achieve sustainable, high-quality development.

REFERENCES

1. 李茹月.美丽乡村建设背景下龙胜乡村旅游发展问题与对策研究 // 南方农机. – 2020. – №51(1). –p.224-226.
2. 冯楠,李绒.«美丽乡村»建设下营山县兴云村乡村旅游发展探讨 //南方农业. – 2021. – №15(19). – p.55-58.
3. 张艳艳.以乡村旅游发展为导向的甘肃省美丽乡村建设的问题及对策研究 // 江西电力职业技术学院学报. – 2021. – №6. – p.12-14.
4. 郝召雷,李玮.乡村振兴背景下美丽乡村建设与乡村旅游发展对策研究 // 农村经济与科技. – 2022. – №33(8). – p.12-14.
5. 赵娜娜,李如跃,巩慧琴. 文旅融合背景下三亚乡村旅游产品开发研究// 西部旅游. – 2024.– №09(18). – p. 32-34.

Сведения об авторах

Чан Сяофан преподаватель

Цзынь Сянфэн

Jin Xiangfeng

ПРИМЕНЕНИЕ ГЛУБОКОГО ОБУЧЕНИЯ ДЛЯ РАСПОЗНОВАНИЯ РЕЧИ НА ХАЙНАНЬСКОМ ДИАЛЕКТЕ ЛИГАО

APPLICATION OF DEEP LEARNING IN SPEECH RECOGNITION OF HAINAN LINGAO DIALECT

Хайньнский институт экономики и бизнеса, г. Хайкоу, Китай

Ural Institute (Hainan Institute of Economics and Business), Haikou, China

Диалект Лингао является уникальным и важным диалектом провинции Хайнань. Разработка системы распознования речи сталкивается со смногими трудностями из-за нехватки речевых ресурсов и отсутствия стандартизации. Чтобы повысить точность распознавания речи на диалекте Лингао. В данной статье собраны многомерные данные о речи на Лингао продолжительностью более 400 часов и создан специализированный набор данных. Используется платформа WeNet3.0 с открытым исходным кодом, при этом используется сверточно-усовершенствованная модель преобразования Conformer для углубленного обучения. Благодаря многократным раундам оптимизации в этой модели распознования речи частоиа ошибок в словах достигает 8,04%.

Lingao dialect is a unique and important dialect in Hainan Province. The development of

speech recognition system faces many challenges due to the scarcity of speech resources and lack of standardization. In order to improve the speech recognition accuracy of Lingao dialect, this paper collects more than 400 hours of Lingao speech data in a multi-dimensional way, and constructs a specialized dataset. On this basis, this paper utilizes the WeNet3.0 open-source platform and adopts the convolutionally enhanced Transformer (Conformer) model for in-depth training. Through multiple rounds of optimization, this speech recognition model achieves a word error rate of 8.04%.

Ключевые слова: искусственный интеллект, глубокое обучение, распознавание речи
Keywords: artificial intelligence, deep learning, speech recognition.

With the rapid development of artificial intelligence and deep learning technology, speech recognition technology has made significant progress in several fields, and it has been widely used in smart home, automatic customer service, voice assistant and other scenarios. However, speech recognition technology still faces great challenges when dealing with some special dialects, such as Lingao dialect in Hainan Province [1]. Lingao is a dialect with unique phonetic characteristics, but the limited resources of its speech library and the lack of standardization make it difficult to develop a high-precision speech recognition system. In order to overcome these obstacles, this paper collects more than 400 hours of Lingao dialect speech data by itself and constructs a specialized dataset. Based on this dataset, this paper conducts multiple rounds of training on the WeNet3.0 open-source platform using the advanced Conformer model, which has the ability to handle long time sequences and complex speech features by combining the convolutional neural network and self-attention mechanism. After several iterations of optimization, the speech recognition model finally achieves a word error rate (WER) of 8.04%, and the result can provide a useful reference and reference for speech recognition research in other low-resource dialects.

1 Core Technology Principles.

1.1 Principles of End-to-End Speech Recognition.

The End-to-End (E2E) model represents a new technological paradigm in the field of speech recognition, which is significantly different from traditional speech recognition systems. Traditional speech recognition systems are usually composed of multiple modules, including pronunciation dictionaries, acoustic models, language models, etc., and each module works independently to accomplish speech-to-text conversion. The independence of each module and the complex engineering design make the training and optimization process of the system cumbersome and time-costly.

The E2E model simplifies the whole speech recognition process by a unified neural network architecture that directly learns the corresponding mapping relationship between the speech features

at the input and the labeled text [2]. In the E2E model, the speech input is directly processed by the deep learning network, and the output is the corresponding text sequence, without relying on a separate pronunciation dictionary or language model. The schematic diagram of the principle is shown in Fig. 1.

