Dialect speech recognition based on pre-trained model

Lu Shenghui, Yao Jianglong, Ren Pengyu, Gong Tao
School of Informatics, Xiamen University
{31520231154303 ,31520231154324 ,31520231154309 ,30920231154346}@stu.xmu.edu.cn

Abstract

In recent years, there has been rapid development in speech pre-training models, substantially improving the performance of speech recognition and reducing the amount of data required to train a speech recognition system. One of the application domains for speech and pre-training models is dialect recognition. This study aims to explore the recognition performance of current popular pre-training models on Chinese dialects. By comparing the recognition performance of different self-supervised upstream models under varying durations of training data, we observed a significant improvement in the effectiveness of pre-trained models in low-resource scenarios. However, in cases where the pre-training data of self-supervised models and the dialect training data domain do not match, simply augmenting pre-training data does not necessarily enhance dialect recognition performance.

Introduction

Automatic Speech Recognition (ASR) is a technology that converts human speech into text. Thanks to the advancements in deep learning, ASR models trained on annotated data have achieved satisfactory performance(Radford et al. 2022). In recent years, pre-trained models have demonstrated a significant enhancement in downstream tasks related to speech(Mohamed et al. 2022). Furthermore, pre-training on speech has further improved the accuracy of speech recognition while reducing the required amount of annotated data.In the context of the wav2vec 2.0 framework, it has been demonstrated that a model pre-trained on a substantial amount of data requires only 10 minutes of fine-tuning with labeled data to achieve a remarkably low error rate(Baevski et al. 2020). This highlights the efficacy of leveraging extensive pre-training data to enhance the efficiency and accuracy of speech recognition models, thereby reducing the dependency on a large volume of annotated data.

An important application scenario for self-supervised pre-trained models is low-resource speech recognition(Mohamed et al. 2022), with dialect recognition being a notable representative case. There are two primary approaches to leveraging pre-trained models for dialect recognition. The first involves fine-tuning the pre-trained model with a downstream model, while the second approach entails freezing the pre-trained model and using the representations it extracts as input features.

Currently, numerous studies have validated the effectiveness of pre-trained models in dialect recognition, as demonstrated by (Hsu et al. 2021; Baevski et al. 2020; Chen et al. 2022). However, the majority of these investigations have predominantly explored the utilization of English pre-trained models for English or closely related dialects. There has been relatively limited exploration of pre-trained models in the context of Chinese dialects. Possible reasons for this disparity include:

- There is a scarcity of publicly available datasets for Chinese dialects.
- There is a limited number of self-supervised speech models trained on Chinese corpora.
- The use of English or multilingual models (with minimal Chinese training data) introduces domain mismatch issues.

In addressing the aforementioned issues and aiming to validate the efficacy of representations extracted by various pre-trained models for enhancing Chinese dialect recognition, as well as demonstrating the potential of pre-trained models in the context of Chinese dialects, this study collected 111 hours of labeled Chinese dialect data from online sources(Zhang et al. 2022). The reliability of the data was verified, and it was subsequently divided into three datasets containing 100, 10, and 1 hour of data, respectively. The study then compared the speech recognition performance of pre-trained models using different training datasets on these data subsets.

In contrast to the remarkable results demonstrated by wav2vec 2.0(Baevski et al. 2020) in pre-training and recognition within the same language, this study observed that, in a new low-resource dialect, pre-trained models indeed enhance speech recognition performance. However, due to a domain mismatch between pre-training data and the target language, a common challenge in practical dialect recognition, the effectiveness of pre-trained representations was found to be less pronounced. We observed improvements in recognition performance across datasets containing 1, 10, and 100 hours of data. However, in the case of the 100-hour dataset, the magnitude of improvement in pretraining representation was relatively modest. Furthermore, pretraining on English data, fine-tuning on Chinese dialects, and increasing the amount of English data did not result in a significant enhancement. This further underscores the impact of pretraining domain mismatch on recognition.

Method

Dataset Collacting

There is no existing large-scale Chinese dialect datasets available online.Therefore, we collects a sufficient amount of dialect audio-Chinese text pairs from online dialect TV dramas and uses them as the training datasets after evaluating their quality. After comparing various languages, it was found that there are numerous Taiwanese dialect TV dramas with high-quality subtitles. Hence, Hokkien is chosen as the target dialect. The data collection process is outlined as follows.

