university of groningen

campus fryslân

Transfer Learning for Sichuan Dialect Automatic Speech Recognition Based on pretrained Wav2vec 2.0 Model

ZiYi Li

university of groningen / campus fryslân

University of Groningen - Campus Fryslân

Transfer Learning for Sichuan Dialect Automatic Speech Recognition Based on pretrained Wav2vec 2.0 Model

Master's Thesis

To fulfill the requirements for the degree of
Master of Science in Voice Technology
at University of Groningen under the supervision of
Dr. Joshua Schäuble (Voice Technology, University of Groningen) (Voice Technology, University of Groningen)
with the second reader being
Dr. Matt Coler (Voice Technology, University of Groningen)

ZiYi Li(S6070310)

July 1, 2025

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor, Dr. Joshua Schäuble, for his invaluable guidance, patience, and encouragement throughout the course of this thesis. His expertise in voice technology and insightful feedback have been instrumental in shaping the direction and quality of my research.

I am also sincerely thankful to my fellow classmates and friends in the Voice Technology program. Your support, stimulating discussions, and willingness to share ideas and resources made this journey not only more manageable but also more enjoyable. In particular, I would like to thank [insert specific names if desired] for their help with data collection, technical troubleshooting, and moral support during challenging times.

Finally, I want to thank myself. Even though a lot of things happened in the second half of my studies, I am grateful that I persisted, even though I had a hard time and suffered from many illnesses.

Abstract

This thesis explores the application of self-supervised pre-trained models to low-resource dialectal speech recognition, using Sichuanese as a case study. We fine-tune the wav2vec2-large-xlsr-53 pre-trained model on a limited amount of manually transcribed Sichuanese speech, aiming to develop a practical automatic speech recognition (ASR) system in a highly resource-constrained setting. Our primary experimental results demonstrate that transfer learning can effectively reduce the character error rate (CER) from over 77% to below 28% using less than 11 hours of diverse training data. We further examine the impact of different training data compositions and propose a multi-source integration strategy that maintains performance while utilizing additional data. In contrast, a naive mixture of heterogeneous datasets significantly degrades model performance. Analysis reveals that data diversity plays a more crucial role than quantity in low-resource ASR, and that dialect-specific phenomena contribute notably to recognition errors. This study highlights the effectiveness of pre-trained models for dialectal ASR and provides practical insights into data selection and fine-tuning strategies. The proposed methodology contributes to the broader goal of enabling speech technologies for underrepresented languages and dialects.

Contents

1 Introduction

1.1 Research Background

Since the 1950s, the Chinese government has aggressively promoted Putonghua (Standard Mandarin) as a national lingua franca, in efforts to unify communication across its vast regions (Chen, 1999; Zhou, 2003). This language policy was enshrined in the 1982 Constitution – Article 19 explicitly mandates that "the state promotes the nationwide use of Putonghua"(Kim, 2017), solidifying Mandarin's status as the common language of China. Over subsequent decades, the uptake of Putonghua has been remarkable: by 2021, more than 80% of China's population could speak Mandarin(The State Council of the People's Republic of China, 2021), a figure that has risen sharply from the turn of the century. In some eastern coastal provinces, The literacy rate of Mandarin has now exceeded 95%(Zhao & Wu, 2020), virtually achieving seamless inter-regional communication. This high adoption of a standard language has yielded significant socio-economic benefits – from nationwide media and advertising markets to large-scale e-commerce and call-center operations – by providing a "linguistic infrastructure" that minimizes communication costs. The network effects of Putonghua adoption are self-reinforcing: as more citizens, businesses, and institutions standardize around Mandarin, public and private investments increasingly prioritize Mandarin-based technologies. For instance, major national innovation funds—such as the China Internet Investment Fund and the National Technology Transfer Fund—have directed billions into AI and speech technologies that rely heavily on Mandarin training data, further entrenching it as the de facto linguistic infrastructure of China.

However, the same nation-building process has put China's regional dialects under unprecedented pressure of marginalization. UNESCO's Atlas of the World's Languages in Danger lists 36 Chinese dialects as endangered or severely endangered (Bradley, 2005; Moseley, 2010). In other words, dozens of traditional Chinese local varieties are at risk of falling out of use within the coming generations. The imbalance has only been exacerbated in the AI era: development of speech technologies requires massive, richly annotated training corpora, yet dialect data are costly and scarce. Collecting and transcribing dialect speech can cost several times more per hour of audio than for Mandarin(J. Li, Zheng, Byrne, & Jurafsky, 2006), while the potential user base for a given dialect may be only a fraction of the national market. For example, producing 1 hour of high-quality transcribed audio in Cantonese or Minnanese can be several times more expensive than in Putonghua, even though speakers of those dialects number less than one-tenth of China's population. Faced with such economics, many commercial AI projects for dialects have been downsized or shelved in favor of Mandarin-centric initiatives. Moreover, rapid urbanization and education policies have accelerated language shift among younger generations. In the past twenty years, China's urbanization rate rocketed from roughly 36% in 2000 to over 65% in 2022(Textor, 2025), meaning tens of millions of youth from dialect-speaking regions have moved to cities and undergone schooling in Putonghua. This has led to "demother-tongue-ization" – younger speakers often have limited proficiency in their local dialects – resulting in even fewer competent dialect speakers and a severe intergenerational break in dialect transmission. Although regional media, viral videos, and cultural preservation campaigns occasionally spark renewed interest in local speech (with some dialect content even being designated as intangible cultural heritage), these sporadic surges of popularity have not translated into sustained support for dialect-focused language technology. In summary, against the backdrop of China's successful Mandarin promotion and its attendant economic boon lies a concerning decline

in dialect vitality and a widening technological gap for non-Mandarin languages.

1.2 Problem Statement

The SiChuan dialect (SiChuanese Mandarin) exemplifies this technology gap. As one of the largest Chinese dialect groups – covering over 100 million speakers in Southwestern China – SiChuanese has distinct phonological characteristics that pose serious challenges for standard speech recognition systems(Y. Li, Best, Tyler, & Burnham, 2020). Linguistically, SiChuan dialect retains features that have been lost in standard Mandarin. For instance, many SiChuan sub-varieties preserve the Middle Chinese entering tone (rùshēng), manifested as short, abruptly stopped syllables that do not occur in Putonghua. SiChuanese also exhibits unique tonal patterns: it generally has four tones like standard Mandarin, but with different pitch contours and even additional tone splits or mergers in certain areas (Y. Li et al., 2020). Notably, there is significant regional variation within SiChuanese – the pronunciation differs markedly between sub-dialects, such as the Chengdu-Chongqing area versus southern SiChuan regions (Cheng, 2022). These divergences in phonology and tone mean that an automatic speech recognition (ASR) model trained on standard Mandarin audio struggles to parse SiChuan dialect input, since it encounters sounds and tone realizations outside its training experience.

Compounding the issue, existing Mandarin ASR systems perform poorly on SiChuan dialect speech, with studies reporting character error rates (CER) of 30–35% when state-of-the-art Mandarin models are applied to SiChuanese input, far exceeding the threshold for practical usability (Q. Li, Mai, Wang, & Ma, 2024). In contrast, CER for standard Mandarin is significantly lower, highlighting the severe performance degradation caused by dialect mismatch. This high error rate underscores the critical need for dialect-specific ASR solutions to enable effective real-world applications. For instance, in domains such as telemedicine (e.g., remote consultations in local dialect), smart home controls, or voice-activated customer services, such error-prone recognition—where nearly one in three characters is misrecognized—renders systems frustrating or unusable. This issue disproportionately affects elderly and rural SiChuanese speakers who primarily use the local dialect, limiting their access to voice-driven digital services and exacerbating the digital divide. In essence, the lack of robust SiChuan dialect ASR not only degrades user experience but also perpetuates techno-linguistic inequality.

From a technical perspective, SiChuan dialect ASR faces several major challenges:

- Distinct phonetic features: The presence of rùshēng (checked-tone) syllables in SiChuanese results in very short, closed syllables that Mandarin-trained models do not handle well. These abrupt end-of-syllable stop consonants often cause recognition errors since standard models lack exposure to such patterns.

- Tonal system differences: SiChuan dialect's tone system, while based on four lexical tones, includes regional tonal variations and contour differences from standard Mandarin. A model trained only on Putonghua's tonal patterns struggles to map SiChuanese tonal cues to correct characters or words.

- Data scarcity: Labeled speech data for SiChuan dialect is extremely scarce. Publicly available SiChuanese speech corpora comprise under 100 hours of transcribed audio in total with various quality – orders of magnitude less than what is available for Mandarin. This paucity of training data severely limits the performance of dialect-specific acoustic models.

- Internal dialect diversity: SiChuanese itself is not monolithic – it consists of sub-dialects with considerable variation in pronunciation and vocabulary. Differences between, say, the ChengYu Area and the GuanChi Area mean that an ASR model may not generalize well across all varieties of "SiChuan dialect," especially if the training data covers only a narrow subset(D. Li, 2017).

These factors have created a formidable technical gap: off-the-shelf Mandarin speech recognizers cannot simply be applied to SiChuan dialect users with acceptable accuracy, yet creating a high-quality SiChuan-specific ASR system is difficult due to limited data and linguistic complexities. Bridging this gap is crucial to ensure that speakers of SiChuan dialect – and other under-resourced Chinese dialects – are not left behind in the advancement of speech technology.

1.3 Research Goals and Contributions

Transfer learning offers an effective pathway to address the above challenges. In the context of speech recognition, transfer learning involves leveraging a model pretrained on a large general dataset (often using self-supervised learning on vast amounts of unlabeled audio) and fine-tuning it on a smaller target-domain dataset. This approach has proven highly successful for low-resource languages, because the pretrained model provides rich latent representations of speech that can be adapted with relatively few labeled examples (Howard & Ruder, 2018; Pan & Yang, 2010). Recent breakthroughs in self-supervised speech models like wav2vec and wav2vec 2.0 have demonstrated the power of this paradigm. Schneider, Baevski, Collobert, and Auli (2019a) first introduced wav2vec, an unsupervised pre-training method that learns speech features from raw audio, and Baevski, Zhou, Mohamed, and Auli (2020) presented wav2vec 2.0, a refined framework using a contrastive objective and transformer architecture to learn robust representations from thousands of hours of speech audio. Notably, wav2vec 2.0 achieved state-of-the-art ASR results by fine-tuning on limited supervised data. Building on this, Conneau, Baevski, Collobert, Mohamed, and Auli (2020) extended the approach to a multilingual setting with the XLSR (cross-lingual speech representations) model, which was pretrained on 53 languages (including Mandarin Chinese) and showed that a single model can be adapted to many languages with excellent performance. These advances imply that a pretrained model can be transfer-learned to a specific dialect like SiChuanese, potentially overcoming the data scarcity by using knowledge learned from other speech data.

The goal of this research is to apply transfer learning with wav2vec 2.0 to significantly improve ASR for SiChuan dialect. Specifically, we fine-tune the pretrained wav2vec2-large-xlsr-53 model (which has been trained on a diverse collection of languages including Chinese Mandarin) on a curated SiChuan dialect speech corpus. By doing so, we aim to reduce the character error rate on SiChuan dialect speech from over 70% to below 30%, bringing it into an acceptable range for real-world applications. In practical terms, a CER under 30% would mark a dramatic improvement, potentially enabling usable voice interfaces for SiChuan dialect speakers. This study operates under an extremely low-resource scenario – our fine-tuning uses 11 hours of transcribed SiChuanese audio – to test the limits of how far transfer learning can go in a data-sparse setting. Achieving the target performance under these conditions would validate the feasibility of dialect ASR with minimal data, and could serve as a blueprint for other under-resourced dialects or languages.

The main contributions of this research are summarized as follows: Now that the motivation for this research has been presented, the structure of this thesis is as follows:

- Construction of a reproducible SiChuan dialect ASR benchmark: We assemble and release a standard corpus for SiChuanese ASR by combining an open-source MagicData SiChuan dialect dataset (6.4 hours) with an additional 45-hour collection of SiChuan narrative speech from local performer Li Boqing on GitHub[1] is a popular platform for code sharing.. We perform careful preprocessing and create a standardized train/validation/test split, along with consistent evaluation protocols. This provides the first publicly documented benchmark dataset for SiChuan dialect ASR, enabling future researchers to reproduce results and compare methods on a common platform.

- Systematic transfer learning methodology for dialect ASR: We propose an end-to-end transfer learning framework tailored to low-resource dialect speech recognition. This includes effective data preprocessing techniques (such as audio augmentation and noise filtering strategies), optimal fine-tuning configurations for the wav2vec 2.0 model on dialect data, and a comprehensive evaluation scheme (using character error rate as the primary metric, alongside analyses by phoneme and tone error). The methodology is presented in a step-by-step manner, providing a reference blueprint for applying self-supervised pretrained models to other dialects or low-resource languages.

- In-depth ablation studies and analysis: We conduct extensive experiments to dissect the impact of various factors on SiChuan dialect ASR performance. In particular, we compare model outcomes under different training data combinations (e.g., using the MagicData subset alone vs. adding the larger narrative corpus) to understand the effect of data diversity. We examine the influence of a high Signal-to-Noise Ratio (SNR) data filtering (e.g., using only clips with SNR $\geq 20\,\mathrm{dB}$) on model generalization. These ablation studies provide empirical evidence on whether common data cleaning practices help or hurt in low-resource dialect scenarios, offering insights for dialect ASR data engineering best practices.

- Demonstration of practical feasibility: By achieving a substantial reduction in CER with only 11 hours of labeled data, we demonstrate that industry-grade ASR for a major dialect can be developed with very limited resources through transfer learning. The system developed in this work shows that, given a powerful pretrained model, even a small dialect corpus can yield a functional speech recognizer. This finding lays a foundation for the commercialization and deployment of dialect ASR technologies – it indicates that tech companies or research labs can realistically build voice interfaces for dialects like SiChuanese without the prohibitive expense of collecting hundreds of hours of training data. Our results thus serve as a proof-of-concept for bringing inclusive speech technology to dialect speakers, ultimately helping to narrow the digital divide discussed earlier.

In summary, our research not only delivers an improved SiChuan dialect ASR model but also provides valuable artifacts (datasets, code, and methodologies) and insights that can accelerate future work on low-resource speech recognition. We show that with modern transfer learning techniques, the unique linguistic features of a dialect and the paucity of training data, while challenging, are not insurmountable obstacles.

[1] https://github.com/lzy13890024272github/Transfer-Learning-for-Sichuan-Dialect-Automatic-Speech-Recognition-Based-on-pretrained-Wav2vec2.0

1.4 Research Questions and Hypotheses

To focus the investigation, this study centers around the following key research questions:

> **To what extent can SiChuan dialect ASR performance be improved through performing transfer learning based on the pretrained Mandarin wav2vec 2.0 model?**

This main question can be decomposed into the following sub-questions, each addressing a specific aspect of the transfer learning process and its contributing factors:

- RQ1: Can transfer learning using a large-scale Mandarin-pretrained model (wav2vec2-large-xlsr-53-chinese-zh-cn[2]) significantly reduce the character error rate (CER) of SiChuan dialect ASR, even under low-resource conditions?
- RQ2: What impact does applying a standard signal-to-noise ratio (SNR) filtering threshold ($\geq 20\,\text{dB}$) have on the model's generalization ability and overall ASR performance? speech alone?
- RQ3: How does the inclusion of single-speaker narrative speech, when combined with multi-speaker conversational data, affect ASR performance compared to using multi-speaker data alone?

To guide empirical exploration, the following hypotheses are proposed in correspondence with the sub-questions:

- Hypothesis 1 (H1): Fine-tuning a large-scale Mandarin-pretrained model on a small amount of SiChuan dialect speech will significantly reduce CER, demonstrating that transfer learning can yield acceptable ASR performance even in low-resource settings.
- Hypothesis 2 (H2): SNR filtering will have only marginal or no positive impact on ASR performance in this context. While filtering improves audio quality, it also reduces dataset size and variability, which may limit the model's ability to generalize to noisy or diverse real-world conditions.
- Hypothesis 3 (H3): Integrating the single-speaker narrative dataset significantly improves model performance by increasing training volume and linguistic coverage, despite its limited speaker diversity.

Together, these sub-questions and hypotheses provide a structured framework for investigating the potential and limitations of transfer learning for dialect-specific ASR, with a focus on practical gains in recognition accuracy under real-world constraints.

[2] https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-chinese-zh-cn

2 Literature Review

Automatic speech recognition (ASR) has evolved significantly, yet substantial challenges remain for dialect-specific ASR due to unique phonetic, tonal, and linguistic features, compounded by limited data availability. This section provides a comprehensive review of research on Chinese dialect automatic speech recognition (DSR), with specific focus on the gap of low-resource dialects and how Mandarin-based transfer learning is being used to bridge that gap. Chinese dialect ASR faces significant challenges due to the linguistic diversity and data scarcity of dialects. By surveying recent literature, we highlight the current state of dialect ASR and underline why leveraging Mandarin (the standard language with abundant resources) is crucial for improving dialect recognition performance.

