Journal of Civil and Hydraulic Engineering
https://www.acadlore.com/journals/JCHE

An Intelligent Recording Method for Field Geological Survey Data in Hydraulic Engineering Based on Speech Recognition

Zuguang Zhang[1], Qiubing Ren[1*], Wenchao Zhao[1,2], Mingchao Li[1], Leping Liu[1], Yuangeng Lyu[1]

[1] State Key Laboratory of Hydraulic Engineering Intelligent Construction and Operation, Tianjin University, 300350 Tianjin, China
[2] Bei Fang Investigation, Design & Research Corporation Limited, 300222 Tianjin, China

* Correspondence: Qiubing Ren (qbren@tju.edu.cn)

Received: 06-30-2024 **Revised:** 09-16-2024 **Accepted:** 09-25-2024

Citation: Z. G. Zhang, Q. B. Ren, W. C. Zhao, M. C. Li, L. P. Liu, and Y. G. Lyu, "An intelligent recording method for field geological survey data in hydraulic engineering based on speech recognition," *J. Civ. Hydraul. Eng.*, vol. 2, no. 4, pp. 220–237, 2024. https://doi.org/10.56578/jche020403.

Abstract: Field data collection is a crucial component of geological surveys in hydraulic engineering. Traditional methods, such as manual handwriting and data entry, are cumbersome and inefficient, failing to meet the demands of digital and intelligent recording processes. This study develops an intelligent speech recognition and recording method tailored for hydraulic engineering geology, leveraging specialized terminology and speech recognition technology. Initially, field geological work documents are collected and processed to create audio data through manual recording and speech synthesis, forming a speech recognition training dataset. This dataset is used to train and construct a speech-to-text recognition model specific to hydraulic engineering geology, including fine-tuning a Conformer acoustic model and building an N-gram language model to achieve accurate mapping between speech and specialized vocabulary. The model's effectiveness and superiority are validated in practical engineering applications through comparative experiments focusing on decoding speed and character error rate (CER). The results demonstrate that the proposed method achieves a word error rate of only 2.6% on the hydraulic engineering geology dataset, with a single character decoding time of 15.5ms. This performance surpasses that of typical speech recognition methods and mainstream commercial software for mobile devices, significantly improving the accuracy and efficiency of field geological data collection. The method provides a novel technological approach for data collection and recording in hydraulic engineering geology.

Keywords: Hydraulic engineering; Geological survey; Intelligent speech recognition; Deep learning; Digital recording

1 Introduction

Field data collection and recording in geological surveys are crucial tasks in hydraulic engineering construction, characterized by their necessity and preeminence [1]. The raw data from hydraulic engineering geological surveys are critical for describing the on-site geological environment and serve as a primary source for geological big data [2], featuring large data volumes and diverse types. The complex work environment of hydraulic engineering geological surveys makes data collection by handwriting or typing in the field inconvenient, hindering the accurate and efficient gathering and recording of geological survey data. Therefore, designing and developing a novel method for field geological data recording that simplifies the data collection process is essential for reducing the difficulty of data acquisition and improving the accuracy and efficiency of hydraulic engineering geological data collection.

Since the 1970s, methods for geological data collection have undergone significant reform and innovation. Traditional methods involved recording data in field notebooks and later storing the collected data in databases for management, which often resulted in issues such as non-standardized recording formats and low collection efficiency [3]. To address these problems, the Queensland Geological Survey combined handheld devices with supporting software to develop a mobile GIS-based field geological data collection technology [4]. This approach offers advantages such as low power consumption and ease of operation, integrating data input, storage, management, and output. However, it also has drawbacks, including insufficient utilization of built-in sensors and cumbersome data collection processes [5].

In recent years, as mobile devices have become increasingly intelligent and hardware performance has rapidly improved, the variety of hardware sensors has also expanded. Utilizing these latest software environments and hardware devices to make field data collection more efficient and accurate is a growing trend in modern information technology. Mobile GIS-based field geological data collection technology has thus become a more efficient collection method [6]. Furthermore, the significant advancements in artificial intelligence (AI) have led to its widespread application across various fields. The use of AI-based mobile devices for data collection is simple and convenient, enhancing the efficiency of fieldwork for geologists [7, 8], and holds substantial significance for revolutionizing traditional field data collection methods.

The complex fieldwork environment in hydraulic engineering geological surveys makes geological investigation and data collection particularly challenging. For intelligent geological data collection, it is necessary to quickly and conveniently record observed data into devices by applying various advanced technologies to achieve efficient and convenient data recording. However, using mobile devices for field data collection and recording often requires one hand to observe samples while the other operates the mobile device. This sometimes requires additional tools such as magnifying glasses, geological hammers, and compasses [9], making single-handed operation challenging, prone to errors, and inefficient. Therefore, a data collection method that avoids occupying both hands is needed to make the geological data collection process more convenient, reducing the complexity and operational difficulty of field geological work.

With the rapid development of deep learning methods, human-computer interaction has gradually evolved from mouse-keyboard interaction and touchscreen interaction to natural language interaction [10, 11]. Using voice control for field geological data collection, where data is collected by voice control of mobile devices without manual operation, allows data to be converted to text through speech recognition, which is faster than manual text entry on mobile devices, significantly improving collection efficiency. Speech recognition technology has been widely applied in areas such as terminal control [12], software interaction [13], and meeting transcription [14]. Currently, open platforms like iFlytek and Baidu Voice provide technologies such as speech recognition, lexical analysis, and speech synthesis, which can support general field geological data observation and recording [15]. However, there are limitations in recognizing specialized geological terminology.

The current implementation methods for speech recognition mainly include offline and online speech recognition, each suited to different application scenarios [16]. Offline speech recognition is characterized by high accuracy and fast recognition speed, but its primary drawback is the limited scope of recognizable speech, as it can only recognize content within its built-in language models and acoustic model libraries [17]. Online speech recognition, on the other hand, has a broader recognition range, supporting multiple languages such as Chinese and English [18]. It is suitable for general tasks such as daily communication, software interaction, and intelligent device control, but its recognition accuracy for specialized geological terminology is relatively lower than offline methods, as it lacks a specialized geological lexicon. Using a general lexicon for field geological data collection often results in high rates of misrecognition. Considering the challenges of weak signals and low bandwidth in fieldwork environments, it is necessary to develop an offline speech recognition method tailored for hydraulic engineering geological survey scenarios. This would enable the verbal description of observed geological information in the field, which is then intelligently converted into text records, improving the efficiency of field recording and addressing the inefficiencies in field geological data collection.

In response to these challenges, this paper proposes an intelligent recording method for field geological survey data in hydraulic engineering based on speech recognition interaction. This method simplifies the data collection process in geological surveys, improving the accuracy and efficiency of data collection in hydraulic engineering geological surveys. Initially, geological survey text data were collected and used to create a geological survey speech dataset through manual recording and speech synthesis. Then, a speech recognition acoustic model tailored for geological surveys was trained based on the Conformer acoustic model architecture [19], and a geological survey-specific language model was trained using the N-gram algorithm [20]. Finally, the recognition performance of the proposed method was evaluated using CER and single-character decoding time as metrics, and its effectiveness was validated through comparative testing. Experimental results demonstrate that the proposed offline recognition method outperforms classical speech recognition models and mainstream commercial software, efficiently and accurately enabling the collection and recording of hydraulic engineering geological survey data, effectively addressing the current inefficiencies and error-prone nature of field geological data collection.

2 Hydraulic Engineering Geological Survey Specialized Speech Dataset

2.1 Data Collection

Hydraulic engineering geological terminology is characterized by its strong domain specificity, high level of specialization, low frequency of common usage, and a high occurrence of uncommon characters, making it difficult to be accurately recognized by mainstream speech recognition models and programs currently available on the market. Therefore, geological-related corpora are used to construct a speech recognition dataset for hydraulic engineering

geological surveys, which is then used to train the speech recognition model. The geological specialized speech data includes audio and its corresponding text, with the text data sourced from various geological professional materials, mainly including hydraulic engineering geological monographs, textbooks on hydraulic engineering geology, fieldwork specifications for geological surveys in water conservancy and hydropower engineering, geological borehole exploration entry materials, and engineering geological survey reports.

To convert the text data into dataset labels for input into the acoustic model, the following four preprocessing tasks are required: (1) Content Extraction: Convert the geological specialized materials into editable text and manually remove information sections with low relevance to geological surveys. (2) Text Cleaning: This includes removing meaningless data such as spaces and punctuation marks, replacing Arabic numerals, letters, and other non-Chinese characters with Chinese characters, and removing meaningless data such as spaces and punctuation marks [21]. (3) Sentence Segmentation: Break down long sentences into multiple short sentences of 2 to 25 characters based on key punctuation marks such as commas, periods, and semicolons, forming 16,352 sentences of text data. (4) Word Segmentation: For each short sentence, segment the sentence into word information based on sentence components and vocabulary attributes, using spaces as delimiters, while retaining specialized geological terms. The overall preprocessing workflow is shown in Figure 1.

