thus improving the performance of the speech recognition system.

4.4.2. Parameter Tuning Experiment

Since the self-built northern Shaanxi dialect data set cannot fully match the end-to-end model of depth, it is necessary to tune the model hyperparameters to perform the best results as far as possible. In this paper, several key parameters of the model: the depth and width (layer number and dimension) of the encoder (Conformer module), the multiple heads and dimension of the self-attention mechanism, and the size of the deep convolution kernel. By selecting these parameters, the best model parameters suitable for the data set of northern Shaanxi dialect.

- Conformer Number of modules

 doi:10.20944/preprints202412.0477.v1

Other parameters remain unchanged, adjust the number and dimension of Conformer modules, set the number of Conformer modules according to 16,12, and 8, and set the corresponding dimensions according to 1024,2048, and 4096.

Experimental results Table 5 shows the influence of the number of encoders (Encode number) and the encoder dimension (Encoder dimension) in different groups on the performance. From the WER index, group 3 achieves the optimal effect. The dimension of the encoder was increased to 2048, providing a larger parameter space to capture the complex features of the audio. Higher encoder dimensions can better represent audio data and learn a richer feature representation, thus improving the accuracy of speech recognition. The poor performance in group 1 is be due to the larger number of encoders, resulting in excessive model parameters and increasing the risk of over fitting. The performance in Group 3 is also poor, probably due to the high dimension of the encoder, which makes the model too complex to fully learn and generalize. Therefore, we find the optimal parameters to achieve balance, thus improving the performance of the speech recognition system.

Table 5. Conformer module change test results.

group	Encoder count	Encode dimension	WER（%）
1	16	1024	31.5
2	12	2048	26.9
3	8	4096	20.4

- The Self-Attention module

Other parameters remain unchanged, the multiple heads of the encoder Self-Attention module are set according to 2,4 and 8, and the corresponding dimensions are set according to 512,256 and 128. The Self-Attention in the corresponding decoder is also set consistently for experiments.

Experimental results Table 6 shows that the influence of Self-Attention multiple heads and each head dimension on speech recognition performance are interrelated, and from the WER index, group 1 achieves the optimal effect. In this group, the dimension per Attention long head is high, allowing each head to better capture the key information in the input sequence, and the multiple heads are balanced with the dimension, thus reducing identification errors. Groups 2 and 3 performed poorly relative to group 1. As the number of Self-Attention multiple heads increases, the dimensionality of each head decreases. The lower head dimension may limit the expression ability of each head to the input sequence, causing the model to accurately learn key features.

Table 6. Attention module change test results.

group	Attention Multiple head number	Each head dimension	WER（%）
1	2	512	26.9
2	4	256	28.8
3	8	128	29.6

- Convolution module

Other parameters are unchanged, the convolution kernel size is singular, and the experiments were conducted according to 3,7 and 15.

Experimental results Table 7 indicates that group 1 had the best results. Generally speaking, the larger the convolution core, the larger the receptive field, the more information the network "sees", and the better the obtained feature representation. However, in the current scale dialect, the use of small convolutional kernel size can better capture local details, which may help to better extract the characteristics of local details. The large convolutional kernel instead blur these detailed features, which makes group 1 better than the other two groups. At the same time, the larger the size, the larger the computation, the greater the computational cost.

 doi:10.20944/preprints202412.0477.v1

Table 7. Convolution module change test results.

group	Convolutional kernel size	WER（%）
1	3	26.9
2	7	27.8
3	15	28.6

4.4.3. Comparison Experiments

Using self-built corpus, some mainstream end-to-end models including WeNet (Conformer), WeNet (Transformer) are identified. Through the results shown in Table 8, it can be seen that this model is better than other mainstream models.

Table 8. Recognition rates of Northern Shaanxi dialect datasets on different models.

model	Error word rate (WER%)
WeNet(Conformer)	36.1
WeNet(Transformer)	37.2
Transformer-CTC	34.5
Conformer-CTC	33.8
Ours	26.9
Ours (-pretreatment module)	30.8

The experimental results show that the Conformer end-to-end dialect recognition model improves the sequence generation ability by introducing Transformer and CTC, while enhancing the ability to model long sequences and the flexible alignment. At the same time, it shows that when the preprocessing module is not used, the recognition performance is lower than that of the preprocessing module. Through the dimension reduction, feature transformation and position modeling, the model is more adapted to the phonetic characteristics of northern Shaanxi dialect and better extract the unique acoustic features and patterns of dialect.

5. Conclusions and Outlook

This paper presents a dialect speech recognition model based on the end-to-end coal mining industry in northern Shaanxi. The Conformer-Transformer-CTC fusion approach utilizes the strengths of each component, including robust feature extraction, accurate sequence generation, and alignment flexibility. Through experimental analysis and comparison, we prove that the proposed dialect speech recognition model is more suitable for northern Shaanxi dialect, thus having lower error rate and better generalization performance. The next step will be

This paper studies how to effectively integrate external language models to improve dialect speech recognition performance, and how to extract more effective pronunciation features of different dialects to achieve better recognition effect and further expand in other official dialects.

