Wenzhou Dialect Speech to Mandarin Text Conversion

Zhipeng Gao, Akihiro Tamura, Tsuneo Kato
Doshisha University
{ctwk0111@mail4, aktamura@mail, tsukato@mail}.doshisha.ac.jp

Abstract

The Wenzhou dialect is a Chinese dialect that is significantly distinct from Mandarin, the official language of China. It is among the most complex Chinese dialects and is nearly incomprehensible to people from regions such as Northern China, thereby creating substantial communication barriers. Therefore, the conversion between the Wenzhou dialect and Mandarin is essential to facilitate communication between Wenzhou dialect speakers and those from other Chinese regions. However, as a low-resource language, the Wenzhou dialect lacks publicly available datasets, and such conversion technologies have not been extensively researched. Thus, in this study, we create a parallel dataset containing Wenzhou dialect speech and the corresponding Mandarin text and build benchmark models for Wenzhou dialect speech-to-Mandarin text conversion. In particular, we fine-tune two self-supervised learning-based pretrained models, that is, TeleSpeech-ASR1.0 and Wav2Vec2-XLS-R, with our training dataset and report their performance on our test dataset as baselines for future research.

1 Introduction

China has 56 ethnic groups over a wide geographical span, resulting in a significant variety of Chinese dialects. In addition to Mandarin, which is the official standard Chinese, active Chinese dialects can generally be divided into nine categories: Wu Chinese[1], Cantonese, Jin Chinese, Min Chinese, Hakka, Xiang Chinese, Gan Chinese, Hui Chinese, and Pinghua (Chinese Academy of Social Sciences and City University of Hong Kong, 2012). Some speakers of these dialects experience difficulty communicating with Mandarin speakers, and these dialects have significant differences between branches. This study focuses on the Wenzhou dialect, a branch of Wu Chinese that does not have a formal writing system (Ethnologue, 2017).

Figure 1: Differences in pronunciation

Figure 2: Differences in word usage

The Wenzhou dialect is considered the most difficult dialect in China because of its unique phonetic system, which differs markedly from those of Mandarin and other Chinese dialects (Steger, 2014; Xu et al., 2012). Figures 1–3 present simple examples illustrating the differences between the Wenzhou dialect and Mandarin. Figure 1 shows the differences in pronunciation between Wenzhou dialect and Mandarin. Although the same Chinese characters are used in Figure 1, the pronunciation of the Wenzhou dialect is completely different from that of Mandarin. Figure 2 illustrates the differences in word usage between the Wenzhou dialect and Mandarin. In the figure, the orange underlined parts are specific to the Wenzhou dialect, with certain Chinese characters no longer used in modern Mandarin. Figure 3 illustrates the differences in

[1] https://www.ethnologue.com/language/wuu/

Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2025), pages 36–43

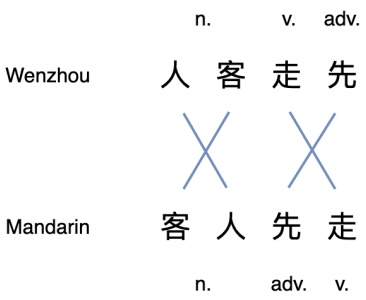

Figure 3: Differences in grammar

terms of grammar between the Wenzhou dialect and Mandarin. In the Wenzhou dialect, the verb is placed before the adverb, which is the opposite to that in case of Mandarin. Moreover, the order of Chinese characters within the noun differs between the two.

Because of these differences, the Wenzhou dialect is typically incomprehensible to people from regions such as Northern China. Therefore, the conversion between the Wenzhou dialect and Mandarin is essential to facilitate communication between Wenzhou dialect speakers and people from other regions. This study focuses on a method for converting Wenzhou dialect speech to Mandarin text.