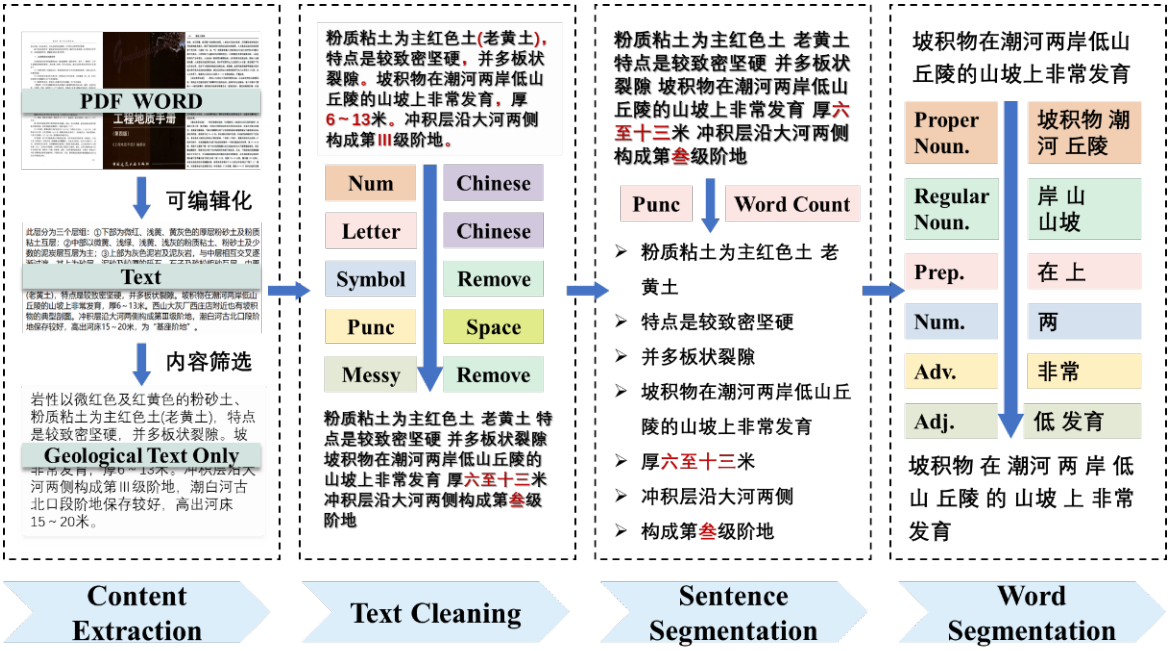

Figure 1. The preprocessing of text information

To obtain audio files matching the text data, audio data is generated through two methods: human voice recording and speech synthesis. Human voice recording is carried out by 15 recording personnel who read the text data content in Mandarin in a quiet environment, recording each line using a mobile phone microphone. Speech synthesis is done by randomly using multiple Text to Speech (TTS) models [22] to generate audio sentence by sentence, simulating multiple different speakers to enhance the diversity of the audio data, thereby ensuring the generalization ability of the speech recognition model. The TTS models used for speech synthesis are shown in Table 1.

Table 1. TTS model for speech synthesis

Model	Speaker Type	Language
Speedyspeech CSMSC	Single	Chinese
Fastspeech2_CSMSC	Single	Chinese
Fastspeech2 LJSPEECH	Single	English
Fastspeech2_AISHELL3	Multiple	Chinese
Fastspeech2 VCTK	Multiple	English
Fastspeech2_MIX	Multiple	Chinese/English
Tacotron2_CSMSC	Single	Chinese
Tacotron2 LJSPEECH	Single	English

2.2 Speech Dataset Construction

Referring to the format of the Aishell dataset [23], the text sequences and audio sequences are mapped to one-to-one relationships, and placed line by line in the record file, constructing the speech recognition dataset for hydraulic engineering geological surveys. This dataset contains 16,532 audio files with a total duration of 13.53 hours. In terms of professional domain recognition, this dataset can ensure recognition accuracy while effectively maintaining the model's generalization performance due to its large volume, thereby ensuring its anti-interference stability during actual use. Considering the large scale of the dataset, it is randomly divided into training, validation, and test sets according to a ratio of 18:1:1 [24] to simulate field geological survey conditions as closely as possible, ensuring the reference value of the model evaluation. The process of constructing the dataset is shown in Figure 2.

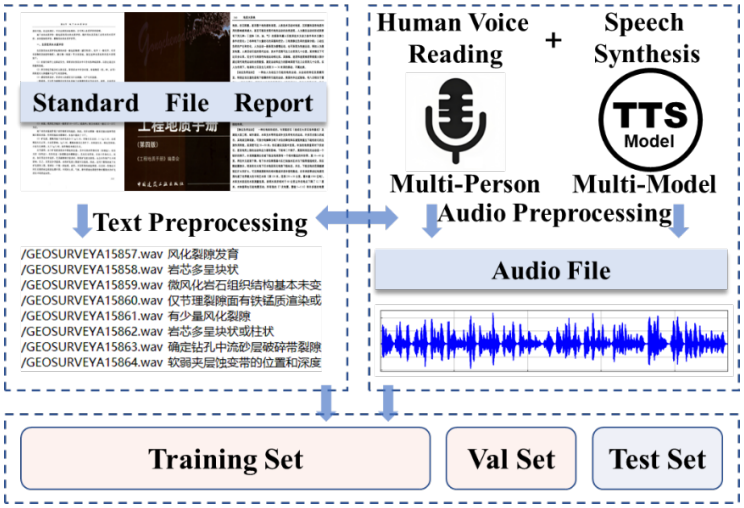

Figure 2. Building of the dataset

3 Hydraulic Engineering Geological Survey Specialized Speech Recognition Model

3.1 Acoustic Signal Conversion

The conversion of acoustic signals is the fundamental task of speech recognition. According to the waveform of the speech signal, acoustic signals are converted into effective acoustic features to describe and capture different speech signals. The effectiveness of feature extraction directly impacts the accuracy of subsequent speech recognition.

To fully extract the features of the human voice in the speech signal while reducing the impact of background noise on linguistic information, this study uses Mel-scale Frequency Cepstral Coefficients (MFCCs) as the features of the input acoustic signal, which are widely used in speech recognition [25]. MFCCs are cepstral parameters extracted in the Mel-scale frequency domain. The Mel scale describes the nonlinear characteristics of human ear frequency perception, and its relationship with frequency is as follows:

$$\mathrm{Mel}(f) = 2595 \times \lg\left(1 + \frac{f}{700}\right) \tag{1}$$

Based on the audio files, MFCC features are obtained through the following eight steps:

(1) **Pre-emphasis.** The high-frequency part of the speech is easily lost after being emitted by the human vocal organs. Pre-emphasis is used to compensate for the amplitude of the high-frequency part of the speech signal. The calculation formula is as follows:

$$x_{emp}(n) = x(n) - \alpha \cdot x(n-1) \tag{2}$$

where, $x(n)$ is the n-th sample point of the input speech signal, and α is the pre-emphasis coefficient.

(2) **Framing.** Speech signals exhibit short-term stationarity, so a segment of speech over a short period is taken as a frame. To ensure continuity of the speech signal after framing, a portion of the previous frame is retained in the next frame, called the frame shift. Typically, the frame shift is set to 10ms, and the frame length is set to 25ms.

(3) **Windowing.** Windowing is associated with framing and smooths the frame through a window function. A Hamming window is used to retain the frequency characteristics of the speech signal effectively, and its function is as follows:

$$w(n) = (1-a) - a \cdot \cos\left(\frac{2\pi n}{N} - 1\right) \tag{3}$$

where, a is 0.46, and N is the window length.

(4) **Fourier Transform.** The Fourier Transform converts the signal from the time domain to the frequency domain for spectrum analysis. The Short-Time Fourier Transform (STFT) is suitable for short-term stationary signals, and its expression is:

$$T_t(k) = \sum_{n=0}^{N-1} x(n) \cdot w(i) \cdot b_k(i) \tag{4}$$

where, x represents the speech signal, t is the sequence number of the frame, w is the window function, and b represents the transformation coefficient at the k-th frequency point.

(5) **Mel Filter Bank.** The Mel filter bank is used to extract frequency bands in the low-frequency region, and its function is represented as shown in Figure 3. The response of the m-th filter at the k-th frequency point can be calculated by the following formula:

$$H_m(k) = \begin{cases} 0, k < f(m-1) \\ \frac{k-f(m-1)}{f(m)-f(m-1)}, f(m-1) \le k < f(m) \\ \frac{f(m+1)-k}{f(m+1)-f(m)}, f(m) \le k < f(m+1) \\ 0, k \ge f(m+1) \end{cases} \tag{5}$$

where, $f(m)$ is the center frequency of the filter.