References

1. WANG Guofa, DU Yibo, CHEN Xiaojing, et al. Development and innovative practice from coal mine mechanization to automation and intelligence: Commemorating the 50th anniversary of the founding of Industry and Mine Automation[J]. Industry and Mine Automation,2023,49(6):1-18.
2. WANG Feng. Eight ministries and commissions: Promote the integrated development of intelligenttechnology and coal industry[J]. China Plant Engineering, 2020(7):1-1.
3. Kingdom law. Interpretation of The Guide to Intelligent Construction of Coal Mine (2021 edition) ——from the perspective of the writing team [J]. Smart Mine, 2021,2 (4): 2-9.
4. WANG Guofa.Interpretation of theguide to intelligent construction of coal mines (2021): From the perspective of the writing group[J].Journalof Intelligent Mine,2021, 2(4):2-9.
5. LIU Zhiqiang MA Zhiqiang ZHANG Xiaoxu et al.IMUT-MC：a speech corpus for Mongolian soeech recognition[J].China Scientific Data,2022,7(2):75-87.

 doi:10.20944/preprints202412.0477.v1

6. Zhang Ce, Wei Pengcheng, Lu Xiaoyan, et al. Design and implementation of the Speech recognition system in Chongqing dialect [J]. Computer measurement and control, 2018,26 (1):256-259,263.
7. Yu Guo, Huan Jin Xia, Liu Xiaofeng, etc. Implementation and application of speech recognition system in Shanxi dialect [J]. Computer and Digital Engineering, 2021,49 (10): 2168-2173.
8. YU Benguo,HUAN Jinxia,LIU Xiaofeng,et al. Implementationand Application of Speech RecognitionSystem inShanxi Dialect[J]. Computer& Digital Engineering,2021,49(10):2168-2173.
9. Yu Heavy, Wu Jiajia, Chen Yunbing, etc. End-to-end Tujia speech recognition based on the multi-head attention mechanism [J]. Computer Simulation, 2022,39 (03): 258-262 + 282.
10. YU Chongchong,WU Jiajia,CHEN Yunbing,et al.End-to-End Speech Recognition Framework for Tujia Languagebased on Multi-head Attention Mechanism[J].ComputerSimulation,2022,39(03):258-262+282.
11. Yu C, Chen Y, Li Y, et al.Yu, Chongchong, et al.Cross-language end-to-end speech recognition research based on transfer learning for the low-resource Tujialanguage.Symmetry, 2019,11(2):179 -193.
12. XU Fan,YANG Jianfeng,YAN Weizhi,et a1.An end-to-end dialect speech recognition model based
13. on self attention[J].Journal of Signal Processing,2021,37(10):1860-1871.
14. Kim S, Gholami A, Shaw A, et al. Squeezeformer: An efficient transformer for automatic speech recognition[J].arXiv preprint arXiv:2206.00888, 2022.
15. Yin W, Kann K, Yu M, et al. Comparative study of CNN and RNN for natural language processing[J].arXiv preprint arXiv:1702.01923, 2017.
16. Ren S , He K , Girshick R ,et al.Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J].IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6):1137-1149.
17. Majumdar S, Balam J, Hrinchuk O, et al. Citrinet: Closing the gap between non-autoregressive and autoregressive end-to-end models for automatic speech recognition[J].arXiv preprint arXiv:2104.01721, 2021.
18. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J].Advances in neural information processing systems, 2017, 30:5998–6008.
19. Gulati A, Qin J, Chiu C C, et al. Conformer: Convolution-augmented transformer for speech recognition[J].arXiv preprint arXiv:2005.08100, 2020.
20. Wang D, Wang X, Lv S. An overview of end-to-end automatic speech recognition[J]. Symmetry, 2019, 11(8): 1018.
21. Li J. Recent advances in end-to-end automatic speech recognition[J].APSIPA Transactions on Signal and Information Processing, 2022, 11(1).
22. Zhang Q, Lu H, Sak H, et al. Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 7829-7833.
23. Chang F J, Radfar M, Mouchtaris A, et al. End-to-end multi-channel transformer for speech recognition[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 5884-5888.
24. Li J, Ye G, Das A, et al. Advancing acoustic-to-word CTC model[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018: 5794-5798.
25. Deng K, Cao S, Zhang Y, et al. Improving ctc-based speech recognition via knowledge transferring from pre-trained language models[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022: 8517-8521.
26. Lee J, Watanabe S. Intermediate loss regularization for ctc-based speech recognition[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 6224-6228.
27. Li B, Chang S, Sainath T N, et al. Towards fast and accurate streaming end-to-end ASR[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 6069-6073.
28. Li S, Dabre R, Lu X, et al. Improving Transformer-Based Speech Recognition Systems with Compressed Structure and Speech Attributes Augmentation[C]//Interspeech. 2019: 4400-4404.
29. Sainath T N, Pang R, Rybach D, et al. Two-pass end-to-end speech recognition[J]. arXiv preprint arXiv:1908.10992, 2019.
30. Zhang B, Wu D, Peng Z, et al. Wenet 2.0: More productive end-to-end speech recognition toolkit[J]. arXiv preprint arXiv:2203.15455, 2022.
31. Peng Y, Dalmia S, Lane I, et al. Branchformer: Parallel mlp-attention architectures to capture local and global context for speech recognition and understanding[C]//International Conference on Machine Learning. PMLR, 2022: 17627-17643.
32. Cui M, Deng J, Hu S, et al. Two-pass decoding and cross-adaptation based system combination of end-to-end conformer and hybrid tdnn asr systems[J]. arXiv preprint arXiv:2206.11596, 2022.

 doi:10.20944/preprints202412.0477.v1

33. Nozaki J, Komatsu T. Relaxing the conditional independence assumption of CTC-based ASR by conditioning on intermediate predictions[J].arXiv preprint arXiv:2104.02724, 2021.
34. Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.