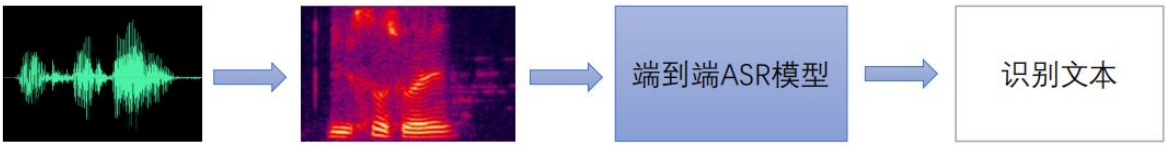

Figure 1–Schematic diagram of the principle of end-to-end speech recognition

1.2 WeNet Platform.

WeNet is an end-to-end speech recognition toolkit for industrial landing applications, providing a complete set of services from training to deployment of speech recognition models [3], with the following main features:

1) Unified streaming/non-streaming speech recognition scheme using conformer network structure and CTC/attention loss joint optimization method, with industry first-class recognition effect;

2) Provide direct deployment solutions on the cloud and on the end, minimizing the engineering work between model training and product landing;

3) Simple framework, model training part is completely based on pytorch ecosystem, does not rely on kaldi and other complex tools;

4) Detailed annotations and documentation, ideal for learning the basics of end-to-end speech recognition and implementation details;

5) Support for timestamps, alignment, endpoint detection, language modeling and other related features.

2 Data sets.

2.1 Data source.

Dialect speech data are collected and provided by Hainan College of Economics and Trade Vocational and Technical College's horizontal project “Hainan Dialect Speech Data Collection Service” (Project No.: hnjmhx2022012).

2.2 Data description.

A total of 34354 voices were collected, with a cumulative length of 408 hours. The total number of speakers is 24, 12 men and 12 women, from 9 towns in Lingao area: Tiao Lou Township (7), Nan Bao Township (2), Duowen Township (1), Heshe Township (4), Dongying Township (1),

Lincheng Township (4), Bo Lian Township (2), Xin Ying Township (2), and Bohou Township (1). By age group: A (16-19 years old) 9 persons; B (20-39 years old) 12 persons; C (40 years old and above) 3 persons.

3 Data Preprocessing.

Setting idx as the unique ID of the corpus, a corpus for data collection consists of two parts:

1) idx.wav audio file: the original audio corresponding to the corpus;

2) idx.txt annotation file: the text annotation corresponding to the corpus.

According to the above file naming convention, we can generate wav.scp, text text, mapping file data.list and lexicon (lang_char.txt) which are needed for the training of WeNet platform.

3.1 wav.scp speech file list.

This file saves the speech number and the absolute position path of the speech in the system, its function is to correspond the speech number and the absolute path of the speech, so that in the phase of acoustic feature extraction and data enhancement can be accessed to the speech, and then process the speech, the format is as follows:

<speech number><absolute path of the speech>.

BAC009S0002W0124 /data_aishell/wav/train/S0002/BAC009S0002W0122.wav

3.2 Text annotation.

This file stores the speech number and the corresponding transcribed text of the speech, its function is to associate the speech number with the corresponding transcribed text of the speech, so that the corresponding transcribed text can be accessed when calculating the CTC or attention loss, and its format is as follows:

<speech number><speech corresponding transcript>.

BAC009S0002W0124 Since the end of June, Hohhot has taken the lead in announcing the lifting of the purchase restriction.

3.3 Mapping file data.list.

This file contains the audio id, audio path and its corresponding transcribed text. The format is as follows:

{

“key“: ‘BAC009S0002W0122’,

“wav“:”/data1/data/aishell/data_aishell/wav/train/S0002/BAC009S0002W0122.wav”,

“txt": ”And the purchase restriction that has the most inhibiting effect on property market transactions”

}

3.4 Dictionary (lang_char.txt).

The unique representation of the number of each character used in the corpus, called token, has the following sample format:

<character><number corresponding to character>

<blank> 0

<unk> 1

<sos/eos> 2

ah 76

Mourning 77

4 Model Training.

4.1 Experimental Environment.

Table 1 contains Experimental environment.

Table 1

Experimental environment

Device	Configurations
Operating System	CentOS 7
Memory	128G
CPU	Intel(R) Xeon(R) Gold 5218 CPU @ 2.30GHz*64
GPU	NVIDIA Corporation TU104GL [Tesla T4]*4
Python	3.8.18
Torch	1.13.0
Training Platform	WeNet3.0

4.2 Data set splitting.

Table 2 contains Data set splitting.

Table 2

Data set splitting

Lingao Dialect Corpus	Number of Participants (persons)	Audio Digits (articles)
Training set (train)	24	32978
Validation set (dev)	4	688
test	4	688

As shown in Table 2, 96% of the Lingao dialect corpus is assigned to the training set, and the remaining 4% is equally distributed to the validation and test sets.

4.3 Feature Extraction.

In this speech recognition task, we perform Cepstrum Mean Normalization (CMVN) on the spectrum of the speech file to make the features obey a Gaussian distribution with mean 0 and variance 1. Such a process allows the neural network to learn the speech features more easily.

global_cmvn sample:

{“mean_stat”: [51409404.0, 60613008.0, 79658952.0, 91865936.0, 105758736.0, 115681168.0, ... , 158674736.0, 156011168.0, 151878496.0, 143558896.0], “var_stat”: [1684987264.0, 1803099008.0, 2221526528.0, 2406902272.0, 2657631488.0, 2849029888.0, ... , 2198164224.0, 2030811264.0], “frame_num”: 22408960}

4.4 Model Core Parameter Configuration.

The core parameter configuration of the model in this paper uses the train_conformer.yaml file provided by the WeNet 3.0 platform to take full advantage of the Conformer model. Specifically, the encoder uses conformer, the decoder uses transformer, the model uses hybrid CTC/attention, and the weight of ctc is 0.3.

4.5 Model Training.

The training command script is launched and after 240 iterations, the training results are visualized as shown in Figure 2.

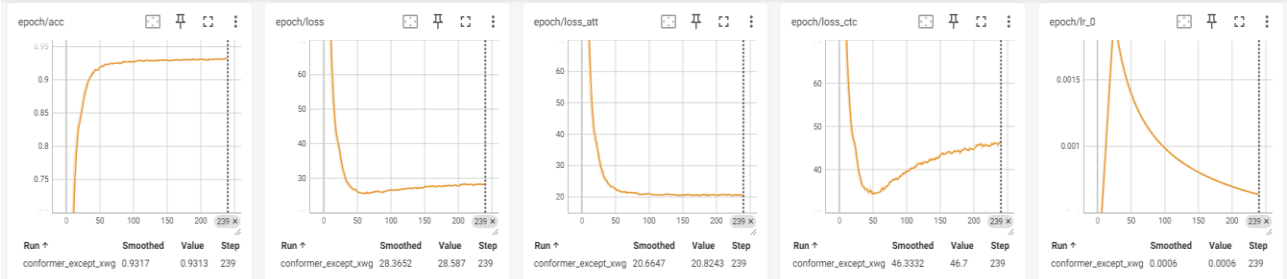

Figure2– Training results

The model ultimately achieves 93.13% accuracy on the validation dataset with a validation loss of 28.587, with a CTC loss of 46.7% and an attention loss of 20.82%.

5 Model Merging and Evaluation.

5.1 Evaluation index selection.

After the models are trained, the 30 models with the smallest validation loss are selected from 240 iterations, and their parameters are averaged to merge the final model final.pt. WeNet provides four decoding modes, which are ctc_greedy_search, ctc_prefix_beam_search, attention and attention_rescoring. we use these four decoding modes to evaluate the model by performing decoding validation on the test set data.

5.2 Model Evaluation.

The final model final.pt was decoded on the test set data using the final model according to the four decoding patterns and their word error rates (WER) were calculated and the test results are shown in Table 3.

Table 3

Four decoding modes WER

model	ctc_greedy_search	ctc_prefix_beam_search	attention	attention_rescoring
Overall	10.87%	10.89%	8.04%	9.98%
Mandarin	10.79%	10.81%	7.96%	9.90%
Other	93.10%	100.00%	96.55%	96.55%

5.3 Analysis of model decoding effect.

From the test results, the decoding mode attention achieves the optimal decoding effect, and the word error rate reaches 8.04%, which is contrary to the traditional attention rescoring as the optimal decoding mode. Considering that the convergence of CTC loss in the training process is worse than that of attention loss, the author believes that the reason for this is that the Mandarin text labeled in the corpus does not match the pronunciation of dialectal speech completely, which is to be verified by further experiments.

6 Conclusion.

In this paper, we used the Lingao dialect speech corpus collected by ourselves and trained themodel based on the Conformer model as an encoder on the WeNet platform. Eventually, the modelachieves 93.13% accuracy on the validation set, and the word error rate (WER) of the test set is 8.04%in attentional decoding mode. Although this result still falls short of the commercial standard, thereis still a lot of room to improve the recognition performance of the model by further increasing thelength of the corpus and optimizing the quality of the corpus. Future research can continue to focuson the expansion of the dataset and optimization of the model [3], gradually narrowing the gap with commercial applications and laying a more solid foundation for the promotion and application of Lingao dialect speech recognition technology.

REFERENCES

1. 余旭文. 基于深度学习的海南方言语音识别 //海南大学. – 2020.
2. 郝焕香. 基于深度学习的方言语音识别模型构建 // 自动化与仪器仪表. – 2022. – №04. – p. 48-51.
3. 鱼昆. 低资源方言语音识别方法研究及应用 //长安大学. – 2021.

Сведения об авторах

Цзын Сянфэн преподаватель