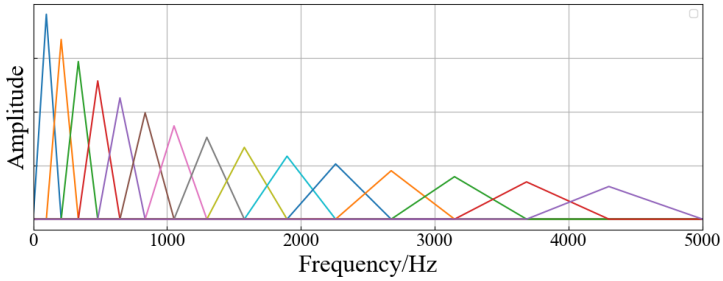

Figure 3. Equal area Mel filter bank

(6) **Logarithmic Power Operation.** The amplitude spectrum obtained from the Fourier Transform is squared to obtain the short-term power spectrum. By multiplying and accumulating through the filter, the logarithm is taken to obtain the logarithmic power spectrum, which relatively amplifies the low-frequency signal. The resulting feature is the F-bank feature of the audio, and its expression is as follows:

$$s(m) = \ln\left(\sum_{k=0}^{N-1} H_m(k) \cdot |T_t(k)|^2\right), 0 < m < M \tag{6}$$

where, M is the total number of filters.

(7) **Discrete Cosine Transform (DCT).** The logarithmic energy obtained above is input into the DCT to concentrate the signal's energy, and the MFCCs are obtained according to the following formula:

$$C(i) = \sum_{m=0}^{N-1} s(m) \cdot \cos\left(\frac{\pi i(m-0.5)}{M}\right), i = 1, 2, \cdots, L \tag{7}$$

where, L represents the order of the MFCCs, usually set to 12∼16.

(8) **Dynamic Differential.** The DCT only obtains static MFCC features. Dynamic MFCC features can be obtained by taking the differential of the static features. The calculation method for the first-order differential coefficient is as shown in Eq. (8), and the second-order differential coefficient is obtained by repeatedly substituting the first-order differential coefficient into the equation. The combination of static, first-order, and second-order coefficients results in the complete MFCC features.

$$d_t = \frac{\sum_{n=1}^{\mathrm{N}} n(C(t+n) - C(t-n))}{2\sum_{n=1}^{N} n^2} \tag{8}$$

where, t is the sequence number of the frame, and N represents the frame sequence difference of the first-order derivative, which can be 1 or 2.

The overall extraction process of MFCCs features is shown in Figure 4.

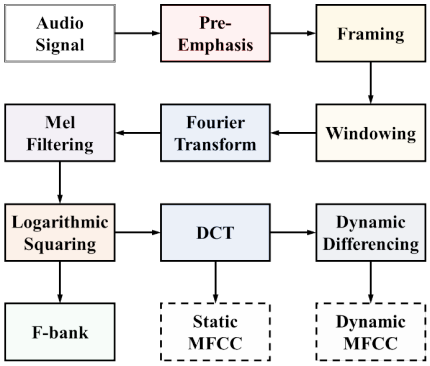

Figure 4. The calculation process for extracting MFCCs

3.2 Conformer Acoustic Model

The acoustic model is one of the most important modules in a speech recognition system, directly affecting the system's performance [26]. The acoustic model takes the vector sequence obtained from feature extraction of the speech signal as input and establishes a mapping relationship between speech features and phonemes to obtain the probability of the speech waveform corresponding to the model's output speech signal.

The Conformer model combines the strengths of the Transformer [27] and Convolutional Neural Network (CNN), leveraging CNN's advantage in capturing local features while retaining Transformer's capability of acquiring long-range dependencies. This enhances the network's ability to model both global and local dependencies simultaneously, demonstrating outstanding performance in recognition accuracy, inference speed, and model parameter size.

The core component of the Conformer model is the Conformer block, which mainly consists of four modules: the feedforward network, multi-head attention mechanism module, convolution module, and the second feedforward network. Its structure is similar to a macaron structure, where the two feedforward networks each contribute half of the Conformer's output. The speech features xi are processed by the Conformer model as follows:

$$\begin{aligned} X_i' &= x_i + 0.5FFN\left(x_i\right) \\ X_i'' &= x_i' + MHSA\left(x_i'\right) \\ X_i''' &= x_i'' + MHSA\left(x_i''\right) \\ Y_i &= LN\left(x_i''' + 0.5FFN\left(x_i'''\right)\right) \end{aligned} \tag{9}$$

where, FFN represents the feedforward network, $MHSA$ stands for the multi-head self-attention module, and LN represents layer normalization. Residual connections are used between each module.

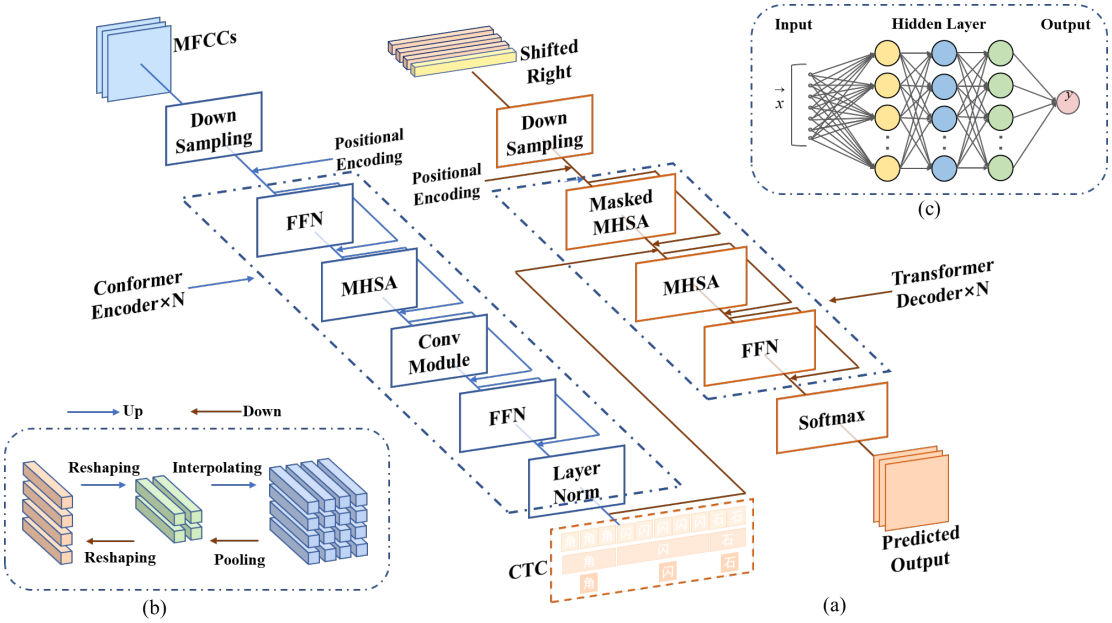

Figure 5. Structure of conformer model

In the encoding stage, the audio is first processed by a convolutional subsampling layer, and then multiple Conformer blocks are used for further processing, as shown in Figure 5. In the figure, (a) represents the Conformer model structure, (b) shows the upsampling and downsampling operations for aligning the input feature space, and (c) illustrates the fully connected layer structure of the feedforward network.

3.2.1 MHSA mechanism

The multi-head attention mechanism [28] module replaces the positional encoding in the Transformer with the relative sinusoidal positional encoding from Transformer-XL, allowing the attention module to generalize better across different input lengths and making the generated encoder more stable to changes in audio sequence length.

The self-attention mechanism is a variant of the attention mechanism, reducing the dependency on external information and being more adept at capturing the internal correlations within data or features. In the self-attention mechanism, the Q (query) vector, K (key) vector, and V (value) vector all originate from the same input sequence X obtained from character encoding. This means that after the model reads the input information, it determines the most important information based on the input itself. The calculation process is shown in Figure 6.

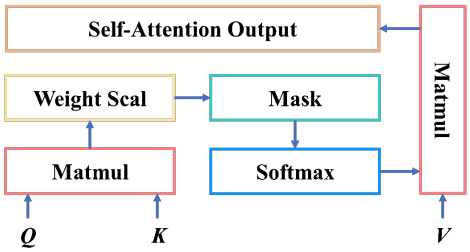

Figure 6. Self-attention mechanism

In this study, the correlation between Q and K is first calculated using scaled dot-product to capture the local relationships within the input sequence. To handle the scale differences between Q and K, their scores are scaled to stabilize the variance of the attention weights. The results are then normalized using the softmax function to highlight the weights of important elements, resulting in the attention weight coefficients. Finally, the V values are weighted and summed according to the attention weight coefficients to obtain the Self-Attention Value, with the calculation formula as follows:

$$Attention(X) = softmax\left(\frac{QK^T}{\sqrt{\dim_k}}\right) \cdot V \tag{10}$$

In the formula, Q, K, and V are obtained from the original input sequence X by multiplying it with the linear transformation matrices W_Q, W_K, and W_V respectively; $\dim_k$ represents the dimension of K and V.