2.1 Search Methodology

We conducted a structured literature search to ensure a broad and relevant coverage of this topic. Major databases and libraries (e.g., Google Scholar, ACL Anthology, CNKI, IEEE Xplore, arXiv) were queried for studies on Chinese dialect speech recognition. The search spanned primarily the last decade (2015–2024) to capture developments in the deep learning era while including earlier foundational works when relevant. Both English and Chinese-language publications were considered (to include important domestic research). The following keyword strategy was used:

- **Chinese Dialect ASR:** "Chinese dialect speech recognition,""Chinese Mandarin ASR," "dialect ASR," "low-resource speech recognition," "Chinese dialect corpus"
- **Transfer Learning & Multilingual ASR:** "Mandarin transfer learning," "multilingual ASR," "cross-lingual speech recognition," "dialect adaptation"

After gathering literature, we applied inclusion and exclusion criteria to select the most relevant works:

- **Inclusion:** Peer-reviewed studies focusing on automatic speech recognition for Chinese dialects, especially those leveraging Mandarin or other high-resource languages to improve dialect ASR (e.g., transfer learning or multilingual models).
- **Inclusion:** Recent works (approximately 2015–2024) from major speech and AI conferences/journals (e.g., Interspeech, ICASSP, IEEE/ACM TASLP) and survey papers, to capture current state-of-the-art methods.

Using this methodology, we identified key themes and approaches in the literature, which we organize and discuss below.

2.2 Challenges and Current Status of Dialect ASR

Dialectal variations in speech present specific difficulties for automatic speech recognition (ASR) systems primarily trained on standard Mandarin. Chinese "dialects" are often linguistically as distinct as separate languages – many are mutually unintelligible and have unique phonological systems

(Chen, 1999; Zhou, 2003). For example, Southwestern Mandarin (which includes the Sichuan dialect) uses different tonal patterns and pronunciations compared to standard Mandarin (Y. Li et al., 2020). These divergences mean that models trained on one variety (e.g. standard Mandarin) do not generalize well to others, resulting in significantly higher error rates when applied to dialect speech. Studies have documented substantial performance degradation when a Mandarin ASR system is used on dialectal input, underscoring the need for dialect-specific modeling (Q. Li et al., 2024). In the following subsections, we discuss two primary challenges in Chinese dialect ASR: key technical hurdles and the pervasive issue of data scarcity.

2.2.1 Technical Challenges in Chinese Dialect ASR

ASR for Chinese dialects faces several technical hurdles due to fundamental linguistic differences between dialects and standard Mandarin. Pronunciation and lexical variations are prominent – dialects often have distinct sound inventories (consonants, vowels, tone systems) that standard Mandarin models do not cover (Q. Li et al., 2024). Unlike many European dialects which share a base phonetic system, Chinese dialects can differ dramatically. For instance, Cantonese has six or more tones whereas Mandarin has four, and on the other hand some Southwestern Mandarin dialects have merged certain tone categories or consonant distinctions (Y. Li et al., 2020). Without explicit handling of these differences, an ASR system may confuse words that are minimal pairs in a dialect but not in Mandarin. It has been observed that a state-of-the-art Mandarin ASR model performs poorly on non-Mandarin speech, with error rates spiking on dialectal utterances (Wu et al., 2024). This performance gap highlights the importance of incorporating dialect-specific knowledge into the acoustic and language models. Past research pointed out these issues early on – for example, accent mismatches and unseen pronunciations were found to severely degrade recognition accuracy(Baevski et al., 2020; Zhou, 2003). Overall, the heterogeneity of Chinese dialects poses unique challenges to ASR, necessitating specialized techniques to handle divergent phonology and pronunciation patterns (Q. Li et al., 2024).

2.2.2 Scarcity of Data Resources

A primary bottleneck for dialect ASR is the severe lack of high-quality training data. Developing robust speech recognition systems requires large volumes of transcribed audio, yet building a sizable dialect corpus is expensive and labor-intensive (Q. Li et al., 2024). Many Chinese dialects are spoken by relatively small populations, sometimes in remote areas, and some are even endangered (Bradley, 2005; Moseley, 2010). This means less existing audio data. Collecting new dialect speech corpora is costly, as it requires recruiting native speakers and expert annotators—and many dialects lack a standardized orthography, making transcription even harder (Besacier, Barnard, Karpov, & Schultz, 2014). For instance, certain regional variants use colloquial words or pronunciations that are not reflected in written Chinese, complicating the transcription process. As a result, most speech resources in China are heavily skewed toward standard Mandarin, with dialect corpora remaining very limited in size and scope (Q. Li et al., 2024). According to recent reports, over 80% of China's population could communicate in Mandarin by 2020 (The State Council of the People's Republic of China, 2021), reflecting national language promotion efforts. While positive for mutual communication, this also implies that daily usage of local dialects has diminished, further reducing the opportunities to gather natural dialect speech data. In summary, data scarcity – in terms of both quantity and

diversity – substantially hinders the progress of Chinese dialect ASR. In particular for Sichuanese ASR, where publicly available SiChuanese speech corpora comprise under 100 hours of transcribed audio in total with various quality and high-quality transcribed speech is even scarcer, any approach aiming to improve dialect recognition must contend with this paucity of labeled examples.

2.3 Transfer Learning in Low-resource Speech Recognition

Transfer learning has emerged as a promising approach to address the data scarcity problem in speech recognition, by leveraging knowledge from resource-rich languages or tasks to improve low-resource scenarios (Yadav & Sitaram, 2022). Instead of training an ASR model from scratch on a limited dialect dataset, a model can be pre-trained on a large amount of data from a related high-resource language (such as standard Mandarin or even other languages) and then fine-tuned on the target dialect. This technique effectively transfers learned representations and acoustic models, providing a head start for the low-resource dialect. Researchers have explored various transfer learning strategies relevant to dialect ASR, including self-supervised pre-training on massive speech corpora, supervised model adaptation and fine-tuning for specific dialects, and multilingual training that allows cross-lingual knowledge transfer. This section reviews these strategies and how they have been applied to Chinese dialect speech recognition.

2.3.1 Rise of Self-supervised Pre-training Models

In recent years, self-supervised learning has revolutionized speech recognition by enabling powerful pre-trained models that do not require transcriptions. In the seminal wav2vec framework, a neural network is first trained on a large quantity of unlabeled audio to learn high-level speech representations, which can later be fine-tuned with a small labeled dataset for ASR (Schneider, Baevski, Collobert, & Auli, 2019b). The second-generation model, wav2vec 2.0, further improved this approach by using a transformer architecture and masked prediction of latent speech units, achieving state-of-the-art results on multiple benchmarks (Baevski et al., 2020). Critically, wav2vec 2.0 learned rich acoustic representations from thousands of hours of audio, making it extremely effective when adapted to a new task with limited data. Following wav2vec, other self-supervised models like HuBERT introduced the idea of predicting clustered latent representations and showed similar gains Hsu et al. (2021). These pre-trained models capture fundamental properties of speech (phonetics, speaker characteristics, etc.) from vast corpora. When fine-tuned on a low-resource dialect, they provide a strong initialization, often outperforming models trained from scratch by a wide margin. For example, even with only a few hours of dialect data, fine-tuning a wav2vec 2.0 base model can yield competitive word error rates, whereas a conventional system would severely overfit or fail to learn (Baevski et al., 2020). The rise of self-supervised pre-training thus offers a powerful transfer learning paradigm: the dialect ASR can inherit knowledge from general speech representations learned on larger datasets (possibly Mandarin or multi-lingual audio) and thereby compensate for its limited training data.

2.3.2 Dialect-specific Transfer Learning Strategies

Dialect-specific automatic speech recognition (ASR) faces significant challenges due to data scarcity and high variability across dialects. To address this, transfer learning has become a crucial strategy,

leveraging pretrained models to adapt to low-resource dialect datasets. This section outlines three common transfer learning approaches—Full Fine-tuning, Partial Transfer, and Multitask Learning—highlighting their applications and distinctions in dialect-specific ASR contexts.

- **Full Fine-tuning** involves adapting all parameters of a pretrained speech model to the target dialect dataset. This approach effectively captures dialect-specific features, making it suitable for dialects like Mandarin Wu or Arabic Levantine with sufficient data. However, it is computationally intensive and risks overfitting in low-resource scenarios. This technique has shown clear improvements: for instance, experiments adapting Mandarin ASR models to Cantonese achieved lower character error rates than training solely on the limited Cantonese data (Q. Li et al., 2024).

- **Partial Fine-tuning** updates only a subset of the pretrained model's parameters, typically higher-level layers (e.g., top Transformer layers), while keeping lower layers frozen to retain robust general representations. This reduces computational costs and mitigates overfitting but may not fully capture prosodic or phonetic variations unique to certain dialects.Researchers have explored which parts of the model to transfer– some studies found transferring lower-layer acoustic representations yields gains, while others also transplant higher layers and even language model components (Xie, Sui, Liu, & Wang, 2022).

- **Multitask Learning** simultaneously optimizes multiple related tasks, such as standard language and dialect recognition, using shared lower-level feature extraction layers and task-specific output heads. This enhances generalization by learning commonalities and distinctions between dialect and standard speech. However, careful task weighting, such as adjusting loss weights, is necessary to prevent one task from dominating training. A study by (Xu, Dan, Yan, Ma, & Wang, 2021a), for example, combined transfer learning with data augmentation to success fully improve recognition of several low-resource Chinese dialects.

Overall, dialect-specific transfer learning strategies– whether simple fine-tuning or more complex adaptive training– have proven effective in boosting ASR performance. They capitalize on the existence of related models (or data) in Mandarin or other languages, bridging the resource gap for dialects by reusing learned knowl edge. This significantly mitigates issues of insufficient data and helps the model handle dialectal pronunciations and acoustics more adeptly.

2.3.3 Cross-lingual Transfer through Multilingual Pre-training

An extension of the transfer learning concept is multilingual pre-training, where a model is trained on multiple languages (or dialects) simultaneously to learn a universal speech representation. The resulting model can then be fine-tuned to a specific target dialect. This approach enables cross-lingual knowledge transfer: features learned from one language may benefit recognition in another, especially if they share some linguistic properties. A notable example is the XLSR (cross-lingual speech representation) model built on wav2vec 2.0, which was pre-trained on 53 languages including Chinese variants (Conneau et al., 2020). XLSR demonstrated that a single model can encode diverse languages and, when fine-tuned, outperform monolingual models on low-resource languages by leveraging shared phonetic patterns across languages (Conneau et al., 2020). For Chinese dialect ASR, multilingual pre-training offers the possibility of implicitly transferring knowledge from both

standard Mandarin and other languages with similar characteristics. Researchers have applied this idea to Tibetan and other minority languages in China – for instance, (Wang et al., 2022) fine-tuned a Mandarin-trained model to recognize Lhasa Tibetan, achieving substantial improvements over training from scratch. In a similar vein, a multilingual ASR model trained on many Chinese dialects and Mandarin can serve as a strong starting point for any specific dialect, as it has already seen a variety of pronunciations and acoustic conditions (Yadav & Sitaram, 2022). Cross-lingual transfer is further facilitated by data augmentation techniques like mixing dialect data or using universal phonemic representations. The success of multilingual pre-trained models underscores a key insight: information learned from other languages or dialects can compensate for data deficiencies in the target dialect. By sharing representational space across languages, the model can generalize better to new dialects. This strategy is highly relevant for Sichuan dialect ASR, since one can leverage models trained on standard Mandarin and perhaps other Chinese dialect corpora to bootstrap recognition for Sichuan speech.

2.4 Data Engineering and Pre-processing Strategies

In low-resource dialect ASR, the quality of the training data is just as important as the quantity. Given the limited size of available Sichuan dialect corpora, effective data engineering and pre-processing are critical to maximize the utility of every recording. This section discusses how careful dataset preparation can improve model performance. We focus on two aspects: (1) Data quality control, which involves cleaning and balancing the corpus to reduce noise and bias; and (2) Data augmentation techniques, which artificially expand the training set to mitigate overfitting and improve robustness.

2.4.1 Data Quality Control

Ensuring high-quality data is a fundamental step before model training. For dialect speech corpora, quality control may include verifying transcript accuracy, normalizing orthography, filtering out audio with excessive noise, and balancing speaker demographics. Even small inconsistencies or errors in transcripts can significantly impact an ASR model trained on a tiny dataset. Therefore, researchers often invest effort in manual review or semi-automatic cleanup of dialect data (Q. Li et al., 2024). Standard practices from corpus design in Mandarin can be applied – for example, ensuring a diverse set of speakers and phonetic coverage, as outlined by (A. Li et al., 2004) in the design of early Chinese speech corpora. However, many dialect corpora are collected in ad-hoc ways (e.g., regional field recordings or crowdsourced data) and may lack uniform standards. It becomes necessary to post-process such data: removing mistranscribed utterances, correcting inconsistent spellings, and segmenting long recordings into manageable utterances. Another challenge is text normalization for dialects: deciding on written forms for purely oral dialect words. Researchers might map dialectal vocabulary to standard Chinese characters, or use Chinese Pinyin, but whichever approach, it should be applied consistently across the corpus (Q. Li et al., 2024). But in our experiments, we did not employ any additional vocabulary mapping or dialect romanization. Instead, we directly used the original transcripts provided in standard Chinese characters, thus maintaining consistency implicitly by relying on standardized transcriptions from the data source. Additionally, to avoid bias, the data should ideally be balanced for different speaking styles, ages, and genders. If one speaker or one type of content dominates the corpus, the model could overfit to that. In summary, rigorous data

quality control increases the effective information content of a low-resource dataset. It reduces the noise the model has to contend with and leads to more reliable and generalizable training. Given the expense of gathering dialect data, maximizing its quality is a cost-effective way to improve ASR performance without needing more data.

2.4.2 Data Augmentation Techniques

Alongside quality improvements, data augmentation is a vital strategy to artificially increase the amount and diversity of training data for dialect ASR. Augmentation involves applying various transformations to existing audio to create new "virtual" training examples, thereby helping the model generalize better. A simple and widely used augmentation is speed perturbation, where the audio is slightly stretched or compressed in time (Ko, Peddinti, Povey, & Khudanpur, 2015). This simulates different speaking rates: for example, an utterance can be made 10% slower or faster, producing a new sample with the same transcript but a different duration and pitch. Speed perturbation effectively increases the dataset size and makes the model more robust to tempo variations (Ko et al., 2015). Other common methods include pitch shifting (altering the fundamental frequency to mimic different speaker intonations), additive noise (mixing background noise at various signal-to-noise ratios to teach the model to handle noisy environments), and reverberation or impulse response filtering (to simulate different room acoustics). More recently, spectrogram-domain augmentations like SpecAugment randomly mask portions of the spectrogram (in time or frequency) during training, forcing the model to learn invariant features (Park et al., 2019). SpecAugment proved highly effective in improving ASR robustness without needing extra data, and it has become a standard component in training pipelines for state-of-the-art systems (Park et al., 2019). For dialect ASR, which often suffers from overfitting due to small data, augmentation is especially useful. By generating plausible variations of the existing recordings, the model is exposed to a broader range of speaking styles and acoustic conditions than the original corpus provides. This leads to improved generalization – the model is less likely to latch onto spurious patterns or noise present in the limited data. In the context of Sichuan dialect, employing a suite of augmentation techniques can help ensure the trained acoustic model is not overly tuned to the specific speakers or recording setup of the training set, thereby improving its performance on new speakers of the dialect.

2.5 Tone and Prosody Modeling in Chinese Dialects

Tone and prosody are central features of Chinese languages, and their proper handling is crucial for accurate speech recognition. Most Chinese dialects are tonal, meaning that pitch patterns (tones) distinguish word meanings. The Sichuan dialect, like other Southwestern Mandarin variants, has its own tonal system and intonation patterns that differ from standard Mandarin (Y. Li et al., 2020). These differences in prosody can pose challenges: an ASR model trained on Mandarin might misrecognize a Sichuan dialect word if it relies on Mandarin's tone-frequency mapping. In this section, we examine how tone and other phonological phenomena specific to dialects are modeled in ASR systems. We first discuss representation of tonal features, then methods to handle dialect-specific phonological variations (such as unique sound mergers or sandhi).

2.5.1 Representation and Modeling of Tonal Features

In tonal dialects, the pitch contour of each syllable carries lexical information, so the ASR system must capture and interpret these tone cues. Standard MFCC or mel-spectrogram features implicitly contain some pitch information, but often not enough for high accuracy on tonal distinctions – especially if the training data is limited. Therefore, researchers have explored explicitly incorporating tone features into the acoustic model. One approach is to append pitch-related features (like fundamental frequency F_0 and delta-F_0) to the input feature vector for each frame of audio (Q. Li et al., 2024). By doing so, the model is informed of the tone height and contour, which can help it differentiate words that differ only by tone. Another approach is to have a separate prediction task or network branch for tone classification in a multi-task learning setup, ensuring the model learns to recognize tones in addition to the main speech recognition task. For Mandarin, such techniques have been shown to improve recognition of tone-heavy contexts (Q. Li et al., 2024). We expect similar benefits for Sichuan dialect ASR, given its tonal nature. The Sichuan dialect's tone system is not identical to standard Mandarin – for instance, it has fewer tonal distinctions in certain positions and different pitch contours for what Mandarin would consider the same tone number (Y. Li et al., 2020). An ASR model that naively assumes Mandarin tone patterns might confuse Sichuan dialect tones. By training the model with dialect-specific tone annotations or ensuring the pitch features are prominent, we can reduce tone substitution errors. Prosodic modeling also extends to intonation and stress patterns over phrases, which can influence recognition (though lexical tone is the primary factor in Chinese dialects). In summary, tone modeling is a key aspect of dialect ASR: including explicit tone information and training the acoustic model to leverage it leads to better disambiguation of homophones and overall improved accuracy for tonal languages.