Neural network-based end-to-end frameworks have achieved remarkable success in speech-to-text tasks, such as automatic speech recognition (ASR) (Chorowski et al., 2015; Zhang et al., 2020; Gulati et al., 2020) and speech-to-text translation (ST) (Berard et al., 2016; Weiss et al., 2017; Li et al., 2021; Tang et al., 2022). Certain advanced models perform very well on high-resource languages owing to the richness and diversity of available data. In contrast, low-resource languages face many challenges, particularly some dialects lacking data support and complex phonetic systems. The Wenzhou dialect is also a low-resource language with no publicly available datasets, and methods for converting Wenzhou dialect speech into Mandarin text are yet to be explored.

In an attempt to address these challenges, this study makes the following three key contributions:

- **Construction of a parallel dataset**: To the best of our knowledge, this study is the first to construct a dataset comprising Wenzhou dialect speech and the corresponding Mandarin text. Specifically, we collect news videos in the Wenzhou dialect from YouTube and then manually annotate their audio with Mandarin text. This effort yields a dataset containing 8,068 parallel samples, which provides a vital resource for the development of methods to convert the Wenzhou dialect speech into Mandarin text. In addition, we believe that this constructed dataset could also be a valuable resource for future research on linguistic and computational challenges associated with the Wenzhou dialect.

- **Development of baselines**: Using the constructed dataset, we establish a benchmark for the conversion of the Wenzhou dialect speech to Mandarin text. We fine-tune two self-supervised learning (SSL)-based pretrained models, namely, TeleSpeech-ASR1.0-large and Wav2Vec2-XLS-R-300M, using our training dataset for this task. Then, we evaluate the fine-tuned models on our test dataset, thereby providing baseline results and identifying areas for improvement for this new challenging task.

- **Open access to resources**: To facilitate further research and development, we share our dataset and models[2].

2 Related Work

This section presents a review of related work on the SSL-based pretrained models, TeleSpeech-ASR1.0 and Wav2Vec2-XLS-R, which form the basis of our benchmark models. Low-resource languages lack the resources to support the scale of annotated data required to train neural network models. SSL has achieved tremendous success across various fields, including natural language processing and speech processing. SSL facilitates neural networks in learning rich feature representations from large amounts of unlabeled multilingual data (Hsu et al., 2021; Chen et al., 2022). Models pretrained by SSL can be fine-tuned on various downstream tasks (Zhang et al., 2023), and the resulting fine-tuned models enable strong downstream performance even with limited access to annotated data.

2.1 TeleSpeech-ASR1.0

TeleSpeech-ASR1.0[3] is a Transformer (Vaswani et al., 2017)-based multidialect ASR model developed by Tele-AI (China Telecom Artificial Intel-

[2] https://gaozhipengcn.github.io/WenzhouDialectSpeech2MandarinText/

[3] https://github.com/Tele-AI/TeleSpeech-ASR

Mandarin	Beijing	Southwest	Zhongyuan	Northeast	Lan-Yin	Jiang-Huai	Ji-Lu	Jiao-Liao
4.61	8.23	8.74	7.62	7.89	9.72	12.89	8.91	9.30

Table 1: Character error rate (CER) (%) of TeleSpeech-ASR1.0-large-KeSpeech on the KeSpeech test dataset

ligence Research Institute)[4]. The ASR model has been pretrained with 300,000 hours of unlabeled Chinese multidialect speech data and fine-tuned using 30 types of internal labeled data, each representing different Chinese dialects, such as Cantonese, the Shanghai dialect, the Sichuan dialect, and the Wenzhou dialect.