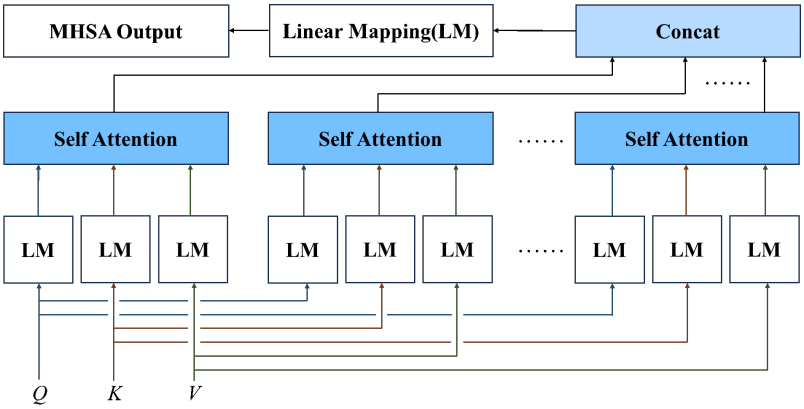

Figure 7. MHSA mechanism

When encoding the information at the current position, the self-attention mechanism tends to focus excessively on its own position, making it less effective than CNN in capturing useful information. To address this issue, the multi-head attention mechanism is used. It applies h (generally h=8) different sets of independently learned linear transformation matrices to the input sequence X to transform Q, K, and V. These h sets of transformed Q, K, and V then undergo self-attention mechanism operations in parallel, resulting in multiple Self-Attention Values. Finally,

the h Self-Attention Values are concatenated and transformed through another learnable linear transformation matrix W_0 to produce the final Multi-Head Self-Attention output. The overall calculation process is shown in Figure 7.

3.2.2 MHSA mechanism

The convolution module consists of five parts: a pointwise convolution layer, a Gated Linear Unit (GLU) activation function, a depth-wise convolution layer, a Swish activation function, and a second pointwise convolution layer. The structure is shown in Figure 8.

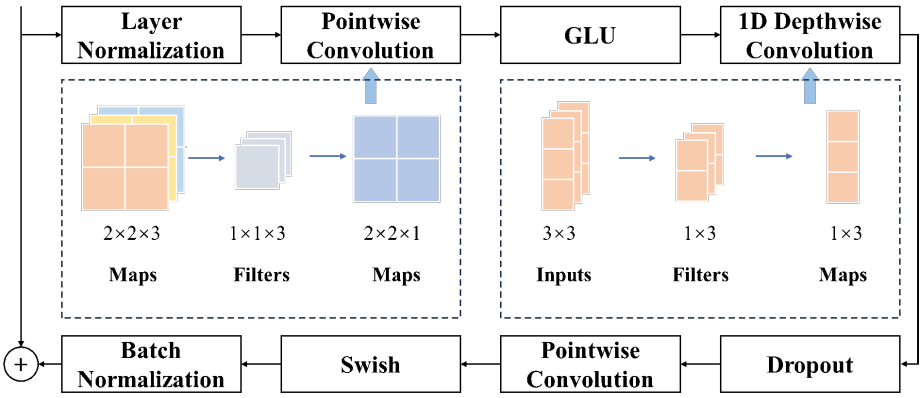

Figure 8. Convolution module of Conformer

Depth-wise convolution and pointwise convolution are used together as an efficient combination to replace traditional CNNs, effectively reducing the number of network parameters and improving computational efficiency. In depth-wise convolution, one convolution kernel is responsible for one channel, meaning that each channel is convolved by only one convolution kernel, focusing solely on the dependencies within the sequence in each channel, without considering dependencies between different channels. On the other hand, pointwise convolution is very similar to regular convolution operations, with a convolution kernel size of $1\times1\times M$, M equals to the number of channels in the previous layer. Its convolution operation weights and combines the feature maps from the previous step along the depth direction to generate a new set of feature maps with the same number as the convolution kernels, thus focusing on dependencies between different channels while ignoring intra-channel dependencies.

The GLU activation function is an activation function used in neural networks that incorporates a gating mechanism, helping the network better capture long-term dependencies in sequence data. The GLU activation function is defined as follows:

$$GLU\left(x_i\right) = x_i \otimes Sigmoid\left(g\left(x_i\right)\right) \tag{11}$$

where, X is the input vector, $\otimes$ represents element-wise multiplication, $g\left(x_i\right)$ is the intermediate vector obtained through the convolution layer, and the Sigmoid function is defined as follows:

$$Sigmoid(x) = \left(1 + e^{-x}\right)^{-1} \tag{12}$$

The Swish activation function has characteristics such as a lower bound, smoothness, and non-monotonicity, effectively addressing the problems of gradient vanishing and neuron death encountered in the ReLU activation function. Its calculation formula is:

$$Swish\left(x_i\right) = x_i \cdot Sigmoid\left(\beta x_i\right) \tag{13}$$

where, β is a trainable parameter.

3.2.3 Feedforward module

A FFN is a fully connected feedforward neural network consisting of two fully connected layers and a nonlinear activation function. The feedforward module in the Conformer model uses the Swish activation function. The first fully connected layer is used for dimensionality expansion, and the second fully connected layer is used for dimensionality reduction. This design aims to perform nonlinear transformations and mappings of the embedding vectors, allowing the model to learn more abstract features. Additionally, layer normalization, the Dropout mechanism, and residual summation are introduced to accelerate network training, improve the model's generalization ability, and alleviate the gradient vanishing problem [29]. The feedforward module's calculation process is shown in Figure 9.

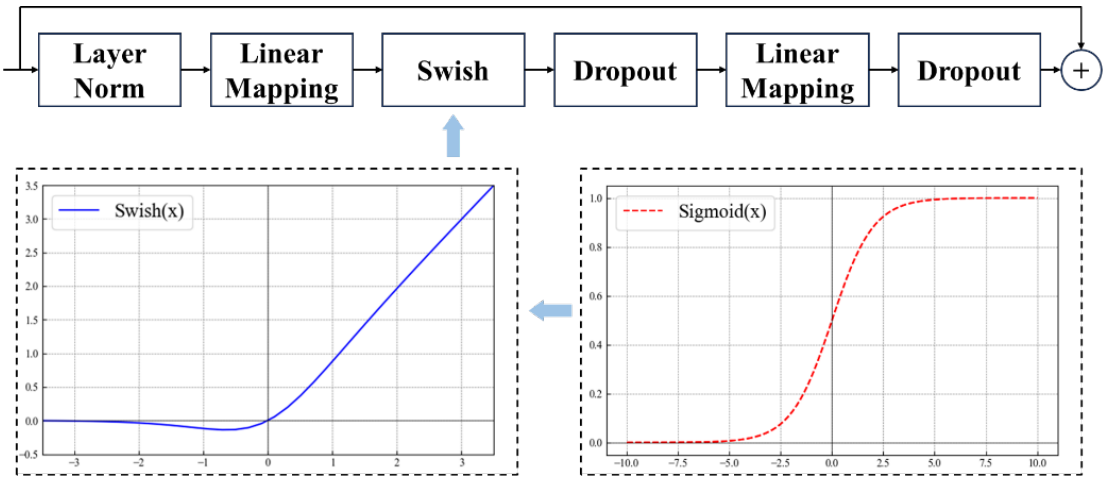

Figure 9. Calculation process of feedforward module

3.2.4 Loss function

During training, the Connectionist Temporal Classification (CTC) loss and attention loss are used to supervise the training of the CNN and Transformer branches, respectively [30], to obtain features that exhibit both CNN and Transformer characteristics. During inference, the outputs of these two classifiers are summed to serve as the prediction results. The cross-entropy function is expressed as follows:

$$L_{CE} = -\sum_{i=1}^{n} T_i \log\left(P_i\right) \tag{14}$$

where, Ti represents the actual probability distribution, and Pi represents the predicted probability distribution, both of which are normalized using the softmax function.

The CTC loss function measures the gap between the predicted result and the actual result when no annotated text is provided. For an input sequence x of length T and an output sequence y of length U, the cross-entropy loss function is:

$$L_{CTC}(x, y) = -\log P(y \mid x) \tag{15}$$

where, $P(y \mid x)$ represents the probability of the output sequence y given the input sequence x.

The attention loss function measures the model's focus on different positions in the input sequence. For an input sequence x of length T, an output sequence y of length U, and corresponding attention weights a, the loss function is expressed as:

$$L_{AT}(x, y, a) = -\sum_{t=1}^{T}\sum_{u=1}^{U} a_{u,t} \log y_u^t \tag{16}$$

where, $a_{u,t}$ represents the degree to which the model focuses on the t-th position in the sequence when predicting y_u; y_u^t represents the predicted probability of the u-th character. The logarithmic likelihood of these two parts is combined in a weighted sum.

The loss function for the joint training process of decoding usually includes both cross-entropy loss and attention loss, calculated as follows:

$$L_{CTC-AT} = \lambda L_{CTC}(x, y) + (1 - \lambda) L_{AT}(x, y, a) \tag{17}$$

where, λ is used to set the weights of the two types of loss, generally set to 0.3.

3.3 N-gram Language Model

The purpose of the language model is to further decode the output of the acoustic model, converting the speech signal into the corresponding text sequence [31], and it is used for model CTC beam search decoding prediction. The CTC beam search method involves having B-Size candidate sequences, and at each time step, generating a new set of the best B-Size candidate sequences. The final result is the sequence with the highest probability among the B-Size candidate sequences. The principle of CTC beam search is shown in Figure 10.