- Identify multiple TV dramas with external subtitles and download them.
- Remove data from the text containing numbers, English, and other meaningless symbols.
- Exclude data that is too long (audios longer than 7 seconds are often music or background noise) and too short (less than 0.3 seconds, mostly silent audio), as well as data where label length severely mismatches speech length.
- Utilize the Wenet model trained with Wenetspeech to recognize the aforementioned data. Through manual listening, establish a strict recognition accuracy threshold (striving to include Mandarin completely) and eliminate Mandarin content.
- Employ the ESPnet toolkit to recognize the obtained data and validate its quality.

Because the selected TV drama subtitles in this work are generally reliable, the quality of the dataset is relatively high after filtering. The paper has gathered 111 hours of data, divided into training and test sets with a 10:1 ratio. The verified recognition accuracy is 42.30%. For a small-scale dialect dataset with 100 hours of noise, this is an acceptable resultU̇nlike transcribing clean speech dedicatedly by others, this work focuses on evaluating self-supervised models on more generalized data. The dataset is sourced from daily dialogues in 17 different types of TV dramas, encompassing various aspects.

SSL Model

In this study, we employed the wav2vec2 model as the upstream pre-training model for extracting audio representations, as illustrated in Figure 1(Baevski et al. 2020). Wav2vec2.0 directly processes audio signals through a CNN convolution to obtain the vector $\mathcal{Z}$. Subsequently, $\mathcal{Z}$ undergoes random masking and is passed through a 24-layer Transformer to yield contextual representations $\mathcal{C}$. Simultaneously, $\mathcal{Z}$ is subjected to a quantization module, wherein the vector corresponding to the codebook is selected to obtain the quantized representation $\mathcal{Q}$. The model is then trained by contrasting $\mathcal{C}$ with the quantized representations $\mathcal{Q}$ of adjacent frames.

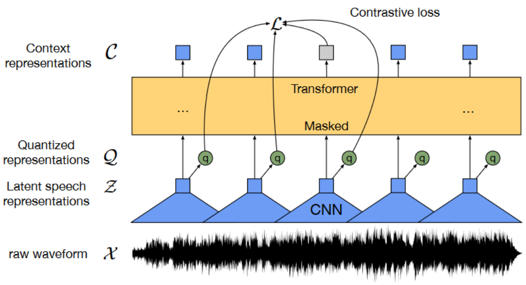

Figure 1: Framework of wac2vec2.0

This study adopts the wav2vec2.0 large model(Baevski et al. 2020), and after pre-training, its parameters are frozen. The model output is employed as acoustic representations inputted into the downstream model(Huang et al. 2021). The pre-trained model consists of 24 Transformer Blocks, and for each frame, it outputs a 1024-dimensional acoustic representation. Due to the diverse semantic features represented at different depths in each of the 24 layers, this paper performs a weighted sum of these acoustic representations(Mohamed et al. 2022). The weights are determined through neural network training, resulting in the ultimate semantic representation. A linear mapping layer is then applied to compress the representation to an 80-dimensional acoustic feature for subsequent recognition tasks.

ASR Model

This work employs an Encoder-Decoder architecture for downstream Automatic Speech Recognition (ASR) tasks, utilizing the popular CTC (Connectionist Temporal Classification)(Graves et al.) and AED (Attention-based Encoder-Decoder) joint decoding approach. Following the encoding stage of the Encoder, the encoded information is subjected to linear mapping for CTC non-autoregressive decoding, and simultaneously, it is fed into a Transformer Decoder for autoregressive decoding. The two results are jointly optimized with a loss function that takes into account both CTC and AED outcomes in comparison to the ground truth labels(Yao et al. 2021). During model inference, the CTC result is reranked using the Transformer Decoder to enhance precision.

For the Encoder, this work adopts the Branchformer model(Peng et al. 2022), as illustrated on the left side of Figure 2. In contrast to the Transformer Encoder, it incorporates an additional branch that employs convolution to capture local information during self-attention, and subsequently, the information from both branches is fused. This structure has been proven to yield superior results compared to single-branch Transformer and Conformer encoders.