2.5.2 Handling Dialect-specific Phonological Phenomena

Beyond tones, Chinese dialects exhibit various phonological phenomena that differ from standard Mandarin – these must be addressed to achieve high recognition performance. The Sichuan dialect, for example, has notable pronunciation differences, such as the merging of certain consonant sounds and vowel variations. One well-known case is the neutralization of retroflex vs. alveolar sibilants: Standard Mandarin distinguishes sounds like "sh" vs. "s", but many Southwestern Mandarin speakers (including in Sichuan) pronounce them the same, effectively merging these phonemes (W. Zhang & Levis, 2021). If an ASR system expects the standard distinction, it may incorrectly decode a Sichuan dialect speaker's input. To handle such differences, dialect-specific lexicons or phoneme sets are often developed. A dialect lexicon would include alternate pronunciations of words as spoken in the dialect. For instance, a Mandarin word with an initial "zh" sound might be listed with a "z" sound in the Sichuan dialect lexicon. During decoding, the ASR system can then consider dialectal pronunciations as valid hypotheses (Q. Li et al., 2024). In end-to-end systems, one cannot directly modify a lexicon, but one can train the model on dialect transcripts written in a way that reflects dialect pronunciation (e.g., using dialect spelling or appropriate phonemic transcription), so that the model learns the mapping from audio to the dialect-specific text. Another phenomenon is dialect-exclusive vocabulary or colloquial expressions that do not appear in standard Mandarin text corpora. This necessitates adapting the language model: incorporating dialect text data or translating common dialect words into characters. Without this, the ASR might force an acoustic decode into the "closest" Mandarin word, yielding transcription errors. Including Sichuan

dialect phrases and pronunciations in the training data and language model has been found important for improving recognition of those dialectal words (Q. Li et al., 2024). In summary, addressing dialect-specific phonology in ASR involves augmenting the system's knowledge of how words are pronounced in the dialect and possibly adjusting the modeling units. Prior research on accented Mandarin speech recognition and dialect adaptation has used methods like pronunciation augmentation, acoustic model adaptation, and even training separate models for each dialectal variant (Bahari, Saeidi, Van hamme, & Van Leeuwen, 2013). Our work will build on these insights, ensuring that the peculiarities of Sichuan dialect speech – from sound mergers to unique vocabulary – are accounted for in the recognition process, either through training data choices or model architecture.

2.6 Evaluation Methods and Metrics

Evaluating the performance of a dialect ASR system requires careful consideration of metrics. Conventional ASR evaluations use metrics like Word Error Rate (WER) or Character Error Rate (CER) to quantify accuracy. However, for Chinese dialects, these traditional metrics have certain limitations. This section first examines why standard evaluation metrics may be insufficient or misleading for dialect ASR, and then discusses more fine-grained evaluation approaches that can provide deeper insight into system performance on dialect speech.

2.6.1 Limitations of Traditional Evaluation Metrics

The most common metric for speech recognition is WER, which calculates the percentage of words that are inserted, deleted, or substituted in the ASR hypothesis relative to a reference transcript. For character-based languages (and character-based ASR systems), a similar metric is CER, computed at the character level. While WER/CER are useful overall indicators, they may not fully reflect the intelligibility or the specific weaknesses of a dialect ASR system. One issue is that dialectal variations can cause systematic transcription differences that inflate WER even if the meaning is preserved. For example, if the reference uses a dialect-specific word but the ASR outputs an equivalent standard Mandarin word, it would count as an error in WER, despite the core meaning being understood. Conversely, an ASR might get the main content right but mis-recognize some dialect-specific particle or tonal inflection, and WER would penalize it without telling us what went wrong. Researchers have noted that in accented or dialectal speech recognition, WER alone doesn't diagnose whether errors come from acoustic confusions, tone mistakes, or language model issues (Bahari et al., 2013). Bahari et al. (2013) used an accent classification accuracy metric to complement ASR error rates, arguing that a system's ability to identify dialectal accent could be relevant when standard metrics fall short. Another limitation is that WER/CER do not account for the significance of errors – a minor functional word error and a major content word error both count equally. In the context of dialect ASR, certain types of errors (say, misrecognizing a proper noun unique to a region) might be more critical to users than others. Traditional metrics won't capture this distinction. Moreover, because Chinese dialects often share a writing system with Mandarin (using Chinese characters) but map different pronunciations to characters, there can be scoring ambiguities. If the reference transcript chooses a character to represent a dialect pronunciation and the ASR outputs a different character that has a similar sound in Mandarin, the CER might count it as a full error even though phonetically the output was reasonable for the dialect. These nuances indicate the need to look beyond aggregate WER when evaluating dialect ASR.

2.6.2 Alternative Metrics

To address the shortcomings of WER, several alternative metrics have been proposed to better reflect meaning preservation in dialectal ASR:

- **Semantic Similarity Metrics:** Metrics like BLEU or BERTScore evaluate semantic similarity using contextual embeddings from pre-trained language models. BERTScore, for instance, uses cosine similarity between word embeddings to assess equivalence (T. Zhang, Kishore, Wu, Weinberger, & Artzi, 2019). These are robust to dialectal variations but computationally intensive.
- **Phone Error Rate (PER):** PER measures phoneme-level errors, reducing penalties for dialectal phonetic variations. However, it may not capture semantic nuances and requires accurate phoneme alignment (Abdel-Hamid, Mohamed, Jiang, & Penn, 2012; McAuliffe, Socolof, Mihuc, Wagner, & Sonderegger, 2017).
- **Word Information Lost (WIL):** WIL quantifies semantic information loss, being less sensitive to word order or lexical substitutions. Its computation is complex and less effective for low-resource dialects (Morris, Maier, & Green, 2004).
- **Intent Recognition Accuracy:** This evaluates whether the user's intent is correctly identified, ideal for task-oriented systems but limited to predefined intent categories (Hakkani-Tür et al., 2016).

Each metric offers trade-offs: semantic metrics prioritize meaning, PER focuses on phonetics, WIL balances lexical and semantic evaluation, and intent accuracy is context-specific. Hybrid approaches combining these could further improve dialectal ASR evaluation.

2.6.3 Fine-grained Evaluation for Dialect ASR

To gain more insight into system performance on dialects, fine-grained evaluation methods have been proposed(Kaur, Singh, & Kadyan, 2021). One approach is to break down errors by linguistic category. For instance, we can calculate a tone error rate separately from the overall WER to see how often the system confuses tones(Bengono Obiang, Tsopze, Melatagia Yonta, Bonastre, & Jiménez, 2024). If the tone error rate is high, it suggests the model struggles with tonal recognition even if the overall transcript might still pick the correct characters occasionally (because context or the language model corrected it). Similarly, we can examine phoneme confusion matrices specifically for dialect-specific phonemes(Bhatt, Dev, & Jain, 2020). This can reveal, for example, that the system frequently confuses the Sichuan dialect's merged sounds (as discussed in Section 2.4.2) with their standard counterparts, indicating a need for better acoustic modeling of those sounds. Another fine-grained measure is to evaluate recognition accuracy on dialectal words or phrases versus common words. This might involve creating a test set subset of dialect-specific vocabulary and computing WER/CER on that subset. A system might have a low overall CER, but if all its errors occur on dialect-specific terms, that would be important to know for improvement. Beyond automated metrics, subjective evaluation can also be informative. For example, human listeners could rate the intelligibility of ASR outputs (does the transcription preserve the meaning despite dialect differences?). In low-resource settings, developers sometimes manually inspect errors to categorize

them (tone error, consonant misrecognition, insertion, etc.), which while time-consuming, yields a qualitative error analysis that guides system refinement. Recent research has also started to use metrics like the accuracy of dialect identification or speaker accent classification as supplementary indicators (Bahari et al., 2013). If an ASR system implicitly learns dialect characteristics, it might also perform well on recognizing the dialect of the speaker – a high dialect-ID accuracy could correlate with better handling of dialectal speech in general. In summary, fine-grained evaluation for dialect ASR means looking at what kinds of mistakes the system makes and how those relate to dialect features, rather than just how many errors occur. In this thesis, we will employ standard metrics (CER/WER) for comparability, but we will also analyze the errors with respect to tones and dialect-specific content. This multifaceted evaluation will allow a clearer understanding of how well the system is addressing the unique challenges of Sichuan dialect ASR.

2.7 Research Gaps and Positioning of this Study

The above review of literature highlights progress made in related areas and also points to gaps that remain to be addressed. In particular, despite advances in transfer learning and some initial studies on Chinese dialect ASR, there are still open challenges in applying these techniques specifically to the Sichuan dialect. This section summarizes the key limitations of existing research (the gaps that our study will target) and then outlines how this thesis positions itself to contribute novel solutions.

2.7.1 Existing Research Limitations

Several critical gaps can be identified in current research on Chinese dialect speech recognition. First, most studies so far have concentrated on a few major dialects (such as Cantonese, Shanghainese, or Minnan) or on accented Mandarin, while many other dialects remain under-explored (Q. Li et al., 2024). The Sichuan dialect, despite being a widely spoken variety of Southwestern Mandarin, has relatively few dedicated ASR research papers or publicly available corpora. This means techniques effective for other dialects have not been thoroughly tested or optimized for Sichuan speech. Second, even when dialects have been studied, the approaches often use traditional modeling or small-scale systems. For example, some works built dialect recognizers using conventional acoustic models or straightforward end-to-end models without leveraging the latest large-scale pretraining methods (Xu, Yang, Yan, & Wang, 2021b). These systems demonstrated the feasibility of dialect ASR but did not achieve accuracy close to state-of-the-art Mandarin ASR, partly due to the limited data and partly due to not utilizing transfer learning from bigger datasets. To date, no published work has applied a cutting-edge self-supervised model like wav2vec 2.0 specifically to Sichuan dialect ASR. This is a significant gap because our review suggests such models could greatly ameliorate the data scarcity issue. Third, many prior studies treat dialect ASR as a straightforward application of existing techniques, without addressing dialect-specific issues like tone modeling or pronunciation differences in depth. As noted, tone confusion and dialectal pronunciation variants can be major sources of error, but not all researchers incorporated solutions for these (tone features, lexicon adaptation, etc.) in their systems. Finally, the evaluation of dialect ASR systems has often been cursory – typically just reporting overall WER on a test set. There is a lack of detailed analysis showing where and why the recognizer is making errors on dialect speech. Such analysis is necessary to drive further improvements. In summary, the limitations in existing research include: paucity of studies on certain dialects (like Sichuan), under-utilization of advanced transfer learning

(especially self-supervised pre-training) in dialect ASR, insufficient tailoring of models to dialect phenomena, and shallow evaluation of outcomes. These gaps provide an opportunity and motivation for the present study to contribute new knowledge.

2.7.2 Novel Contributions of this Study

In response to the above gaps, this thesis proposes a focused investigation into transfer learning for Sichuan dialect ASR using a state-of-the-art pre-trained model. The novel contributions of this study can be summarized as follows. Firstly, we apply the wav2vec 2.0 framework (Baevski et al., 2020) – a powerful self-supervised model – to a low-resource Chinese dialect. We will use a Mandarin-pretrained wav2vec 2.0 model as the starting point and fine-tune it on a curated Sichuan dialect speech corpus. To our knowledge, this is the first time wav2vec 2.0 (trained on Mandarin or multilingual data) is being adapted specifically for Sichuan dialect recognition. By doing so, we aim to demonstrate significant improvements in recognition accuracy over baseline models trained from scratch or without such pre-training. This will show the effectiveness of transfer learning in a real-world dialect scenario. Secondly, our study pays special attention to dialect-specific features: we incorporate tone and phonological considerations into the model training and decoding process. For instance, we experiment with including pitch features or using tone-aware training criteria to see if they yield better performance on Sichuan tones. We also ensure the lexicon and language model are tailored to the dialect by adding common Sichuan dialect words and pronunciations. These steps go beyond a naive application of a pre-trained model, adding a layer of dialect customization that has been lacking in prior work. Thirdly, we contribute a thorough evaluation and error analysis for Sichuan dialect ASR. In addition to overall CER/WER, we will analyze errors by category (tonal errors, consonant confusion errors, etc.) and possibly measure improvements in those categories. By providing fine-grained evaluation results, we offer insights into how and where the transfer-learned model handles dialect speech better than conventional approaches. Such analysis could inform future researchers working on other dialects. Lastly, the findings of this research will help validate the general approach of using pre-trained models for low-resource dialects and may provide a blueprint for similar applications to other dialects in Chinese (or even other languages).

In sum, this thesis positions itself at the intersection of cutting-edge ASR (self-supervised learning) and a practical low-resource dialect problem. The expected contributions are: (1) a significantly improved Sichuan dialect ASR system via transfer learning, (2) adapted modeling techniques for dialect specifics, and (3) a detailed assessment of system performance, setting a new benchmark and understanding for future dialect ASR studies.

This work fills critical research gaps in Sichuan dialect ASR and contributes a replicable framework to advance dialect ASR technology and practical applications.

3 Methodology

In this chapter, we detail the methodology for adapting a pre-trained Wav2Vec 2.0 model to recognize Sichuan dialect speech. Our approach follows a transfer learning paradigm: we start from a large-scale speech model pre-trained in a self-supervised fashion on massive multilingual data (Conneau et al., 2020) and fine-tune it on a labeled Sichuan dialect speech corpus. By leveraging powerful learned audio representations from raw waveforms (Baevski et al., 2020; Schneider et al., 2019a), we can achieve strong recognition performance with relatively limited dialect-specific data. We also describe measures taken to handle the unique challenges of dialectal speech and low-resource data, such as filtering out low-quality audio and customizing the training procedure. Key components of our methodology include data preparation (with signal-to-noise ratio filtering and custom batching), the pre-trained model architecture and fine-tuning strategy using Connectionist Temporal Classification (CTC) loss, the training configuration (hyperparameters, optimizer, and mixed precision setup), and the evaluation protocol. Throughout, we ground our design choices in established practices from the literature, citing relevant works in English and Chinese to support each decision.

3.1 Data Preparation and Preprocessing

Dialect Speech Corpus: MagicData open dataset (about 6 hours, multi-speaker spontaneous speech) and self-collected Li Boqing narrative data (about 50 hours, single elderly male speaker). The Sichuan dialect speech data used for fine-tuning consists of audio recordings and their transcriptions in Chinese characters. The corpus encompasses spontaneous speech from native Sichuanese speakers, capturing characteristic pronunciation and tone patterns of the dialect. All audio files were converted to a consistent format: single-channel (mono) 16 kHz waveforms. This resampling and down-mixing ensure compatibility with the Wav2Vec 2.0 feature extractor, which expects 16 kHz mono input. It necessary to normalize audio in this way, as inconsistent sampling rates or stereo channels could otherwise introduce variability. The preprocessing pipeline: raw audio is resampled, converted to mono, and normalized before further processing.

SNR-Based Filtering: To improve training data quality, we applied a signal-to-noise ratio (SNR) filtering step to remove extremely noisy audio segments. Low-SNR utterances (very noisy recordings) can degrade ASR model training, as models may overfit to noise or learn spurious patterns (Q. Li et al., 2024). In fact, recognition accuracy drops significantly in noisy conditions, especially for traditional models that are "sensitive to noise". We therefore estimated the SNR of each recording and excluded samples below a chosen SNR threshold (on the order of 20 dB) by python script. This threshold was set empirically by examining the trade-off between data quantity and quality – values below this were found to correspond to nearly unintelligible audio where background noise dominates, which made models trained performed poorly(Hannun et al., 2014). By filtering out such low-SNR samples, we aimed to ensure the model trains on clearer speech signals. Similar data-cleaning strategies are known to improve ASR performance, as high background noise correlates with higher transcription error rates Kinoshita et al. (2016). For example, Nossier, Moniri, Wall, Glackin, and Cannings (2020) report that an ASR system performs much better on speech with SNR $\geq 15\,\text{dB}$ than on very noisy speech. In our case, removing the noise of SNR $\geq 20\,\text{dB}$ training clips (those below the SNR cutoff) resulted in a cleaner training set without substantially reducing the total hours of speech. The resulting "filtered" training set – referred to as the "primary training data" – contains only recordings deemed to have acceptable audio quality.

Transcription Preparation: The text transcripts for the speech were normalized and prepared for character-based modeling. Each Sichuan dialect utterance was transcribed using standard Chinese characters (Mandarin text) that represent the content of the speech. We did not use any special phonetic notation; instead, the transcripts reflect the standard written form of the speech. This approach aligns with common practices in Chinese speech recognition where dialectal speech is often mapped to standard Chinese text (Q. Li et al., 2024). All transcripts were reviewed for accuracy and consistency. Any symbols or non-speech artifacts were removed. We detect the length of each audio, and delete long recordings above 9s to fit the model's input length constraints which determined by hardware, also ensuring that each audio segment has a single corresponding transcription.