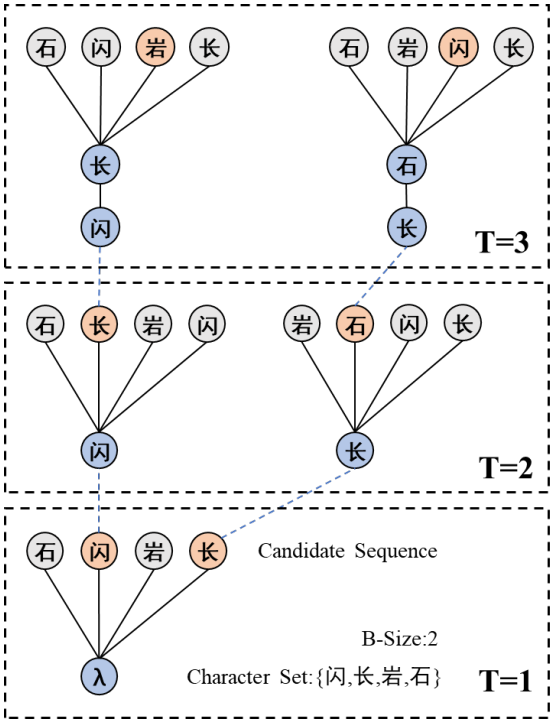

Figure 10. CTC beam search

A language model is a knowledge representation of a sequence of words. By using a statistical-based language model and processing a large specific text corpus, the probability distribution of a given word sequence can be obtained, further determining the likelihood of a text sequence and finally obtaining the accurate prediction result of the text sequence with the highest probability.

Currently, the N-gram algorithm is the most commonly used language model in speech recognition systems. Its basic idea is to introduce the Markov assumption, considering that the probability of each word occurring is only related to several preceding words, and it ignores longer-distance context. Based on this assumption, the joint probability of words in a sentence is calculated to determine the likelihood of the sentence. N is usually an integer between 1 and 5, and when N is 3, the corresponding Tri-gram algorithm is as follows:

$$\begin{aligned} P(S) &= P\left(w_1, w_2, \cdots, w_n\right) \\ &= \prod_{i=1}^{n} P\left(w_i \mid w_{i-1} \cdots w_1\right) \\ &= P\left(w_1\right) P\left(w_2 \mid w_1\right) \cdots P\left(w_n \mid w_{n-1} w_{n-2}\right) \end{aligned} \tag{18}$$

where, S represents the sentence, and w_i represent the word sequence forming the sentence.

In scoring the candidate results of decoding, in addition to the score of the acoustic model, there are also additional language model scores and length penalty scores. Let W be the decoding result and X be the input speech. The final score is calculated as follows:

$$score = P_{am}(W \mid X) \cdot P_{lm}(W)^{\alpha} \cdot |W|^{\beta} \tag{19}$$

where, am represents the acoustic model score, lm represents the language model score, α and β are the set hyperparameters.

Field geological survey text data in hydraulic engineering involves many specialized terms and rare characters. The current N-gram language models are mostly trained based on common Chinese text [32], such as the large-scale Giga Chinese model (zh_giga.no_cna_cmn.prune01244.klm, 2.75 GB) and the lightweight People's Daily 2014 corpus model (people_2014_corpus_char.klm, 0.14 GB). These language models, due to their corpora not fully encompassing the content of the hydraulic engineering geological domain, result in higher CER in recognition and may lead to situations where specific characters cannot be displayed in the recognition results, as shown in Figure 11.

To solve the problem of missing characters and further optimize the speech recognition results, geological specialized text data was converted to form a geological specialized Chinese text corpus. Additionally, by combining the ToRCH2009 Modern Chinese Balanced, ToRCH2014 Modern Chinese Balanced, ToRCH2019 Modern Chinese

Balanced, and The BFSU DiSCUSS four open-source general Chinese corpora, the KenLM toolkit was used to train the model and compress it into binary format, forming a specialized language model for hydraulic engineering field geological surveys. To ensure that the model accurately and comprehensively learns the probability distribution of text sequences, the N value of the N-gram model was set to 5 [33].

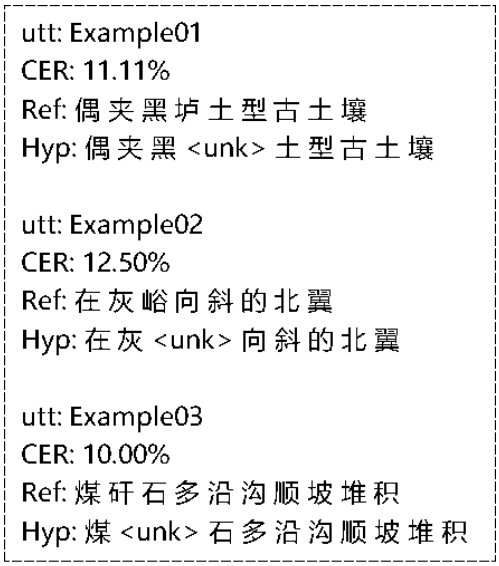

Figure 11. Missing characters in recognition results

3.4 Model Fine-Tuning

Model fine-tuning is an important technique for addressing the challenge of limited computational power under large-scale data conditions [34]. It allows a general model to quickly adapt to the specific needs of a domain using a small amount of labeled data, thereby improving model performance, increasing recognition speed, and enhancing the accuracy of recognizing targets in specific fields. Additionally, training on top of an existing large-scale dataset helps the model fully learn acoustic features, enhancing its generalization capability and anti-interference stability [35].

The task of specialized speech recognition for hydraulic engineering field surveys involves numerous geological terms and related expressions that are rarely encountered in everyday language. The current mainstream speech recognition models are trained on datasets mostly derived from publicly available common language materials, where geological survey-specific vocabulary and speech account for a small proportion, resulting in weak recognition capabilities for geological domain-specific speech. Therefore, it is necessary to fine-tune the existing mainstream speech recognition models for the geological domain to improve their performance in this field.

When fine-tuning the geological-specific speech recognition model, the WenetSpeech [36] large-scale Chinese speech dataset is first used to train the model as a base model for transfer learning. Then, fine-tuning is performed on the specialized dataset for geological field surveys, updating all parameters of the entire network during the fine-tuning process. Cepstral Mean and Variance Normalization (CMVN) features [37] are used to normalize the speech signal and remove noise, while CTC loss and attention loss are used for supervision.

3.5 Evaluation Metrics

To ensure the recognition performance of the model, this study uses decoding speed and CER as metrics to evaluate the performance of the proposed speech recognition model. The time taken to decode a single character represents the time required by the model to decode the audio corresponding to a single character, and it is used to assess the model's recognition efficiency. The smaller the unit decoding time, the higher the model's recognition efficiency. CER is used to evaluate the degree of difference between the predicted text and the original reference text, reflecting the accuracy of text recognition. The lower the CER, the better the model's recognition performance. The calculation method is as follows:

$$\mathrm{CER} = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C} \tag{20}$$

where, S represents the number of substitutions in the predicted result relative to the original text, D represents the number of deletions in the predicted result, I represents the number of insertions in the predicted result, N represents the total number of characters in the original text, and C represents the number of correctly recognized characters in the predicted result.

4 Model Validation

4.1 Data Preprocessing

The audio editing software WavePad was used to trim all audio files, removing silent sections. FFmpeg was used to modify the sampling rate of all audio to 16 kHz, lock the channel to mono, and convert the files to WAV format. The parameter information is shown in Table 2.

Table 2. Speech data parameter information

Sampling Rate	Bit Depth	Channel	File Format
16 kHz	16 bit	Mono	WAV

After processing the audio, a file-by-file reading and matching method was used to generate a data list file, which was used to index the audio files and their corresponding text information. Each line of data includes the relative path of the audio file and the annotated content corresponding to that audio file. The format is shown in Figure 12.

```
/GEOSURVEYA15857.wav 风化裂隙发育
/GEOSURVEYA15858.wav 岩芯多呈块状
/GEOSURVEYA15859.wav 微风化岩石组织结构基本未变
/GEOSURVEYA15860.wav 仅节理裂隙面有铁锰质渲染或矿物略有变色
/GEOSURVEYA15861.wav 有少量风化裂隙
/GEOSURVEYA15862.wav 岩芯多呈块状或柱状
/GEOSURVEYA15863.wav 确定钻孔中流砂层破碎带裂隙密集带
/GEOSURVEYA15864.wav 软弱夹层蚀变带的位置和深度
```

Figure 12. Format of the data list file

Afterward, all characters contained in the dataset are counted to generate a vocabulary file. Finally, the number of frames, mean, and standard deviation are calculated for CMVN operations, with the default being to use all speech data to calculate the mean and standard deviation.

Finally, feature extraction is performed on all audio files to obtain their MFCC features, and JSON-formatted data feature list files are generated for both the training set and the test set for model training. In practice, the features of the same phoneme may differ due to the influence of different microphones, recording environments, and audio channels. Through CMVN operations, standard features with a mean of 0 and a variance of 1 can be obtained, thereby improving the anti-interference stability of the acoustic features. The MFCCs spectrogram is shown in Figure 13.