Concerning the Decoder, this work combines CTC Decoder and Transformer Decoder (Zhang et al. 2020)to jointly optimize the model. CTC decoding involves passing the output from the Encoder through a linear layer with a dimension

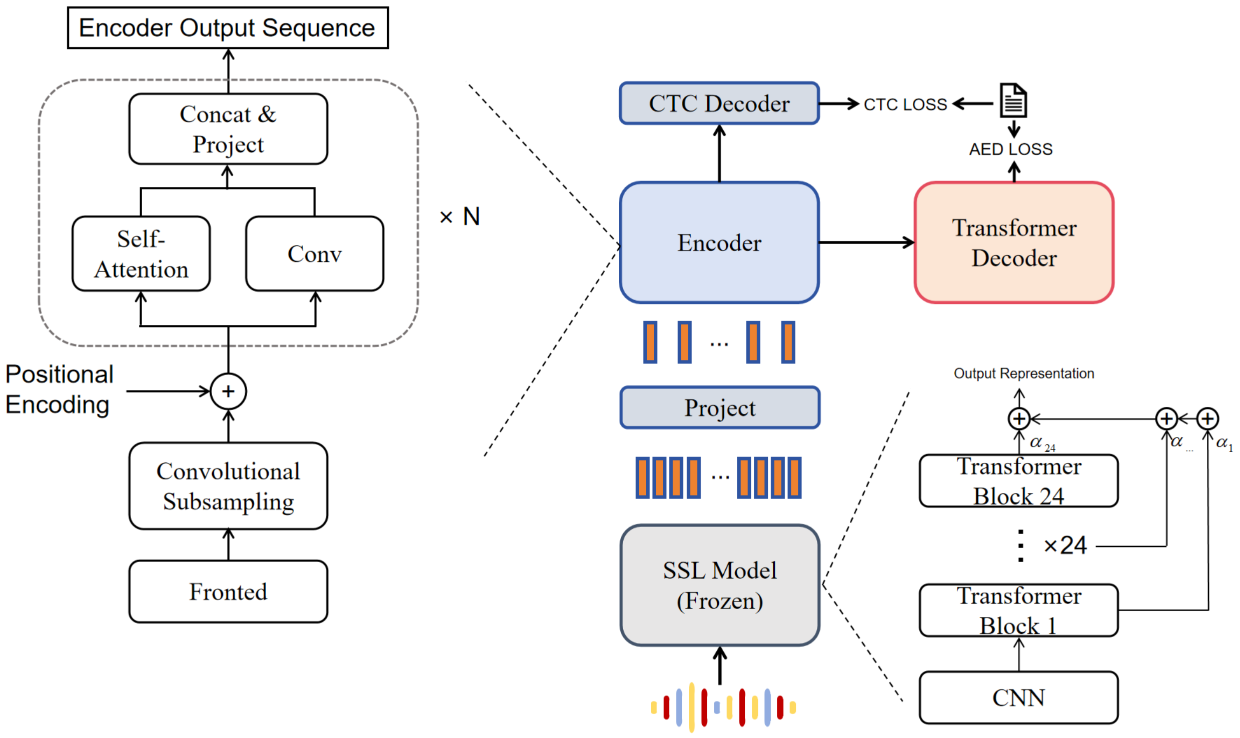

Figure 2: Framework of our model

equal to the dictionary length, followed by standard CTC decoding to obtain recognition results. AED decoding follows the same process as the Transformer Decoder, generating output results through self-attention layers and cross-attention in an autoregressive manner.

Experiment

Table 1: Dataset Statistics

Hours	utts	Avg time(secs)	Avg characters
100	151453	2.377	7.7905
10	15180	2.372	7.814
1	1521	2.367	7.788
0.1	158	2.278	7.665

Datasets

To ensure data diversity, this study selected 17 different types of TV dramas, encompassing genres such as blogger exploration, family stories, and youth idols, to cover a wide range of aspects. A total of 111.1 hours of labeled data were collected. For training and testing, the data were randomly divided into independent datasets of 100, 10, 1, and 0.1 hours. Using the same configurations, models were trained on datasets of 100, 10, and 1 hour, maintaining a 10:1 ratio between training and testing sets in each experiment. Table 1 presents basic information about the audio data used in the experiments.