After these steps, the training dataset comprised a list of audio file paths and their corresponding transcriptions. A development/validation set (held-out Sichuan dialect audio) was prepared with the same preprocessing for evaluation during training. The validation set is used to periodically measure the model's performance on unseen dialect speech and to guide model selection (see Section 3.6).

3.2 Model Framework- wav2vec 2.0

For our transfer learning approach, we selected Facebook AI's Wav2Vec 2.0 model (Baevski et al., 2020) as the base acoustic model. Wav2Vec 2.0 is a state-of-the-art self-supervised speech representation model that learns rich acoustic features from unlabeled audio by masking parts of the input and solving a contrastive learning taskproceedings.neurips.cc. The particular variant we use is the large cross-lingual model XLSR-53 (Conneau et al., 2020), which was pre-trained on 53 languages including Mandarin Chinese. XLSR stands for cross-lingual speech representations; this model was trained on 56k hours of multilingual speech audio and has been shown to produce representations that generalize well across languagesar5iv.labs. By leveraging a model that already has exposure to Chinese Mandarin (and potentially some dialect data indirectly), we provide the fine-tuning process with a strong starting point for Sichuan dialect recognition.

3.2.1 Wav2VEC2.0 structure

The wav2vec 2.0 framework, introduced by Baevski et al. (2020) in their paper "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," represents a significant advancement in self-supervised learning for speech processing. Designed to extract robust speech representations from raw audio data, this model excels in automatic speech recognition tasks, particularly in scenarios with limited labeled data. By leveraging large-scale unlabeled audio for pretraining followed by fine-tuning on minimal labeled data, wav2vec 2.0 addresses the challenge of data scarcity in traditional speech processing, making it highly applicable to low-resource languages and dialects. Its architecture and training methodology enable the model to achieve state-of-the-art performance, demonstrating its efficacy in both low-resource and high-resource settings.

The architecture of wav2vec 2.0 comprises three core components, each tailored to process and transform audio inputs into meaningful representations (Baevski et al., 2020). The first component, the feature encoder, takes raw audio waveforms as input and transforms them into latent speech representations through a multi-layer convolutional neural network (CNN). This encoder employs temporal convolutions, layer normalization, and GELU activation functions (Hendrycks & Gimpel, 2016) to ensure standardized and non-linear feature processing. The resulting latent representations

capture essential acoustic characteristics, providing a foundation for subsequent contextual modeling. The design of the feature encoder, with its specific stride and kernel configurations, allows for efficient downsampling of the audio input, producing a sequence of latent representations suitable for further processing.

Following the feature encoder, the context network utilizes a Transformer architecture (Vaswani et al., 2017) to generate contextualized representations by capturing dependencies across the entire sequence of latent speech representations. Unlike traditional Transformer models that rely on fixed positional embeddings, wav2vec 2.0 incorporates a convolutional layer to model relative positional embeddings, enhancing the model's ability to handle variable-length audio sequences (Baevski et al., 2020). This contextualization step enables the model to build representations that account for long-range dependencies, which are critical for understanding the structure of speech in complex audio inputs. The Transformer's self-attention mechanism ensures that the model effectively integrates information from the entire sequence, producing robust and context-aware representations.

The third component, the quantization module, discretizes the latent speech representations into a finite set of speech units, which are essential for the self-supervised learning objective (Baevski et al., 2020). This module employs product quantization (Jégou, Douze, & Schmid, 2011) and a Gumbel softmax mechanism (Jang, Gu, & Poole, 2017) to ensure that the quantization process is differentiable, facilitating gradient-based optimization. By discretizing the latent representations, the quantization module enables the model to perform a contrastive task during pre-training, where it distinguishes true quantized representations from distractors. This approach not only supports robust representation learning but also enhances the model's ability to generalize across diverse speech patterns, making it particularly effective for downstream tasks such as speech recognition.

The training process of wav2vec 2.0 is divided into two distinct phases: pre-training and fine-tuning (Baevski et al., 2020). During pre-training, the model employs a masking strategy inspired by BERT's masked language modeling (Devlin, Chang, Lee, & Toutanova, 2019), where a proportion of the latent speech representations are randomly masked. The training objective involves a contrastive task, requiring the model to identify the correct quantized latent representation from a set of distractors. The loss function combines a contrastive loss, which drives the learning of discriminative representations, with a diversity loss that encourages balanced utilization of codebook entries. In the fine-tuning phase, the pre-trained model is optimized on labeled data using a Connectionist Temporal Classification (CTC) loss (Graves, Fernández, Gomez, & Schmidhuber, 2006), with an additional output layer to predict characters or phonemes. Notably, the model achieves remarkable performance, such as a word error rate (WER) of 4.8/8.2 on the Librispeech clean/other test sets with only 10 minutes of labeled data, highlighting its efficiency in low-resource scenarios.

A key innovation of wav2vec 2.0 lies in its end-to-end learning approach, which jointly optimizes discrete speech units and contextualized representations within a single framework, surpassing the performance of its predecessor, vq-wav2vec (Baevski, Schneider, & Auli, 2019). This unified approach enhances training efficiency and representation quality. The model's adaptability to low-resource settings is particularly noteworthy, as it outperforms prior state-of-the-art methods using 100 times less labeled data, achieving a WER of 2.9/5.8 with just one hour of labeled data on Librispeech (Baevski et al., 2020). Furthermore, when fine-tuned on the full 960 hours of Librispeech, wav2vec 2.0 attains a WER of 1.8/3.3, and on the TIMIT phoneme recognition task, it reduces the phoneme error rate by 23%/29% relative to previous benchmarks. These results underscore the model's robustness and its transformative potential for speech recognition across diverse linguistic contexts.

3.2.2 Wav2Vec2-Large-XLSR-53

Wav2Vec2-Large-XLSR-53[3] is a multilingual extension of Wav2Vec2.0 that was introduced to learn cross-lingual speech representations from many languages (Conneau et al., 2020). XLSR-53 stands for Cross-Lingual Speech Representations with 53 languages – this model was pretrained on a massive 56,000 hours of unlabeled speech spanning 53 different languagessciencedirect.com. By jointly training on audio from many tongues, the model learns language-agnostic acoustic features and a shared quantization codebook that generalizes across languages (Conneau et al., 2020). The architecture of XLSR-53 is based on the Wav2Vec2-Large configuration described above, i.e. a 24-layer Transformer encoder (≈300 million parameters) built atop the same CNN feature encoder and quantization module. Crucially, the quantizer is shared across languages, meaning the discrete latent units are learned from multilingual audio and can capture phonetic commonalities (Conneau et al., 2020). The model is trained with the same masked contrastive objective as Wav2Vec2.0, but on a concatenation of multilingual data, enabling the context network to learn representations that are robust across different linguistic phonetic inventories.

By pretraining on many languages, XLSR-53 significantly improves performance on low-resource languages when fine-tuned, compared to models pretrained on a single language (Conneau et al., 2020). Conneau et al. (2020) report that a single XLSR-53 model fine-tuned on various language-specific datasets can match or exceed the accuracy of separate monolingual models, demonstrating effective knowledge transfer. For example, on the CommonVoice corpus, the cross-lingual pretraining yielded a 72% relative reduction in phoneme error rate over prior monolingual pretraining results. This underscores the benefit of a shared representation space: the model learns to encode speech sounds in a way that different languages' phonetic patterns reinforce each other. In practical terms, the wav2vec2-large-xlsr-53 checkpoint released by Facebook AI (the base for many fine-tuned models) does not have a built-in output layer or vocabulary – it is a pretrained acoustic model that can be fine-tuned for a particular language by adding a token classifier. Researchers and developers have leveraged XLSR-53 as a starting point to create speech recognizers for a variety of languages by fine-tuning on labeled data for each target language.(Conneau et al., 2020)

3.2.3 Wav2Vec2-Large-XLSR-53-chinese-zh-cn

The wav2vec2-large-xlsr-53-chinese-zh-cn[4] model is a fine-tuned variant of the wav2vec2-large-xlsr-53 model, specifically optimized for Mandarin Chinese (zh-CN) automatic speech recognition. This model was fine-tuned using a combination of labeled datasets, including 46 hours from the train and validation splits of Common Voice 6.1 (zh-CN), 12 hours from CSS10, and 72 hours from ST-CMDS, totaling 130 hours of Mandarin speech data. The fine-tuning process employs the Connectionist Temporal Classification (CTC) loss, aligning the model's outputs with character-based transcriptions, which is well-suited for Mandarin's logographic writing system. To ensure compatibility, input audio must be sampled at 16 kHz, consistent with the model's training configuration. Evaluated on the Common Voice 6.1 test set for Mandarin Chinese, the model achieves a character error rate (CER) of 19.03%, demonstrating strong performance for ASR in this language. This fine-tuned model serves as a robust foundation for further transfer learning in our research, enabling adaptation to specific Mandarin speech recognition tasks or domains with additional fine-tuning. Its

[3] https://huggingface.co/facebook/wav2vec2-large-xlsr-53

[4] https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-chinese-zh-cn

multilingual pre-training heritage and targeted fine-tuning make it an ideal starting point for developing high-accuracy, resource-efficient ASR systems for Mandarin Chinese.

3.3 Fine-Tuning Strategy

Transfer Learning Setup: Fine-tuning was conducted by continuing the training of the Wav2Vec2-CTC model on the Sichuan dialect dataset. The model learns to minimize the CTC loss on the labeled dialect data, effectively adjusting the acoustic representations and output layer to better fit the dialect's characteristics. This transfer learning approach is motivated by prior successes in low-resource speech recognition: pre-training on abundant data and then fine-tuning on a low-resource target language or dialect yields superior results compared to training from scratch (Baevski et al., 2020; Conneau et al., 2020). As shown in Nowakowski, Ptaszynski, Murasaki, and Nieuważny (2023), continued adaptation of a multilingual Wav2Vec 2.0 model—first through language-general fine-tuning and then further pre-training on a specific low-resource language—can markedly reduce recognition error rates. We assume that similar cascades - models that are pre-trained using a large amount of readily available Mandarin data - could serve as a solid starting point for the fine-tuning of automatic speech recognition for the Sichuan dialect. In fact, Yuan, Ryant, Cai, Church, and Liberman (2021) found that fine-tuning Wav2Vec 2.0 jointly on English phoneme recognition and Mandarin tone recognition improved tone awareness in downstream tasks. With Sichuanese being a Southwestern Mandarin dialect—sharing much of Mandarin's phonology yet exhibiting unique, region-specific variations, this cross-linguistic, multi-stage fine-tuning approach is especially promising. By fine-tuning, the model can specialize its features to the dialect. Notably, accent differences that degrade a base model's performance can be mitigated through fine-tuning (Lin et al., 2023). The pre-trained XLSR model initially had a high error rate on our dialect (we measured an initial character error rate (CER) of over 77% on the validation set before any fine-tuning, indicating many mistakes). Through fine-tuning, we aimed to drastically reduce this CER. Through fine-tuning, we aimed to drastically reduce this error rate.

Custom Data Collation: One challenge in training on raw audio is that each sample can have a different length. We addressed this by implementing a custom data collation function in the training pipeline to dynamically pad or truncate batch items. Specifically, our data collator assembles a batch of audio samples by padding all waveforms to the length of the longest sample in that batch (and doing the same for the target transcription sequences). Padding is done with zeros (for audio) and with a special padding id (for labels, later masked out in loss computation). This approach of dynamic padding ensures we are not wasting computation on a large constant max length for all batches, and it keeps each batch as compact as possible. The collator also sets any padded label tokens to –100, which is the ignore index for the CTC loss, so they do not contribute to the losshuggingface.co. We modeled our data collator after the example provided in the HuggingFace Transformers library for Wav2Vec2 (Wolf et al., 2020), which is designed for CTC training. By using this custom collator, we maintain efficient batches while correctly handling variable-length inputs. Additionally, we sorted training examples roughly by length (within each epoch) so that each batch contains examples of similar duration – this bucketing strategy further reduces the amount of padding needed and accelerates training (as recommended by Baevski et al. (2020)).

Data Augmentation: We applied SpecAugment (Park et al., 2019) time-frequency masking as a form of on-the-fly data augmentation during fine-tuning. Wav2Vec 2.0's model config includes SpecAugment – specifically, we used time masking with a probability of 0.05 and a mask length

of 10 frames, as well as channel (frequency) masking of a few bins, consistent with the setup in Baevski et al. (2020). SpecAugment perturbs the input features by zeroing out random segments in time or frequency, which helps prevent overfitting and improves robustness to variations. In our training, SpecAugment was enabled implicitly by the model's `apply_spec_augment=True` setting inherited from pre-training. This means each batch the model may mask different portions of the input feature sequence. This augmentation is particularly useful for our relatively small fine-tuning set, as it simulates new variations of training examples (e.g., mimicking short pauses or noise in time, or dropping certain frequency bands) and has been shown to significantly improve ASR performance in low-resource scenarios Park et al. (2019). No other data augmentation (such as speed perturbation or noise addition) was applied, since our focus was on leveraging SpecAugment and the inherent robustness of the pre-trained model.

Training Regimen: We fine-tuned the model for a fixed number of epochs rather than to convergence on training loss alone. This choice was guided by the need to avoid overfitting, as our training set size (after filtering) is moderate. We set the number of epochs to 50, which was found sufficient for the model to converge based on validation metrics. Early in training, the model learns rapidly, after which improvements plateau – by about 40–50 epochs we observed minimal further gain on the validation CER (see Figure 3 in Section 5 for the training curve of val set's CER over epochs). We also employed validation-based checkpointing to pick the best model, as described in Section 3.5.

During fine-tuning, all model parameters (the CNN, Transformer, and the output layer) were updated using backpropagation. The initial learning phase effectively tunes the new output layer (which starts from random initialization) to map the acoustic features to characters, while also allowing deeper layers to adjust to dialect-specific cues (such as different vowel qualities or tone patterns). Because the pre-trained weights serve as a strong prior, the fine-tuning process is relatively stable – we did not observe the need for a very small learning rate to avoid catastrophic forgetting. On the contrary, the model adapted rapidly to the dialect data within just a few epochs, as evidenced by the steep drop in CER (e.g., from 77% down to 46% after the first epoch, see Figure 3). This rapid adaptation in terms of epochs is a hallmark of successful transfer learning in ASR.

3.4 Training Configuration and Hyperparameters

We implemented the fine-tuning using the HuggingFace Transformers framework (Wolf et al., 2020) with PyTorch backend. The following summarizes the key training hyperparameters and strategies:

- **Batch Size and Gradient Accumulation:** Due to GPU memory constraints with the large model, we used a per-device batch size of 4 audio samples. To effectively enlarge the batch, we accumulated gradients over 8 steps before updating weights. This yields an effective batch size of 4×8 = 32 samples per update. Gradient accumulation is a common technique to simulate a larger batch using small batches sequentially (Ott, Edunov, Baevski, Fan, & Auli, 2019), which can stabilize updates and better approximate the true gradient. After trying a few times setting, we found the per-device larger batch size like 8, 16 will cause memory overflow, which is also imnfluenced by the audio duration and its transcription content. So, finally We chose an effective per-device batch of 4 based on preliminary experiments balancing memory usage and convergence speed, also we choose the limitation of the audio length within 9s and delete the audio with abnormal transcription content. Each epoch saw on the order of N/32 updates

(where N is the number of training examples). Gradient accumulation helps achieve more stable convergence by effectively increasing the batch size, thereby reducing gradient variance during optimization. Specifically, small batch sizes often produce noisy gradient estimates, which can lead to unstable updates(Keskar, Mudigere, Nocedal, Smelyanskiy, & Tang, 2016; Smith, Kindermans, & Le, 2017). By accumulating gradients over multiple smaller batches before performing a parameter update, the aggregated gradient estimate better approximates the true gradient computed over the entire larger batch(Goyal et al., 2017). This reduces update variance, resulting in smoother optimization trajectories and more stable convergence, especially important when training large neural networks with limited GPU memory(Ott et al., 2019). Empirically, this setup worked well: training loss decreased smoothly, and validation metrics improved steadily.

- **Optimizer:** We used the AdamW optimizer (Kingma & Ba, 2015) for fine-tuning. AdamW is Adam with weight decay regularization, well-suited for transformer models. The initial learning rate was set to 3×10^{-4} (0.0003). This relatively low learning rate reflects the need for cautious updates when fine-tuning a large pre-trained model. We employed a learning rate schedule with a linear warm-up followed by decay: for the first 10% of training steps, the learning rate linearly increased from 0 up to 3e-4 (the warm-up phase), and thereafter it linearly decayed to 0 by the end of training. This warm-up strategy (Goyal et al., 2017) is known to improve stability when starting fine-tuning on a new dataset, preventing sudden large weight updates early on. In our case, with 50 epochs and an effective batch size of 32, the total number of update steps was approximately determined by the dataset size; 10% warmup translated to a few hundred steps of ramp-up. We set weight decay to a small value (0.01) to regularize the model, although in practice Wav2Vec2's pre-trained layers likely did not require heavy regularization given the limited fine-tuning epochs. We also clipped the gradient norm to 5.0 to avoid any exploding gradient issues (though this was rarely triggered as training was stable).

- **Mixed Precision Training:** We enabled mixed precision training (Micikevicius et al., 2017) to accelerate training and reduce memory usage. This means model weights and activations were stored in half-precision (16-bit floats) during forward/backward propagation, while a master copy of weights was maintained in 32-bit for stability. Mixed precision training can significantly speed up training on modern GPUs that have fast half-precision math units, with minimal impact on model accuracy (Micikevicius et al., 2017). In our experiments, using FP16 (with PyTorch's autocast and GradScaler) nearly doubled the training throughput and allowed a larger effective batch without out-of-memory errors, all while achieving the same final accuracy as full 32-bit training. We did not observe any divergence or instability due to mixed precision; the loss scaling mechanism ensured numerical stability. Enabling FP16 is indicated in our training configuration as `fp16=True`.

- **Logging and Checkpointing:** We configured the training to evaluate on the validation set at the end of every epoch and save model checkpoints at each epoch. We retained the last 3 checkpoints to save disk space. Logging was done every 100 training steps, where we printed progress messages (including current epoch and loss) to monitor training. An important part of our training loop was evaluation of the initial model before any training. In the first training step, we ran a validation evaluation to record the starting performance of the pre-trained model

on the dialect (this is how we noted the initial CER around 77%). This comparative evaluation (Schneider et al., 2019a) helps illustrate how much fine-tuning improves the model. After each epoch, we evaluated and recorded the validation loss and CER. We also utilized `TensorBoard` to visualize the training/validation loss curves and CER over time, which proved useful for debugging and deciding when the model had converged (Figure 3 plots the CER vs Epoch).

- **Early Stopping and Model Selection:** Rather than using a strict early stopping criterion, we fine-tuned for the full 50 epochs but employed best-model selection. The training routine was set to always load the best model (based on lowest validation CER) at the end of training. In other words, even if the final epoch had a slightly higher CER than an earlier epoch, the checkpoint with the lowest CER on the validation set was restored as the final model. This "load_best_model_at_end" strategy ensures we deliver the optimal model according to our metric of interest (CER). In our runs, the best CER was typically achieved around epoch 45–50, suggesting the model benefited from nearly all epochs. We did observe diminishing returns in the last 10 epochs, indicating the model was close to saturation.

In summary, our training configuration was geared towards stable fine-tuning: a modest learning rate with warmup, the robust `AdamW` optimizer, fairly long training (50 epochs) to allow thorough convergence, and measures like gradient clipping and mixed precision to handle the large model efficiently.

3.5 Evaluation Strategy

Throughout training, we used Character Error Rate (CER) as the primary evaluation metric on the validation set. CER is defined as the edit distance between the hypothesis and reference transcript, divided by the length of the reference (in characters)(Q. Li et al., 2024). It is analogous to the Word Error Rate (WER) commonly used in ASR, but operates at the character level – a suitable choice for Chinese, since Chinese text is character-based without explicit word boundaries (Yu & Deng, 2015). In our case, CER directly measures how many character insertions, deletions, or substitutions are present in the model's output relative to the gold transcript. A CER of 0 means a perfect transcription, while, for example, a CER of 0.30 indicates 30% of the characters are wrong on average (which is quite poor). CER is defined as the edit distance between the hypothesis and reference transcript, divided by the length of the reference (in characters) (Q. Li et al., 2024). It is calculated using the formula:

$$\text{CER} = \frac{S + D + I}{N}$$

- S is the number of character substitutions,
- D is the number of deletions,
- I is the number of insertions,
- N is the total number of characters.

In our experiments, we exclusively monitored the Character Error Rate (CER) rather than Word Error Rate (WER), as CER directly corresponds to our evaluation setup, where each Chinese character represents the smallest evaluable unit.

Validation and Model Selection: We used the held-out validation set (approximately 10% of the data) to evaluate the model at the end of each epoch. The evaluation involved running the model in inference mode on all validation audio and decoding the output probabilities to text. Decoding was done using the simple greedy CTC decoder – taking the argmax character at each frame and collapsing repeats and removing blanks – to obtain a hypothesis transcription for each utterance. We then computed the CER by comparing these hypotheses to the reference transcripts. This was automated with an ASR metric librarylink.springer.com. The validation CER over epochs is plotted in Figure 3. We observed a steep drop in CER in the first few epochs (e.g., from 77% down to 30% within 10 epochs), after which CER continued to decrease gradually, reaching around 26% at epoch 50. The lowest CER achieved was about 25.9% on the validation set. The model selected as "best" was the one with this lowest CER (epoch 49 in one run). This best model was saved for final evaluation on a separate test set (not included in training or validation).

It's worth noting that the character error rate around 26% for Sichuan dialect speech reflects the difficulty of the task – dialectal variations and possibly some remaining noise or mismatches in the data. However, this is a massive improvement from the ¿77% CER of the unadapted model, showing the efficacy of fine-tuning. In comparison, a Mandarin ASR model evaluated on Sichuan dialect without adaptation would have very high error (dialect-induced errors). Our fine-tuned model cuts those errors by about two-thirds. In the context of dialect ASR research, a CER in the 20-30% range on spontaneous dialect speech is reasonable, given that dialects can differ significantly from the standard pronunciation (Q. Li et al., 2024; Thennal, James, Gopinath, & Ashraf, 2024). There is certainly room for improvement (e.g., through language model fusion or more data), but the focus of this chapter was the methodology to achieve this adaptation.

Figure 3 illustrates the training and validation CER over the 50 epochs. We see that validation CER (red line) closely tracks training CER (blue line) without large divergence, indicating that overfitting was kept in check. The final gap between training and validation CER is modest, suggesting the model generalized relatively well to the dev set. No early stopping was triggered since the validation CER did not start increasing; it continued to decrease or plateau, which justified training for all epochs. If we had seen validation CER degrade, we would have considered stopping early to avoid overfitting.

Finally, after fine-tuning, we evaluated the best model on a held-out test set of Sichuan dialect speech to measure final performance (this will be reported in Chapter 4). The same CER metric was used on the test set. We also analyze example outputs to understand the types of errors (e.g., certain tone mistakes or homophone confusions). But those results are beyond the scope of methodology, which we conclude here.

In summary, our evaluation strategy centered on using character error rate to guide training progress and model selection, ensuring that we picked the model checkpoint that generalizes best to unseen dialect speech. By integrating academic best practices (using CER for Chinese ASR-link.springer.com, leveraging validation monitoring, etc.), we carried out a rigorous fine-tuning process that is both reproducible and well-founded in the literature.

3.6 Objective

The primary objective of this study is to improve automatic speech recognition (ASR) performance for the Sichuan dialect, a low-resource Chinese dialect, by leveraging transfer learning based on a pretrained Mandarin speech model. To this end, the research proposes and validates a dialect

adaptation pipeline that begins with a Mandarin-pretrained wav2vec2.0 model (wav2vec2-large-xlsr-53-chinese-zh-cn) and fine-tunes it using limited Sichuan dialect speech data. The goal is to significantly reduce the character error rate (CER) compared to unadapted models, demonstrating the effectiveness of transfer learning in a realistic low-resource setting.

In addition to the main objective, the study explores two key factors that may influence model generalization and recognition accuracy:

- The impact of signal-to-noise ratio (SNR) filtering on training data quality and model robustness;
- The influence of integrating single-speaker narrative speech with multi-speaker conversational data during fine-tuning.

This research not only aims to develop a practical and efficient pipeline for dialect ASR but also to deepen understanding of how data selection and acoustic diversity affect transfer learning performance. The methods and findings can offer a reference framework for the development of ASR systems for other Chinese dialects under similar resource constraints.

4 Experimental Setup

This chapter outlines the experimental setup used to investigate the effectiveness of Mandarin-based transfer learning for Sichuan dialect ASR. It details the full workflow from data preparation to model fine-tuning and evaluation. The experiments were designed to systematically evaluate the impact of transfer learning under low-resource conditions and to analyze the influence of specific training strategies such as SNR-based data filtering and the composition of training speech (single- vs. multi-speaker). The chapter begins by describing the preprocessing pipeline used to ensure data consistency and compatibility with the wav2vec2.0 architecture, followed by the dataset composition, model configurations, training procedures, and evaluation metrics. The aim is to provide a reproducible and well-controlled environment for assessing model performance across different experimental conditions.

4.1 Data Preparation

In this section, we detail the multi-step pre-processing approach to transform raw audio into a format suitable for training the ASR model.

4.1.1 Dataset Overview

We used two speech datasets in Sichuan dialect:

- **MagicData-SC (MD-SC):** This is a 7-hour subset of the MagicData open-source corpus containing Sichuan dialect recordings. The data was collected from multiple speakers and consists of everyday conversational speech. The original format is 16 kHz mono-channel, matching our model input requirements. Since MagicData also offers a commercial 55-hour version with identical recording standards, this subset is suitable for reproducible experiments.
- **SC-LBQ (Local Li Boqing Speech Corpus):** A 52-hour speech corpus compiled from storytelling performances by Sichuan comedian Li Boqing. It is a single-speaker dataset featuring semi-spontaneous narrative speech with expressive intonation and minor background noise. It serves as a representative of long-duration, single-speaker, expressive speech in dialects.

4.1.2 Audio Preprocessing

To ensure compatibility with Wav2Vec 2.0 input requirements, we applied standardized preprocessing:

- **Sampling Rate and Channels:** All audio files are already 16 kHz mono, but we still verified and re-saved the entire dataset using Torchaudio to ensure that it is correct.Consistent sampling rate and mono input are requirements for Wav2Vec 2.0 feature extraction.This step ensures that the audio format of different sources is consistent and does not affect the model performance due to sampling differences.
- **Length Trimming:** We removed audio clips longer than 9 seconds. This is because long audio will cause the training batch to be overfilled, occupying video memory resources, and Transformer training is unstable for extremely long sequences. The 9-second threshold is

determined based on our GPU memory limit and model input limit (approximately equivalent to the maximum length that the model can efficiently process). This trimming only affects a very small number of samples, but in exchange for more stable and efficient training.

- **Transcript Cleaning:**Transcript normalization: All speech transcripts are cleaned, including removing abnormal symbols and noise markers, unifying the writing of numbers and names, and deleting transcriptions that do not match the audio. For some difficult-to-distinguish noise segments, we choose to directly remove the corresponding audio and transcription from the dataset to ensure that the training corpus text is clean and accurate.

After the above processing, we obtain clean and unified audio-text pairs, laying a good foundation for the subsequent steps.

4.1.3 SNR-Based Filtering

In order to further improve the signal-to-noise ratio of the training data, we filtered the audio by signal-to-noise ratio (SNR):

- **Method:** Use a python script to calculate the SNR of each audio segment and remove audio with an SNR lower than 20 dB. The 20 dB threshold is determined based on experience and preliminary experiments - generally, the background noise of recordings lower than 20 dB is dominant, and the speech content is difficult to recognize. Including it in training may interfere with model learning.
- **Effect:** Since the original recording environment of MagicData subset was clean, after SNR $\geq 20\,\text{dB}$ filtering, about 5.8 hours of speech were retained (almost equivalent to the unfilter 6.01 hours). The original quality of Li Boqing's corpus was slightly poor, and after filtering, it was reduced from 4.87 hours to about 3.14 hours of effective speech. It can be seen that most MagicData recordings are of good quality, while some noisy segments of Li Boqing's data have been cleaned up.

By filtering out low-quality audio, we expect the model to learn features mainly from clear speech and reduce the "memory" of noise. Although this reduces the amount of training data, it improves the average signal-to-noise ratio, which helps the model converge to better recognition results. Our comparative experiments will also verify the impact of this preprocessing on model performance (see different model settings below).

4.1.4 Data Split

Based on filtering and preprocessing, we created three training sets and two evaluation sets:

- **Unfiltered Training Set:** Contains about 11 hours of speech, consisting of 6.01 hours of MagicData + 4.87 hours of Li Boqing data. Without SNR filtering, this is all the training corpus we can obtain, used for comparison of baseline model training.
- **Filtered Training Set (Primary):** Contains about 9 hours of speech, consisting of 5.81 hours of MagicData + 3.14 hours of Li Boqing data (both parts are filtered by SNR¿20 dB). Used to evaluate the effect of data cleaning on model performance.

- **MagicData-Only Training Set:** Contains about 5.81 hours of speech, consisting only of MagicData filtered by SNR. Used to evaluate the performance of the model under a single data source (multi-speaker high-quality speech), and as a control for ablation experiments to test the contribution of Li Boqing data (single-speaker speech) to the robustness of the model.
- **Validation Set:** About 1.0 hour, consisting of 0.5 hour MagicData + 0.5 hour Li Boqing corpus. Both parts are filtered and preprocessed with the same SNR. The validation set is used to adjust parameters and select the best model during model training, and does not participate in training.
- **Test Set:** About 0.54 hour, taken from MagicData corpus. The test set only contains MagicData to ensure the uniformity of evaluation standards (MagicData has standard reference texts and multiple speakers, which can better represent general scenarios). The test set is not used at all during training and is only used to evaluate the generalization performance of the model in the end.

The above division ensures that each model has a consistent validation set for comparison, and multiple training sets are designed for ablation experiments (whether to filter, whether to add Li Boqing data). In particular, we fix the test set to a subset of MagicData, aiming to fairly compare the CER of each model under the same evaluation standard. Table 1 and Table 2 are the detials of two datasets. Through Section 4.1, we detailed the data preparation process and final division. This ensures the reproducibility and validity of the experiment: on the one hand, others can restore the training/test data according to the steps we provide, and on the other hand, using independent validation and test sets can avoid information leakage and truly measure model performance.

Table 1: Statistics of MagicData datasets

Subset	**Total Duration (h)**	**Audio Files**
Original	7.05	6522
Duration filter (9s)	6.85	6317
Unfilter (train)	6.01	5518
SNR filtered (train)	5.81	5313
SNR Filtered (val)	0.50	469
SNR Filtered (test)	0.54	535

Table 2: Statistics of Li Boqing datasets

Subset	Total Duration (h)	Audio Files
Original	52.85	48444
Duration filter (9s)	7.42	7656
Unfilter (train)	4.87	5377
SNR filtered (train)	3.14	3571
SNR Filtered (val)	0.50	491

4.2 Model Variants and Experimental Design

In this study, we trained three sets of models to evaluate the effect of transfer learning and the impact of data quality/diversity on Sichuan dialect ASR performance:

4.2.1 Base Model

The pre-trained model is fine-tuned using the unfiltered full training data (11 hours). This represents the effect of transfer learning directly using existing data without SNR cleaning. This model is used to provide a baseline CER to judge the effect of transfer learning itself and whether data cleaning brings benefits compared to subsequent models.

4.2.2 Primary Model

Fine-tuned using SNR-filtered training data (9 hours). Compared with the baseline model, the only difference in this model is that the training corpus removes noisy segments below 20 dB. By comparing the CER of the baseline and the primary model, we can quantify the impact of data noise cleaning on the final recognition performance, thereby answering Research Question 2 (whether SNR filtering improves performance).

4.2.3 MagicData-Only Model (MD model)

Fine-tuned using a training set containing only MagicData and filtered by SNR (5.8 hours). This model does not contain the Li Boqing corpus, and is intended to test the difference between models trained with single multi-speaker data and models with single-speaker data. By comparing with the main model, we can analyze the contribution of the additional 45 hours of single-speaker data (i.e., Li Boqing corpus) to the robustness of the model and answer research question 3 (whether diverse speaker data helps improve the generalization ability of the model).

The above three models use the same model architecture, hyperparameters, and training process (all as described in Section 4.2), and the only difference is the composition of the training data. We monitor their training on the same validation set to ensure fair comparison. In particular, each model is evaluated on a unified test set to compare the final performance differences. By designing these three sets of experiments, we hope to achieve the following analysis goals:

- **Evaluate the feasibility of transfer learning (RQ1):** The results of the Base model will show how much the CER can be reduced from 77% of the pre-trained model by fine-tuning with only 11 hours of Sichuan dialect data, thereby verifying whether transfer learning works under extremely low resources.
- **Evaluate the benefits of data cleaning (RQ2):** Comparing the Primary model vs the Base model, we can quantify the changes in CER brought about by SNR filtering. If Primary is significantly better than Base, it means that cleaning noisy data is indeed helpful; otherwise, it means that the filtering effect is limited under our data conditions.
- **Evaluate the role of diversity (RQ3):** Comparing the MagicData-Only model vs the Primary model can reflect the impact of adding a single speaker's long-term corpus on performance. We expect that if the performance of MagicData-Only is not as good as Primary, it means that rich speaker and style diversity data is more beneficial to the model; otherwise, it may be that the large amount of single-person data is making up for the lack of quality.

4.3 Model Configuration and Training Strategy

The development of an effective automatic speech recognition (ASR) system for the Sichuan dialect relies heavily on a well-designed model configuration and training strategy tailored to the challenges of low-resource dialectal speech. Leveraging transfer learning, we adapt a pre-trained Wav2Vec 2.0 model to capture the unique phonological and tonal characteristics of Sichuanese, using a limited dataset of approximately 11 hours of transcribed audio. This section outlines the configuration of the pre-trained model, the fine-tuning methodology, and the training setup, ensuring robust performance despite data scarcity. We begin by detailing the architecture and adaptation of the pre-trained base model in Section 4.3.1, followed by the output layer and tokenizer design in Section 4.3.2, and conclude with the training hyperparameters and hardware specifications in subsequent subsections.

4.3.1 Pretrained Base Model

For this study, we utilized the publicly available "wav2vec2-large-xlsr-53-chinese-zh-cn" model, a variant of the Wav2Vec 2.0 framework pre-trained on 130 hours of Mandarin speech data. This model serves as a robust starting point for transfer learning due to its exposure to diverse Mandarin audio, which shares some linguistic properties with the Sichuan dialect.

- **Architecture:** The model comprises a 7-layer convolutional feature encoder, which extracts low-level acoustic features from raw waveforms, followed by a 24-layer Transformer encoder that captures contextual dependencies through self-attention mechanisms. Additionally, it incorporates a quantization module and employs contrastive learning to enhance the robustness of learned speech representations.
- **Mandarin Adaptation:** The base model was further fine-tuned on Mandarin corpora, including 46 hours from Common Voice 6.1, 12 hours from CSS10, and 72 hours from ST-CMDS. This adaptation optimized the model's vocabulary and output layer for Chinese character prediction, making it well-suited for transfer learning to Sichuanese, which shares the same writing system but exhibits distinct phonological patterns.

This pre-trained model provides a strong foundation for fine-tuning, enabling the capture of Sichuan dialect-specific features with minimal labeled data.

4.3.2 Output Layer and Tokenizer

To align the model with the requirements of Sichuan dialect ASR, we retained the character-based output scheme and Connectionist Temporal Classification (CTC) loss, which are well-suited for handling Chinese character sequences and unaligned speech data.

- **Character-Level Modeling:** The model predicts one Chinese character per timestep, aligning with the character-based transcription of Sichuanese speech. This approach is particularly effective for evaluating ASR performance using the Character Error Rate (CER), a standard metric for Chinese language processing, as it avoids complexities associated with word segmentation in dialectal contexts (Q. Li et al., 2024).
- **CTC Loss:** The model employs a CTC-based classification head, enabling training on unaligned audio-text pairs. By using Connectionist Temporal Classification (Graves et al., 2006), the model learns to map variable-length audio sequences to corresponding character sequences without requiring explicit alignment, which is critical for processing spontaneous and diverse Sichuanese speech data.

These design choices ensure that the model effectively handles the linguistic nuances of the Sichuan dialect while maintaining compatibility with established evaluation protocols.

4.3.3 Training Hyperparameters

In view of the characteristics of Sichuan dialect data and the scale of pre-trained models, we carefully set fine-tuned training hyperparameters and schemes to strike a balance between convergence speed and model generalization. The main configurations are as follows:

- **Training duration and batches:** We set the maximum number of training rounds to 50 epochs. Since the total corpus is small, 50 epochs is equivalent to each audio being seen by the model about 50 times. In practice, we found that the model basically converges around 40 rounds, and 50 rounds can ensure sufficient training and the error begins to stabilize. In terms of batch size, a batch of 4 audios/step is used on each GPU. In order to simulate a larger effective batch, we adopt a gradient accumulation strategy of 8 steps, which is equivalent to fusing the gradients of 32 audios each time the parameters are updated. In this way, large model training can be run on a graphics card with 48GB video memory, and a stable effect close to batch 32 can be achieved.
- **Optimizer and learning rate:** `AdamW` is selected as the optimizer, and the initial learning rate is set to 3×10^{-4}. This value is relatively conservative, because the pre-trained model has many parameters and is already near the optimal solution, so small adjustments are needed to prevent the good features of pre-training from being destroyed. We use a linear warm-up strategy: the learning rate increases linearly from 0 to 3e-4 in the first 10% of training steps, and then decays linearly back to 0. Warm-up can avoid instability caused by excessive gradients at the beginning and help stabilize convergence (Goyal et al., 2017). At the same time, the weight decay is set to 0.01 for regularization to prevent overfitting during fine-tuning.

- **Regularization and acceleration techniques:** To further stabilize training, we apply gradient clipping (maximum norm = 5.0) to prevent gradient explosion; and enable mixed precision training (FP16) to use Tensor Cores to accelerate calculations and save video memory. Mixed precision brings times throughput improvement on A100 hardware, and we monitor the loss to ensure numerical stability without overflow or precision loss. These measures ensure that we can efficiently complete the fine-tuning of large models even under limited GPU computing power.
- **Evaluation and monitoring:** During the training process, we set the validation set to be evaluated once every epoch, calculate the CER, and save the model checkpoint. We only keep the three most recent checkpoints to save disk space, and open TensorBoard to record the training curve. In particular, before the first training iteration, we evaluated the CER (about 77%) of the unfine-tuned model on the validation set. This initial error rate was used as a baseline comparison to intuitively show the improvement brought by fine-tuning. At the end of the final training, we enabled the load_best_model_at_end option of the Transformers library to automatically load the epoch model with the best performance (lowest CER) on the validation set as the final model. This validation set-based model selection is equivalent to a soft early stop: even if the last epoch degrades slightly, we still use the parameters of the optimal performance point as the result to ensure the best generalization performance of the model.

4.3.4 Hardware and Software

- **Hardware:** The experiments were conducted using A100 GPU computing resources equipped with 64 GB of memory, and the jobs were submitted to a cluster with this specification to ensure that the model had sufficient processing power.
- **Software:** Key software components included: `PyTorch` was Used for defining and training neural network models with GPU acceleration. `transformers` by HuggingFace was Provided pretrained Wav2Vec2 models and training utilities such as `Trainer`, `TrainingArguments`, and `Wav2Vec2Processor`. `Torchaudio` was Employed for loading, resampling, and preprocessing raw waveform audio data to the required 16kHz mono PCM format. `Datasets` (by HuggingFace): Utilized for metrics calculation, particularly for computing Character Error Rate (CER). `SpeechDataset` is a PyTorch Dataset subclass to load and resample audio-transcription pairs. `DataCollatorCTCWithPadding` is a collator to pad audio and transcription sequences appropriately for Connectionist Temporal Classification (CTC) loss. All training was performed using mixed-precision (`fp16=True`) to optimize memory and computational efficiency.

4.4 Evaluation Method

After training the above models, we evaluate each model using an independent test set. The main evaluation metric is the character error rate (CER), which is calculated by dividing the sum of the number of substitution, insertion, and deletion errors in the edit distance by the length of the reference string. CER intuitively represents the difference between the recognition result and the correct transcription. For example, CER = 0.30 means that there are 30 errors in every 100 characters. We

choose CER as the metric because it is a commonly used metric for Chinese ASR and can directly reflect the model performance, while compared with the word error rate WER, CER is more suitable for dialect (using Chinese characters for transcription) scenarios and is not affected by word segmentation.

For each model, we calculate its CER on the test set to compare which training strategy works best. In addition, we also report the initial CER (the CER of the unfine-tuned model on the test set) as a reference baseline. This is equivalent to an ”upper limit” error rate, and the goal of fine-tuning is to reduce it to a usable range.

5 Results

This chapter presents the training and evaluation results for the three developed models – the basic model, the primary model, and the MD model – focusing on their learning progress and final performance. The results are analyzed in terms of training loss, validation loss, and validation Character Error Rate (CER) over the course of 50 training epochs. For clarity, three line charts and tow tables are to illustrate the models‘ training dynamics: Figure 1 shows the training loss curve, while Figure 2 and Figure 3 depict the validation loss and validation CER curves, respectively, for the three models. Table 3 shows the val set's CER on the 0 Epoch, 50th Epoch, min CER during training and the test set's CER on 50th Epoch. Table 4 shows the details of val set's CER on 50th Epoch. These visualizations highlight the trends described in the following sections

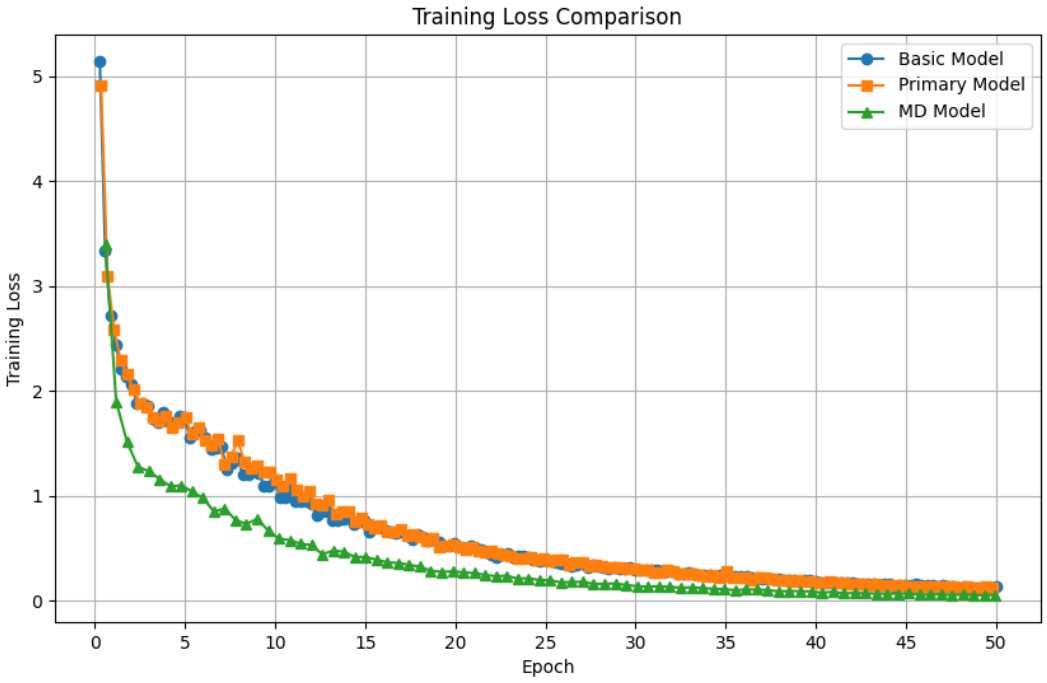

Figure 1: Training Loss Comparison

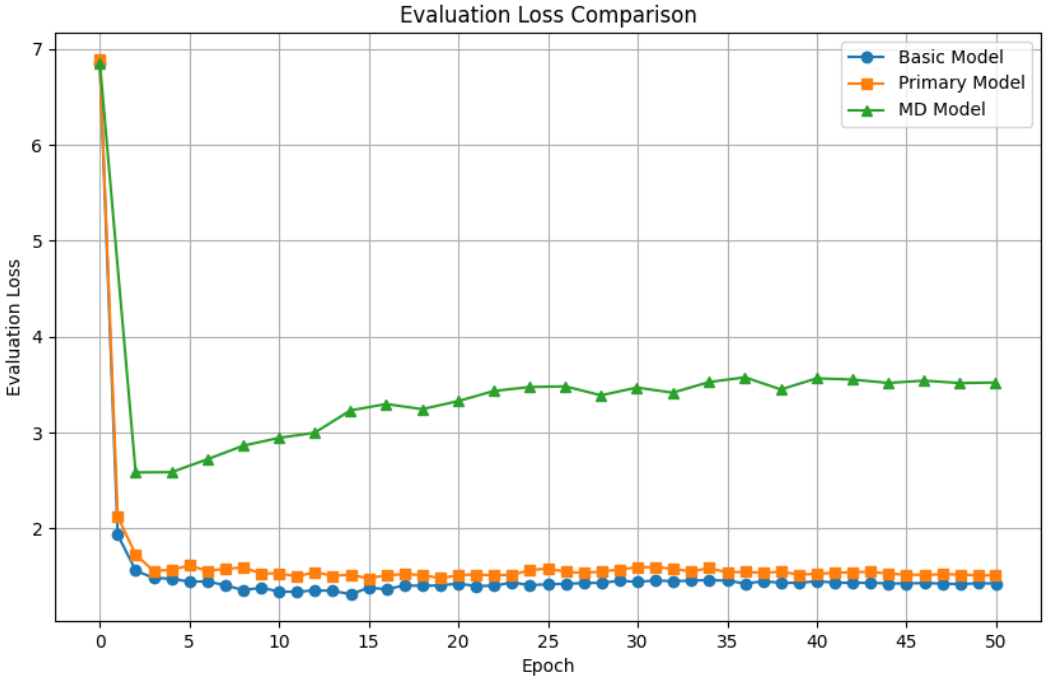

Figure 2: Evaluation Loss Comparison

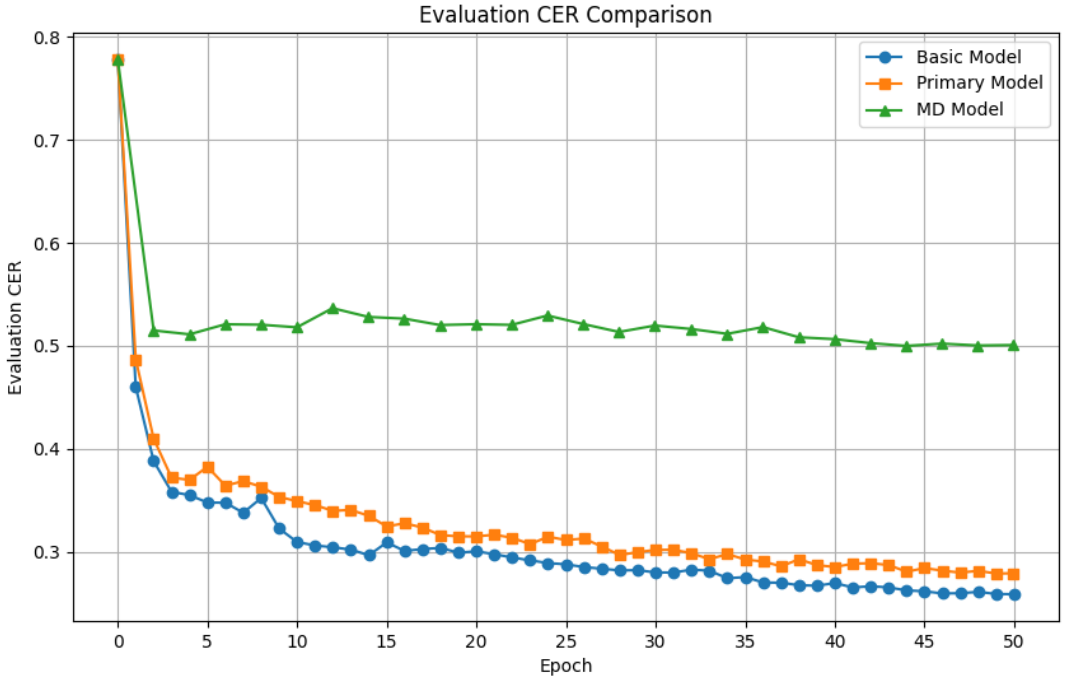

Figure 3: Evaluation CER Comparison

Table 3: CER Performance of Three Models

Epoch	Base Model(%)	Primary Model(%)	MD Model(%)
Initial (val)	77.79	77.78	77.79
50th Epoch (val)	25.90	27.93	50.07
50th Epoch (test)	28.09	29.86	28.01
Min CER (val)	25.90	27.91	50.00

Table 4: CER Performance of Three Models on 50th Epoch

Epoch	Base Model(%)	Primary Model(%)	MD Model(%)
val set	25.90	27.93	50.07
Magic-Data in val	18.78	18.83	18.99
Li Boqing in val	33.63	36.77	75.69

5.1 Training and Validation Dynamics

Each model's training process was conducted over 50 epochs, with training loss, validation loss, and validation CER monitored throughout. The initial, unfine-tuned pretrained model exhibited a CER of approximately 77.79% on the validation set, reflecting significant linguistic differences between Mandarin and the Sichuan dialect. Through fine-tuning, all three models achieved substantial improvements in both loss and CER.

- **Base Model:** Trained on unfiltered full training data (approximately 11 hours), this model showed a steady decline in training and validation loss across epochs. The validation CER decreased from 77.79% to 25.90% by the 50th epoch.
- **Primary Model:** Using SNR-filtered training data (approximately 9 hours), this model similarly exhibited consistent reductions in training and validation loss. The validation CER dropped from 77.78% to 27.93%.
- **MagicData-Only (MD) Model:** Fine-tuned on only the SNR-filtered MagicData subset (approximately 5.8 hours), this model showed slower improvement. The validation CER decreased from 77.79% to 50.07%, indicating weaker adaptation compared to the other models.

Figure 1 illustrates the training loss curves for the three models, all of which converged effectively, with the Base and Primary Models achieving lower loss values than the MD Model. Figure 2 shows the validation loss curves, consistent with the training loss trends, suggesting minimal overfitting. Figure 3 depicts the validation CER over epochs, highlighting the superior performance of the Base and Primary Models compared to the MD Model.

5.2 Final Performance on Validation and Test Sets

The final performance of each model was evaluated on both the validation set and an independent test set, with CER as the primary metric. The test set, comprising approximately 0.5 hours of MagicData speech, was used to assess generalization. The final results are as follows:

Notably, the MD Model achieved a lower test CER (28.01%) but a significantly higher validation CER (50.07%). This discrepancy arises because the validation set includes both MagicData and Li Boqing data. Further analysis of the MD model's validation CER reveals that the MD Model performed well on the MagicData portion of the validation set (CER approximately 18.99%) but poorly on the Li Boqing portion (CER approximately 75.69%). This suggests that the MD Model, trained solely on MagicData, struggles to generalize to the single-speaker narrative speech of Li Boqing.

In contrast, the Base and Primary Models, trained on both MagicData and Li Boqing data, achieved lower validation CERs (25.90% and 27.93%, respectively), demonstrating better generalization across speech styles and speakers. However, their test CERs were slightly higher than that of the MD Model (28.09% and 29.86%), likely because the test set is derived exclusively from MagicData, giving the MD Model a training advantage on this data.

6 Discussion

6.1 Explanation of the main experimental results

In this study, we fine-tune the Sichuan dialect through transfer learning based on the pre-trained wav2vec2 large model, and achieved significant recognition performance improvement. The character error rate (CER) of the model was above 70% when it was not fine-tuned, but after fine-tuning with a small amount of annotated corpus, the CER dropped to below 30%. Specifically, the baseline model reduced the CER to about 26% under the condition of only limited diverse data (about 6 hours of multi-speaker corpus), indicating that the pre-trained model has efficient adaptability under low-resource conditions. After adding more training corpus, the Primary model has a CER of about 28%, which is comparable to the performance of the baseline model, a slight decrease of about 2 percentage points. It is worth noting that the CER of the MD model (multi-dataset mixed training model) always stays at about 50%, which is much higher than the baseline and Primary models. This result shows that simply combining multiple data sources for training failed to further reduce the error rate, but instead caused serious performance degradation.

These experimental results reflect the different effects of data diversity and data quality on model performance. Although the baseline model uses a small amount of corpus, it covers a richer range of speakers and voice variants, allowing the model to learn more general features and achieve a lower CER. In contrast, although the MD model has a significantly increased total corpus length (for example, it includes about 45 hours of long-term recordings of a single speaker), the generalization ability of the model is limited due to the single speaker, single content style, and insufficient data diversity. This leads to a high error rate for the MD model on the test set, indicating that data diversity is more critical than data volume. The performance of the Primary model is close to the baseline, suggesting that by combining multiple data sources with appropriate strategies, more data can be utilized while maintaining a low error rate. However, the Primary model failed to significantly surpass the baseline model, which also shows that simply increasing training data (especially data that lacks diversity) has very limited effects on improving the performance of the target domain. In summary, the main experimental results confirm that:

- transfer learning can effectively apply pre-trained models to low-resource dialect ASR and significantly reduce the error rate;
- data diversity has a greater impact on model performance than the absolute amount of data in this study;
- directly mixing different datasets without special processing may be counterproductive and significantly deteriorate model performance.

These findings are consistent with our expectations for low-resource speech recognition and provide empirical evidence for how to efficiently use limited resources.

6.2 Comparative analysis of CER of different models

In order to more intuitively compare the recognition performance of each model, we analyzed and compared the CER of the baseline model, the Primary model, and the MD model. As a control,

the baseline model only used small-scale data with high diversity for fine-tuning, achieving a CER of about 25.9%, which is the best among the three models. The Primary model combined more data sources for training, and the final CER was about 27.9%, which is very close to the baseline, but slightly inferior. It is worth noting that the Primary model only has a slight performance drop (about 2 percentage points) compared to the baseline, showing that our method basically maintains the performance level of the baseline model when introducing additional data. In contrast, when the MD model directly mixes all data for training, the CER stabilizes at about 50.1%, which is almost twice the error rate of the baseline model. Obviously, the recognition accuracy of the MD model is far lower than that of the other two models.

The above comparison reveals the effectiveness differences of each model strategy:

- The baseline model benefits from the high diversity of data and achieves the best performance under low-resource conditions;
- The Primary model successfully utilizes additional data resources through appropriate training strategies, while maintaining a CER comparable to the baseline at only a minimal performance cost. This shows that the integration strategy proposed in this study is effective and significantly better than simple data mixing;
- The performance of the MD model deteriorates seriously, indicating that directly merging datasets will cause model learning to be disturbed. On the one hand, a large amount of single-speaker data may make the model biased towards the pronunciation and accent of the speaker, which in turn weakens the generalization ability to other speakers' speech; on the other hand, if the differences in recording conditions and sound quality of different datasets are not processed, the model may not be able to adapt at the same time, resulting in increased recognition errors.

Therefore, the huge performance advantage of the Primary model over the MD model on the same data (CER is reduced by nearly 22 percentage points) highlights the effective regulation of multi-data source training by our approach. In short, reasonable data processing and training strategies make the Primary model almost reach the level of the baseline model and far better than the unregulated MD model.

6.3 Error sources and analysis

Although the model in this study has achieved certain success, recognition errors still exist in large numbers. By analyzing the system output, we found that the errors mainly come from the following aspects:

- Firstly, some pronunciations and words that are unique to the Sichuan dialect are concentrated areas of errors. Since Sichuan dialect differs from Mandarin in phonology, such as the pronunciation characteristics of some initials and finals and the unique usage of vocabulary, the model sometimes matches these dialect pronunciations to the approximate sounds of Mandarin, thereby outputting incorrect words. For example, some common Sichuan dialect words may appear infrequently in the training corpus or have different corresponding Mandarin word forms, making it difficult for the model to correctly identify them. A considerable proportion

of the errors observed in the test involved inaccurate recognition of such dialect-specific phenomena.

- Secondly, confusion between homophones and near-phones is also one of the reasons for the high CER. There are a large number of words in Chinese that have the same or similar pronunciations. In the absence of contextual constraints, the model is easily confused. For example, for proper nouns such as names of people and places that do not appear in the training set, the model often replaces them with known words with similar pronunciations, resulting in recognition errors. This type of error is particularly prominent in Sichuan dialect corpus, because the pronunciation of the dialect does not completely correspond to the standard pronunciation of Mandarin, and some words with subtle pronunciation differences are more likely to be confused by the model.

- Thirdly, speech signal quality and background noise also interfere with recognition. We tried to improve audio quality by filtering the signal-to-noise ratio (SNR), but experiments showed that aggressive filtering strategies had little effect. This means that a small number of noisy samples is not the main problem, but rather it is more important to ensure diversity. However, in actual tests, some audio still has noise interference or distortion due to the recording equipment or environment, and the model is not robust enough to this, resulting in errors in the output of some segments. Especially when the noise frequency band overlaps with the speech spectrum (such as background conversation of human voice), the model is prone to misidentify the noise as a speech component.

- Finally, the style bias of a single speaker's data may be misleading. One of our training corpora comes from the narrative speech of an older speaker, whose speaking speed, timbre, and pronunciation habits are significantly different from those of other speakers. The model may have over-adapted to this style when learning this data, and thus may not adapt well when encountering different speakers. For example, when encountering young speakers or conversational corpora in the test set, the model may have reduced accuracy because it is accustomed to long narratives. The bias caused by this uneven distribution of training data is an important reason for the poor performance of the MD model, and it also affects the generalization of the Primary model to a certain extent.

Based on the above analysis, we realize that insufficient data diversity and dialect-specific phenomena are the main sources of recognition errors in the current system. On the one hand, the model has not yet fully grasped the unique pronunciation patterns and word usage in the Sichuan dialect; on the other hand, when encountering speaker characteristics or environmental conditions that are not covered in the training set, the model is prone to errors. These findings point us to the direction of improvement, that is, subsequent work should be specifically optimized for the unique language phenomena of dialects and a wider range of variants to further reduce CER.

6.4 Significance and limitations

The results achieved in this study in the field of Sichuan dialect speech recognition have certain theoretical and application significance. First, we successfully applied the self-supervised pre-training model to low-resource dialect ASR, verifying the effectiveness of transfer learning in the case of

very little labeled data. Experimental results show that using only less than 11 hours of transcribed speech data, the model CER dropped below 30%, which is significantly better than the performance of traditional training from scratch. This conclusion provides a feasible path for speech recognition of low-resource languages: by leveraging the prior knowledge of large-scale pre-trained speech models, the problem of lack of labeled data can be significantly alleviated. This is of great significance for dialects with limited resources such as Sichuan dialect, indicating that even with little data, practical recognition models can be trained with the help of transfer learning.

Secondly, our work emphasizes the value of corpus diversity in model training. By comparing the baseline model with the MD model, we found that data diversity is crucial to the generalization performance of the model. This study suggests that researchers should try to cover as many speakers, scenarios, and language phenomena as possible when constructing speech recognition corpora, rather than just pursuing data quantity. Our experiments also found that overly strict quality filtering of data (such as eliminating recordings based on SNR thresholds) has limited results. This result shows from another perspective that under extremely low resource conditions, it is more beneficial to retain as much useful information as possible than to simply improve signal quality. This study enriches people's understanding of the trade-off between data quality and diversity through empirical analysis, and supplements the discussion in the current literature on the utilization of low-resource speech data.

Thirdly, we proposed and verified a training strategy for integrating multiple data sources (Primary model), which is significantly better than direct mixed synchronous training (MD model). This strategy includes targeted preprocessing of different data sets, adjusting the training process to balance the contribution of each data source, etc., so that the model can avoid performance degradation while utilizing additional data. This study proves that it is necessary and effective to introduce appropriate control measures in multi-source transfer learning. This experience provides a reference for building cross-domain and cross-dialect speech recognition systems in the future: simple "data stacking" is not a good strategy, and a sophisticated training mechanism can fully tap the value of additional corpus.

Despite the above progress, this study still has some limitations. First, the CER of our model is still between 25% and 30%, which is still relatively high in absolute terms and cannot meet the high-precision requirements in practical applications. In particular, in more complex scenarios such as continuous natural conversations, the inference error rate may further increase. Therefore, the current model is more of a verification prototype and is still far from large-scale deployment. Second, the scale and scope of our training data are still limited. There are also different sub-dialects and complex linguistic phenomena within the Sichuan dialect, and our corpus mainly comes from limited sources and may not represent all variants. This limits the adaptability of the model to a wider Sichuan dialect environment. Third, we did not specifically model some unique features of Sichuan dialect (such as special tone changes, modal particles, etc.), and the model still uses the modeling assumptions of Mandarin, which may not fully capture the characteristics of the dialect. This is also a major reason for the high CER. In addition, when fusing multiple data sources, our strategy is still relatively preliminary, although effective. For example, there is no fine-tuning method using speaker adaptation technology or domain adaptation, so the model is still sensitive to differences between different speakers/data domains. Finally, this study only focused on one evaluation metric, the character error rate, and did not conduct an in-depth analysis of the actual usability of the recognition results. In some application scenarios, metrics such as keyword recognition rate and semantic accuracy may be more valuable for reference, but they were not covered in this study.

In general, the significance of this study is that it provides new ideas and empirical results for speech recognition of low-resource dialects, but there is still room for improvement in terms of accuracy and generalization. The above limitations point out the direction that we need to overcome in subsequent research.

7 Conclusion

This paper focuses on the topic of automatic speech recognition for Sichuan dialect. We explore the use of self-supervised pre-training models (wav2vec2.0 XLSR) under extremely low resource conditions to build an efficient Sichuan dialect recognition system. By combining small-scale multi-speaker corpora with single-speaker long-term corpora for transfer learning, we evaluate the impact of data diversity and quality on recognition performance. Based on the previous discussion, this paper summarizes this research and looks forward to future work and impact.

7.1 Summary of the Main Contributions

The main contributions and innovations of this paper are as follows:

- **Model and task design:** We proposed a model solution for low-resource Sichuan dialect ASR, which fully utilized the advantages of pre-trained cross-language speech models. By fine-tuning the wav2vec2.0 large pre-trained model for Sichuan dialect, we built a recognition model that can work efficiently with a small amount of labeled data. In terms of model design, we retained the powerful representation ability of the pre-trained model and adjusted the output layer according to the characteristics of the dialect to adapt it to the Chinese character-level recognition task, successfully migrating the pre-trained knowledge to the target dialect.

- **Training methods and strategies:** We innovatively adopted a phased, multi-data source integrated training strategy to improve model performance. First, we used small-scale data with high diversity to fine-tune the model to quickly converge its basic recognition capabilities, and then gradually introduced large-scale single speaker data to continue training. At the same time, we dynamically balanced and screened samples from different data sets during the training process (such as moderate signal-to-noise ratio filtering and data enhancement) to avoid the model's over-reliance on a certain data source. This training scheme effectively alleviates the performance degradation problem caused by direct mixed training, and maximizes the use of available corpus while ensuring stable convergence of the model.

- **Data processing and resource construction:** This study developed a dialect-oriented corpus preprocessing strategy. We unified and localized the transcription formats of different datasets, cleaned up the noisy segments, and standardized the transcribed texts to ensure that the character representations of various data sources were consistent during model training. In addition, we constructed a reproducible Sichuan dialect ASR benchmark dataset, including a 6-hour open multi-speaker corpus and a 45-hour single-speaker narrative corpus, and opened up the relevant processing scripts and configurations. This provides a valuable reference resource for subsequent research.

- **Experimental evaluation and results:** We verified the effectiveness of the above method through systematic experimental evaluation. The results show that using less than 11 hours of annotated speech, the CER of the transfer learning model is reduced from more than 77% without fine-tuning to about 26%. Compared with naive mixed training, our method significantly reduces the CER of multi-source training from 50% to about 28%, which is almost the same as

the baseline model trained only with high-quality small data. We also conducted ablation experiments to quantify the contribution of each part, including with or without signal-to-noise ratio filtering, with or without additional single speaker data, etc. The comparison results support our key hypothesis that data diversity is more important for improving performance than strict data screening. These rich experimental results consolidate the reliability of the conclusions of this article.

- **Theoretical and practical significance:** This work empirically expands the knowledge boundaries in the field of low-resource speech recognition. We demonstrate the applicability of pre-trained models on low-resource dialects and emphasize the important role of data diversity. This finding corrects the industry's conventional wisdom that "the more data, the better" and provides a new perspective for low-resource learning in theory. In practice, our research provides specific methodological guidelines for building ASR systems for similar dialects or low-resource languages, including model selection, training strategies, and data processing details, which has practical significance for the universal application of speech technology.

7.2 Future Work

Although preliminary results have been achieved, there are still many directions for further improvement in this study. In future work, we plan to start from the following aspects:

- **Improve model recognition accuracy:** The error rate of the current model is still relatively high, and the CER needs to be further reduced in the future to make the system reach a practical level. To this end, language model fusion or re-ranking technology can be introduced to use large-scale text corpora in the decoding stage to improve the coherence and accuracy of output sentences. At the same time, more advanced acoustic model architectures (such as introducing Transformer-Transducer or deeper Conformer networks) can be tried to enhance the model's expressiveness. In response to the common homophone confusion problem in Sichuan dialect, it is also possible to consider integrating pronunciation dictionary constraints into the model or increasing the modeling of phonemes/tones to reduce errors caused by pure acoustic discrimination.

- **Expand and diversify training data:** The limitation of data resources is one of the bottlenecks of the current system performance. In the future, we will seek to expand the Sichuan dialect corpus, including Sichuan dialect sub-dialects from different regions, voices of more speakers, and corpus that is closer to daily conversations. This will help improve the model's generalization ability to various accents and scenarios. In addition, we are considering using semi-supervised and self-supervised learning to obtain more training materials: for example, automatically generating transcriptions from unlabeled Sichuan dialect audio (inferred by existing models and then manually corrected) to expand the size of the training set; or using pre-trained models to continue pre-training on Sichuan dialect unlabeled speech to better capture the dialect-specific acoustic features.

- **Special modeling of dialect features:** Sichuan dialect has some systematic differences in pronunciation compared to Mandarin, such as aspirated sounds, erhua sounds, tone changes, etc. Future work will conduct special modeling for these dialect phenomena. On the one

hand, we can try to add feature extraction modules for tone and intonation to the front end of the model, or introduce multi-task learning to simultaneously predict tone categories, so as to help the model distinguish dialect-specific pronunciation differences. On the other hand, we can combine the correspondence between Sichuan dialect and Mandarin to introduce cross-dialect phoneme mapping or pronunciation correction mechanism, so that the model can more effectively map dialect speech to the correct characters. By explicitly characterizing the characteristics of dialects, it is expected that recognition errors caused by dialect pronunciation differences can be further reduced.

- **Model domain adaptation and personalization:** Due to the diversity of dialect user groups and application scenarios, it is difficult for a unified model to perfectly adapt to all situations. In the future, speaker adaptation technology can be explored to add speaker embedding vectors to the model or use a small amount of target user voice to quickly fine-tune the model to improve the recognition accuracy of specific users. Similarly, environmental adaptation methods can be studied, such as learning unique parameter adjustments for different recording devices and background noise conditions, so as to improve the robustness of the model in real environments. These adaptive capabilities will make the model closer to actual application needs.

- **More comprehensive evaluation indicators:** Subsequent research will introduce a variety of evaluation indicators to comprehensively measure system performance. For example, the accuracy of single words/words, the recognition rate of named entities, or the semantic error rate can be calculated to evaluate the effectiveness of the model in actual interactions. At the same time, subjective evaluation is also an important direction. We plan to use user dictation tests to determine the usability of recognition results for end users. By enriching the evaluation methods, we hope to discover the weak links in the current model performance and make targeted improvements.

References

Abdel-Hamid, O., Mohamed, A.-r., Jiang, H., & Penn, G. (2012). Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition. In *2012 ieee international conference on acoustics, speech and signal processing (icassp)* (p. 4277-4280). doi: 10.1109/ICASSP.2012.6288864

Baevski, A., Schneider, S., & Auli, M. (2019). vq-wav2vec: Self-supervised learning of discrete speech representations. *CoRR*, *abs/1910.05453*. Retrieved from http://arxiv.org/abs/1910.05453

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations.* arXiv. Retrieved from https://arxiv.org/abs/2006.11477 doi: 10.48550/ARXIV.2006.11477

Bahari, M. H., Saeidi, R., Van hamme, H., & Van Leeuwen, D. (2013). Accent recognition using i-vector, gaussian mean supervector and gaussian posterior probability supervector for spontaneous telephone speech. In *2013 ieee international conference on acoustics, speech and signal processing* (p. 7344-7348). doi: 10.1109/ICASSP.2013.6639089

Bengono Obiang, S. G. B., Tsopze, N., Melatagia Yonta, P., Bonastre, J.-F., & Jiménez, T. (2024, November). Improving tone recognition performance using wav2vec 2.0-based learned representation in yoruba, a low-resourced language. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, *23*(12). Retrieved from https://doi.org/10.1145/3690384 doi: 10.1145/3690384

Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, *56*, 85-100. Retrieved from https://www.sciencedirect.com/science/article/pii/S0167639313000988 doi: https://doi.org/10.1016/j.specom.2013.07.008

Bhatt, S., Dev, A., & Jain, A. (2020). Confusion analysis in phoneme based speech recognition in hindi. *Journal of Ambient Intelligence and Humanized Computing*, *11*(10), 4213–4238. Retrieved from https://doi.org/10.1007/s12652-020-01703-x doi: 10.1007/s12652-020-01703-x

Bradley, D. (2005). Introduction: language policy and language endangerment in china. *International Journal of the Sociology of Language*, *2005*(173), 1–21. Retrieved 2025-06-30, from https://doi.org/10.1515/ijsl.2005.2005.173.1 doi: doi:10.1515/ijsl.2005.2005.173.1

Chen, P. (1999). *Modern chinese: History and sociolinguistics*. Cambridge, UK: Cambridge University Press. Retrieved from https://books.google.com/books/about/Modern_Chinese.html?id=wKMZmdxVj9gC (Internet Archive https://archive.org/details/modernchinesehis00chen)

Cheng, W.-l. (2022). A preliminary analysis of culinary cultural differences based on chengdu–chongqing regional dialects. (33), 55–57. Retrieved 2025-07-01, from https://oversea.cnki.net/KCMS/detail/detail.aspx?dbcode=CJFD&dbname=CJFDLAST2022&filename=WHCC202233019 (In Chinese)

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., & Auli, M. (2020). Unsupervised cross-lingual representation learning for speech recognition. *CoRR*, *abs/2006.13979*. Retrieved from https://arxiv.org/abs/2006.13979

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio

(Eds.), *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N19-1423/ doi: 10.18653/v1/N19-1423

Goyal, P., Dollár, P., Girshick, R. B., Noordhuis, P., Wesolowski, L., Kyrola, A., ... He, K. (2017). Accurate, large minibatch SGD: training imagenet in 1 hour. *CoRR*, *abs/1706.02677*. Retrieved from http://arxiv.org/abs/1706.02677

Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on machine learning* (p. 369–376). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/1143844.1143891 doi: 10.1145/1143844.1143891

Hakkani-Tür, D., Tur, G., Celikyilmaz, A., Chen, Y.-N. V., Gao, J., Deng, L., & Wang, Y.-Y. (2016). Multi-domain joint semantic frame parsing using bi-directional rnn-lstm. In *Proceedings of interspeech 2016*. ISCA. Retrieved from https://www.isca-archive.org/interspeech_2016/hakkanitur16_interspeech.pdf

Hannun, A. Y., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., ... Ng, A. Y. (2014). Deep speech: Scaling up end-to-end speech recognition. *CoRR*, *abs/1412.5567*. Retrieved from http://arxiv.org/abs/1412.5567

Hendrycks, D., & Gimpel, K. (2016). Bridging nonlinearities and stochastic regularizers with gaussian error linear units. *CoRR*, *abs/1606.08415*. Retrieved from http://arxiv.org/abs/1606.08415

Howard, J., & Ruder, S. (2018). Fine-tuned language models for text classification. *CoRR*, *abs/1801.06146*. Retrieved from http://arxiv.org/abs/1801.06146

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (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. doi: 10.1109/TASLP.2021.3122291

Jang, E., Gu, S., & Poole, B. (2017). *Categorical reparameterization with gumbel-softmax*. Retrieved from https://arxiv.org/abs/1611.01144

Jégou, H., Douze, M., & Schmid, C. (2011). Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(1), 117-128. doi: 10.1109/TPAMI.2010.57

Kaur, J., Singh, A., & Kadyan, V. (2021). Automatic speech recognition system for tonal languages: State-of-the-art survey. *Archives of Computational Methods in Engineering*, *28*, 1039–1068. Retrieved from https://doi.org/10.1007/s11831-020-09414-4 doi: 10.1007/s11831-020-09414-4

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., & Tang, P. T. P. (2016). On large-batch training for deep learning: Generalization gap and sharp minima. *CoRR*, *abs/1609.04836*. Retrieved from http://arxiv.org/abs/1609.04836

Kim, S. S. (2017). China's long struggle for linguistic unification. *Global Asia*, *12*(2). Retrieved from https://www.globalasia.org/v12no2/feature/chinas-long-struggle-for-linguistic-unification_samuel-s-kim (Accessed July 2025)

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *International conference on learning representations (iclr)*. Retrieved from https://arxiv.org/abs/

1412.6980

Kinoshita, K., Delcroix, M., Gannot, S., Habets, E. A. P., Haeb-Umbach, R., Kellermann, W., ... Yoshioka, T. (2016). A summary of the REVERB challenge: State-of-the-art and remaining challenges in reverberant speech processing research. *EURASIP Journal on Advances in Signal Processing*, *2016*. Retrieved from https://asp-eurasipjournals.springeropen.com/articles/10.1186/s13634-016-0306-6 doi: 10.1186/s13634-016-0306-6

Ko, T., Peddinti, V., Povey, D., & Khudanpur, S. (2015). Audio augmentation for speech recognition. In *Proceedings of interspeech 2015* (pp. 3586–3589).

Li, A., Yin, Z., Wang, T., Fang, Q., Hu, F., & Qian, Y. (2004). Rasc863: A chinese speech corpus with four regional accents. In *Proceedings of the icslp/cocosda workshop on chinese spoken language processing* (pp. 1–8). Retrieved from http://paslab.phonetics.org.cn/wp-content/files/research_report/2004/2004_15.pdf

Li, D. (2017). *Research on the leshan dialect phonetic system of sichuan* (Ph.D. dissertation, Jiangxi Normal University, Nanchang, China). Retrieved from https://oversea.cnki.net/KCMS/detail/detail.aspx?dbcode=CMFD&dbname=CMFD201801&filename=1017084925.nh (In Chinese)

Li, J., Zheng, T. F., Byrne, W., & Jurafsky, D. (2006). A dialectal chinese speech recognition framework. *Journal of Computer Science and Technology*, *21*(1), 106–115. Retrieved from https://jcst.ict.ac.cn/cn/article/id/1204

Li, Q., Mai, Q., Wang, M., & Ma, M. (2024). Chinese dialect speech recognition: A comprehensive survey. *Artificial Intelligence Review*, *57*(2), 25. Retrieved from https://link.springer.com/article/10.1007/s10462-023-10668-0 doi: 10.1007/s10462-023-10668-0

Li, Y., Best, C. T., Tyler, M. D., & Burnham, D. (2020). Tone variations in regionally accented mandarin. In *Proceedings of interspeech 2020* (pp. 4158–4162). Shanghai, China: International Speech Communication Association. Retrieved from https://www.isca-archive.org/interspeech_2020/li20ja_interspeech.pdf doi: 10.21437/Interspeech.2020-1235

Lin, Y., Zhang, S., Gao, Z., Wang, L., Yang, Y., & Dang, J. (2023). Wav2vec-MoE: An unsupervised pre-training and adaptation method for multi-accent asr. *Electronics Letters*, *59*(11), 580–583. doi: 10.1049/el.2023.0461

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal forced aligner: Trainable text–speech alignment using kaldi. In *Proceedings of interspeech 2017*. ISCA. Retrieved from https://www.isca-archive.org/interspeech_2017/mcauliffe17_interspeech.pdf

Micikevicius, P., Narang, S., Alben, J., Diamos, G. F., Elsen, E., García, D., ... Wu, H. (2017). Mixed precision training. *CoRR*, *abs/1710.03740*. Retrieved from http://arxiv.org/abs/1710.03740

Morris, A. C., Maier, V., & Green, P. D. (2004). From WER and RIL to MER and WIL: Improved evaluation measures for connected speech recognition. In *Proceedings of interspeech 2004* (pp. 2765–2768). ISCA. Retrieved from https://www.isca-archive.org/interspeech_2004/morris04_interspeech.html

Moseley, C. (Ed.). (2010). *Atlas of the world's languages in danger* (3rd ed.). Paris: UNESCO Publishing. Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000187026 (: http://www.unesco.org/culture/en/endangeredlanguages/atlas)

Nossier, S. A., Moniri, M., Wall, J., Glackin, C., & Cannings, N. (2020). Enhancing automatic speech recognition quality with a second-stage speech en-

hancement generative adversarial network. In *Proc. of the ieee international conference on tools with artificial intelligence (ictai)*. IEEE Computer Society. Retrieved from https://repository.uel.ac.uk/download/c2e33ff687bb68226084f4cf1211492f10daf1a6e327dd8ddecc537da28736e1/1974844/Nossier_ICTAI_317.pdf (Accessed: 2025-07-01)

Nowakowski, K., Ptaszynski, M., Murasaki, K., & Nieuważny, J. (2023, March). Adapting multilingual speech representation model for a new, underresourced language through multilingual fine-tuning and continued pretraining. *Information Processing amp; Management*, *60*(2), 103148. Retrieved from http://dx.doi.org/10.1016/j.ipm.2022.103148 doi: 10.1016/j.ipm.2022.103148

Ott, M., Edunov, S., Baevski, A., Fan, A., & Auli, M. (2019). FAIRSEQ: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 4: Demonstrations* (pp. 48–53). Association for Computational Linguistics. Retrieved from https://aclanthology.org/N19-4009.pdf

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345-1359. doi: 10.1109/TKDE.2009.191

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019, September). Specaugment: A simple data augmentation method for automatic speech recognition. In *Interspeech 2019*. ISCA. Retrieved from http://dx.doi.org/10.21437/Interspeech.2019-2680 doi: 10.21437/interspeech.2019-2680

Schneider, S., Baevski, A., Collobert, R., & Auli, M. (2019a). wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

Schneider, S., Baevski, A., Collobert, R., & Auli, M. (2019b). wav2vec: Unsupervised pre-training for speech recognition. *CoRR*, *abs/1904.05862*. Retrieved from http://arxiv.org/abs/1904.05862

Smith, S. L., Kindermans, P., & Le, Q. V. (2017). Don't decay the learning rate, increase the batch size. *CoRR*, *abs/1711.00489*. Retrieved from http://arxiv.org/abs/1711.00489

Textor, C. (2025). *Degree of urbanization in china in selected years from 1980 to 2024*. Retrieved 2025-06-30, from https://www.statista.com/statistics/270162/urbanization-in-china/

The State Council of the People's Republic of China. (2021). *China sees rising number of mandarin speakers*. Retrieved from https://english.www.gov.cn/news/topnews/202106/02/content_WS60b6c3c9c6d0df57f98df0d5.html (Accessed July 2025)

Thennal, D. K., James, J., Gopinath, D. P., & Ashraf, K. M. (2024). *Advocating character error rate for multilingual asr evaluation*. Retrieved from https://arxiv.org/abs/2410.07400

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In I. Guyon et al. (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Wang, Z., Zhao, Y., Wu, L., Bi, X., Dawa, Z., & Ji, Q. (2022). Cross-language transfer learning-based lhasa-tibetan speech recognition. *Computers, Materials & Continua*, *73*(1), 1405–1421. Retrieved from https://doi.org/10.32604/cmc.2022.027092 doi: 10.32604/cmc.2022.027092

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... Rush, A. (2020, October). Transformers: State-of-the-art natural language processing. In Q. Liu & D. Schlangen (Eds.), *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations* (pp. 38–45). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.emnlp-demos.6/ doi: 10.18653/v1/2020.emnlp-demos.6

Wu, J., Wen, Z., Huang, H., Su, H., Liu, F., Wang, H., ... Wu, Q. (2024, March). A reweighting method for speech recognition with imbalanced data of mandarin and sub-dialects. *Service Oriented Computing and Applications*, *18*, 145–152. Retrieved from https://doi.org/10.1007/s11761-024-00384-0 doi: 10.1007/s11761-024-00384-0

Xie, X., Sui, X., Liu, X., & Wang, L. (2022). *Investigation of deep neural network acoustic modelling approaches for low resource accented mandarin speech recognition.* Retrieved from https://arxiv.org/abs/2201.09432

Xu, F., Dan, Y., Yan, K., Ma, Y., & Wang, M. (2021a, October). Low-resource language discrimination toward chinese dialects with transfer learning and data augmentation. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, *21*(2). Retrieved from https://doi.org/10.1145/3473499 doi: 10.1145/3473499

Xu, F., Yang, J., Yan, W., & Wang, M. (2021b). An end-to-end dialect speech recognition model based on self-attention. *Journal of Signal Processing*, *37*(10), 1860–1871. Retrieved from https://signal.ejournal.org.cn/en/article/doi/10.16798/j.issn.1003-0530.2021.10.009 doi: 10.16798/j.issn.1003-0530.2021.10.009

Yadav, H., & Sitaram, S. (2022). *A survey of multilingual models for automatic speech recognition.* Retrieved from https://arxiv.org/abs/2202.12576

Yu, D., & Deng, L. (2015). *Automatic speech recognition: A deep learning approach*. London / New York: Springer. Retrieved from https://doi.org/10.1007/978-1-4471-5779-3 doi: 10.1007/978-1-4471-5779-3

Yuan, J., Ryant, N., Cai, X., Church, K., & Liberman, M. Y. (2021). Automatic recognition of suprasegmentals in speech. *CoRR*, *abs/2108.01122*. Retrieved from https://arxiv.org/abs/2108.01122

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). Bertscore: Evaluating text generation with BERT. *CoRR*, *abs/1904.09675*. Retrieved from http://arxiv.org/abs/1904.09675

Zhang, W., & Levis, J. M. (2021). The southwestern mandarin /n/-/l/ merger: Effects on production in standard mandarin and english. *Frontiers in Communication*, *Volume 6 - 2021*. Retrieved from https://www.frontiersin.org/journals/communication/articles/10.3389/fcomm.2021.639390 doi: 10.3389/fcomm.2021.639390

Zhao, E., & Wu, Y. (2020). *Over 80 percent of chinese population speak mandarin.* People's Daily Online. Retrieved from https://en.people.cn/n3/2020/1016/c90000-9769716.html (Accessed: 2025-06-29)

Zhou, M. (2003). *Multilingualism in china: The politics of writing reforms for minority languages, 1949–2002*. Berlin: Mouton de Gruyter. Retrieved from https://books.google.com/books/about/Multilingualism_in_China.html?id=joE5ZASNCGYC

Appendices

A AI Declaration for Master's Thesis

I hereby affirm that this Master's thesis, titled "Transfer Learning for Sichuan Dialect Automatic Speech Recognition Based on Pretrained Wav2vec 2.0 Model," was composed by myself, and that the work herein is my own except where explicitly stated otherwise in the text. This work has not been submitted for any other degree or professional qualification except as specified, nor has it been published. Where other people's work has been used (from any source: printed, internet, or other), this has been carefully acknowledged and referenced in accordance with academic standards.

During the preparation of this thesis, I used Grok 3, created by xAI, All content generated or assisted by Grok 3 was subsequently reviewed, verified, and substantially modified by me to ensure accuracy, alignment with my research objectives, and adherence to academic standards. I understand the limitations of Grok 3, particularly its potential for generating generic or contextually inaccurate outputs in specialized domains like automatic speech recognition, and I critically evaluated its suggestions to ensure they were appropriate for the Sichuan dialect ASR context. I chose Grok 3 over other AI tools due to its robust natural language processing capabilities and ability to provide structured technical summaries, which were suitable for my needs in drafting and refining technical content.

No AI tools were used for generating research hypotheses, experimental methodology, data analysis interpretations, conclusions, or any content where independent reasoning and domain expertise were being assessed, as such uses are prohibited under the University of Groningen AI policy.

ZiYi Li July 3, 2025