A total of three models related to TeleSpeech-ASR1.0 are open-sourced, including two pretrained models, TeleSpeech-ASR1.0-base and TeleSpeech-ASR1.0-large, and a fine-tuned model of TeleSpeech-ASR1.0-large with the KeSpeech dataset (Tang et al., 2021), TeleSpeech-ASR1.0-large-KeSpeech. The KeSpeech dataset is an open source speech dataset that contains 1,542 hours of speech audio recorded by 27,237 speakers from 34 Chinese cities. The pronunciation in the dataset includes standard Mandarin and its eight subdialects: Beijing, Southwest, Zhongyuan, Northeast, Lan-Yin, Jiang-Huai, Ji-Lu, and Jiao-Liao. Tele-AI used 1,396 hours of training speech data from the KeSpeech dataset as supervised data to fine-tune TeleSpeech-ASR1.0-large to obtain TeleSpeech-ASR1.0-large-KeSpeech and calculated the character error rate (CER) (Graves et al., 2006) of the fine-tuned model on the KeSpeech test dataset.

Table 1 summarizes the CER (%) of TeleSpeech-ASR1.0-large-KeSpeech for standard Mandarin and its eight subdialects. The table shows that TeleSpeech-ASR1.0-large-KeSpeech performs well in the ASR task for standard Mandarin as well as its eight subdialects. This is attributed to the fact that these subdialects are not significantly different from standard Mandarin. Thus, the training data for standard Mandarin can benefit the subdialects, particularly similar subdialects, and vice versa. It should be noted that TeleSpeech-ASR1.0-large-KeSpeech has not been evaluated for dialects that differ significantly from Mandarin or other dialects, including the Wenzhou dialect. Furthermore, the pretrained models, TeleSpeech-ASR1.0-base and TeleSpeech-ASR1.0-large, have not yet been fine-tuned for the task of converting the Wenzhou dialect to Mandarin.

[4]While TeleSpeech-ASR1.0 is open-sourced, its detailed model architecture is not fully documented and remains unclear.

	Hours	Utterances	Label
Train	10.78	5037	Mandarin text
Dev	1.86	1004	Mandarin text
Test	5.19	2027	Mandarin text

Table 2: Statistics of the Wenzhou Dialect Speech to Mandarin Text Dataset

2.2 Wav2Vec2-XLS-R

Wav2Vec2-XLS-R (Babu et al., 2022) is Facebook AI's large-scale multilingual pretrained model for speech processing. The model is pretrained on 436K hours of publicly available speech audio in 128 languages based on wav2vec2.0 (Baevski et al., 2020b). The training data related to the Chinese language includes Cantonese, Chinese CN, Chinese HK, and Chinese TW but excludes the Wenzhou dialect. Pretrained Wav2Vec2-XLS-R models with different numbers of parameters (300M, 1B, 2B) can be fine-tuned for downstream tasks, such as ASR, ST, and speech classification.

3 Benchmark Dataset for Converting the Wenzhou Dialect to Mandarin

This study constructs a benchmark dataset for converting the Wenzhou dialect to Mandarin because currently there is no publicly available dataset for this task.

Although we have presented the Wenzhou dialect texts for convenience in the three examples (Figures 1–3) used for comparisons of the Wenzhou dialect and Mandarin in the introduction section, the Wenzhou dialect does not have a formal writing system. Even native Wenzhou dialect speakers might not know the meaning of the orange-underlined Chinese character in Figure 2, some of which are no longer in use in modern Mandarin. Therefore, we construct a parallel dataset comprising the Wenzhou dialect speech (not text) and the corresponding Mandarin text.

We construct the dataset from a local news video program in the Wenzhou dialect as per the following two steps: (1) extraction of the Wenzhou dialect audio and (2) annotation of the corresponding Mandarin text. In Step 1, we first collect 144 news videos hosted and broadcast in the Wenzhou dialect