As the data were sourced from online videos, the audio often contained random noise. Unlike other works that test on clean speech, the results of this experiment align more closely with real-world speech recognition scenarios. To assess data quality, the espnet baseline configuration was employed to recognize the 100-hour dataset, yielding a verified recognition accuracy of 42.30%. For a small-scale dialect dataset with 100 hours of noisy data, this result is deemed acceptable.

Experiment seeting

For the pre-training phase, this study utilized three publicly available models from the s3prl toolkit(Huang et al. 2021): librispeech960, librilight60k, and xlsr53k, all based on the wav2vec2.0 Large version.

For the encoder, the branchformer configuration from the espnet toolkit(Watanabe et al. 2018) is adopted. Subsequently, the input is passed through 24 branchformer blocks, featuring 4 attention heads and a dropout rate of 0.1.

Decoding involves the use of both CTC and Transformer decoders, with a 1:1 weighting of CTC and AED losses to obtain the final loss for model optimization. The Adam optimizer is employed for model optimization, with a learning rate set to 0.001 and weight decay at 0.000001. To explore the relative improvements offered by pre-trained models, results obtained from recognizing Fbank features are used as a baseline, and the effectiveness of different pre-trained mod-

Table 2: Character Error Rate (CER) of ASR using different upstream speech self-supervised learning (SSL) models on test set, trained with 1,10,100 hour. Snt, Wrd means the number of test sentence and test Characters. Sub, Del, In, Err means the ratio of substitution, Delete, Insert and Total error rate.

Training Dataset	Representation	Snt	Wrd	Sub	Del	In	Err
1h	fbank	158	1211	64.49%	25.68%	4.46%	94.63%
	ls960	158	1211	59.37%	30.47%	4.38%	94.22%
	ll60k	158	1211	60.12%	29.40%	3.14%	92.65%
	xlsr	158	1211	55.24%	34.27%	2.23%	91.74%
10h	fbank	1521	11845	61.63%	18.83%	7.42%	87.88%
	ls960	1521	11845	39.16%	16.08%	7.45%	62.70%
	ll60k	1521	11845	39.65%	14.39%	8.73%	62.76%
	xlsr	1521	11845	36.01%	14.51%	8.14%	58.66%
100h	fbank	15180	118609	23.11%	11.23%	7.96%	42.30%
	ls960	15180	118609	21.79%	11.14%	7.73%	40.66%
	ll60k	15180	118609	21.38%	10.36%	8.24%	39.97%
	xlsr	15180	118609	18.42%	10.02%	7.35%	35.79%

els is analyzed.

In our experiments, we attempted to fine-tune the upstream self-supervised pre-trained model on a small dataset without freezing its layers. However, we observed that this approach led to training instability and often resulted in a decline in model performance. Therefore, this work did not extensively explore fine-tuning pre-trained models for speech recognition. Instead, we only present various experimental results using frozen pre-trained models.

Experimental Results

We extracted four different features from training sets of 1h, 10h, and 100h, respectively, and trained networks on each feature set. Fbank, a traditional acoustic feature commonly used in tasks like speech recognition, is one of the features. The other three features, ls960, ll60k, and xlsr, are obtained by pre-training the wav2vec2.0 model on librispeech, librilight, and a multilingual dataset of 60,000 hours, respectively. Our evaluation is divided into two aspects: one is the impact of pre-trained models on dialect training sets of different sizes, and the other is the evaluation of different upstream pre-trained models on the selected Chinese dialect. We believe that English and Chinese, including Chinese dialects, belong to different language families, and there are differences in the domains of speech representation.

This study found that the incorporation of pre-trained models more or less improved dialect recognition results, especially evident in the 10h dataset. Training on a 1-hour dataset resulted in a very high error rate exceeding 90%, indicating that the recognition system was not effectively trained, as expected, due to the limited 1-hour speech data for a model with a 24-layer encoder. For this dialect, obtaining an effective speech recognition system requires at least several hours of labeled data. Training on a 10-hour dataset, the introduction of pre-trained models resulted in an improvement of over 30% in recognition performance. However, the improvement in recognition performance on the 100-hour dataset was relatively modest. Comparing different-sized training datasets, it was evident that the impact of incorporating pre-trained models was more pronounced in low-resource speech recognition, with the relative improvement diminishing as the training set size increased.