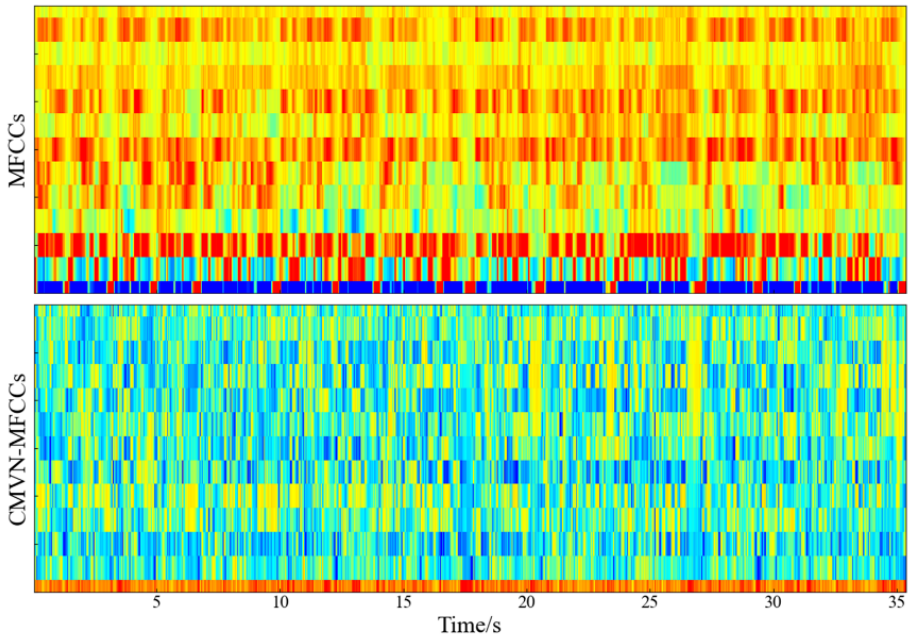

Figure 13. MFCC spectrogram

To further enrich the diversity of the data and thereby enhance the model's generalization and anti-interference capabilities, small random perturbations were added to the original audio during model training to generate new audio for data augmentation. The augmentation methods include noise perturbation, speed perturbation, volume perturbation, and SpecAugment [38]. The overall data preprocessing process is shown in Figure 14.

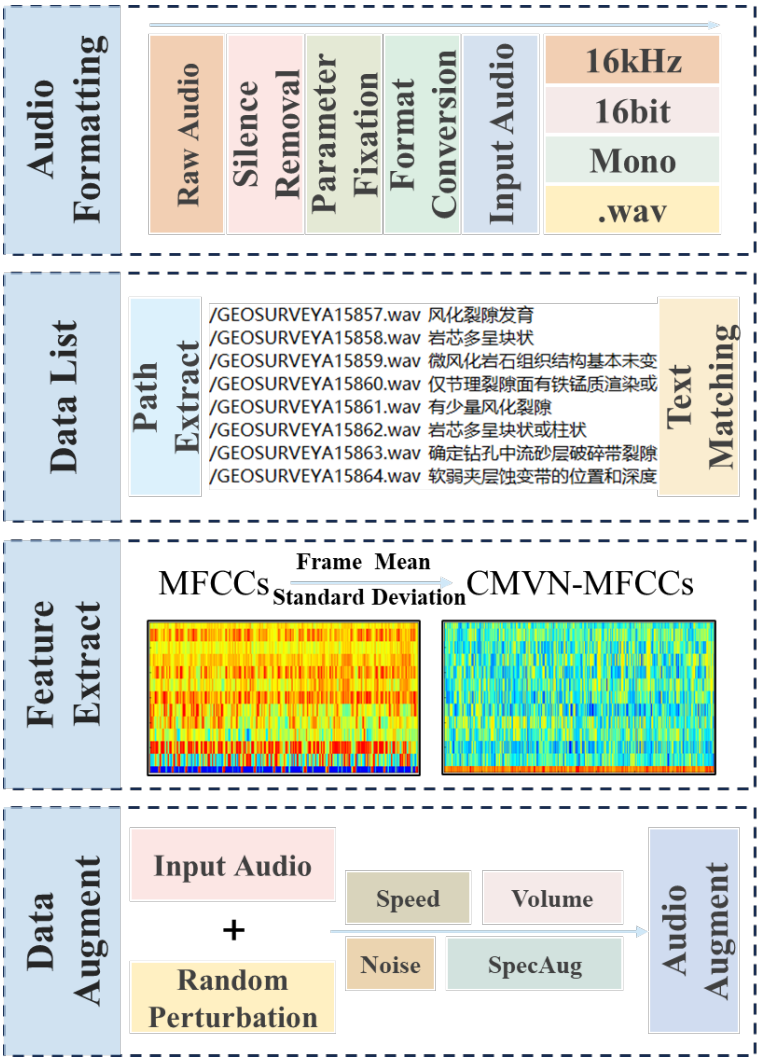

Figure 14. Data preprocessing process

4.2 Hyperparameter Adjustment

Table 3. Model training hyperparameter settings

Parameter	Value
Number of epochs	200
Batch size	8
Gradient accumulation	4
Optimizer type	Adamm [39]
Initial learning rate	0.001
Learning rate decay	0.1
Minimum learning rate	1.0e-5
Weight decay coefficient	1.0e-6
MFCC size	40
Input audio length	0.5∼20

All model experiments were conducted on the same computer equipment, with the operating system being Windows 10 Professional, the programming language Python 3.11, the CPU being Intel® Core™ i5-13490F, and the GPU being NVIDIA® GeForce RTX™ 3070.

In this study, the configuration parameters for the training and validation stages of the Conformer acoustic model were determined based on relevant literature in the field [40]. The specific parameter information is shown in Table 3. The number of epochs represents the total number of training cycles for the model; the batch size is the number of audio samples the model processes simultaneously in one cycle; gradient accumulation is used to achieve the effect of expanding the batch size; the initial learning rate and its decay determine the convergence speed of the objective function; and MFCC size represents the dimensions of the MFCCs.

4.3 Recognition Result Analysis

The Conformer model was trained for 200 epochs, with validation after each epoch. The model converged at the 87th epoch and achieved the best validation performance at the 194th epoch, with a CER of 4.38%, as shown in Figure 15. After combining with the geological specialized language model, the CER was reduced to 2.36%, and the single-character decoding time was 15.5 ms.

To further evaluate the performance of the proposed model, the constructed hydraulic engineering field geological survey speech dataset was used as an example, with other advanced deep learning speech recognition models selected for comparison under the same configuration parameters as the hydraulic engineering geological survey speech recognition model. CER and decoding time were used as evaluation metrics. The decoding times and CERs of each model are listed in Table 4. The comparison results show that the proposed language model, due to its specialization in the geological domain, effectively reduced the CER compared to the Giga Chinese model. It can be seen that this method shows no significant difference in single-character decoding time compared to other methods, meeting the requirements for real-time speech recognition; the CER metric is superior to other methods, indicating that it is more suitable for the speech recognition tasks involved in hydraulic engineering field geological survey data collection. Additionally, fine-tuning and loading the specialized language model significantly improved the recognition performance of the Conformer model, further demonstrating the advantages of the proposed method.

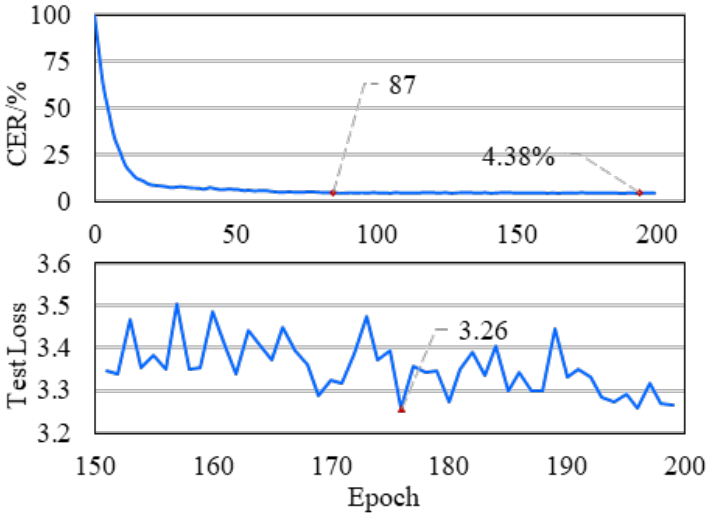

Figure 15. Validation effect and test loss of Conformer model

Table 4. Comparison results of speech recognition models

Acoustic Model	Fine-Tuning	Language Model	Single Character Decoding Time/ms	CER/%
Conformer	Yes	Geological Survey	15.5	2.6
Conformer	Yes	Giga Chinese	15.4	3.9
Conformer	No	Giga Chinese	15.8	3.7
Deepspeech [41]	No	Giga Chinese	29.5	5.2
Transformer	No	Giga Chinese	19.1	8.1
Efficient Conformer [42]	Yes	Giga Chinese	15.1	3.3
Squesezeformer [43]	Yes	Giga Chinese	21.6	17.1

4.4 Comparison with Mainstream Commercial Software Recognition Performance

To directly demonstrate the recognition performance of the proposed model, actual field data recording speech was used as a sample, comparing the predicted results of this model with the output of mainstream commercial software. This speech sample contains a large number of geological terms, with a duration of 23 seconds and a text length of 95 characters. The software and their recognition results, along with the CER, are shown in Table 5.

Table 5. Comparison of recognition performance between our model and mainstream commercial software

Model/Software	Network Status	Recognition Result	CER/%
Actual Corpus	-	花岗岩 浅肉红色 粗粒结构 块状构造 表层岩体呈强风化状 岩体强度较低 锤击易碎 矿物成分以长石 石英 云母和角闪石为主 坡面覆盖层为碎石土 碎石含量百分之十至十五 厚零点二至一米 植被发育 边坡自然坡度四十至四十五度	-
Our Model	Offline	花岗岩 浅肉红色 粗粒结构 块状构造 表层岩体呈强风化状 岩体强度较低 锤击易碎 矿物成分以长石 石英 云母和角闪石为主 坡面覆盖层为碎石土 或石含量百分之十至十五 至零点二至一米 六被发育 边坡自然坡度四十至四十五度	**3.16**
iOS-15.3.1 Native Keyboard	Offline	花岗岩 切肉红色 处理结果 块状构造 表层人体盛强风化妆 人体强度较低 锤击一岁 矿物成分隐藏式 石英 云母和角闪石为主 和面覆盖成为岁时图 岁时含量10%至15 号0.2日一鸣 植被法语 边和自然坡度40-45度	28.42
iOS-15.3.1 Native Keyboard	Offline	花岗岩 浅RAW红色 处理结果 块状构造 表层严提成强风化妆 掩体强度较低 锤击一岁 矿物成分隐藏式 石英 云母和角闪石为主 和面覆盖成为岁时图 睡狮含量10%至15 号0.2日一鸣 植被法语 边和自然过渡40-45度	29.47
Baidu Custom Input Method v8.2.39.795	Offline	花岗岩 前肉红色 初历结构 块状构造 表层盐体城强风化状 颜体强度较低 垂肌易碎 矿物成分已常十 十英 云母和脚产十为主 剖面覆盖曾为碎石土 碎石含量10%至15 后0.2至一米 植被发育 边坡自然过度40至45度	20.00
Baidu Custom Input Method v8.2.39.795	Offline	花岗岩 浅肉红色 粗粒结构 块状构造 表层岩体呈强风化状 岩体强度较低 锤基易碎 矿物成分以常识 石英 云母和角闪石为主 坡面覆盖曾为碎石土碎石含量10%至15% 后0.2至1米 植被发育 边坡自然坡度40至45度	5.26
Sogou Custom Input Method V8.31.22	Offline	八钢研 潜入红色 出力结构 会撞构造 表曾岩体城墙丰华庄 岩体强度叫爹 锤基易碎 矿物成分异常是 石英 韵母和脚闪视为主 画面覆盖层为说实图 碎石含量百分之十至十五 号零点二之一米 职位范玉 斌和自然过度四十至四十五度	41.05
Sogou Custom Input Method V8.31.22	Offline	花岗岩 浅肉红色 粗粒结构 块状构造 表层岩体呈强风化状 岩体强度较低 锤击易碎 矿物成分以常识 石英 云母和角闪石为主 泼面覆盖成为碎石土 碎石含量10%～15 后0点2～1m 植被发育 边坡自然坡度40 45度	6.32
Deepspeech2-Aishell	Offline	华东+ 铁路红鹤 湖里结构 坏狗狗咬 表成人体横行中包括 人体强度较低 惠及一组 矿物成分+行石 石英 云+和较少食为主 国灭复盖成为废食土 退食含量百分之十至十五 后零点二至一米 直被八日 将+自然过渡四十至四十五部	51.58
Conformer-Wenetspeech	Offline	花冈岩 浅肉红色 粗+结构 +状构造 表层岩体++风化妆 掩体强度较低 垂++碎 矿物成分以长时 +英 云母和脚++为主 ++覆盖成为碎石土 碎石含量百分之十至十五 后零点二十+米 +被发育 ++自然坡度四十至四十五度	25.26

It can be seen that the CER of the proposed model is 3.16%, which is 2.10% lower than that of the Baidu Input Method in online mode. Compared to the offline versions of Baidu, iOS, and Sogou Input Methods, the CER is reduced by 16.84%, 25.26%, and 37.89%, respectively, showing better performance than the existing mainstream commercial software. For mainstream commercial software, their online recognition is more geared toward everyday conversational dialogue. Although their dataset size is larger than that used in the proposed method, the total number of geological terms is low, and their proportion is small during training, resulting in poor model performance in the related field, leading to more recognition errors, especially when geological terms are misrecognized as common

conversational words. Notably, the proposed model shows a significant advantage in offline recognition, effectively recognizing geological terms, meeting the needs for recording speech data into text during field geological data collection.

5 Conclusions

This study proposed a method for the collection and recording of hydraulic engineering geological survey data in the field, based on speech recognition interaction, to meet the demands for digital and intelligent data collection in hydraulic engineering geological surveys. The proposed method effectively improves the efficiency and accuracy of field data collection for hydraulic engineering geological surveys. The main conclusions are as follows:

(1) Various geological professional materials were compiled to form a specialized speech dataset for hydraulic engineering geological field surveys. This dataset includes 16,352 sentences of geological survey professional text data, totaling 218,498 characters, and 13.53 hours of speech data generated through both human voice recording and speech synthesis.

(2) Based on the self-made dataset and fine-tuning on the WenetSpeech large-scale dataset model, a Conformer acoustic model was trained using CTC-Attention joint loss supervision. Compared to the WenetSpeech pre-trained model, the CER was reduced to 3.9%.

(3) A speech recognition model for hydraulic engineering geological field surveys was developed. Using tokenized text data combined with open-source Chinese corpora, an N-gram algorithm was used to train and generate a specialized language model for hydraulic engineering geological field surveys, which was used for CTC beam search decoding. The proposed method achieved a single-character decoding time of 15.5 ms and a CER of only 2.6%, a reduction of 1.3%, outperforming other speech recognition methods such as Deepspeech2, Transformer, Efficient Conformer, and Squeezeformer.

(4) By combining the Conformer acoustic model with the N-gram language model, an intelligent speech recognition and efficient recording method tailored for hydraulic engineering geological field surveys was proposed. Compared with mainstream commercial software, this method achieved better results in terms of CER and showed significant advantages in offline recognition, making it suitable for mobile data environments in field geological surveys.

The proposed method performed well on the self-made dataset and in practical speech tests for geological surveys, but there is still room for improvement in recognition accuracy. Future work will include further expanding the dataset, such as collecting geological survey data for bridges, underground projects, and increasing the number of recording personnel. Additionally, the model architecture will be improved using the EfficientNet network to further optimize performance. Moreover, the acoustic and language models developed in this study are still relatively large after export, so future research will explore model lightweighting using knowledge distillation techniques.

Funding

This work was supported by the National Natural Science Foundation of China (Grant No.: 52179139) and the Major Science and Technology Projects of the Ministry of Water Resources (Grant No.: SKS-2022147).

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] X. Zhang, Y. Zhao, J. Xie, C. Li, and Z. Hu, "Geological big data acquisition based on speech recognition," *Multimed. Tools Appl.*, vol. 79, pp. 24 413–24 428, 2020. https://doi.org/10.1007/s11042-020-09064-5

[2] D. F. Merriam, L. L. Brady, and K. D. Newell, "Kansas energy sources: A geological review," *Nat. Resour. Res.*, vol. 21, pp. 163–175, 2012. https://doi.org/10.1007/s11053-011-9164-y

[3] M. Castellini, S. Di Prima, D. Moret-Fernández, and L. Lassabatere, "Rapid and accurate measurement methods for determining soil hydraulic properties: A review," *J. Hydrol. Hydromech.*, vol. 69, no. 2, pp. 121–139, 2021. https://doi.org/10.2478/johh-2021-0002

[4] J. Safari Bazargani, A. Sadeghi-Niaraki, and S.-M. Choi, "A survey of GIS and IoT integration: Applications and architecture," *Appl. Sci.*, vol. 11, no. 21, p. 10365, 2021. https://doi.org/10.3390/app112110365

[5] Y. H. Weng, F. S. Sun, and J. D. Grigsby, "GeoTools: An android phone application in geology," *Comput. Geosci.*, vol. 44, pp. 24–30, 2012. https://doi.org/10.1016/j.cageo.2012.02.027

[6] C. Saint-Martin, P. Javelle, and F. Vinet, "DamaGIS: A multisource geodatabase for collection of flood-related damage data," *Earth Syst. Sci. Data*, vol. 10, no. 2, pp. 1019–1029, 2018. https://doi.org/10.5194/essd-10-1019-2018

[7] S. Han, H. Li, M. Li, and X. Luo, "Measuring rock surface strength based on spectrograms with deep convolutional networks," *Comput. Geosci.*, vol. 133, p. 104312, 2019. https://doi.org/10.1016/j.cageo.2019.104312

[8] M. H. Stephenson, "The uses and benefits of big data for geological surveys," *Acta Geol. Sin. (Engl. Ed.)*, vol. 93, no. s3, pp. 64–65, 2019. https://doi.org/10.1111/1755-6724.14247

[9] T. L. Pavlis, R. Langford, J. Hurtado, and L. Serpa, "Computer-based data acquisition and visualization systems in field geology: Results from 12 years of experimentation and future potential," *Geosphere*, vol. 6, no. 3, pp. 275–294, 2010. https://doi.org/10.1130/GES00503.1

[10] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," *Speech Commun.*, vol. 56, pp. 85–100, 2014. https://doi.org/10.1016/j.specom.2013.07.008

[11] S. Agarwalla and K. K. Sarma, "Machine learning based sample extraction for automatic speech recognition using dialectal Assamese speech," *Neural Networks*, vol. 78, pp. 97–111, 2016. https://doi.org/10.1016/j.neunet.2015.12.010

[12] M. Fujimoto, K. Takeda, and S. Nakamura, "CENSREC-3: An evaluation framework for japanese speech recognition in real car-driving environments," *IEICE Trans. Inf. & Syst.*, vol. 89, no. 11, pp. 2783–2793, 2006. https://doi.org/10.1093/ietisy/e89-d.11.2783

[13] M. L. Rohlfing, D. P. Buckley, J. Piraquive, C. E. Stepp, and L. F. Tracy, "Hey Siri: How effective are common voice recognition systems at recognizing dysphonic voices?" *Laryngoscope*, vol. 131, no. 7, pp. 1599–1607, 2021. https://doi.org/10.1002/lary.29082

[14] M. M. Morgan, I. Bhattacharya, R. Radke, and J. Braasch, "Automatic speech emotion recognition using deep learning for analysis of collaborative group meetings," *J. Acoust. Soc. Am.*, vol. 146, no. S4, pp. 3073–3074, 2019. https://doi.org/10.1121/1.5137665

[15] S. Alharbi, M. Alrazgan, A. Alrashed, T. Alnomasi, R. Almojel, R. Alharbi *et al.*, "Automatic speech recognition: Systematic literature review," *IEEE Access*, vol. 9, pp. 131 858–131 876, 2021. https://doi.org/10.1109/ACCESS.2021.3112535

[16] S. Peivandi, L. Ahmadian, J. Farokhzadian, and Y. Jahani, "Evaluation and comparison of errors on nursing notes created by online and offline speech recognition technology and handwritten: An interventional study," *BMC Med. Inform. Decis. Mak.*, vol. 22, no. 1, p. 96, 2022. https://doi.org/10.1186/s12911-022-01835-4

[17] L. Y. Guo, S. N. Mu, Y. J. Deng, C. F. Shi, B. Yan, and Z. L. Xiao, "Efficient binary weight convolutional network accelerator for speech recognition," *Sensors*, vol. 23, no. 3, p. 1530, 2023. https://doi.org/10.3390/s23031530

[18] G. Sterpu and N. Harte, "Taris: An online speech recognition framework with sequence to sequence neural networks for both audio-only and audio-visual speech," *Comput. Speech Lang.*, vol. 74, p. 101349, 2022. https://doi.org/10.1016/j.csl.2022.101349

[19] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *INTERSPEECH 2020*. ISCA-INT Speech Communication Assoc, 2020, pp. 5036–5040. https://doi.org/10.48550/arXiv.2005.08100

[20] X. Yan, Z. Fang, and Y. Jin, "An adaptive n-gram transformer for multi-scale scene text recognition," *Knowl.-Based Syst.*, vol. 280, p. 110964, 2023. https://doi.org/10.1016/j.knosys.2023.110964

[21] T. Liu, S. R. Zhang, C. Wang, Z. H. Li, W. Guan, and X. H. Wang, "Text intelligent analysis for hydraulic construction accidents based on BERT-BiLSTM hybrid model," *J. Hydroelectr. Eng.*, vol. 41, pp. 1–12, 2022. https://doi.org/10.11660/slfdxb.20220701

[22] M. Soleymanpour, M. T. Johnson, R. Soleymanpour, and J. Berry, "Accurate synthesis of dysarthric speech for ASR data augmentation," *Speech Commun.*, vol. 164, p. 103112, 2024. https://doi.org/10.1016/j.specom.2024.103112

[23] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*. IEEE, 2017, pp. 1–5. https://doi.org/10.1109/ICSDA.2017.8384449

[24] E. Lieskovská, M. Jakubec, R. Jarina, and M. Chmulík, "A review on speech emotion recognition using deep learning and attention mechanism," *Electronics*, vol. 10, no. 10, p. 1163, 2021. https://doi.org/10.3390/electronics10101163

[25] Y. Atmani, S. Rechak, A. Mesloub, and L. Hemmouch, "Enhancement in bearing fault classification parameters using gaussian mixture models and mel frequency cepstral coefficients features," *Arch. Acoust.*, vol. 45, no. 2, pp. 283–295, 2020. https://doi.org/10.24425/aoa.2020.133149

[26] L. Coppieters de Gibson and P. N. Garner, "Training a filter-based model of the cochlea in the context of pre-trained acoustic models," *Acoustics*, vol. 6, no. 2, pp. 470–488, 2024. https://doi.org/10.3390/acoustics6020025

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*. La Jolla, CA: NIPS, 2017, pp. 1–11.

[28] F. J. Rendón-Segador, J. A. Álvarez-García, F. Enríquez, and O. Deniz, "Violencenet: Dense multi-head self-attention with bidirectional convolutional LSTM for detecting violence," *Electronics*, vol. 10, no. 13, p. 1601, 2021. https://doi.org/10.3390/electronics10131601

[29] B. Joshi, V. K. Singh, D. K. Vishwakarma, M. A. Ghorbani, S. Kim, S. Gupta *et al.*, "A comparative survey between Cascade Correlation Neural Network (CCNN) and Feedforward Neural Network (FFNN) machine learning models for forecasting suspended sediment concentration," *Sci. Rep.*, vol. 14, no. 1, p. 10638, 2024. https://doi.org/10.1038/s41598-024-61339-1

[30] S. Liang and W. Q. Yan, "A hybrid ctc+ attention model based on end-to-end framework for multilingual speech recognition," *Multimed. Tools Appl.*, vol. 81, no. 28, pp. 41 295–41 308, 2022. https://doi.org/10.1007/s11042-022-12136-3

[31] G. Hou, Y. Jian, Q. Zhao, X. Quan, and H. Zhang, "Language model based on deep learning network for biomedical named entity recognition," *Methods*, vol. 226, pp. 71–77, 2024. https://doi.org/10.1016/j.ymeth.2024.04.013

[32] A. Mukhamadiyev, M. Mukhiddinov, I. Khujayarov, M. Ochilov, and J. Cho, "Development of language models for continuous uzbek speech recognition system," *Sensors*, vol. 23, no. 3, p. 1145, 2023. https://doi.org/10.3390/s23031145

[33] W. Naptali, M. Tsuchiya, and S. Nakagawa, "Class-based n-gram language model for new words using out-of-vocabulary to in-vocabulary similarity," *IEICE Transactions on Information and Systems*, vol. 95, no. 9, pp. 2308–2317, 2012. https://doi.org/10.1587/transinf.E95.D.2308

[34] R. Nahar, S. Miwa, and A. Kai, "Domain adaptation with augmented data by deep neural network based method using re-recorded speech for automatic speech recognition in real environment," *Sensors*, vol. 22, no. 24, p. 9945, 2022. https://doi.org/10.3390/s22249945

[35] Y. Liu, X. Yang, and D. Qu, "Exploration of whisper fine-tuning strategies for low-resource ASR," *EURASIP J. Audio Speech Music Proc.*, vol. 2024, no. 1, p. 29, 2024.

[36] B. Zhang, H. Lv, P. Guo, Q. Shao, C. Yang, L. Xie *et al.*, "Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6182–6186. https://doi.org/10.1109/ICASSP43922.2022.9746682

[37] G. Farahani, "Autocorrelation-based noise subtraction method with smoothing, overestimation, energy, and cepstral mean and variance normalization for noisy speech recognition," *EURASIP J. Audio Speech Music Proc.*, pp. 1–16, 2017.

[38] D. S. Park, W. Chan, Y. Zhang, C. C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *INTERSPEECH 2019*, 2019, pp. 2613–2617. https://doi.org/10.48550/arXiv.1904.08779

[39] P. D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015, pp. 1–15.

[40] P. Jiang, W. Pan, J. Zhang, T. Wang, and J. Huang, "A robust conformer-based speech recognition model for mandarin air traffic control," *Comput. Mater. Continua*, vol. 77, no. 1, pp. 911–940, 2023. https://doi.org/10.32604/cmc.2023.041772

[41] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case *et al.*, "Deep speech 2: End-to-end speech recognition in English and Mandarin," in *International Conference on Machine Learning*, 2016, pp. 173–182.

[42] S. Li, M. Xu, and X. L. Zhang, "Efficient conformer-based speech recognition with linear attention," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2021, pp. 448–453.

[43] S. Kim, A. Gholami, A. Shaw, N. Lee, K. Mangalam, J. Malik *et al.*, "Squeezeformer: An efficient transformer for automatic speech recognition," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 9361–9373, 2022.