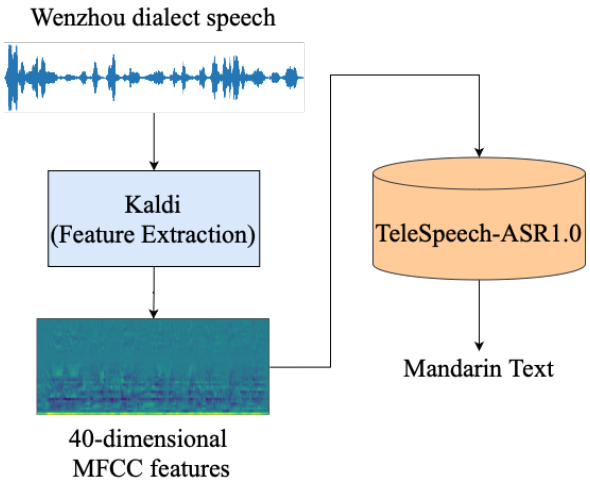

Figure 4: Data preprocessing for TeleSpeech-ASR1.0-large

uploaded on YouTube[5]. Subsequently, we extract the Wenzhou dialect audio from these videos and manually segment the audio into sentences. In Step 2, we extract the Mandarin subtitles embedded in the videos using Paddle optical character recognition (OCR) tools[6], followed by manual correction to ensure quality.

After the two steps, we obtain approximately 18 hours of the Wenzhou dialect speech and the corresponding Mandarin text. The dataset comprises 8,068 audio files, with an average duration of approximately 8 seconds per file. To divide the data into training, validation, and test datasets, we shuffle the order of video IDs and split the dataset based on the video ID as follows: 76 videos for training, 33 videos for validation, and 35 videos for testing. Of the 35 videos for testing, only audio files with durations greater than 5 seconds are extracted as test samples. This corresponds to 5,037, 1,004, and 2,027 audio files for training, validation, and testing, respectively, as summarized in Table 2. All audio is mono with a sampling rate of 16,000Hz and saved as a wav file.

4 Experiments

In this section, we present two benchmark models for the Wenzhou dialect speech-to-Mandarin text conversion and evaluate their performance on the constructed benchmark dataset.

[5]The videos can be downloaded from https://www.youtube.com/playlist?list=PLqP_o2kuQ2LrzYSJoXKgME3i3x_8yDXsX

[6]https://github.com/PaddlePaddle/PaddleOCR/blob/main/README_en.md

4.1 Benchmark Models

As the benchmark models, two pretrained models are fine-tuned: (1) TeleSpeech-ASR1.0-large and (2) Wav2Vec2-XLS-R-300M. In the fine-tuning process, the "train set" and "dev set" (see Table 2) of the constructed dataset were employed as the training and validation data, respectively. The fine-tuning was performed using two NVIDIA TITAN RTX (24GB) GPUs.

4.1.1 Benchmark Model1: TeleSpeech-ASR1.0-large

We fine-tuned the pretrained TeleSpeech-ASR1.0-large model using the fairseq framework (Wang et al., 2020). Hereafter, the fine-tuned model is referred to as TeleSpeech-ASR1.0-large-Wenzhou.

Data Preprocessing The input of the TeleSpeech-ASR1.0-large model must be 40-dimensional mel-frequency cepstral coefficient (MFCC) (Davis and Mermelstein, 1980) features extracted from a 16,000-Hz sampling rate audio. As illustrated in Figure 4, Kaldi was used to extract 40-dimensional MFCC features from all audio in our dataset. Consequently, a ".list" file containing the extracted features of the Wenzhou dialect speech and their corresponding Mandarin text labels was prepared for model training. In addition, we constructed a character-based vocabulary file in fairseq format, named dict.ltr.txt. In this file, each entry comprised a token (i.e., character in Mandarin text) and its corresponding frequency count.

Model Fine-tuning The model was fine-tuned for up to 20K updates with a batch size of 10,000 tokens. We employed the Adam optimizer with a learning rate of 0.00002 and a three-stage learning rate scheduler with phase ratios of 0.1, 0.4, and 0.5. We used the connectionist temporal classification (CTC) (Graves et al., 2006) loss with the `zero_infinity` option, which enables to prevent numerical instability caused by infinite loss values during training. In addition, a masking probability of 50% was applied for both the time and channel dimensions to improve robustness. The layer dropout rate was set to 0.05, and the activation dropout rate was set to 0.1 to prevent overfitting. For the first 10,000 updates, the gradient of the feature extractor was frozen by setting the feature gradient multiplier to 0.

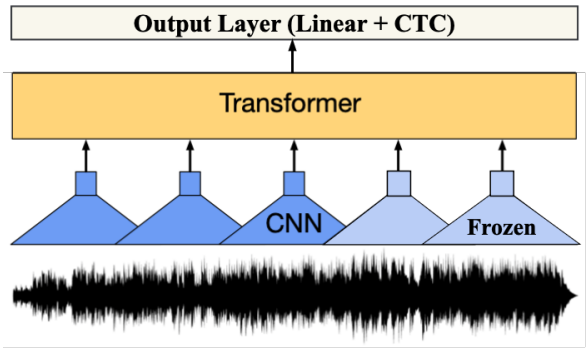

Figure 5: Fine-tuning architecture of Wav2Vec2-XLS-R-300M

4.1.2 Benchmark Model2: Wav2Vec2-XLS-R-300M

We fine-tuned the pretrained Wav2Vec2-XLS-R-300M model published on the Hugging Face Hub[7]. Hereafter, the fine-tuned model is referred to as Wav2Vec2-XLS-R-300M-Wenzhou.

Data Preprocessing The Wav2Vec2-XLS-R-300M model expects input in the format of a one-dimensional array with a 16,000-Hz sampling rate audio. We converted audio samples into one-dimensional arrays using the `Audio` feature from the `datasets` library. In addition to processing the speech signal into the input format of the model, we constructed a character-based vocabulary file in json format called vocab.json. In this file, each entry comprised a token (i.e., character) and a unique index assigned sequentially.

Model Fine-tuning A pretrained Wav2Vec2-XLS-R-300M model is an encoder, which converts an input Wenzhou dialect speech signal to a sequence of contextual representations. For our speech-to-text conversion task, we added an output layer with a linear layer to the top of the transformer block of the pretrained model to map the sequence of contextual representations to its corresponding Mandarin text, as illustrated in Figure 5. In particular, the output layer classifies each context representation into a token class, representing a Chinese character, by first transforming contextual representations into logits by the linear layer and then decoding the Mandarin text based on the logits by CTC. We fine-tuned the pretrained Wav2Vec2-XLS-R-300M model along with the output layer using our dataset.

[7] https://huggingface.co/facebook/wav2vec2-xls-r-300m

Model	CER
`TeleSpeech-ASR1.0-large-KeSpeech`	85.73
`TeleSpeech-ASR1.0-large-Wenzhou`	65.42
`Wav2Vec2-XLS-R-300M-Wenzhou`	29.94

Table 3: Evaluation results (CER (%))

The first component of the Wav2Vec2-XLS-R-300M model comprises a stack of convolutional neural network (CNN) layers, which are used to extract acoustically meaningful but contextually independent features (Baevski et al., 2020a). Although this part of the model has already been sufficiently trained during pretraining, there is no Wenzhou dialect speech in the pretraining data. Therefore, the last three CNN layers were unfrozen and trained during fine-tuning, whereas the other CNN layers of the feature extractor were frozen. This strategy leverages the pretrained knowledge while reducing computational overhead.

The CTC loss function was used as the objective function. We trained the model with a learning rate of 0.0001, employing a linear warmup of 1,000 steps followed by a gradual decay. Training was performed over 60 epochs with a per-device batch size of 8 and gradient accumulation across 2 steps, yielding an effective batch size of 16. We adopted mixed-precision (FP16) training to optimize memory usage. Furthermore, we employed a time masking probability of 0.05 as a regularization technique while not using layer dropout.

4.2 Evaluation and Results

In addition to our fine-tuned benchmark models, TeleSpeech-ASR1.0-large-Wenzhou and Wav2Vec2-XLS-R-300M-Wenzhou, we evaluated the TeleSpeech-ASR1.0-large-KeSpeech model open-sourced by Tele-AI (see Section 2.1). The model performance was evaluated in terms of CER. A lower CER indicates better conversion performance.

Table 3 summarizes the model performance on our test dataset. The table shows that TeleSpeech-ASR1.0-large-Wenzhou and Wav2Vec2-XLS-R-300M-Wenzhou outperform TeleSpeech-ASR1.0-large-KeSpeech. This demonstrates the effectiveness of the fine-tuning using our training dataset, thereby indicating the usefulness of our dataset.

The poor performance of TeleSpeech-ASR1.0-large-KeSpeech can be primarily attributed to it being fine-tuned using the KeSpeech dataset without

```
"prediction": "以盗到的美食物集畲乡人民为间远感来的客送央的满祝福充满浓很用的畲乡风情这些广大的游客吃望后都纷纷水疗",
"reference": "一道道的美食汇聚畲乡人民为远道而来的客人们献上的满满祝福充满很浓的畲乡风情令广大游客吃了以后都赞不绝口",
"cer": 0.47058823529411764
```


Figure 6: Example of output of Wav2Vec2-XLS-R-300M-Wenzhou on the test dataset

the Wenzhou dialect. However, it is not completely incapable of performing any conversion. This may be because the model was pretrained on data that included the Wenzhou dialect as described in Section 2.1.

Table 3 also indicates that Wav2Vec-XLS-R-300M-Wenzhou achieves the best performance among the three models. One reason why the CER of TeleSpeech-ASR1.0-large-Wenzhou on the test set is higher than that of Wav2Vec-XLS-R-300M-Wenzhou is that the predicted sequence of TeleSpeech-ASR1.0-large-Wenzhou was NULL for approximately 30% of the test samples. This may be attributed to the insufficient hyperparameter tuning during the training of TeleSpeech-ASR1.0-large-Wenzhou.

4.3 Discussion

Although Wav2Vec2-XLS-R-300M-Wenzhou achieves the best performance among the three models, its CER is not impressive. In this section, we discuss the areas for improvement of Wav2Vec2-XLS-R-300M-Wenzhou based on its actual output example.

Figure 6 shows an example of the output of Wav2Vec2-XLS-R-300M-Wenzhou on the test dataset. In the figure, "prediction" indicates the prediction sequence output by the model, and "reference" indicates the correct output sequence (i.e., label). The black boxes indicate the incorrect outputs of the model to be focused on.

In the first pair of black boxes, the same characters are used in each box, indicating that the model can approximately address the phonetic differences between the Wenzhou dialect and Mandarin. However, the order of Chinese characters is reversed, indicating that the model faces difficulty in handling the word order difference, similar to the phenomenon shown in Figure 3. The second pair of black boxes shows that the model is not that good at handling the differences in word usage between the Wenzhou dialect and Mandarin, which is similar to the phenomenon shown in Figure 2.

5 Conclusion

This study constructed the first publicly available dataset for converting the Wenzhou dialect speech-to-Mandarin text. Further, we also developed benchmark models, namely, the fine-tuned models of SSL-based pretrained models, and evaluated and reported their performance on the constructed dataset as baselines for future research. The evaluations demonstrated the effectiveness of fine-tuning the SSL-based pretrained models using our dataset. However, the results revealed significant challenges, particularly in terms of capturing the phonetic and syntactic intricacies of the Wenzhou dialect.

To address these limitations, we plan to explore strategies such as the incorporation of decoding architectures to better capture syntactic structures and account for word order differences in future work. In addition, we plan to leverage scaling pretrained models with larger cross-lingual capabilities, such as XEUS (Chen et al., 2024), to enrich representations for handling complex phonetic systems of the Wenzhou dialect.

Limitations

This study has several limitations that should be acknowledged. First, owing to the relatively small size of the constructed dataset, there may be a lack of speech diversity. Second, one of our benchmark models, TeleSpeech-ASR1.0-large-Wenzhou, is based on the TeleSpeech-ASR1.0 model. However, a paper detailing the TeleSpeech-ASR1.0 model has not been published, which significantly limits the clue to address the problem of TeleSpeech-ASR1.0-large-Wenzhou model predicting null sequences in certain test samples. Finally, the hyperparameter tunings of our fine-tuned models can be improved further.

References

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. Xls-r: Self-supervised cross-lingual speech represen-

tation learning at scale. In *Interspeech 2022*, pages 2278–2282.

Alexei Baevski, Steffen Schneider, and Michael Auli. 2020a. vq-wav2vec: Self-supervised learning of discrete speech representations. In *International Conference on Learning Representations*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020b. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Alexandre Berard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. In *NIPS*.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 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.

William Chen, Wangyou Zhang, Yifan Peng, Xinjian Li, Jinchuan Tian, Jiatong Shi, Xuankai Chang, Soumi Maiti, Karen Livescu, and Shinji Watanabe. 2024. Towards robust speech representation learning for thousands of languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10205–10224, Miami, Florida, USA. Association for Computational Linguistics.

Chinese Academy of Social Sciences and City University of Hong Kong. 2012. *Language Atlas of China*. The Commercial Press.

Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

S. Davis and P. Mermelstein. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366.

Ethnologue. 2017. How many languages in the world are unwritten? https://web.archive.org/web/20170227130129/https://www.ethnologue.com/enterprise-faq/how-many-languages-world-are-unwritten-0.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 369–376, New York, NY, USA. Association for Computing Machinery.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech 2020*, pages 5036–5040.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Xian Li, Changhan Wang, Yun Tang, Chau Tran, Yuqing Tang, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. Multilingual speech translation from efficient finetuning of pretrained models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 827–838, Online. Association for Computational Linguistics.

Isabella Steger. 2014. Do you dare try the devil-language? china's 10 hardest dialects. *The Wall Street Journal*.

Yun Tang, Hongyu Gong, Ning Dong, Changhan Wang, Wei-Ning Hsu, Jiatao Gu, Alexei Baevski, Xian Li, Abdelrahman Mohamed, Michael Auli, and Juan Pino. 2022. Unified speech-text pre-training for speech translation and recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1488–1499, Dublin, Ireland. Association for Computational Linguistics.

Zhiyuan Tang, Dong Wang, Yanguang Xu, Jianwei Sun, Xiaoning Lei, Shuaijiang Zhao, Cheng Wen, Xingjun Tan, Chuandong Xie, Shuran Zhou, Rui Yan, Chenjia Lv, Yang Han, Wei Zou, and Xiangang Li. 2021. Kespeech: An open source speech dataset of mandarin and its eight subdialects. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Sravya Popuri, Dmytro Okhonko, and Juan Pino. 2020. fairseq s2t: Fast speech-to-text modeling with fairseq. *arxiv:2010.05171*.

Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-to-sequence models can directly translate foreign speech. In *Interspeech*.

Xiaoying Xu, Xuefei Liu, Jianhua Tao, and Hao Che. 2012. Pitch and phonation type perception in wenzhou dialect tone. In *3rd International Symposium on Tonal Aspects of Languages (TAL 2012)*, pages paper P2–02.

Qian Zhang, Han Lu, Hasim Sak, Anshuman Tripathi, Erik McDermott, Stephen Koo, and Shankar Kumar. 2020. Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7829–7833.

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, Zhong Meng, Ke Hu, Andrew Rosenberg, Rohit Prabhavalkar, Daniel S. Park, Parisa Haghani, Jason Riesa, Ginger Perng, Hagen Soltau, Trevor Strohman, Bhuvana Ramabhadran, Tara Sainath, Pedro Moreno, Chung-Cheng Chiu, Johan Schalkwyk, Françoise Beaufays, and Yonghui Wu. 2023. Google usm: Scaling automatic speech recognition beyond 100 languages. *arxiv:2303.01037*.