In comparing features extracted from different pre-trained models, there was not a significant difference in the performance between models pre-trained on 960 hours of English data and 60,000 hours of English data. This indicates that increasing the scale of the dataset has limited impact on recognition performance when there is a domain mismatch between pre-training and training data. On the other hand, xlsr, which used a multilingual dataset, including a small amount of Chinese speech similar to dialects, consistently enhanced model performance across various datasets. Therefore, we suggest that when increasing pre-training speech data in a domain mismatch scenario, efforts should be made to include diverse multilingual data, especially those closely related to the target domain.

Conclusion

This study explores the recognition performance of popular publicly available pre-trained models on the Min Nan language. It compares the enhancement effects on dialect recognition achieved by self-supervised models trained on different datasets, revealing significant improvements in low-resource scenarios. It is observed that solely augmenting the model with mismatched domain data yields limited performance improvement. These findings provide guidance for the future utilization of pre-trained models on Chinese dialects.

Due to the limited availability of Chinese pre-trained models, a more comprehensive evaluation involving the integration of Chinese pre-trained models or direct self-supervised training on dialect data remains unexplored. Future research can incorporate Chinese pre-trained models or directly utilize dialect data for self-supervised training to further assess the recognition potential of pre-trained models on Chinese dialects.

References

Baevski, A.; Zhou, H.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. arXiv:2006.11477 [cs, eess].

Chen, S.; Wang, C.; Chen, Z.; Wu, Y.; Liu, S.; Chen, Z.; Li, J.; Kanda, N.; Yoshioka, T.; Xiao, X.; Wu, J.; Zhou, L.; Ren, S.; Qian, Y.; Qian, Y.; Wu, J.; Zeng, M.; Yu, X.; and Wei, F. 2022. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE Journal of Selected Topics in Signal Processing* 16(6):1505–1518. arXiv:2110.13900 [cs, eess].

Graves, A.; Fernandez, S.; Gomez, F.; and Schmidhuber, J. CTC-Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks.

Hsu, W.-N.; Bolte, B.; Tsai, Y.-H. H.; Lakhotia, K.; Salakhutdinov, R.; and Mohamed, A. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. arXiv:2106.07447 [cs, eess].

Huang, W.-C.; Yang, S.-W.; Hayashi, T.; Lee, H.-Y.; Watanabe, S.; and Toda, T. 2021. S3prl-vc: Open-source voice conversion framework with self-supervised speech representations.

Mohamed, A.; Lee, H.-y.; Borgholt, L.; Havtorn, J. D.; Edin, J.; Igel, C.; Kirchhoff, K.; Li, S.-W.; Livescu, K.; Maaløe, L.; Sainath, T. N.; and Watanabe, S. 2022. Self-Supervised Speech Representation Learning: A Review. *IEEE Journal of Selected Topics in Signal Processing* 16(6):1179–1210. arXiv:2205.10643 [cs, eess].

Peng, Y.; Dalmia, S.; Lane, I.; and Watanabe, S. 2022. Branchformer: Parallel MLP-Attention Architectures to Capture Local and Global Context for Speech Recognition and Understanding. arXiv:2207.02971 [cs, eess].

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2022. Whisper: Robust Speech Recognition via Large-Scale Weak Supervision. arXiv:2212.04356 [cs, eess].

Watanabe, S.; Hori, T.; Karita, S.; Hayashi, T.; Nishitoba, J.; Unno, Y.; Soplin, N. E. Y.; Heymann, J.; Wiesner, M.; Chen, N.; Renduchintala, A.; and Ochiai, T. 2018. Espnet: End-to-end speech processing toolkit.

Yao, Z.; Wu, D.; Wang, X.; Zhang, B.; Yu, F.; Yang, C.; Peng, Z.; Chen, X.; Xie, L.; and Lei, X. 2021. Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit.

Zhang, Q.; Lu, H.; Sak, H.; Tripathi, A.; McDermott, E.; Koo, S.; and Kumar, S. 2020. Transformer Transducer: A Streamable Speech Recognition Model with Transformer Encoders and RNN-T Loss. arXiv:2002.02562 [cs, eess].

Zhang, B.; Lv, H.; Guo, P.; Shao, Q.; Yang, C.; Xie, L.; Xu, X.; Bu, H.; Chen, X.; Zeng, C.; Wu, D.; and Peng, Z. 2022. Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition.